Will IT matter?
- *the role of IT in development*

Proceedings of the 7th International Information Technology Conference
IITC 2005

Colombo
Sri Lanka

9th November - 10th November 2005
Trans Asia Hotel, Colombo, Sri Lanka

Disclaimer

The views expressed in the papers published in these proceedings are solely those of the authors and they do not necessarily represent the views of the Infotel Lanka Society Ltd.

Proceedings of the 7th International Information Technology Conference

Conference Website : http://www.iitc.lk

ISBN : 955-8974-03-X

Printed from the camera ready copy supplied by the University Of Colombo School Of Computing

Preface

The International Information Technology Conferences (IITC) has matured over the past 6 years into the key ICT event in the Sri Lankan calendar. This can be amply evidenced by the increasing numbers of paper submission readily received.

The IITC focuses on ICT research directions and thus complements other local conferences such as the National IT Conference organized annually by the CSSL which have an industry applications orientation. As such the IITC attempts to showcase future trends in ICT by attracting papers on current and cutting–edge research.

This year's conference themed " Will IT Matter ? – The Role of IT in Development" attracted some 50 paper submission of which a total of 25 have been accepted for presentation and publication by a panel of referees comprising of both international and local academics & professionals. In addition 6 papers have been accepted as Posters. These papers cover a variety of relevant ICT topics ranging from Secure Computing, Web Services and applications, Networking and Internet Technologies, Software Engineering & Processes, Language Processing Techniques, Algorithms and Computational Intelligence and e-Learning.

This 7th International IT Conference comes at an opportune moment with Sri Lanka moving forward with its ambitious e-Sri Lanka programme aimed at making the benefits of ICT felt in every village and to every citizen of the country. As such, the themes chosen for IITC is a bold challenge to us all to show that IT does matter.

In addition to the conference, the IITC workshops provide an excellent opportunity to industry and academia alike to be exposed to high quality in-depth insights into the most relevant and state-of- the-art areas of ICT around the globe. This is a forum for technology experts to provide the industry with updates on technologies that are emerging in the global scene. This year's workshops cover areas such as Wireless Networks – Standards, Protocols and Applications, Computer Security Application Hands-On, e-Learning and Learning Management Systems, Computer Forensics, Re-engineering of Organizations using ICT for Superior Performance, Creating and Remastering GNU/Linux Live CDs and Governance of IT & Measuring the Value of IT.

The Conference and its associated workshop program provide a unique opportunity for initiating and enhancing industry-research partnerships and collaboration both locally and with potential international partners. This is very much at the center of the overall long-term objectives of the IITC and is expected to play a key role in the development of Sri Lanka as an internationally branded country for ICT products and services in the years ahead.

We would like to acknowledge the ready support of the local ICT industry in coming forward with generous sponsorship of the entire IITC 2005 program. We are convinced that theirs is a significant investment towards making IT matter. We also thank all paper authors and presenters as well as workshop and tutorial resource persons for helping make IITC 2005 a fruitful experience for all participants. Finally we wish to record our genuine appreciation of all conference and workshop delegates without whose presence and participation IITC 2005 would not be able to achieve its objectives.

Main Organizing Committee

List of Papers

Secure Computing

Web Services and Applications

Networking and Internet Technologies

Software Engineering & Processes:

Language Processing Techniques

Algorithms and Computational Intelligence

e- Learning

Watermarking Method for 3D Triangular Mesh Models.

Md. M. Rahman[1], K. Kaneda and K.Harada
1-7-1 kagamiyama, Higashi-Hiroshima 739-8521,
Hiroshima University,
Japan
Email:[1]mahfuzurr@yahoo.com

Abstract

A new topology based watermarking method is proposed to embed information in objects with layered 3D triangular meshes such as those reconstructed from CT or MRI data. The main idea of the method is to compare height of the vertices of a triangle lying in the same layer. A watermark message is converted into a binary bit sequence, and then embedded into the model in either way that the first vertex of a triangle in the upper level or in the lower level, that caries information 1 or 0, respectively. For experimental purpose, a watermark message is embedded in a mouse embryo model consisting of 61 layers, 3 elements, 15776 vertices, and 31318 triangles. The method is effective, computationally faster and inexpensive. It is robust against translation, rotation, re-sectioning, local deformation and scaling. It left some artifacts after re-arrangement of local or global numbering. It is very useful for shape sensitive 3D geometric models.

Keywords: watermarking, layered 3D triangular mesh model, topology based embedding, computer graphics.

1. Introduction

In digital media environment, it is very easy to copy, modify and distribute various kinds of digital data such as texts, images, audios, videos and recently 3D geometric models through Internet, CD-ROM, etc. In this environment an important research is watermarking of these data for Intellectual property protection as well as for efficient data management and content labeling. The bulk of the research of watermarking of digital media has focused on media such as texts, images, videos, and audios [1, 2]. Recently 3D geometric models have been widely used in the area of CAD, CAM and Computer Graphics as well as in medical applications and have been recognized as important models. Telemedicine, robot assisted operation, virtual operation, etc. are very promising future trend of research in medical science where there is a lot of use of reconstructed 3D geometric models of different objects. So, watermarking of these 3D geometric models is very important in these days.

Geometry, topology and attributes can be a target of embedding watermark in 3D geometrical models. Researchers have mainly focused on embedding watermark in the geometry of a 3D model [3, 6, 8, 11, 12, 13, 14], while a few have targeted parametric curves and surfaces [10]. Others have targeted movement of 3D models, that is, MPEG4 facial animation parameters [5] or attributes of 3D models. R. Ohbuchi, et al. presented some idea about topology based watermarking of 3D models [9]. However, 3D geometric models, such as for medical applications, rarely tolerate its changes in geometry. They are very geometry sensitive and this ultimately indicates the importance of topology based watermarking method. Unfortunately, so far my knowledge goes, to date no effective topology based watermarking method has been developed.

This paper's contribution is in topology based watermarking of 3D triangular mesh model. Here a topology based watermarking method is developed to embed information in objects with layered 3D triangular meshes such as those reconstructed from CT or MRI data. It is successfully applied to mouse embryo model. Embedding and extraction process are done successfully. It is fast, efficient and effective. It is invisible and robust against translation, rotation, scaling and re-sectioning. It is useful for content labeling, ownership data assertion, efficient data management, detection of change of data etc. prevailing the way for copyright protection of precious 3D model data.

Watermarking loosely refers to the use of steganography in the application areas of ownership assertion, authentication, content leveling, content

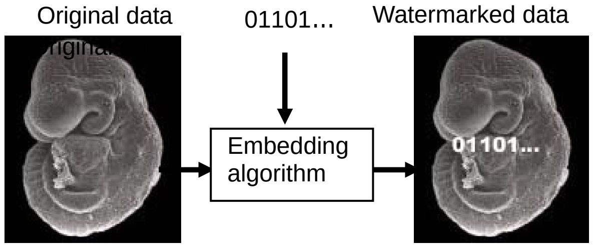


Fig. 1 (a): Watermark embedding process

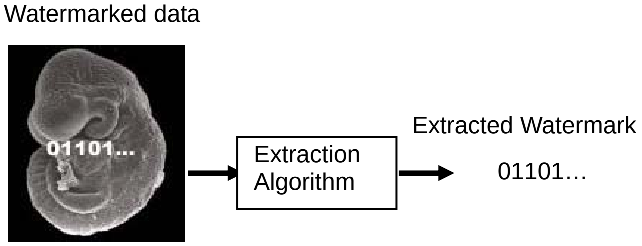


Fig. 1 (b): Watermark extraction process

protection, distribution channel tracing etc. while steganography addresses the problem of hiding information within digital data. In watermarking process, an invisible digital watermark such as important data identification information, copyright or ownership data is embedded into the original data. Then, the author distributes this watermarked data instead of the original data. When necessary, the author can extract the embedded watermark by using the appropriate extraction algorithm. The process is shown in Fig. 1. Not only copyright protection, identification of important data i.e. content labeling, efficient data management as well as detection of change of data are also important roles of watermarking.

2. Background and main idea

Layered data is commonly used in medical field. In medical applications there are many 3D triangular mesh models reconstructed from CT (Computer Tomography), MRI (Magnetic Resonance Imaging) data which are having layered construction. However, these models, used for medical applications, are very sensitive to geometrical error. So, topology based algorithm is necessary to watermark these models. We found that it is possible to exploit the layered construction of these models to embed data into it topologically.

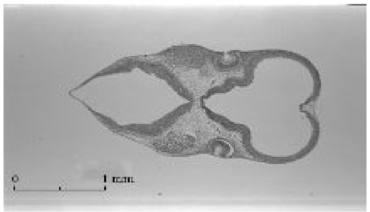


Fig. 2(a): Mouse Embryo cross-section

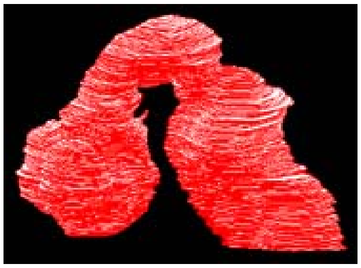


Fig. 2(b): Stack of extracted parallel contours

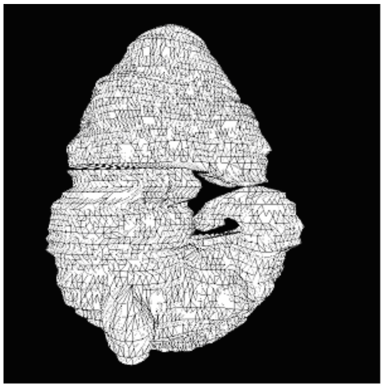


Fig.2(c): Reconstructed mouse embryo model

A mouse embryo model [4] is used, as an example, to embed a watermark. The model is reconstructed from serial sections taken from an optical microscope. One such section is shown in Fig. 2(a). Sections are further processed and finally they become digital picture elements, the pixels. Contour-based reconstruction is used for surface formation of the model. In contour-based reconstruction the idea is to compute the structure boundary in each individual slice as contour lines and to estimate the 3D boundary surfaces as a set of patching elements (usually triangles) that envelops a given set of contours. These contours are all parallel as shown in Fig. 2(b). The reconstructed mouse embryo model, having layered construction, is shown in Fig. 2(c). The developed algorithm aims at embedding watermark topologically to the layered construction of a model keeping the geometry unchanged. The basic idea of the algorithm is to compare height of the vertices of a triangle lying in the same layer. A watermark message is converted into a binary bit sequence, and then embedded into the model in either way that the first vertex of a triangle in the upper level or in the lower level, that caries information 1 or 0, respectively (see Sec. 4). The extraction process is the reverse process of embedding.

3. Requirements of watermarking

The requirements of a watermarking method are always application specific. Still there are some general requirements, what majority of the watermarking methods should satisfy, can be categorized as follows:

3. 1. Perceptually invisibility

The embedded watermark must be perceptually invisible and unnoticeable for a third party in terms of model's intended use.

3. 2. Capacity of the watermark

The amount of watermark embedded should be large enough to record important data identification, author's information etc.

3. 3. Robustness

The embedded watermark should be robust against 3D affine transformations, such as translation, rotation and scaling, which are frequently applied to 3D geometric models. For application purpose the model may need to be cut into several pieces or local deformation could be done to reshape a model. So, the embedding should withstand re-sectioning and local deformation. Some watermarking methods also need to be robust against intentional attacks such as filtering, compression, re-meshing and noise addition to the watermarked model or these operations should degrade the quality of the model regarding its intended use.

3. 4. Geometrical error

The geometrical error caused by the watermark (if any) should be restricted within the specified tolerance according to the application. However, many of the 3D geometric models are error sensitive and rarely tolerate any error.

4. Description of the algorithm

The basic idea of the algorithm is to compare height of the vertices of a triangle lying in the same layer. A watermark message is converted into a binary bit sequence, and then embedded into the model in either way that the first vertex of a triangle in the upper level or in the lower level, that caries information 1 or 0, respectively. The extraction process is the reverse process of embedding. The main algorithm is discussed in section 4.1 where a group of consecutive triangles carry binary information maximizing the capacity of the watermark and its extension is discussed in section 4.2 where binary bits are sparsely embedded among triangles providing low hit rate.

4.1. Main algorithm

4.1.1. Step 1: Preparing binary watermark message

It produces a binary representation of a textual watermark message data. Prior to application, the binary bits are encrypted to an unintelligible format with well-known and powerful RSA encryption algorithm.

4.1.2. Step 2: Message package setup (See Fig. 4.1 (a))

Flag bit is set to 1 to check upside-down rotation. Header stores the total bit number in a message package and footer stores the total bit numbers written in a model. If the model is rotated upside down, the flag bit embedded at the starting point, will be set to 0 and the algorithm will change its extraction strategy.

Flag bit	Header	Message	Footer

Fig. 4.1 (a): Structure of message package

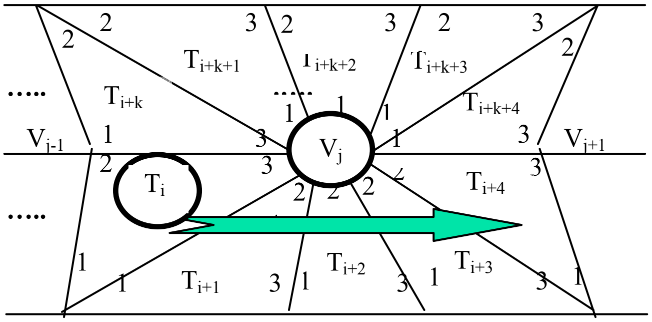


Fig. 4.1 (b): Starting triangle

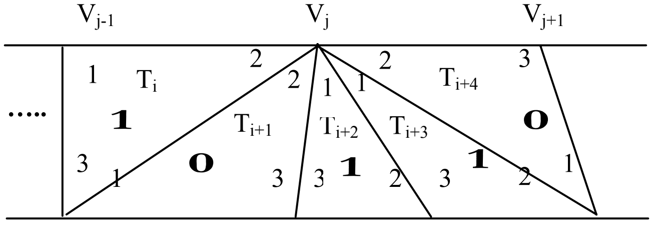


Fig. 4.1 (c): Embedding a binary sequence

Footer is useful for checking whether the watermark message is destroyed or not by cut off operation, local deformation, and so on, when reading the watermark.

4.1.3. Step 3: Selection of the starting point

The vertex that the maximum numbers of triangles are connected with is found in a triangular mesh model, and the triangle with the smallest id number that is connected with the vertex is decided as the starting triangle. As shown in Fig. 4.1 (b) vertex V_j is connected with the maximum number of triangles in a model. Triangle T_i is the smallest in id number among triangles connected with vertex V_j, and hence it is selected as the starting triangle.

4.1.4. Step 4: Data embedding

Local numbering of the vertices of a triangle is changed to embed data. The first vertex in the upper and lower levels carries information 1 and 0, respectively. A binary bit pattern "10110" is embedded in the mesh shown in Fig. 4.1(c). As the algorithm changes only the local numbering of the triangular patch of the model, the geometry of the model remains totally same. In this way almost all of the triangles can be used for data embedding maximizing the capacity of the watermark and the data is embedded sequentially. If capacity doesn't matter and hit rate is the vital factor, then every bit of information is embedded in a scattered manner with a random interval. If consecutive two triangle carries information 1 it is 1 else 0.

4.2. Extension of the algorithm

With an intension to make it difficult for intruder to detect the change in local numbering, the method is further extended embedding the information at random intervals. It provides low hit rate embedding i.e. it lowers the possibility of illegal detection of embedded information without proper knowledge of embedding.

4.2.1. Step 1: First divide the total triangles by the total number of bits in watermark message package producing consecutive segment of triangles (i.e. n triangles each) following normal numbering of the triangular mesh.

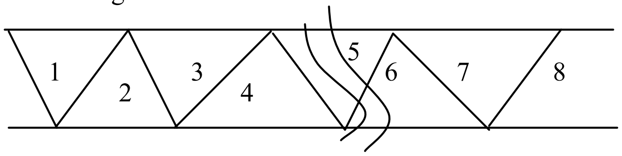


Fig. 4.2 (a) Regular interval division of mesh

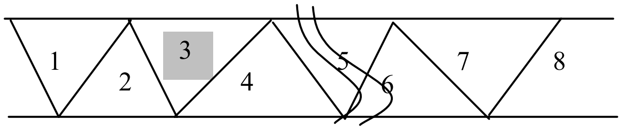


Fig. 4.2 (b) Random selection of a triangle

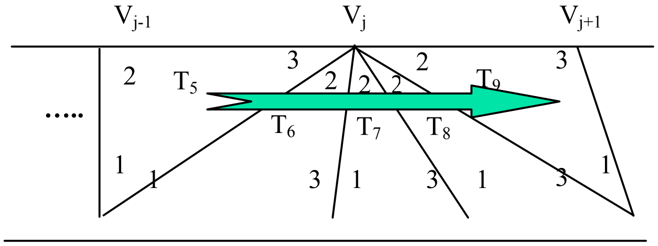


Fig. 4.2 (c): Mesh without any embedded information

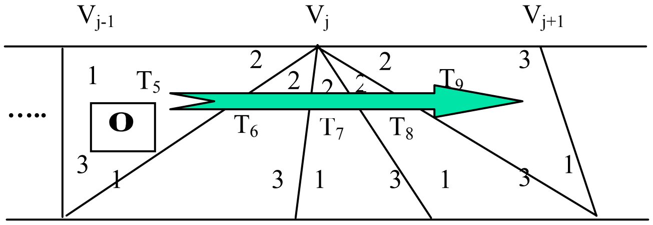


Fig. 4.2 (d): Mesh with embedded information 0

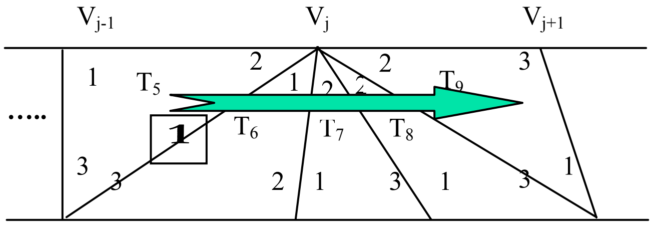


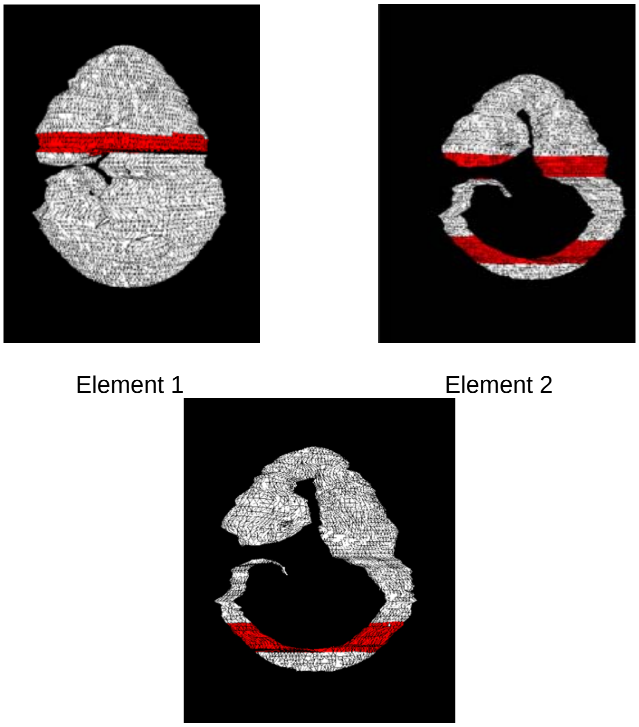
Fig. 4.2 (e) Mesh with Embedded information 1

4.2.2. Step 2: Randomly select any digit out of first (n-1) triangles to embed one bit of information there.

4.2.3. Step 3: To embed information 0 change local numbering of only one triangle and to embed 1 change local numbering of consequent two triangles as described in Sec. 4.1.3. This is an extension at step 4 of the main algorithm. After first three steps of the main algorithm, if capacity of watermark is found very small and hit rate is found important than the extended version of the method is applied.

5. Experimental results and discussion

Watermark is embedded in a mouse embryo model consisting of 61 layers, 3 elements, 15776 vertices, and 31318 triangles. The message package composed of a flag, header, message and footer. To improve the robustness against local deformation and re-sectioning the same binary sequence is embedded several times changing a starting point. The second, third and other starting points are found in the same way described above, removing the vertex with the maximum number of triangles in the previous embedding process. In the experiment, a watermark message composed of 115 characters, i.e., 920 bits of information is embedded. Including 1 bit flag, 10 bits header and 16 bits footer, it becomes 947 bits and the same watermark is embedded for four times. So, the total number of bits written in the model is 3788 bits. Retrieval of the watermark is also done successfully. Watermarked area of the model is visualized with red colored triangles as shown in fig. 5. It is important to mention here that the watermark is visualized only for the purpose of demonstration. In practical application the watermark remains totally invisible. The algorithm is robust against major unintentional attacks such as translation, rotation, scaling, re-sectioning and left artifacts even after some intentional attacks such as local number re-arrangement and global number re-arrangement.



Element 1

Element 2

Element 3

Fig. 5: Mouse embryo model after embedding

These artifacts are unique starting points and the number of its associated triangles. From these artifacts the owner of the model can claim its originality.

5.1. Comparison with an existing method

The proposed method fulfills majority of the general watermarking requirements. It is invisible, capacity is high enough and it can withstand translation, rotation and scaling. It is also robust against re-sectioning. The method is similar to triangle strip peeling symbol sequence embedding (TSPSSE) [9], but it has more advantages as shown in table 1.

With proper knowledge of embedding process, practically every watermark is possible to destroy. In our case if local numbering of the triangles are re-arranged then the embedded watermark will be destroyed but still it left some identifying mark i.e. unique starting triangle as well as the vertex connected with maximum number of triangles.

5.2. Different Modes of Embedding

There are three different modes of Embedding information made the algorithm useful for practical purposes with different application scenarios.

5. 2. 1. Embedding Information Sequentially: In this approach binary bit stream is embedded sequentially following normal numbering of triangles starting from the selected unique starting point and continue embedding till the end of message package. In case of layered triangular meshes, triangles are numbered horizontally. So vertical re-sectioning can destroy the embedded information.

Method	**TSPSSE [9]**	**The proposed method**
Bounding edge crossing problem	Has problem	No problem
Applicable to data type	Any type of triangular mesh	Only layered triangular mesh
Space efficiency	Lower	Higher
Non- uniform scaling	Not robust	Robust
Starting point	Not uniquely determined and not topological	Uniquely determined and topological
Mesh type	Mesh should be orientable	No such requirements

Table 1: Comparison of the proposed method

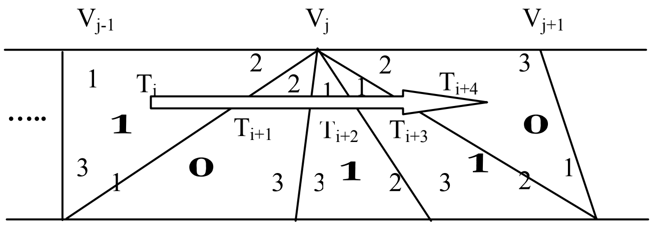


Fig. 5.2 (a): Sequential Embedding

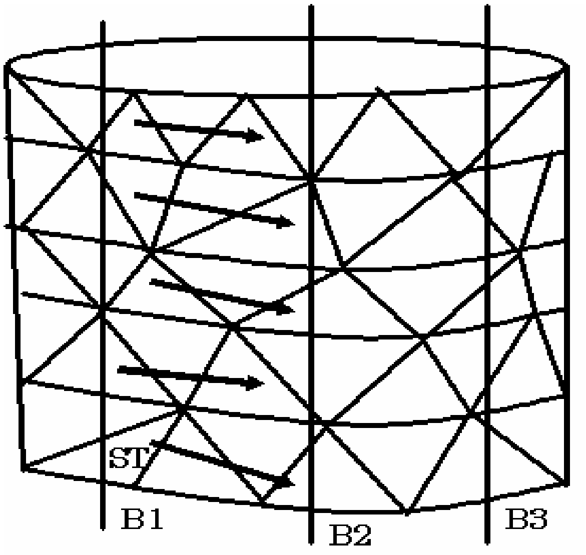


Fig. 5.2 (b): Vertical direction embedding

Two approaches are proposed to make the method robust against vertical re-sectioning. One is embedding information several times in the vertical direction and the other is the spot area embedding discussed in the following section. In vertical direction embedding, some specific boundary is selected in the vertical direction and the same information is embedded for several times as shown in Fig. 5.2 (c).

5. 2. 2. Embedding Information in circular fashion (Spot area embedding):

In spot area embedding, the same message is cut into several pieces and each piece is placed in different spots. If some part of the message is missing, because of re-sectioning, the same information is retrieved from other spot by error correct decoding. It is shown in Fig. 5.2 (d). This approach is complex but more effective.

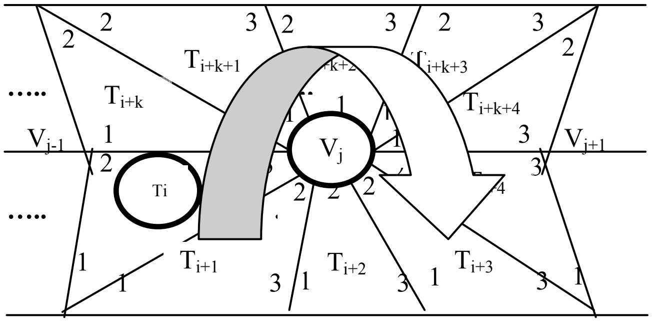


Fig. 5.2 (c): Starting point for circular fashion embedding

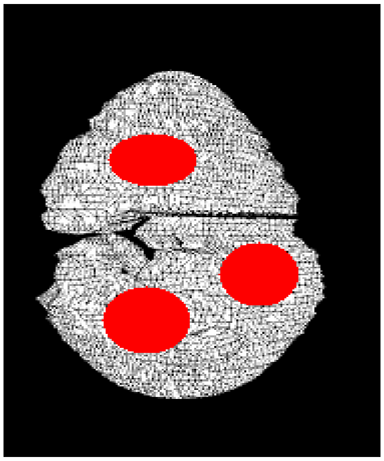


Fig. 5.2 (d): Embedding information in circular fashion (Spot area embedding)

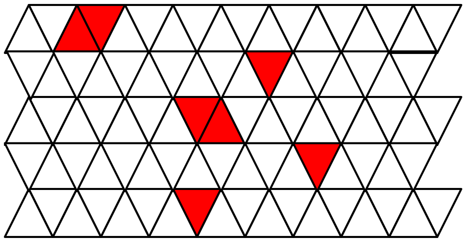


Fig. 5.2 (e): Random interval embedding

5. 2. 3. Embedding Information in random intervals:

As the binary bits are embedded at random interval and only one or two triangle's local numbering are changed in one place, it is less susceptible for the intruder about the availability of a watermark.

5.3. Other data type, hit rate, capacity and secrecy:

To apply the method to irregular data type (i.e. not layered data) it is proposed that the irregular data is first converted to layered one and then the proposed method is applied.

As in the first step of the method the bit stream is encrypted to an unintelligible format with famous and powerful RSA encryption algorithm, even if the intruder detects the watermark, he cannot read it because the message is in encrypted format. From the starting triangle the binary bit stream is embedded sequentially covering almost all the triangles maximizing capacity of the watermark. Alternatively, options are kept to embed information in random regular interval to lower hit rate i.e. it lowers the possibility of illegal detection of embedded information without proper knowledge of embedding.

6. Conclusions

The proposed method is simple, effective, computationally faster, and inexpensive. It is robust against unintentional attacks translation, rotation, re-sectioning, scaling etc. and left artifact after intentional attacks of local and global number re-arrangement. It is applicable for important data identification, detection of change of data, content labeling, ownership assertion, etc. of layered 3D triangular mesh models. It is very useful for watermarking of geometry sensitive 3D triangular mesh model. Further extension of the method is possible making a embedded program that will destroy the model in-terms of its intended use if re-meshing or mesh simplification is applied to it.

Acknowledgements

The first author would like to thank Japanese Ministry of Education, Science, Sports and Culture as well as “Gakushu Shoreihi” scholarship provider for sponsoring the research work.

References

[1] W. Bender, D. Gruhl, N. Morimoto and A. Lu, Techniques for data hiding, IBM Systems Journal, Vol. 35, Nos. 3&4, 1996, pp. 313-336.

[2] I. J. Cox and M. L. Miller, A review of watermarking and the importance of perceptual modeling, Proc. SPIE Conference on Human Vision and Electronic Imaging II, Vol. 3016, February 1997, pp. 92-99.

[3] O. Benedens, Geometry-Based Watermarking of 3D Models, IEEE CG & A, Vol. 19, No. 1, pp. 46-55, 1999

[4] Roman Durikovic, Kazufumi Kaneda, Hideo Yamashita, Imaging and Modelling from Serial Microscopic Sections for the study of Anatomy, Medical & Biological Engineering & Computing, Vol. 36, No. 3, pp. 276-284, 1998

[5] F. Hartung, P. Eisert and B. Girod, Digital Watermarking of MPEG-4 Facial Animation Parameters, Computer Graphics, Elsevier, Vol. 22, No. 4, pp. 425-435, 1998

[6] S. Kanai, H. Date, and T. Kishinami, Digital Watermarking for 3D Polygons using Multiresolution Wavelet Decomposition, Proc. of the Sixth IFIP WG 5.2 International Workshop on Geometric Modeling : Fundamentals and Applications, pp. 296-307, Japan, 1998

[7] S. Katzenbeisser, F. A. P. Petitcolas, Information Hiding, Techniques for steganography and Digital Watermarking, Artech House, London, 2000

[8] R. Ohbuchi, H. Masuda, and M. Aono, Watermarking Three-Dimensional Polygonal Models, Proceedings of the ACM Multimedia `97, Seattle, Washington, USA, pp. 261-272, 1997

[9] R. Ohbuchi, H. Masuda, and M. Aono, Watermarking of 3D Polygonal Models Through Geometric and Topological Modifications, IEEE JSAC,98, pp. 551-560

[10] R. Ohbuchi, et al., A shape-Preserving Data Embedding Algorithm for NURBS Curves and Surfaces, Proc. Computer Graphics International` 99, pp. 177-180, Canada, 1999

[11] R. Ohbuchi, A. Mukaiyama, S. Takahashi, A Frequency-Domain Approach to Watermarking 3D Shapes, in Proc. EUROGRAPHICS 2002, Saarbrucken, Germany, Sept., 2002

[12] Ryutarou Ohbuchi, Hiro Ueda, Shu Endoh, Robust Watermarking of Vector Digital Maps, in proceedings of the IEEE Conference on Multimedia and Expo 2002 (ICME 2002), Lausanne, Switzerland, August 26-29, 2002.

[13] R. Ohbuchi, Akio Mukaiyama, Shigeo Takahashi, Watermarking 3D shape model defined as a point set, Proceedings of the 2004 international conference on cyberworlds, pp. 392-399, Tokyo, Japan, 2004.

[14] E. Praun, H. Hoppe and A. Finkelstein, Robust Mesh Watermarking, MSR-TR-99-05, Microsoft Research, SIGGRAPH, 1999.

Workplace Communication Privacy in the Digital Age

P. Mahanamahewa* and R. Dayarathna**
Attorney-at-Law, Senior Lecturer in Law, Faculty of Law, University of Colombo, Research fellow, T.C Beirne School of Law, University of Queensland, Australia.

**PhD candidate, Stockholm University, Sweden

Abstract

This paper attempts to lay the foundation for future research into an area that has been called the "hottest workplace privacy topic of the next decade." The existing empirical studies and the literature reviewed of this area suggest that the latest intrusive monitoring technologies which have been introduced to the current workplace has undoubtedly created an unwanted and unexpected imbalance and developed a wide gap in the 21st century employer/employee relationship. The paper argues for the introduction of privacy enhancing technologies empowered with legal instruments in protection of workplace privacy. In addition, the paper is of the view that employees' awareness and training on workplace privacy policy developments are decisive factors to achieve this objective and this in turn creates trust and confidence and beneficial to both employees and employers in the current workplace. The paper proposes a contractarian framework to protect employers' interests and employees' on-line rights.

This paper suggests that employees' views and opinions are more important in computer monitoring to develop a privacy policy in the workplace. To attain these objectives an empirical survey was conducted in five government sector organizations in Sri Lanka to gather factual information and to examine attitudes, beliefs and opinions on computer monitoring. The results of the study could be used as a guide for policy-makers and for legislatures involved in drafting privacy legislation, and associated policies relevant to the Sri Lankan workplace.

1. Introduction

The focus of this paper is on employee electronic mail (e-mail) and Internet monitoring in the public sector workplace in Sri Lanka and its impact on their workplace privacy rights. The central research question to be addressed here relates to the absence of specific laws governing this area as how to balance employers' interest and employees' rights in such organizations. The paper proposes a contractual framework to establish a balanced approach which protects employees' rights and employers' interest in the workplaces. The study evolved from examining public sector employees', and managers' views, opinions and attitudes to form an important foundation for subsequent policy recommendations for the implementation of workplace privacy protection in Sri Lanka.

In recent years, the use of electronic on-line communication has increased substantially. Increasing numbers of employees have access to this form of communication in their jobs, mainly through e-mail and Internet. With increasing use of electronic communication, the number of controls available to the employer has also grown. This raises numerous issues for employers and employees in terms of the relationship between workers' privacy and employers need to control and monitor the use of ICT.

2. Literature review

According to International Labour Organisation (ILO, 1993) and Electronic Privacy Information Centre (EPIC, 2004) monitoring of employees by managers is undertaken in many ways and for many reasons. Workplace practices which may affect employee privacy essentially fall into four categories: (a) monitoring and surveillance (including e-mail and Internet, phone and Close-Circuit Television) (b) physical and psychological testing (including pre-employment testing, drug-testing and the use of DNA data) (c) searches of employees and their property and (d) the collection, use and disclosure of workers information.

Due to the advancement of technology, in recent years, e-mail and Internet usage has become a significant concern in the workplace. With the increasing use of e-mail and Internet, the number of controls available to the employer has also grown (Marx and Sherizen, 1986; Griffith, 1993; Websense, 2002; e-policy institute, 2004). This raises significant issues for employers and employees in terms of the relationship between workplace privacy and employers that are needed to control and monitor the use of this facility. The growth of electronic surveillance in the workplace has been phenomenal and has indeed created a global problem (EPIC, 2004).

A recent worldwide survey on e-mail and Internet usage at workplaces conducted by the Privacy Foundation in 2001 revealed that the usage of e-mail and Internet increased by 25% on one hand and on the other hand monitoring increased by 30%. Similarly, another study by the Society for Human Resource Management (2002) found that 74% of the 722 companies surveyed said that they monitored workers Internet use and 72% said they checked on employees e-mail. The latest American Management Association (AMA, 2004), e-mail and Internet monitoring study revealed that almost 80% of the largest companies in the USA had engaged in some form of electronic surveillance over the previous years. This figure is more than double the rate recorded only five years ago: 35.3% in 1997 when AMA began its surveys. In short, as long as there has been employment, employees have been monitored (Nebeker and Tatum, 1993). However, in recent years, with an environment of affordable technology, the availability of less easily observable or detectable monitoring devices, and lack of adequate regulation, has been an explosion in the electronic monitoring and surveillance in the workplace. The balance of the traditional workplace has eroded with this artificial control and unreasonable process (Geist, 2002).

For managers, monitoring is necessary because it is argued that workplace e-mail and Internet monitoring is the most effective means to ensure a safe and secure working environment and to protect their employees. In addition, some contend that monitoring may boost efficiency, productivity and customer service and allows to more accuracy evaluate performance (Detienne, 1993; Sipior and Ward, 1995; Orthmann, 1998). The impact of the monitoring of these workplace relationships is the focus of this paper. If used reasonably it may enhance efficiency without “trenching on” employees rights.

However, critics of monitoring point to research evidencing a link between monitoring as well as psychological and physical health problems, increased boredom, high tension, extreme anxiety, depression, anger serve fatigue and musculoskeletal problems (Kidwell and Bennett, 1994; Chalykoff and Kochan, 1989; OTA, 1987; working Women 9 to 5, 1990). More seriously, critics point to violations of their fundamental right to privacy (Stone et al., 1983; Bylinsky, 1991; Culnan, 1993; Smith, 1993; Vest, Perry and O’Brien, 1995; Sipior and Ward, 1995). Unless an acceptable remedy is soon found, workplace productivity may rapidly deteriorate and employee morale may disintegrate.

These issues are of particular concern for the Sri Lankan workplace as a result of the rapid increase in use of Internet and e-mail by the Sri Lankan public/private organizations over the past few years and also recently launched electronic government project under the auspices of World Bank (e-Sri Lanka, 2003). Most government institutions in Sri Lanka are connected to the Internet and e-mail with associated high computer use. Sri Lanka has also recognized the need to review and reform laws to face new challenges posed by electronic revolution (E-G policy preliminary draft, 2003; Draft ICT policy for Government, 2005).

Sri Lanka, like other countries, has to act quickly, as facts and figures reveal an increased dependence on information technology communication, increased number of users and services available due to fully computerization of public sector organizations and increased coverage of access to e-mail and Internet at work under this electronic government project.

Accordingly, the present study is to propose a sound policy that strikes a philosophical and pragmatic balance between protecting the private life of employee, and legitimate control by the employer. The desire here is to safeguard the basic right of employees to have their private life respected in the work environment, by determining the purposes and conditions for controlling electronic online communication data, with the needs of the good operation of the organization taken into account. This has been endorsed by the (Article 29 EU Working Party, 2003) in their statement specifically:

> “A blanket ban on personal use of the Internet by employees may be considered to be impractical and slightly unrealistic as it fails to reflect the degree to which the Internet can assist employees in their daily life.”

In short, these competing agendas between the employee’s right to privacy and the employers’ right to control the use of their resources is a crucial issue which will influence e-mail and Internet privacy debates at workplace long into the future. This study identifies the beliefs and attitudes of executives and employees on e-mail and Internet monitoring in the public sector workplace in Sri Lanka. Based on these findings a variety of strategies to improve workplace privacy policy is presented.

3. Aims and methodology

This study aims to find out whether employees are fully aware of their on-line rights in the public sector workplace arising from extensive monitoring. Secondly, it seeks to find out the level and manner of e-mail, Internet monitoring in the public sector workplaces. The study also analyses the prevailing attitudes of employees towards e-mail and Internet surveillance implementation. Finally, it attempts to identify the influences likely to determine the future and nature of workplace e-mail and Internet monitoring.

In investigating the workplace privacy in the public sector organizations in particular e-mail and Internet monitoring in Sri Lanka and the predicaments they have encountered, the researcher decided to use a combination of research methods for data collection. For a basic understanding of how public sector employees coped with their jobs and private life at the workplace in Sri Lanka, the researcher used semi-structured questionnaire for these employees and managers in order to gain a detailed picture of a respondent's belief about, or perceptions or accounts of workplace privacy. Semi-structured questionnaires give the researcher and respondent more flexibility than the conventional structured interviews, questionnaires or surveys. It further gives respondents the freedom to depart from the schedule at will and explain their views in the manner best suited to them, while maintaining the skeleton of themed questions. To seek amplification and clarification of the answers provided by the respondents, the questionnaire provided open-ended questions in each section of the questionnaire to capture any additional issues or comments that might be relevant to this research.

The combined use of both quantitative and qualitative methods in the study yielded a comprehensive and balanced perspective in presenting the problems faced by the actors involved that would contribute to a better understanding of the frequently controversial issue of workplace privacy. Five public sector organizations in telecommunication, educational, public administration and research sectors were selected in this survey since they have been using IT for quite a long time. Unsurprisingly, all of the organizations responding to the survey provide their employees with access to Internet, email or other communication technology at work.

Three hundred and twenty five survey questionnaires were equally distributed among the public sector organizations in between June 2003 and May 2004. Two hundred and fifty positive replies received giving a response rate of 77%. Out of them, hundred were at managerial level, others were at clerical level. 40 % of the respondents in the clerical staff worked as data entry operators, others were supporting staff. There were 80% human resource managers and 20% senior level executives in the managerial category. Considering educational level of the sample, 15% have completed BSc and 10% have completed MSc, 40% have diplomas and certificates in the field of IT. Out of the sample, hundred and fifty eight were male and ninety two were female.

4. Results

4.1 E-mail and Internet at the workplace

The respondents were asked: *What percentage of employees at your department has internet access at work?*

Percentage of organizations	Percentage of staff having access to the Internet
5%	<10%
50%	11%-49%
15%	50%-99%
30%	100

Table 1 Internet access at the public work place.

Interestingly, The Sri Lankan government has pledged to make all government services available electronically by 2007, establishing the office of Information Communication Technology Agency under the electronic government project (fully computerization of all government sector organizations) to deliver all government services online to every citizen (ICTA, 2004; E-Sri Lanka, 2004).

4.2 E-mail and Internet Policy

To identify any public sector workplace electronic policy, the respondents were asked: *Does your department have an established and codified e-mail policy?*

Considering both the vital role e-mail plays in the transfer of information around inter governmental departments and global, and the fact that all the respondents have utilized the medium for a considerable period of time, this is a very surprise statistic. It provides the first indication that government organizations are long way behind best practice controlling e-mail use by their employees, and are slow to realize the danger inherent in a lack of such controls.

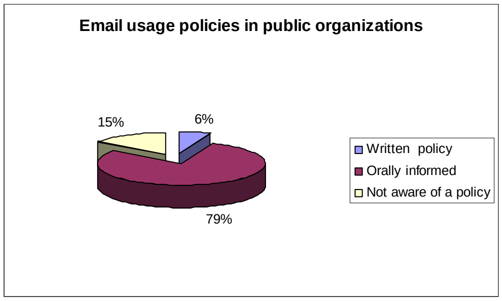


Figure 4.1

4.3 Communication of Policy

The respondents were asked: *Have your Department employee been informed of the fact that their e-mails can be monitored in some way?*

5% of the respondents replied with the answer "Yes" and interestingly 70% of respondents indicated with the answer "No".

Among the sample who said 'yes'. Were employees views sought before implementation?

10% of the respondents of this category raised their voice and said that their employees know that they can be monitored, there is no such rule existence presently but may be a possibility in future, there is a proposal to appoint a committee on departmental level to investigate this matter.

This statistical data further illustrated in the depth interviews; Respondents in the current study were asked to what extent employees were informed of the organisation's monitoring activities. There was an almost unanimous response in favour of employees being informed of any significant moves towards or away from the monitoring of the business e-mails. Only one respondent argued against such decisions being disclosed to staff. This represents a comprehensive rejection of "covert e-mail surveillance", and points to why effective communication on this matter between executive management and employees is so important.

4.4 E-mail and Internet surveillance

Managers often seemed interested in the potential of e-mail and internet monitoring and surveillances as a management tool, believing that it would improve their ability to track events and thus protect the organisation.

The respondents were asked: *Does your department scan or monitor employee e-mails in any way (including checking e-mails) for potential viruses?*

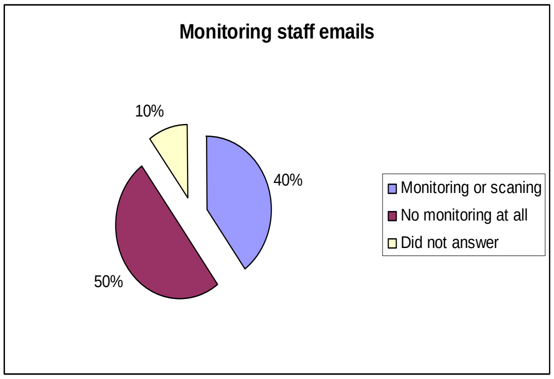


Figure 4.2

In particular in the in-depth interviews on Email and Internet surveillance respondents further elaborated their views in the following manner. The view that monitoring is essential to the conducting of work practices in these offices was repeatedly emphasized in the managerial interviews. They believed that email and Internet monitoring was the only method that could really assess whether employees were "doing their job properly".

4.5 Surfing non-work related sites

4.5.1 Appropriateness of surfing non-work related sites from managers' point of view

The respondents were asked: In your opinion, is it appropriate for employees to surf non-work-related sites? If so, what is the maximum amount of time that should be permissible?

Significantly, 90% of respondents indicated that it is appropriate to surf non-work-related sites at the working hours and 10% of respondents indicated it is inappropriate to surf non-work-related sites at the workplace. With regard to maximum time that should be permissible at work place respondents indicated different time slots.

4.5.2 Preferred Time period for web surfing

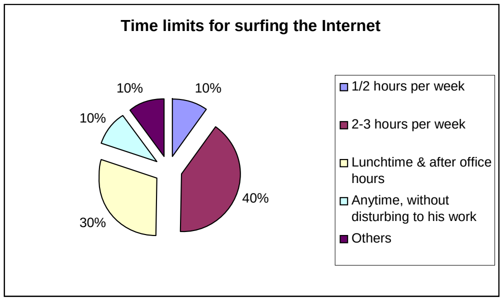


Figure 4.3

This reveals around 80% of respondents would like to use the Internet facility during non working hours.

4.5.3 Actual hours with the Internet

The respondents were asked: Approximately how many hours do you spend accessing non-work-related sites during each week?

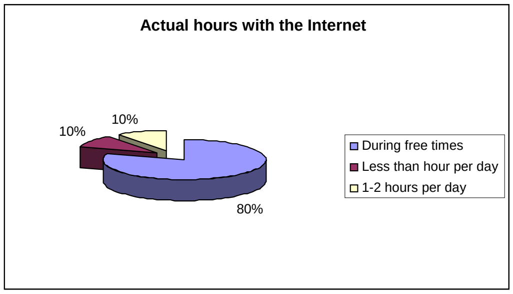


Figure 4.4

Here, during the free times is meant for after office hours and during lunch breaks. This further shows the employees use IT facility during working hours too.

4.5.4 Allowable time period for non work related activities

With regard to maximum time that should be permissible at work place respondents indicated different time slots.

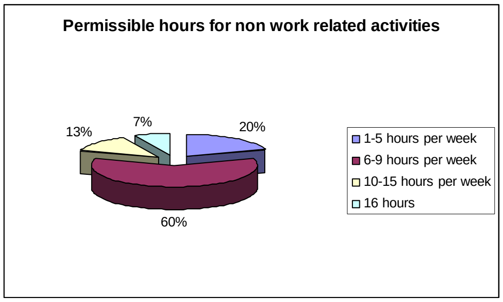


Figure 4.5

The figure 5 and 4 together with figure 3 show that employers would like to provide opportunities to use the IT resources for non work related activities while actual usage is far behind the permissible period. Quite interestingly, employees are also willing to use the facility without disturbing their day to day work.

4.6 Monitoring of the contents of e-mail

The respondents were asked: Which of the following most accurately summarises your attitudes towards potentially having your e-mails examined for content?

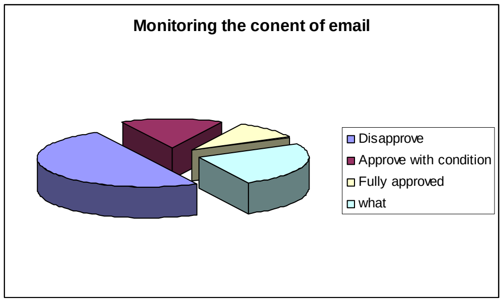


Figure 4.6

The most common sentiment expressed was a totally reject of the prospect of having corporate e-mails monitored, divulged by 50 percent of respondents. Only 15 percent of respondents' replies approved with condition, while only 10 percent of those surveyed fully approved of e-mail surveillance implementation. Combining the above results in general "approval" and "disapproval" camps shows that the disapprovers outnumber the approvers by 50 percent to 10 percent.

4.7 Personal e-mails

The respondents were asked: on average, how often do you send non-work related e-mails per working day?

80% of respondents indicated that they send 2-3 e-mails per day, 8% indicated that they send 4 to 5 e-mails and 7% indicated they cannot specify the non-work related e-mails they send.

4.8 Privacy at the workplace

The respondents were asked: *By monitoring employer infringing your right to privacy in the workplace?*

80% of the respondents replied with "yes" and 20% replied with "no".

5. Analysis of findings

This research found the remedy for secret monitoring that is by disclosing the way of monitoring to fellow employees. This survey suggests that HR policies on staff use of the Internet and e-mail are now widespread, but that better communication of those policies to employees might be needed to limit misuse as far as possible. This survey found that the privacy awareness and attitudes of individuals towards personal privacy issues are largely consistent with the findings of previous surveys.

This study tested the employees and managers attitudes, opinions and beliefs on e-mail and Internet monitoring in the public sector workplace. Results confirm that employee opinions are an influential factor developing a workplace privacy policy. They also confirmed the need to consult with employees and employees before introducing an e-mail and Internet monitoring policy.

6. Practical implications and future

6.1 Practical Implications

The research has a number of important practical implications for public sector organizations. Firstly, survey findings confirm the importance of employee personal life at workplace. If managers consider workplace privacy to be useful, the results show that they are more likely to use it. Moreover, in the subsequent interviews with managers they highly appreciated this in their privacy policy.

Perhaps the most significant implication of this research is the personal space at workplace agreed by both employees and managers at public sector. The assumption that all (or almost all) employees and managers are derive by the same set of values and goals in public sector. Policy makers need to recognize this diversity in their policy setting up process on e-mail and Internet monitoring.

Management should discuss employees' needs and consider their aspirations on ongoing basis so that realistic solution may build up as the findings suggest in this study.

The results of this study demonstrate that employee's acceptance of e-mail and Internet monitoring is significantly greater where there exist a monitoring policy that has been communicated to employees. This is not intended to suggest, however that all employers should actively engage in e-mail monitoring practices, Rather the key point for managers is that they need to establish clear employee expectations with regard to e-mail and Internet surfing.

The findings show that employers must acknowledge that staffs have to organize elements of their private lives in the workplace, but employees in turn must refrain from abusing this privilege. Evidence from the manager's questionnaires and interviews indicates that this recognition on behalf of managers materializing. The results of the study suggest that rules and policies alone will ensure neither the effectiveness of monitoring nor the minimization of misuse of the facility. For a larger portion of employees, the negative impacts of monitoring can be mitigated by managerial attention to good principles and educate employees with their participation of implementation.

The implications of this study are encouraging for organizations that intend to adopt electronic monitoring. These results suggest that deploying these techniques is not a guaranteed recipe for unhappy workers or unfairness in workplace. Attention to organizational justice factors can avoid these negative outcomes.

In setting up an electronic monitoring policy, attention should be paid to the impact that several factors have on the way employees form opinions about the fairness of policy. These factors include the consistency of the system across individuals and time, the potential bias of the system, the accuracy of information obtained, the flexibility of the system to correct mistakes, the compatibility of the system with employee moral and ethical values and the voice employees have in setting up the system.

Perhaps the foremost way for managers to tie all of these fairness antecedents together is by giving employees a role in setting up the electronic monitoring, or at the very least by communicating the purpose of the system explicitly to employees who will be monitored (Grant , 1992). Many misunderstandings about the purpose of the electronic monitoring and how it will aid the organization can be avoided if employee voice and communication are made centrepiece of any new computer-monitoring effort. In the present study, negative reactions to the electronic monitoring by some of the public sector employees may have stemmed from the failure of management to clearly communicate the role of the system, or lack thereof, in the promotion evaluation process.

The findings of this study contribute to a better understanding of the construct structure of the model variables and of the antecedent factors that can facilitate internet usage. As the internet becomes an almost indispensable part of work, it is important for all managers, especially information technology managers, to be cognisant of both individual and organizational factors associated with internet usage. It appears that internet usage could have positive results in terms of enhanced job characteristics, overall job satisfaction, and productivity, though perhaps at the cost of increasing inefficiency. This specially challenges information technology managers to formulate policies and procedures that control, but do not discourage, internet work usage by all employees. The ultimate goal is legitimate usage of the internet while at work, but organizations may need to encourage any usage of the internet for a time if its full potential is to be used for competitive advantage.

Furthermore, organizational internet policies will need to outline clearly the disciplinary consequences that employees will have to face if they flout the guidelines stated explicitly in the policy. As with all other types of organizational policies, disciplinary actions must be meted out accordingly; otherwise, the purpose of having the policy in the first place would have been defeated.

Employees report increased levels of stress when computer monitoring conducted in secret manner without prior notification to them. Therefore, managers should attempt to make "objective" measures of performance as fair as possible. The results show that the usage of e-mails significantly influenced by organisational commitment; higher organisational commitment implies that employees use e-mail more for work-related rather than personal communication at work. This means that it could be more productive for managers (who want their employees to use e-mail only for work related purposes) to try to create such an environment in which employees may feel committed than try to enforce their usage of e-mail with stringent policies. Therefore, managers must pay careful attention to this finding when they try to design an e-mail policy.

The results from this study also imply that management can motivate productive use of the Internet through attitudinal changes and workplace behavioural norms. Since the attitude toward Internet usage is a major predictor of personal Internet usage behaviours, management can reduce the negative effects of personal web usages behaviour and through changing employee attitudes by clearly and openly communicating to them what management views as proper organisational Internet usage.

In short, from a practical perspective this paper believes that the evidence of employee resistance to the privacy policy described in this paper underscores the point that the introduction of monitoring and/or surveillance into an organization most "profitably" occurs within the context of a negotiatory process that brings management and employees to the same table. Without recognition of employee opinion and attitudes the likelihood of effective, beneficial use of organizational monitoring and/or surveillance seems low as the study evidenced.

7. Recommendations

We would like to make the following recommendations based on our study to have a better privacy practice in the public sector.

This survey shows that more than 50% of the employees do not have adequate Internet access at their workplaces (Figure 4.1). Therefore, ICTA (Information and Communication Technology Agency) should expedite it's activities to reach every citizen with ICT facilities by 2007. On the other hand, it is interesting to know that 30% of the respondents have indicated that their staff has unlimited access to the Internet. Further researches should be conducted to identify the usage of technology in the public sector. It would be interesting to know the impact on the official works in the sector.

It is surprising to notice that around 90% of organizations do not have a written policy on email and Internet usage (Figure 4.2). The Sri Lankan government has encouraged using email as a medium of communication but it has not realized the importance of having a written policy on email usage. One alternate solution is to have a common email usage policy for government agencies but we would like to suggest publishing a template which can be used as the basis for implementing email policies in the government agencies. This would facilitate to eliminate inconsistencies among the policies of the government agencies to some extent while giving opportunities to incorporate their own requirements.

We do not oppose monitoring email and internet activities of employees but they should have been informed before monitoring takes place. Monitoring email and the Internet activities is necessary in some cases such as preventing virus attacks and misuse of the system, protecting intangible assets of the company, matters relating to national security etc.. It is also important to respect individuals' right to privacy. Some procedures are given to achieve both the objectives.

There are many definitions for privacy and it has different shapes. But the widely accepted one is "the right to be let alone" (Warren & Brandeis 1890). Even though, the privacy is not recognized constitutionally in some countries, its significant impact on communities cannot be ignored. Some countries have recognized privacy in different ways such as implicit provisions, traditions, customs etc. For example, India has not explicitly recognized privacy in its constitution but a decision given by the Indian Supreme Court in 1964 (*Kharak Singh vs State of UP*) has first recognized that the Indians have the right to privacy, according to the Article 21 of the Constitution, which states, "No person shall be deprived of his life or personal liberty except according to procedure established by law."

The European Union has formulated directives (95/46/EC) as a basis for implementing legislations on data protection for its member countries. The Data Protection Act of the United Kingdom has categorically stated, an individual is entitled to be informed about the description of the data to be monitored, purpose and recipient of those prior to processing of an individual's data (section 7of 95/46/EC). It also recognizes and provides additional safeguards in processing sensitive personal data. (Section 2 of 95/46/EC). There are similar provisions in other legislation in other European countries.

The employees must understand their privacy rights and the employers must respect privacy of their employees. Therefore, before monitoring email and web activities, the employees must be given a clear notice in advance. Then the employees can decide whether to use the Internet for their personal work or not.

Figure 4.4 together with section 4.7 show that employees are using email for non work related activities. This further reveals they are using the ICT to strength their relationships with family members and friends but it is a surprise to note that 20% of respondents do not mind their emails to be read. This may be due to being not concerned about their privacy or their lack of knowledge about the email mechanism. A further research should be conducted to verify this fact. It is interesting to know whether they like to reveal the content of their postal mails or not.

7.1 Have a clear line between official activities and personal activities.

There must be a separate place within the official premises for personal web browsing and email facilities. If resources are not enough to have a separate place, machines used for day to day activities can be used for the purpose in prescribed periods. The Important point is, it should be clear to the employees that they are allowed to use these machines for their personal activities during the given period without being monitored. This gives a win-win situation for both the employees and mangers. If it is possible to have a separate place for personal use, the system for official work is more secure against malicious attacks.

7.2 Privacy policy for Government web sites:

All the Sri Lankan government web sites were searched with the keyword “privacy policy” using the Google search engine. No single government agency has ever published its privacy policy. Two entries were found with the search phrase, describing importance of having a privacy policy. At the moment it is not a significant issue since most of the agencies just provide information without facilitating actual participation. Visiting these sites reveals very limited information such as IP address, client’s operating system, browser’s information etc. which could not be used to identify the user without additional information.

It is important to look at the privacy issues in the initial stage; therefore all government agencies must publish their privacy policies. A privacy policy describes what type of personal information is collected, why those data is collected, how the collected data is to be used, whether these data is shared with others, how the collected data is protected, how does an individual access the collected data related to him, how to contact the collecting authority in the case of dispute etc. (Helena Lindskog, Stefan Lindskog, 2003). It is not sufficient to provide a privacy policy only in English. The policy must be published in machine readable format, in compliance with W3C standard. The machine readable policy can be read by software agents to match with user’s privacy preferences. This makes it easy for the users to decide whether to reveal his/her personal information or not.

7.3 Privacy policy for internal workers and third parties:

It requires further studies to identify the content of email messages in public organizations. Sometime, trade secrets, personal and sensitive personal information may have been transferred without any security. These messages are vulnerable to eavesdroppers and attackers at different levels such as network handlers at the organization and the internet service providers, network maintenance officers etc.. Therefore, it is necessary to have privacy and security agreement with those parties who handle communication networks.

New employees must adhere to the privacy policy of the organization and this clause must be expressly stated in their employment agreement. The existing employees must sign a new agreement with the company. Not only that, this clause must be stated in the contractual agreements with the third parties, especially who have access to the department’s network.

References

International Labour Organisation. (1993) *Workers' Privacy Part II: Monitoring and Surveillance in the Workplace. Conditions of Work Digest* (ILO: Geneva).

Electronic Privacy Information Center (EPIC), '*Privacy and Human Rights'* (2004). <http://www.epic.org>

Marx, G. T. and Sherizen, S. (1986) "Monitoring on the job: How to protect privacy as well as property" *Technology Review,* Vol. 28, p. 63.

Griffith, T. L. (1993) "Monitoring and performance: A comparison of computer and supervisor monitoring" *Journal of Applied Social Psychology,* Vol. 23, p. 549.

Websense, (2002) *‘Cyber-Addiction in the Workplace’* <http://www.websense.com/company/news/research/webatwork2002.pdf>

The e-policy Institute, ‘*Workplace E-mail and Instant messaging Survey* ‘(2004). <http://www.epolicyinstitute.com/survey>

Privacy Foundation, '*Workplace Surveillance Project'* (2001). <http://www.privacyfoundation.org/workplace/technology/tech_show.asp?id=69&action=o> at 18 June 2001.

Society of Human Resource Management, '*Technology and Privacy use'* (2002). <www.shrm.org/trends/visions/default.asp?page=0300c.asp>

American Management Association (AMA) and E-Policy Institute, '*Workplace E-mail and Instant Messaging Survey Summary*' (2004). <http://www.epolicyinstitute.com/survey/sureyp4.pdf>

Nebeker, D. M. and Tatum, B. C. (1993) "The effects of computer monitoring, standards and rewards on work performance, job satisfaction and stress" *Journal of Applied Social Psychology,* Vol. 23, p. 508.

Geist, M., '*Computer and E-mail Workplace Surveillance in Canada: The Shift from reasonable Expectation of Privacy to Reasonable Surveillance'* (2002). <http://www.cjc.com.gc.ca/cmslib/general/Geist_report.en.pdf> at March 2002.

DeTienne, K. B. (1993) "Developing and employee-centered electronic monitoring system" *Journal of Systems Management,* Vol. 44, p. 12.

Sipior, J. C. and Ward, B. T. (1995) "The ethical and legal quandary of email privacy" *Communications of the Association for Computing Machinery,* Vol. 38, p. 8.

Orthmann, R. (1998) "Workplace computer monitoring" *Employment Testing - Law and Policy Reporter,* Vol. 12, p. 182.

Kidwell, R. E. and Bennett, N. (1994) "Employee Reactions to Electronic Control Systems" *Group and Organization Management,* Vol. 19, p. 203.

Chalykoff, J. and Kochan, T. A. (1989) "Computer-aided monitoring: Its influence on employee job satisfaction and turnover" *Personnel Psychology,* Vol. 42, p. 807.

Office of Technology Assessment [OTA]. (1987) *The electronic supervisor: New technology, new tensions* (U.S Government Printing Office: Washington DC).

Working Women Education Fund 9to5. (1990) *Stories of mistrust and manipulation: The electronic monitoring of the American workforce* (Author: Cleveland, OH)

Stone, E. F., Gardner, D. G., Gueutal, H. G. and McClure, S. (1983) "A Field Experiment Comparing Information-Privacy Values, Beliefs, and Attitudes Across Several Types of Organizations" *Journal of Applied Psychology,* Vol. 68, p. 459.

Bylinsky, G. (1991) "How companies spy on employees" *Fortune,* Vol. 124, p. 131.

Culnan, M. J. (1993) "How did they get my name? An exploratory investigation of consumer attitudes toward secondary information use" *MIS Quarterly,* Vol. 17, p. 341.

Smith, H. J. (1993) "Privacy policies and practices: Inside the organizational maze" *Communications of the ACM,* Vol. 36, p. 105.

Vest, J. M., Vest, M. J., Perry, S. J. and O'Brien, F. (1995) "Factors influencing managerial disclosure of AIDS health information to co-workers" *Journal of Applied Social Psychology,* Vol. 25, p. 1043.

Sipior, J. C. and Ward, B. T. (1995) "The ethical and legal quandary of email privacy" *Communications of the Association for Computing Machinery,* Vol. 38, p. 8.

E-Sri Lanka*, 'E Sri Lanka Road Map, An Action-Oriented Plan'* (2003). http://www.e-srilanka.lk

Article 29 - EU Data Protection Working Party, *'Working document on the surveillance of electronic communications in the workplace'* (2002). <http://europa.eu.int/comm/privacy> at 29th May 2002.

Information and Communication Technology Agency of Sri Lanka (ICTA), *'Draft ICT Policy for Government'* (2005) < www.ICTA.lk>

ICTA (2005) *Draft ICT policy for Government 2005* (ICTA: Colombo)

Lim, V. K. G., Teo, T. S. H. and Loo, G. L. (2002) "How do I Loaf here? Let me count the ways" *Communications of the ACM,* Vol. 45, p. 66.

Stanton, J. M. and Julian, A. L. (2002) "The impact of electronic monitoring on quality and quantity of performance" *Computers in Human Behavior,* Vol. 18, p. 85.

Grant, H. (1992) "Monitoring service workers via computer: The effect on employees, productivity, and service" *National Productivity review,* Vol. 8, p. 101.

Kharak Singh vs State of UP, 1 SCR 332 (1964)

Helena Lindskog, Stefan Lindskog (2003) "Web Site Privacy with P3P)

SIM: Secure Instant Messaging for Yahoo

N. Abeywardana[1] and K. De Zoysa[2]

University of Colombo School of Computing,
35, Reid Avenue, Colombo 7,
Sri Lanka.
Phone: +94-71-4823857, +94–11-2581245/8, Fax: +94-11-2587239
E-mail: [1]nirodha23@yahoo.com, [2]kasun@cmb.ac.lk

Abstract

Instant Messaging has become an important means of communication among people around the globe by providing an alternative to telephone and email conversations. The number of users using Instant Messenger (IM) products has been increasing over recent years.

This document describes a Secure Instant Messaging (SIM) client for Yahoo (SIM client for Yahoo) to provide with secure conversation. Most of the current IM systems are inadequately secured and some enterprises are exposed to serious security and economic breaches. SIM supports secure dynamic group establishment, distributions of group keys among group members using a group key server. According to my knowledge, SIM is the first such IM client which supports secure communication.

Keywords: Encryption, Digital Signature, Group Key Agreement Protocols, Secure Instant Messaging

1. Introduction

Instant messaging is the ability to exchange messages in real time with other people over the Internet [1]. The IM usage is rapidly increasing all over the world. With over 130 million IM users today, this is a technology that can not be ignored. International Data Corporation (IDC) [2] indicates that IM users will exceed 300 million users by 2005. There are many free public domain instant messaging services. The most popular are AOL Instant Messenger (AIM), ICQ, MSN Messenger (Microsoft Messenger in XP), and Yahoo! Instant Messenger (YIM).

Yahoo! Messenger is a very popular instant messaging client provided by Yahoo Corporation. YIM is provided free of charge and can be downloaded and used with a generic "Yahoo! ID" which also allows access to other Yahoo! services. Yahoo is very popular among internet users, because of its reliability and user friendliness.

In general, IM is being used by individuals for simple leisure communication and it has evolved to be the most critical communication application within and outside a corporate organization. Millions of IM users believe that IM provides the security services stated below.

- Confidentiality – message content is exposed to the intended person only.
- Access control – Limited access to information and IM system resources.
- Availability – IM servers and directories meet the service level agreements.
- Authentication – Verify the identity of the organization or user.
- Integrity – Confirm that the data received is same as the data sent.

However, any of the IMs, available today do not provide above mentioned security features [3] [4]. Therefore, in this research we have implemented an IM client that provide advanced security (end to end security instead of channel level security) features.

As earlier discussed, IM is now used by various business communities as a viable communication tool to share sensitive information. However, businesses require more security than they get from the consumer versions today. For example, there's chance sensitive information will be transmitted in plain text over the public Internet. Thus, message contents can be intercepted by attackers. Security features in channel level (e.g. using SSL connections) incorporate IM systems are inadequate to address security threats [4]. Therefore, it is clear that the basic protocols currently used in the public IM systems are open to many security threats. Because of above reasons we should have a secure instant messenger client to exchange sensitive information.

We mainly focused on implementing the secure chatting system among group of people which termed as secure conference chatting. The problem comes whenever disseminating keys among the group members. This problem has been solved using modern cryptographic techniques.

2. Review on Modern Cryptography

In this section, we briefly discuss the modern cryptographic techniques used in our work. The word ‘cryptography’ is derived from the Greek language and it means ‘hidden writing’ or ‘secret writing’ [5]. Historically it has been the study of algorithms that were used to secure messages during transmission or storage.

Cryptography can also be supported to ensure security requirement such as integrity, authentication, authorization, data confidentiality and non-repudiation [6].

Confidentiality: Protection from disclosure to unauthorized persons

Integrity: Maintaining data consistency

Authentication: Assurance of identity of person or originator of data

Non-repudiation: Inability to deny the communications already made by the originator.

2.1 Encryption and decryption

Data that can be read and understood without any special measures is called plaintext or cleartext. The method of disguising plaintext in such a way as to hide its contents is called encryption. Encrypting plaintext results in unreadable gibberish called ciphertext. We use encryption to ensure that information is hidden from anyone for whom it is not intended, even those who can see the encrypted data. The process of reverting ciphertext to its original plaintext is called decryption.

2.2 Symmetric Key Cryptography

In some encryption algorithms, the encryption key and the decryption key is the same, or the decryption key can be calculated from the encryption key. These algorithms are known as secret key algorithms (or private key algorithms or symmetric key algorithms). The encryption key must be kept secret and the sender and receiver must coordinate the use of their keys. Following are some popular secret key algorithms:

AES: This is the current data encryption standard algorithm.

DES: This is the former data encryption standard algorithm. There are known ways to attack this encryption.

3DES: This is also known as Triple-DES or multiple-DES. This algorithm uses multiple DES keys to perform three rounds of DES encryption or decryption. The added complexity increases the time required to break the encryption as well as the time required to encrypt or decrypt data.

2.3 Public Key Cryptography

The concept is simple and elegant, yet it has had far-reaching effects on the science of cryptography and its applications. Public key cryptography is based on the notion that encryption keys come in related pairs, private and public. The private key remains concealed by the key owner, while the public key is freely disseminated. The premise is that it is computationally infeasible to compute the private key by knowing the public key data encrypted using the public key can only be decrypted using the associated private key. Figure 1 represents how public key cryptography works.

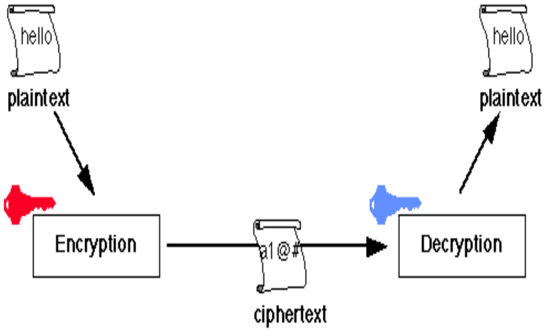


Figure 1: Asymmetric key cryptography

Following is the popular public key algorithm.

RSA: The most famous public key algorithm developed in 1977. It can be used for both encryption and digital signatures which will be discussed later.

2.4 Message Digests

Message digests are used to secure data integrity. In other words, to detect whether data has been modified or replaced. A message digest is a special kind of function referred to as a one-way (hash) function. A one-way function is easy to calculate, but difficult to reverse. Message digests take messages or data as inputs and compute values referred to as hash values that are used as fingerprints to the messages. SHA1 and MD5 are the popular message digest algorithms.

2.5 Hybrid cryptography

The advent of public key cryptography did not signal the end of secret key cryptography. Rather, one cryptographic method complements the other. Public and secret key cryptography together form most cryptographic protocols in use today. These are called hybrid cryptographic systems. A public key system is used for the distribution of a secret key, which can be a long-term key or specific to a particular communications session. Thereafter, the securely distributed secret key is used to encrypt and decrypt a communications channel between two ends of a security protocol. The performance of secret key cryptography over that of public key, and the appeal of key distribution inherent to public key cryptography, are the main reasons for the wide adoption of these hybrid systems.

2.6 Digital Signatures

Digital signatures are mainly used to prove that a message sent is created by a particular individual or an organization. If the receiver can verify the digitally signed message from the sender, then the receiver can make sure that the contents of the message are correct and authentic. Digital signatures have the following properties similar to real world signatures.

- Unforgeability – because the signer uses his private key to sign, only he can sign with that key
- Verifiability – because the signer's public key is openly available, anyone with access to the message and signature can verify that the message was signed by the signer and that neither the message nor the signature has been altered.
- Single use – A signature is unique to a particular message.
- Non repudiation – After a signer has signed a message and the message and signature have been sent to others, the signer cannot claim that he did not sign the message.
- Sealing – A signed message is digitally sealed. It can not be altered without invalidating the signature.

2.7 Digital Certificates

Digital Certificates are the electronic counterparts to driver licenses, passports and membership cards. You can present a Digital Certificate electronically to prove your identity or your right to access information or services online. A Digital Certificate makes it possible to verify someone's claim that he has the right to use a given key, helping to prevent people from using phony keys to impersonate other users. Used in conjunction with encryption, Digital Certificates provide a more complete security solution, assuring the identity of all parties involved in a transaction [7].

3. Review on Group Key Agreement Protocols

Several group key management approaches have been proposed in the last decade. These approaches generally fall into three categories:

1) Centralized
2) Distributed
3) Contributory

Centralized group key management is conceptually simple as it involves a single entity (or a small set of entities) that generates and distributes keys to group members. We claim that centralized group key management is not appropriate for peer group communication since the central key server must be, at the same time, continuously available.

Distributed group key management is more suitable to peer group communication, especially over unreliable networks. It involves dynamically selecting a group member that acts as a key server. Although robust, this approach has a notable drawback in that it requires the key server to maintain long-term pair wise secure channels with all current group members in order to distribute group keys.

In contrast, contributory group key agreement requires each group member to contribute an equal share to the common group key (computed as a function of all members' contributions). This approach avoids the problems with the single points of trust and failure.

As can be expected, the cost of group key management protocols is largely determined by two dominating factors: communication and computation. Typically, efficiency in one comes at the expense of the other. Protocols that distribute computation usually require more communication rounds or messages, whereas, protocols minimizing communication require more computational effort.

3.1 Group Key Agreement Protocols

Group key agreement is a fundamental building block for secure peer group communication systems. Several group key agreement protocols were proposed in the last decade, all of them assuming the existence of an underlying group communication infrastructure [8] [9] [10].

3.1.1 GDH Protocol: Cliques GDH IKA.3 is a contributory key agreement protocol which is essentially an extension of the two-party Diffie-Hellman protocol. The basic idea is that the shared key is never transmitted over the network. Instead, a list of partial keys (that can be used by individual members to compute the group secret) is sent. One member of the group – group controller– is charged with the task of building and distributing this list. The controller is not fixed and has no special security privileges [11].

The protocol works as follows. When a merge event occurs, the current group controller generates a new key token by refreshing its contribution to the group key and then passes the token to one of the new members. When the new member receives this token, it adds its own contribution and passes the token to the next new member. Eventually, the token reaches the last new member. This new member, who is slated to become the new group controller, broadcasts the token to the group without adding its contribution.

Upon receiving the broadcast token, each group member (old and new) factors out its contribution and unicasts the result (called a factor-out token) to the new group controller. The new group controller collects all the factor-out tokens, adds its own contribution to each of them, builds the list of partial keys and broadcasts it to the group. Every member can then obtain the group key by factoring in its contribution.

3.1.2 CKD Protocol: CKD protocol is a simple centralized group key distribution scheme. The group key is not contributory, but it is always generated by one member, namely, the current group controller. The group controller establishes a separate secure channel with each current group member by using authenticated two-party Diffie-Hellman key exchange.

Each such key stays unchanged as long as both parties (controller and the member) remain in the group. The controller is always the oldest member of the group. Whenever a group membership is changed, the group controller generates a new secret key and distributes it to all the group members using the long-term pair-wise key. In case of a merge, the controller, in addition, establishes a secure channel with each new member [11].

3.1.3 GSAKMP: The Group Secure Association Key Management Protocol (GSAKMP) [12] is a general protocol for creating and managing cryptographic groups on a network. A cryptographic group is a logical association of users or hosts which shares cryptographic keying material. While GSAKMP provides mechanisms for cryptographic group creation, other protocols may be used in conjunction with GSAKMP to allow various applications to create groups according to their application-specific requirements.

3.1.4 GSAKMP Light: The specification of GSAKMP-Light (GL) profile is a way to shorten the number of messages exchanged during secure group establishment. The GSAKMP protocol assumed that group members joining a secure group had no information about the specific security mechanisms used by the group (for example, the key length, encryption protocol, etc). GSAKMP-Light provides a profile for the case where group members have been previously notified of these security mechanisms, used for joining a group, during the group announcement or invitation. However, this profile does not sacrifice any of the security properties of the full protocol.

The GSAKMP protocol includes mechanisms for group policy dissemination, group key dissemination, and group rekeys operation. The GL profile shortens the policy and key dissemination steps, however does not limit or decrease the security of either of these operations. GL performs following message sequence (see figure 2) when establishing a group member.

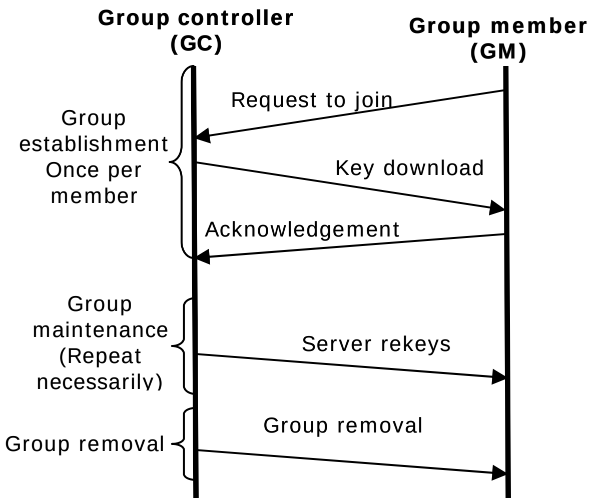


Figure 2: Message Passing Structure (GSAKMP)

Sequence of events

The sequence of events for GL is straightforward [13]. The sequence is:

1. Security suite definition is transmitted outside the protocol.
2. Light Request to Join (L-RTJ)
3. Light Key Download (L-KD)
4. Light Acknowledgement (L-ACK)
5. Server ReKey operation
6. Group removal

Member will initiate the following series of three messages for group establishment.

- Light Request to join initiates the GL group establishment portion of the protocol. L-RTJ contains a key creation field for use in group establishment.
- Light Key Download contains a key creation field and encrypted Policy Token and Key Download payloads.
- The Light Acknowledgement message completes the authentication of the GCKS for the member.

The GL profile is provably secured, supports distributed architectures, allows multiple data sources within a single cryptographic group, and provides group management mechanisms. GL is used as the base protocol in SIM. The reasons for the use of GL as the base protocol are:

- PKI infrastructure can be easily applied to the system.
- Members are previously understood about the policy token.
- Since GL defines complete message sequence of events it can be easily applied to the client / server architecture.
- Easy implementation.

4. System Architecture

4.1 System Overview

As mentioned, the “Secure Instant Messenger (SIM)” is a simple secure chatting system implemented for Yahoo IM server. It enables the secure conversation among members. Chatting group can be established using Yahoo IM server and any of the group members can start the secure chat whenever they want a secure conversation. Other members should be responsible for downloading the group key from the GL Key Server by exchanging some basic massages between them and the Key Server. The figure 3 shows the overall system architecture.

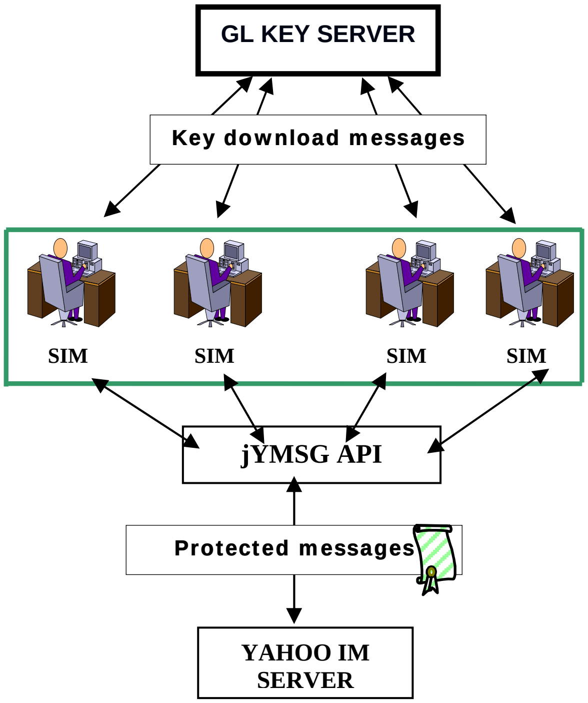


Figure 3: System Overview

In order to implement SIM, jYMSG API was used. jYMSG is a properly documented API which is written in purely Java language[14]. This API supports most of the Yahoo IM features.

As shown in the figure 3, key server was implemented based on GL protocol. Key server is responsible for following events,

- Initiating secure groups
- Generating secret keys
- Authenticate group members
- Perform rekey event

Following messages are exchanged between each client and the key server:

- Initiating a secure group message
- Fetching the group key message
- Leaving the group and rekey message

4.2 Initiation of a Secure Group

While performing an insecure communication among group of people, anybody can initiate a secure group for secure communication. The initiator of the secure conversation creates the message for “Initiate a secure group”. This message consists of message header, digital certificate information and the member information. This message is digitally signed and sent to the GL key server. The key server is responsible for verifying the signature and the content of the message. If the message is verified successfully, the generated group Id and member information are stored on the server for authenticate group members. Then the secret key is generated and stored in the key server. The group key is transmitted to the initiator as an encrypted message by using the initiator’s public key. The initiator has the corresponding private key, so he /she can decrypt the encrypted group key by using the private key.

4.3 Fetching the Group Key

After initiated the secure group, all the other members should be responsible for fetching the corresponding group key for secure chatting. For that, each group members creates a message “Fetch the group key”. Fetch group key message consists of message header, digital certificate information, the member Id information and group Id information. This message is digitally signed and sent to the key server. The key server is responsible for verifying the signature and the content of the message. If the message is verified successfully then the group Id and member Id are also verified in order to authenticate the group member. The corresponding group key is transmitted to the group member by encrypting the group key using the group member’s public key. The encrypted group key is transmitted to the group member. Since, the group member has the corresponding private key he/she can decrypt the encrypted group key.

4.4 Leaving the Group and ReKey process

Group members can leave the conference while performing the secure conference with other group members. Whenever a member left, the group key should be changed because the member who left knows the group. The rekey message triggers the new group key creation process. The rekey message consists of a message header, the member Id information and group Id information. The rekey message is digitally signed and sent to the key server. The key server is responsible for verifying the signature and the content of the message. If the message signature is verified successfully then the group Id and member Id are also verified. If it is also verified, the member Id is removed from the particular group and a new group key is generated. The newly generated group key is stored in the key server. Finally, the existing group members fetch the new group key.

In our system, whenever a member leaves a group, the rekey message is triggered automatically and the other group members automatically fetch the new group key from the key server.

5. Features of the System

The system is fully functional and can be used commercially if required. There are no complexities in the system integration. After the initial installation of the key server, all the other users can access the system through the Internet via any type of connection.

Since this system is developed using Java, it is platform independent. It can be used in any operating system including Unix, Linux, Windows etc. However, the system requires JDK 1.4, since the earlier versions of Java do not contain the necessary cryptographic methods.

All graphical user interfaces have been developed in a user friendly manner (see figure 4). Even though the system provides better security capabilities, the user can freely use the system without an additional effort. Once the user fetches the key, he/she can carry out the conversation in a usual manner. Rekey and fetch the new key operations are transparent from the user.

Each component of the system was tested and debugged for accuracy at the time when they were being developed. After completion, a system test was also carried out to identify further errors and they were corrected accordingly. However, it must be noted that further testing would improve the system. Following are some important features of the SIM system.

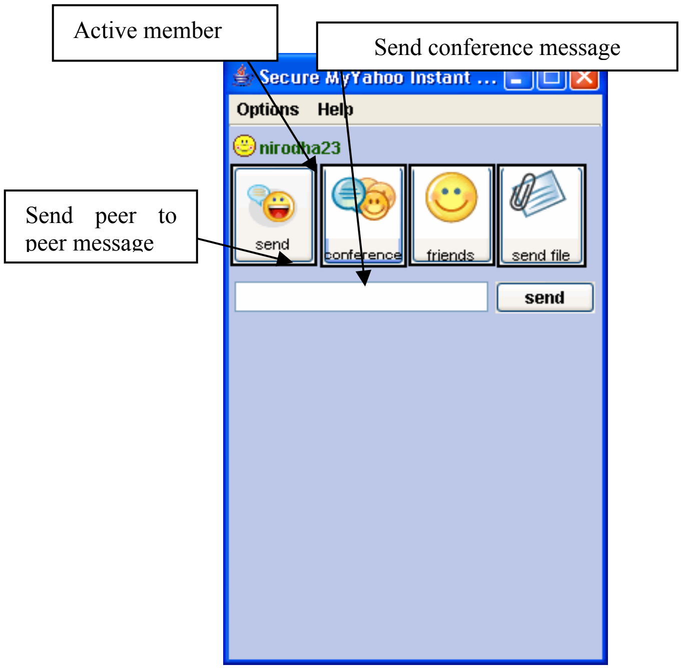


Figure 4: Main Window of SIM

Initiate a secure group: In this system, anybody of the group can initiate a secure group. The group can be initiated with any number of group members.

Dynamic group establishment: With this system, conference secure chatting can be carried out with group of online people. The events such as joining the group and leaving the group can occur dynamically, and it is visible to all the group members.

Fetch a group key: When a secure group is initiated, others can fetch the group key. When a group member leaves the group, the new group key is fetched automatically.

Leaving and rekey: When a group member leaves the group, the rekey operation is triggered automatically.

6. Conclusions

In this research, we have proved that the concept of secure group conversation can be achieved with public IM servers. At present, SIM works only with Yahoo IM server but it can be extended to support the other public IM servers. SIM supports end to end security instead of channel level protection provided by Secure Socket Layer (SSL).

The group key server, which we implemented, supports distribution of the group key among group members. However, it is still in a primary level.

As mentioned, the system can be extended to support other public chat servers such as ICQ and AOL. In addition, SIM can be extended to support smart cards which protect private key of a user [9]. We also can introduce secure file sharing facilities among clients. The group key server can also be improved by providing user-friendly administration facilities.

References

[1] "Yahoo! Messenger Features", [Online] Available at http://messenger.yahoo.com

[2] "Instant Messaging: Next best thing since e-mail?", Canada Law Book Inc., Technology Feature Article, July, 2001 [Online] Available at http://www.canadalawbook.com/headlines/headline128_arc.html

[3] "Securing Instant Messaging", white Paper, Symantec, [Online] Available at securityresponse.symantec.com/avcenter/reference/secure.instant.messaging.pdf

[4] Mohammad Mannan, P.C. Van Oorschot "Secure Public Instant Messaging: A Survey", School of Computer Science, Carleton University

[5] Scott Oak's, "Java Security", O'Reilly & Associates, 1998

[6] Bruce Schneier, "Applied Cryptography", John Wiley & Sons Inc., November, 1995

[7] "Introduction to Certificate", [Online] Available at http://www.verisign.com.au/repository/tutorial/digital/intro1.shtml

[8] Yang Richard Yangy, Simon S. Lam, "A Secure Group Key Management
Protocol Communication Lower Bound", Department of Computer Sciences, The University of Texas, Austin, July, 2000

[9] Xukai Zou, "A block-free TGDH key agreement protocol for secure group communications", Department of Computer and Information Sciences, Indiana University, USA

[10] L. R. Dondeti, S. Mukherjee, and A.
Samal, "Disec: a distributed framework for scalable secure many-to-many
communication," In Proceedings of 5th IEEE Symposium on Computers and
Communications (ISCC 2000), pp. 693–698, July 2000

[11] Y. Amir, Y. Kim,C. Nita-Rotaru and G. Tsudik "On the performance of group key agreement protocols", Proceedings of the 22nd IEEE International Conference on Distributed Computing Systems, Viena, Austria June 2002.

[12] H Harney, U Meth, A Colegrove and G Gross, GSAKMP: "Group Secure Association Group Management Protocol", Internet Internet Engineering Task Force
Internet-Draft, May 2005

[13] Sead Muftic, Gernot Schmölzer, D. Sierra, Kasun De Zoysa, "A Survivable Group Security Architecture", NSA/LUCITE Project Report, CSPRI/GWU, December 2002

[14] jYMSG API - Yahoo IM and Chat for Java, [Online] Available at http://sourceforge.net/projects/jymsg9/

Securing Mobile Agents for Survivable Systems

J. Mwakalinga and L. Yngström

Department of Computer and System Sciences,
Royal Institute of Technology/Stockholm University, Kista, Sweden
Fax: +46 8 703 9025 Tel: +46 8 16 1721

[jeffy@dsv.su.se, louise@dsv.su.se]

ABSTRACT

We have what we have today because of the decisions and actions that we made in the past. Our lives and computer technology in the future will depend on the decisions and actions we make today about them. In future, it is very likely that we will be walking with Web servers in mobile phones, PDAs or MP3 players or in whatever devices. There will be so much information from banks, insurance, government, health, nursery and schools requiring instant response that will necessitate people to carry Web servers. People will be required to make different authorization and privacy decisions which can't wait. The amount of information and actions can necessitate the need for helping hands in the form of mobile software agents, which are forms of non-human computer secretaries. These can be used in diverse business areas like auctions, contract negotiations, stock trading, and money transfer. These agents will need to carry information and perform transactions securely. How do we secure software mobile agents? In this paper we describe ways of securing mobile agents for survivable systems. We describe ways of protecting mobile agents and the information that they carry

Keywords: Software mobile agents, survivable systems, agent platforms, agent certifier and accountability.

1 Introduction

The aim of this work is to study ways of securing software agents which are used to perform different tasks during deterrence, protection, detection, response and recovery services in the survivable systems. According to [14], "*An agent is an encapsulated computer system situated in some environment and capable of reactive, pro-active, and autonomous action in that environment in order to meet its design objective*". An agent consists of three main components [3]: header, code, and a database. The header contains identity of the agent, agent attributes, signatures, travel paths, level of trust, ownership and other related information. The code section contains a system of programs performing the specific tasks of the agent. The database contains internal and the collected data while traversing in different environments. Agents are generated from an agent platform like Java Agent Development Framework (JADE) [15].

There are already software agents for different purposes. When one wants to find the best ticket through the Internet to fly to a specified location it can take a lot of time and energy. To save time and energy one can send a software mobile agent instead to do the job. Manufacturers of different products can negotiate prices, delivery of goods, terms of delivery and other services with supplies through their respective agents [4]. Other services suitable for mobile agents include network management, intrusion detection, testing security of networks and so on. The use of mobile agents reduces network traffic because they perform actions at agent servers reducing the request/reply messages in traditional client-server transactions. Mobile agents have to perform transactions and carry information securely. In this paper we describe how to secure mobile agents for survivable systems. Survivable systems are those that are required to run all the time like air traffic systems, banking systems, medical systems, radars, and different business systems. To be able to run all the time they are required to have fault-tolerance measures. The methodology for building security in survivable systems is described in [13]. The Systemic-holistic approach [2] and the Immune system [1] paradigms are used as foundations in building security in survivable systems. The Systemic-holistic paradigm is used for studying security of a system as a whole by considering the system, the environment of the system and also by considering technical and non-technical factors. The Immune system is used for protecting human bodies from different viruses and helps humans to survive in different environments. We study how living systems, particularly humans, survive in open environments and apply the features of the immune system to make systems survive. We use mobile agents in survivable systems and there are a number of security threats for mobile agents.

1.2 Security Threats for Mobile Agents

Before addressing security we need to understand the different security threats for mobile agents. The parties that are involved in transactions include agents and agent servers (platforms). An agent can attack an agent server, an agent server can attack an agent, an agent can attack another agent, an agent server can attack another agent server, and other outside attackers can cause security threats to the agents and agent servers [4].

Attacks from agents to agent servers include masquerading, denial of service and unauthorized access. Masquerading is when an agent pretends to be another agent in order to gain unauthorized access of resources or to damage the reputation of the other agent and the owner of the agent. Denial of service is when an agent disrupts the services offered by the agent server by running programs that heavily exploit system vulnerabilities of an agent server to degrade the performance of the agent server. Agent servers accommodate many mobile agents from different organizations. Some of the agents may try to access information on the agent server that they are not authorized to.

Attacks from agent to agent include masquerading, denial of service, repudiation and unauthorized access. An agent can exploit the weaknesses of another agent and steal its identity. The agent can then masquerade and perform any actions under this other's agent's identity. Agents can launch denial of service attacks against each to intentionally prevent them from finishing their tasks [4]. An agent can cheat another agent to sign a bad contract and then repudiate later from having done that. An agent can change the information or programs in another agent if they are not secured. An agent can even call another agent's methods in an attempt to change the behavior of the agent.

An agent server can attack a visiting agent in many forms: by masquerading, by denial of service, by reading agents information or by modifying agent's information and programs. An agent can be cheated into paying higher prices for items that are being sold by an agent server. Outsiders can attack agent servers and agents by masquerading, unauthorized access, denial of service and by coping agents or parts of the agent messages and replaying them. After discussing security threats we will discuss security requirements for mobile agents.

1.3 Security Requirements for Mobile Agents

According to [4] security requirements on agent frameworks include confidentiality, integrity, accountability, availability and anonymity. Confidentiality is required so that all the classified information can be kept secret at agent platforms and while being carried by the agents. Communications between agents and between agents and agent servers should also be confidential. All messages' flow should be kept secret so that the listeners should not be able to find out the number of messages nor analyze the traffic between agents and platforms. Even the location of agents should be confidential. Agents can choose to be public and in such cases they should be allowed to be. The activities of agents should also remain confidential so the audit logs of their activities must be protected.

Integrity of agents' code, state, internal data and collected data should be provided to ensure that unauthorized modification of code, state and data is not done. Agents should be able to detect when modification of their code, state and data is done. The agent server must also be provided with integrity. Access control should also be addressed so that only authorized agents should be able to access and perform the tasks on agent servers. Changes to agent servers should be made only by authorized users.

Accountability, according to [4], includes identification, authentication and audit of human users, agents and agent servers. This includes maintaining records of security related events of user/agent name, access to objects, time of access, type of event, success or failure of event. Audit logs will force users and agents to be accountable for their actions making it difficult for them to deny having performed the actions. Audit trails of agents should also be kept to help tracing activities in case of errors. Agents and agent server must authenticate each other before performing any transactions. Authentication could be strong or simple, depending on the classification of transactions. When agents are accessing public information, agent servers may not require any verification of identities of agents.

Availability of information and services to mobile agents must be ensured. The agent servers must support simultaneous access, allocate resources fairly, be able to recover from different failures and so they should have fault-tolerance measures. Agent servers should scale and be able to handle requests from many agents. When the agent servers are not able to provide this service they should notify agents about it. Denial of services attacks from malicious agents or other sources on the agent servers should addressed.

Anonymity is another security requirement for mobile agents. This requirement is challenging to meet since some transactions require participants to be strongly authenticated before performing them. The agent server should have a balance of the need for the agent to be anonymous and the need for the platform to hold the mobile agent accountable for its actions. The agent server can keep the identity of an agent and its actions secret from other agents as long as the agent is behaving in accordance to the policies and security requirements of the agent server but when the agent crosses the red line it will be revealed to other agents.

1.4 Organization of Sections

Section two covers related work. Section three describes the security architecture of survivable systems. Section four discusses agent security. Section five briefly discusses conclusions.

2 Related Work

In [3] a comprehensive security infrastructure for mobile agents is described. The infrastructure provides authentication, authorization, integrity, accountability and non-repudiation. Authenticity of agents is provided by giving identities to agents. The Agent identity has static, dynamic identity and other specific identities. Static identity comprises of agent author's ID (author's certificate), agent owners ID (owners certificate) and agents name. Dynamic identity consists of agent home ID and time of launch. To verify identities one verifies the certificates. Authorization of agents is provided through agent attributes which contain level of trust, agent task specifications, constraints of agents, agent owner credentials. Constraints on agents include expire-time, maximum size, whether an agent can create children, and others extensions. Integrity is a security service for making sure that information is not modified when on storage or on transmission, Integrity of agents and agent servers is provided through digital signatures. The signatures that every agent must have include agent authors, agent owners, trusted appraisal's, privilege authority's, sender's, agent server's signatures. Confidential information that is carried by agents is kept secret from other agents.

The lifecycle of an agent includes creation, owning, launching, traversing, hosting and returning home as shown in figure 4. The author creates an agent, signs it and attaches the digital certificate. The agent is then sent to the trust appraisal that verifies the signatures, tests the agent and then puts a level of trust on the agent. She signs the agent and puts her certificate. The agent is then sent to the owner who had requested it. The owner verifies the signatures of the author and the trust appraisal. If successful she accepts the agent. The owner assigns agent identity to distinguish it from other agents.

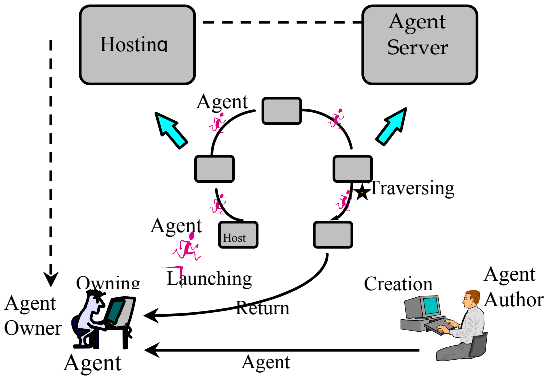


Figure 4: Mobile Agent Computing Model [3]

Before launching the agent, the owner writes specifications on the agent, gives constraints to the agent of life time, maximum size and other specified properties. The owner then assigns the home address, destination server, time of launch. She then signs the agent and seals it with the destination server's public key and sends the agent.

The destination server opens the seal, verifies the signatures of the author, trust appraisal and of the owner. If verification is successful she accepts the agent. The server hosting agent protects its information that is classified from the visiting agent. The server monitors the actions of mobile agent. Information collected from the agent server is sealed by the owner's public key and then it is signed. The hash of the state of the agent is sent to the state server. The agent can then be sent home or to another agent server and the procedure before sending an agent is the same as when the owner was sending the agent to the destination server. When the agent arrives home to its owner the signatures are verified the state of the agent is checked. If something has gone wrong the owner extracts hashes of states from the state server and traces the whole communication. This system provides most of security services in accordance to the security requirements, which were discussed in sections 1.3, confidentiality, integrity and accountability. The limitation of this system is that it does not provide anonymity and availability requirements.

3 Security Architecture for Survivable Systems

In [13] we developed a methodology for security survivable systems and the architecture for these systems is shown is figure 1. The components in the architecture of survivable system include the deterrence, protection, detection, response and recovery sub-systems. It also includes an administration component containing the agent generation library, a system manager, a database, an integrated security system, special analysis component and the system fault tolerance manager. The fault tolerance manager detects errors, assesses the damage, and confines the damage, performs error recovery measures, does fault treatment measures, locates the errors and performs measures for continued service. Every sub-system has sections: inputs, process, outputs, and fault tolerance manager. The sub-systems also have memory and feedback mechanisms for analyzing and modifying inputs when necessary.

3.1 Deterrence Sub-system

The deterrence sub-system is aimed at scaring off attackers (like how a cat scares off attackers by increasing its size and through fierce screams). When criminals plan to rob a bank in the physical world they do surveillance of the bank to determine whether it is possible to attack, take what they want and get out without being caught and without living evidence. In the digital world the attackers do more or less the same. Before would be attackers intrude a system, they do some kind of scanning to determine the operating systems and their versions, the ports that are open, the applications and versions that and on the victim's system.

Then the attackers do possibly also social engineering to understand the architecture of the system inside. There are many ways of doing this, from just asking the people working there to listening to conversations of system administrators there or secretaries working there. It is surprising how employees like to talk about their jobs during lunches and even dinners! From the results of scanning and social engineering the criminals decide whether it is possible to attack the system, and get out without being caught and without living evidence. The attackers will not attack a system if it is considered to risky. The functions of the deterrence sub-system include: adapting to the new and unknown surveillance methods; organizing training to prevent social engineering; monitoring surveillance attempts; redirecting attacks to specialized environments (like honey pot system); handling replies to scanners (returning nothing, a warning, etc); auditing; tracing scanning sources.

3.2 Protection Sub-System

Protection sub-system has measures for guiding the territory of a system and its entities. Home cats establish territories, a special place on a sofa, and put rules. Wild cats mark territories by using peculiar identifying items like natural scents. The protection sub-system provides security services: authentication, integrity, confidentiality, non-repudiation and authorization of entities and information during storage, transmission, processing, collection and display. Other features of this sub-system include: adaptability in which the system learns new protection ways by applying the latest standards; organizational, like configurations in accordance to the security policy; semi-autonomy in which the system makes some decisions without involving the management of the system, but the critical decisions must involve the system management; multi-layer protection , where protection is provided at the boundary of a system and inside the system and sub-systems; partial distribution – in some cases protection is done locally while in some cases protection is coordinated.

3.3 Detection Sub-System

This sub-system is responsible for detecting the abnormalities, storing and protecting the log of events, analyzing the events, monitoring, managing and interacting with other subsystems. Other features include multiple-layer detection, adaptability of new ways of monitoring and detecting, semi-autonomous, and dynamic coverage, sending reports to the database and the administration. The normal behaviors of outgoing and incoming messages are defined. Software agents are used to detect the abnormal behaviors of incoming and outgoing messages, as cells are used to detect foreign cells in immune systems. All the entities that belong to a system are labeled as 'self' by being given special identities and being registered in a database. Software agents monitor a system to discover the non-self entities in a system.

3.4 Response Sub-System

This sub-system is responsible for incident management. It classifies incidents into false alarms, minor and major incidents in accordance with the security policy of the system. The response and speed of reaction depends on the classification. It makes decisions on how to respond for every incident. The decisions include disconnecting the affected sub-system from others, slowing, shutting down or restarting the affected system, etc. The sub-system also sends reports to the affected users, to the database and to the administration. Other functions of this sub-system include managing patches and adaptability, tracing the attack, mitigation of the attack and so on.

3.5 Recovery Sub-system

The recovery sub-system is for bringing an attacked system back to normal. The functions of this sub-system include managing back-ups, re-installing the programs, periodic and emergency vulnerability testing, restoring a system from back-ups, collecting and protecting evidence, fixing the vulnerabilities. The agents can help to define and test business continuity plans. This process can be very expensive and takes much time if done manually. At every moment three types of the state of system and sub-systems and operations are stored: the original state; the intended state; and the actual state. When an incident occurs the system can go back to the original state and flush all the rest. This feature can be partially or wholly implemented depending on the current technology and other back-up resources.

3.5 Other Components

The integrated security system is used for certificate management, managing authorization and provides smart cards, database and information protection services. The special analysis component is used for analyzing inputs and other objects that are not understood by the sub-systems. The system fault-tolerance manager is responsible for the overall fault-tolerance of the whole system. It also controls the fault tolerance managers of the sub-systems. The system manager is responsible for managing all the operations of the system. This includes configurations, communications with other systems, controlling the all the components. All these sub-systems have fault tolerance managers which have error detection measures; damage assessment measures; damage confinement measures; error recovery measures; fault treatment and locator and continued service measures.

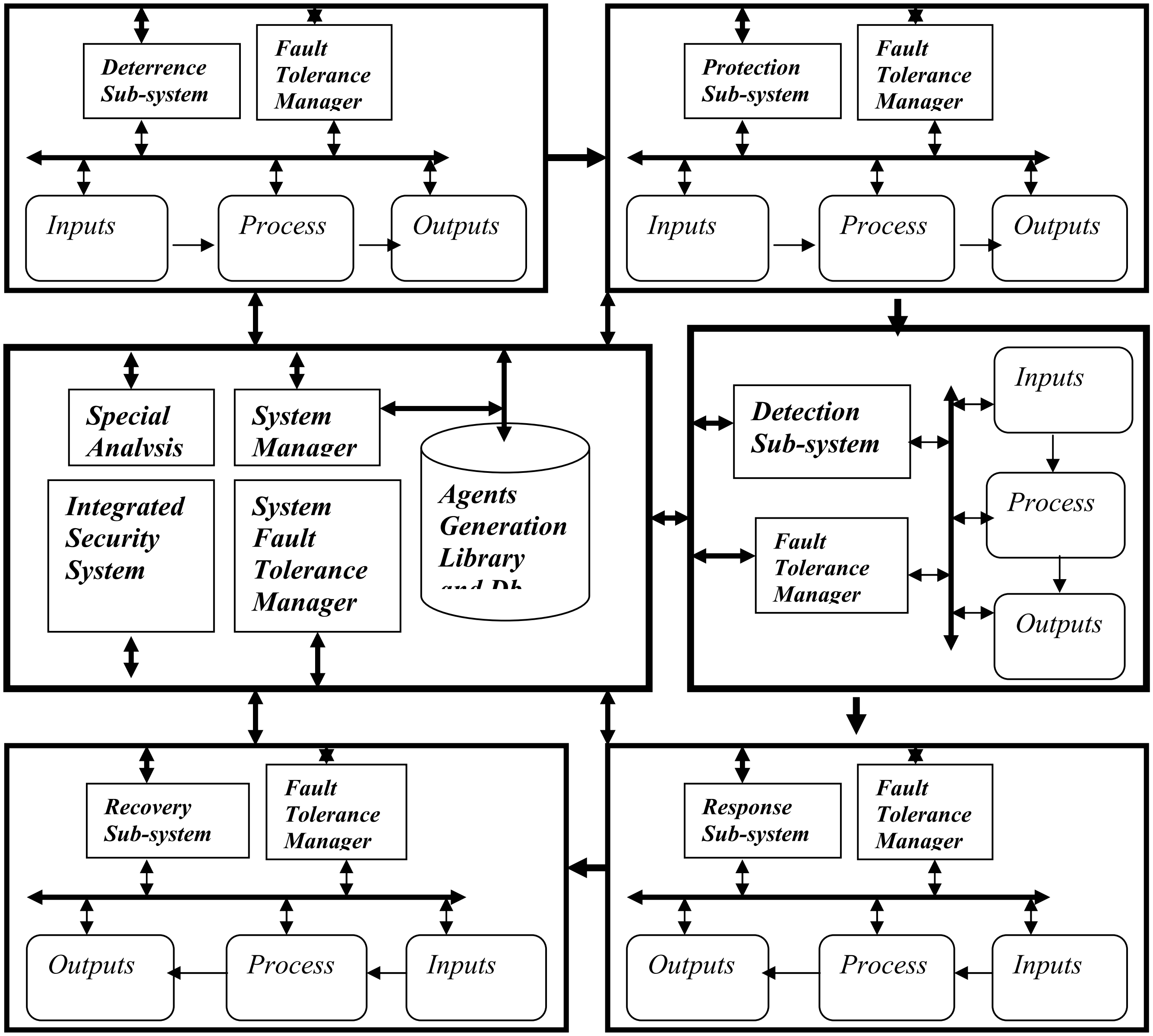


Figure 1: Generic Survivable Systems

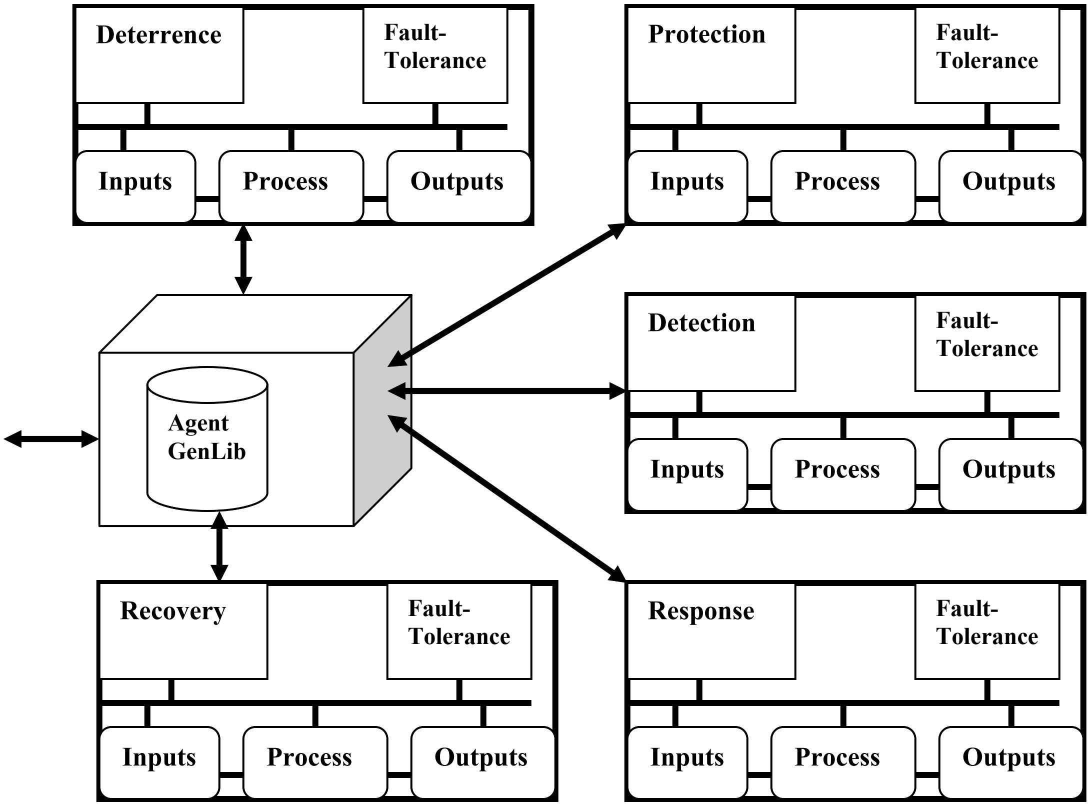


Figure 2: Agents

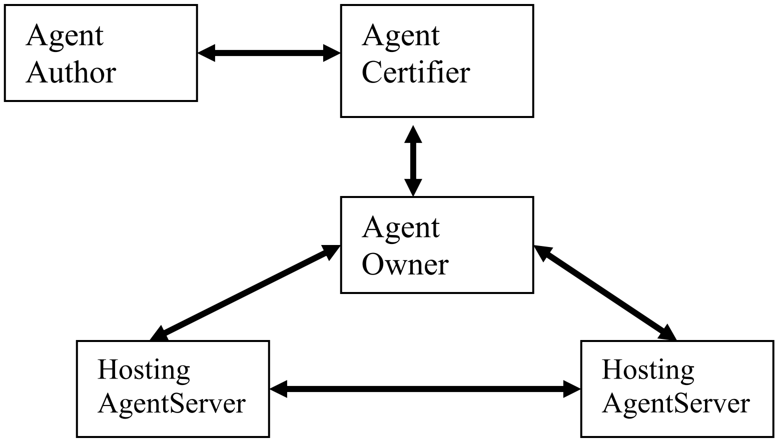


Figure 3: Agent Stations

4 Agents Security

4.1 Overview

The lifecycle of software mobile agent starts at an agent author, then it goes to an agent certifier, thereafter it goes to an owner, and then it is launched to different servers to perform the tasks specified [3] as shown in figure 3. The agent author and owner could be the same but the agent certifier and author/owner are not the same. An owner sends a request to an author with task specifications of the agent. The author creates the agent, comprising of a header, code and data [3]. The header contains the identity, attributes, recipient information, travel path and signatures. The data section is divided into internal data and collected data. The identity of the agent has three main parts [3]: static identity; dynamic identity; other specific identities. The static identity contains the author's ID, owner's ID and the ID of the agent. The dynamic identity contains the agent home and time of launch. Other specific identities can include digital certificates and other tokens. Attributes of the agent include level of trust, task specifications of the agent, constraints of the agent, and credentials of the agent owner.

Signatures include agent authors, certifier's, privilege authority's (this authority issues agent's security credentials to users), agent owners, agent sender's and agent server's signatures. Collected information include is information from different servers where an agent has been visiting. Agents are created by an author according the task specifications. The tasks are specified by the one requesting the services who then signs the agent

4.2 Protecting the Agent server

There are a number of technologies [4] for protecting agent servers. One of this is called Software-based Fault Isolation (sandboxing) [5]. This is when untrusted agents are isolated and monitored in a special environment. When other agents, which are not part of the protection system, come to an agent server they will be authenticated and put in different domains or sandboxed depending on the trust level of the agents. The second technology is known as Safe code Interpretation [4] which means that a command that is harmful can be made safe or denied execution. Many agents today are created in using interpretative programming languages, like java, that are platform independent and scripts to be able to run in all platforms. Another technology is called signed code in which agents and other objects are signed digitally by private keys. A digital signature enables the agent server to verify the identity of an agent, the origin of the agent and its integrity. Java applets can be signed, which enables them to perform actions in a wider range of platforms.

Another technique is called State Appraisal [6], which is a way of verifying the correct state of an agent before accepting the agent and before authorizing the agent to access objects. Path Histories [7] is another technology which aims at making sure that the agent servers that were visited before the current platform are authentic and have good reputation. This is done by having the agent servers sign the information collected by the agent. Another technology is known as proof carrying code [8] which is a way of forcing authors of agents to prove that they have included safe measures in designing and creating agents. The proof and the code are sent together to the consumer where it can be verified in a simple way without using complicated cryptographic measures and without needing any help.

In this work we apply Signed Code, Path Histories, a form of State Appraisal and a form of Sandboxing. The agents are signed by both the creator of the agent, the verifier, the owner and the sender of the agent. In this way we can verify the identity of the agent, the home platform, the sender and the verifier of the code. Path Histories' method is used by having the servers, where the agent is visiting, sign the information collected. State appraisal is done not by the agent server but by the certifier of the agent where a trust level is specified so that the hosting agent can decide in which category to put the agent. Sandboxing is applied to agents that are not from the protecting system. Next we discuss how agents are protected.

4.3 Protecting the Agent

Protecting agents is different from protecting agent servers [4] because the agents don't have their own processors and they can't extend the home platform, but have to rely on the environments provided for them. Protecting agents is more of a detective and deterrent manner while protecting agent servers is preventive, detective and deterrent. There exist a number of technologies for protecting agents [4]. One of them is called Partial Result Encapsulation, in which the results from each visited agent server are encapsulated. This can be done by the agent or by the agent server. But it is recommended to be performed by the agent itself. One way that can be applied is called sliding encryption [9] in which the agent seals information every time it collects it. The agent can use the public key of the owner to seal the information, so that only when the agent returns home that the collected information is unsealed.

Another technology is known as Mutual Itinerary Recording [10] in which two cooperating agents record and tracks each other's movements by sending to each through a secure channel the last agent server, the current and the next agent server. The next technique is called Itinerary Recording with Replication and Voting [11] which is similar to Path Histories [7] but it has been extended with fault-tolerant measures. There are multiple copies of an agent doing the same tasks. This method is resource demanding. The next technology is called Environmental Key Generation [12]. This is a way of protection in which an agent generates a key and protects all the executables if some environmental conditions are true. In this work we use Partial Result Encapsulation as described in section 4.7. Details of security services in different scenarios are described in the following sections.

The mobile agents that are performing fault-tolerance tasks have special security properties. They are authorized to access agents and inspect the agent headers, agent codes and data to detect errors, assess damages and so on.

4.4 Security Services during Agent Creation

For survivable systems the agent generation library, shown figure 1, is the agent author. In future, agents could be purchased / requested from other qualified authors. The sub-systems deterrence, protection, detection, response and recovery are the agent owners. These sub-systems request agents from the agent generation library in accordance to their specifications as shown in figure 2. The special analyzer acts as an agent certifier, but in future there could be an independent body for certification of agents.

The sub-system, for instance Deterrence, verifies the agent generation library before requesting a mobile agent. After successful authentication the denial of service cookies will be shared between the sender and the destination. These are functions of an address and a secret key. These will be part of all the communications between these parties. The aim of denial of service cookies is to reduce denial of service attacks. Communications that don't have denial of service cookies attached with specified properties are ignored. These cookies are not like the normal cookies that servers give to client browsers when visiting their sites.

The agent generation library verifies the identity of the particular sub-system. If the verification is successful the sub-system requests the required agent for its tasks from the agent generation library. Every sub-system has many different agents for doing diverse kinds of tasks in this sub-system. The agent generation library composes the code. The agent generation library calculates the integrity of the code and then the separate integrities of the header and the data and attaches its digital certificates. Note that the private keys of the agent's author, certifier and owner are never stored in the agent. To provide authenticity of agents the agent generation library signs the agent. To provide confidentiality requirement, the agent generator seals the agent by using the public key of the special analyzer which is acting as the agent certifier. The agent is then sent to the special analyzer. The special analyzer opens the message by the private key and verifies the signatures and the integrity of the agent. If successful the special analyzer checks whether the agent is behaving in accordance to the specifications. The certifier puts a trust level and its digital certificate. The analyzer signs the agent, protects it and sends it to the sub-system.

The sub-system opens the message using its private key. It verifies the signatures of the agent generator and of the special analyzer. The sub-system then notifies the agent generator and the certifier that it has received the agent. The sub-system also puts authorization attributes like mobility, expiration time, size limit of data it can collect and whether the agent can create (spawn) children. The sub-system's controller acts as a privilege authority and issues credentials like roles, group membership and monitoring attributes.

4.5 Security Services during Agent launching

The agent can be operating locally or it can be sent to deter, detect, protect at other locations of the system. Before being sent to the location the sender does the following procedure:

1) Specifies the tasks of the agent.
2) Assigns the dynamic identity by adding agent home ID and time of launch for authentication purposes
3) Attaches the digital certificate of the sender
4) The owner's signature of the agent is added for providing integrity.
5) Sends the state of the agent to the controller of the sub-system for audit trails.
6) The signature of the sender is calculated by putting the receiver's address, adding the hash of the agents' state, adding the timestamp and a random number. All this information is put in the recipient information field and is then signed.
7) To provide confidentiality the whole message is sealed by the destination server's public key.
8) The sender and the receiver authenticate each other before sending the agent. After successful authentication the denial of service cookies will be shared between the sender and the destination.

4.6 Security Services during Agent Hosting

According to [4] the agent server should provide separate domains for each agent that it is hosting, but in this work we don't provide separate domains for the security agents because they are deterring, detecting and protecting the system and they are supposed to move freely. When the destination server receives the agent it does the following procedure:

1) Opens the agent using its private key.
2) Verifies the agent generation library's and certifier's signatures to check for integrity.

3) Verifies agent owner's digital certificate
4) Verifies agent owner's signature
5) Verifies agent sender's signature
6) Checks the time stamp, hash of state and intended recipient in the recipient's information.
7) If successful the server accepts the agent
8) Monitors the agent to provide accountability requirement and the audit logs are protected by the agent server.
9) When the agent has done the tasks the agent server will sign the information and send the agent home or to the next agent server as described in section 4.7

The agents that are not from the security system will be sandboxed if they are not fully trusted. This will reduce the denial of service threat from these agents. In cases where denial of service is launched by outside attackers the address of the agent platform will be temporarily changed until the problem has been solved. The agents that require anonymity will have their identities hidden from other agents.

4.7 Sending an Agent for Cloning

If the agent is very successful in deterring, protecting, detecting intrusions and other tasks in accordance to the specified criteria the agent will be sent to the agent generation library for cloning. In sending the agent the following procedure will be followed:

1) The agent generation library and the sender will authenticate each other before sending the agent.
2) Attach owners digital certificate
3) Assign agent home ID and a time stamp which is the dynamic identity for authenticity purposes.
4) Create the owner's signature for integrity requirement.
5) Create sender's signature by putting the receiver's address, adding the hash of the agents' state, adding the timestamp and a random number. All this information is put in the recipient information field and is then signed.
6) The whole message is sealed by the agent generation's library's public key to provide confidentiality and then it is sent.
7) When the agent generation library receives the agent it will perform the procedure in section 4.6 and will then clone the agent and will send the agent back though the agent certifier as described in section 4.4. A copy of the agent is stored in the database of the agent generation library.

5 Conclusions

In this work we have provided ways of securing agents for survivable systems. The security requirements confidentiality, integrity, accountability are met. Information carried by agents and that which is stored at agent servers is kept confidential. Communications between agents and agent owners and agent servers are protected. Integrity of agents and data is provided through signatures. Accountability is provided through monitoring, signatures and log files protection. Denial of service is partially addressed by using denial of service cookies and by sandboxing untrusted agents. Anonymity is not complete; a mobile agent can be anonym to other agents but not to the agent server. An agent server has the right to monitor an agent. Limitation is that availability and anonymity requirements are partially met. Future work will be to implement the agent security. Agents that are used for fault-tolerant have all the authority to access agents for inspection purposes.

References

[1] A. Somayaji, S. Hofmeyr and S Forrest. Principles of Computer Immune System, 1997 *New Security Paradigms Workshop, ACM p75-82*

[2] Louise Yngström. A systemic-Holistic Approach to academic programs in IT Security, Ph. D thesis, Stockholm University / Royal Inst. of Technology ISRN SU-KTH/DSV/R--96/21--SE, 1996.

[3] Yi Cheng. A comprehensive Security Infrastructure for Mobile Agents. ISRN SU-KTH/DSV/R—97/13--SE

[4] Jansen, W. Karygiannis, T. National Institute of Standards and Technology Special Publication 800-19 – Mobile agent Security.

[5] Wabhe R, Lucco S, Anderson T. Efficient Software based Fault Isolation. Proceedings fo the Fourteenth ACM Symposium on Operating Systems Principles. 1993. URLhttp://www.cs.duke.edu/~chase/vmsem/readings.html

[6] Farmer W, Guttman J, Swarup V. Security for Mobile agents: Authentication and State Appraisal. Proceedings of the 4th European Symposium on Research in Computer security, 1996

[7] Ordille J. When agents Roam, Who Can You trust? Proceedings of the First Conference on Emerging Technology and Applications in Communications, Portland, Oregon, 1996

[8] Necula G, Lee P. Safe Kernel Extensions without Run-Time Checking. Proceedings of the 2nd Symposium on Operating System Design and Implementation. Seattle, Washington. 1996. URL:http://www.cs.ucsb.edu/~vigna/listpub.html

[9] Young A, Yung M. Sliding Encryption: A cryptographic Tool for Mobile Agents. Proceedings of the 4th International Workshop on Fast software Encryption. FSE'97. 1997.

[10] Roth V. Secure Recording of Itineraries Through Cooperating Agents. Proceedings of the ECOOP Workshop on Distributed Object Security and the 4th Workshop on Mobile Object Systems: Secure internet Mobile Computations. INRIA, France 1998. URL:www.igd.fhg.de/www.igd-a8/pub/#Mobile Agents

[11] Schneider F.B. Towards Fault-Tolerant and Secure Agentry. Proceedings of 11th International Workshop on Distributed Algorithms, Saarbucken, Germany 1997.

[12] Riordan J. Schneier B. Environmental Key Generation Towards Clueless Agents. Vinga G (Editor). Mobile Agents Security, Springer-Verlag, Lecture Notes in Computer Science No. 1419, 1998.

[13] Mwakalinga J, Yngström L. Generic Methodology for Creating Survivable Systems based on Systemic-Holistic Paradigm and the Immune System, Proceedings of NORDSEC 2005.

[14] N.R. Jennings. Agent-Based Computing: Promise and Perils. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Stockholm, Sweden. Pp.1429-1436, 1999.

[15] F. Bellifemine, T. Trucco. Java Agent Development Framework: http://jade.tilab.com/index.html.(15-04-2005).

User Friendly Authentication Mechanism for Rural Communities

R. Dayarathna [1] H. Ekanayake [2]

[1] PhD Candidate,
Stockholm University,
Sweden

[2] Department of Computation and Intelligent Systems
University of Colombo School of Computing
Colombo, Sri Lanka.

E-mail: [1] rasika@cmb.ac.lk, [2] hbe@ucsc.cmb.ac.lk

Abstract

This paper proposes a simple, user friendly but more secure mechanism for user authentication. Developing countries should make necessary changes to the technologies in developed countries before using them in developing countries without blindly using technologies in developed countries. If they realize value of their cultural aspects and value systems they can have state of the art systems with fewer resources. This paper addresses how the traditional value system facilitates to have a user friendly but more secure user authentication system.

Keywords: DigiPass, User authentication, One-time Password, Cultural issues, E-Governance, Rural community, Security

1. Introduction and Motivation

Information security is based on three building blocks, namely confidentiality, integrity and availability [1]. Authentication is one of the most important aspects in providing confidentiality. Authentication verifies identity of the claimant [2]. Identity card is one method used in proving authenticity of individuals. In the digital world, authentication can be done at various levels. One level is client authentication; in this case the identity of the user is verified. Another level is device authentication which verifies the identity of the device.

There are a number of existing user authentication systems but the most popular and widely used mechanism is based on user name and password. Even though, **it's the** widely used mechanism, it has certain limitations. There is a battle between user friendliness and security in this method,. since more complex passwords are tend to be forgotten. On the other hand simple passwords are vulnerable to guessing attacks [3]. The other problem is, it is not wise to use the same passwords over and over again since it would make easy for attackers [4]. There are other user authentication systems, such as, certificate based authentication, smart card based authentication, bio-metric authentication etc.. But they are complex and very expensive in terms of setting up and maintenance.

A better mechanism, called DigiPass, which gives `one-time password`, was introduced to overcome these limitations. The proposed mechanism made further enhancements to the existing DigiPass system especially targeting at people in developing countries. It is a simple, easy to understand and user-friendly system for IT-illiterate people. The new feature of the DigiPass system called Category password facilitates to share the device with owner's close associates.

There is a trend to bride the digital divide by introducing the ICT (Information Communication Technology) to rural communities but we have to realize that, generally, people are reluctant to switch to new technology until they are confident with it. Security is one of the features people are worried about. Therefore, it is necessary to provide understandable and user friendly security mechanisms to build confidence over new IT systems [2].

Numbers of researches are being done to assess the impact of traditions, value systems and customs in communities on using ICT [5]. They reveal that It would be easy for people to understand something they have already experienced with instead of having a totally new stuff. One of the stuff people are familiar with is, keys used to access doors, boxes etc. The users of keys have realized the value of them and also established certain practices in managing their keys. Hardware token has a number of similarities to the keys in certain ways. Therefore, people would easily understand the importance of having hardware token rather than a mere password written on a paper.

People share not only tangible things but also intangible assets in order to have an overall advantage. But, no one has addressed how to share a password. Theoretically, a password should not be shared with anyone. It can be shown that sharing a password gives certain benefits to the owner. Sharing has a positive correlation with trust. The trust levels depend on various aspects such as economical, cultural values, relationships among them etc. [6]. It has shown that in India the trust level of the society is much higher than that in the USA [7]. The password can also be shared with others depending on their trust status, situations and the value of the assets accessed by the given password.

2. ICT Issues for E-Governance

The electronic money order system [8], which is an alternative to credit card payment, was introduced to Sri Lanka in 2004. Statistics reveals, this system was becoming a very popular payment method in Sri Lanka. Since this is the only option for the poor people to do business transactions over the Internet, this was known as poor man's credit card. Minimizing time and saving the cost are major benefits of the system. However, lack of IT literacy is one of the major obstacles to further progress of providing necessary ICT infrastructure [9]. Therefore, IT literate persons have to provide their assistance to others to make use of the system. But, there is a security risk in getting help from others. This enhanced DigiPass mechanism facilitates to obtain assistance from others without compromising security.

3. DigiPass Framework

3.1. What is DigiPass?

DigiPass is a hand-held electronic device, having a keypad and a display in its face as depicted in Figure 1. A user can feed values to the device. The device then uses its embedded algorithms to process the input and finally gives a 'one time password'.

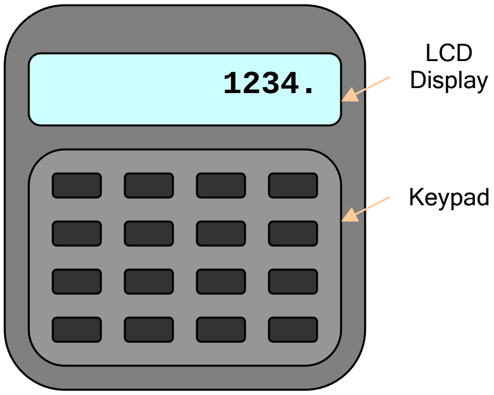


Figure 1: The DigiPass Device

3.2. DigiPass Mechanism

Categorizing User Transactions

Today, citizens interact with their government institutions and other private agencies via electronic means, such as visiting their web sites, participating in eDemocratic processes, paying for various services provided by those agencies etc.. These transactions do not require the same level of security for users' data. Therefore, it's possible to divide transactions into different categories according to their severity, for instance, paying fees for a service provided by an agency requires higher security than visiting a web site for gathering information. Based on that, different security levels could be assigned to these categories. Table 1 gives some identified transaction categories, the associated security levels and instances for each category.

Transaction Category	Security Level	Transaction Instances
Logging in to websites to get information	1	Getting applications for a driving license, passport etc.
Viewing details of utility bills	2	Viewing outstanding amount of water bills, electricity bills, etc.
Getting bank balance	3	Getting bank balance of saving's accounts, current accounts
Paying utility bills	4	Paying a telephone bill, water bill, electricity bill etc.
Transferring money among owner's accounts	5	Transferring money from a saving account to a current account etc.
Transferring money to a third party account	6	

Preparing with the DigiPass

A user is given a DigiPass with a default master password for the device. As soon as the user gets the device, the default password must be changed to a new password. This password is called master password which is a 4 digit number. Once the initial password is changed, a notice must be given to the issuing agency. Until the notice is given, the device cannot be used. The master password of the device is required to assign category passwords. If the owner forgets the master password, there is no way to retrieve the lost password. Then the device must be discarded and a new device be obtained from the issuing agency.

Once the master password has been set up, the device becomes ready to use. Now the owner could login to the device using the master password and setup category passwords for different transaction categories. It is not necessary to set up category passwords for all categories at the same time; passwords can be assigned when necessary. After category passwords are setup, there are two options to set the life time of the password. One option is to limit the life time for a given period of time varying from one minute to a number of days. In this case, the assigned password is valid only for a given period of time. Another option is to restrict the lifetime to a number of attempts. The process of setting category passwords is depicted in Figure 2 and 3.

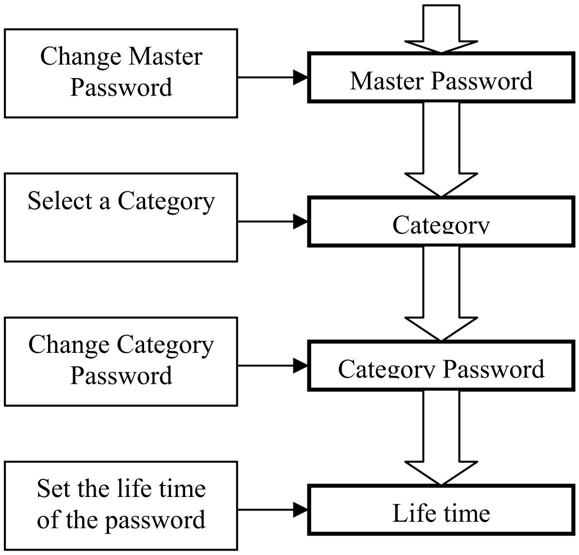


Figure 2: The owner has to first change the master password of the DigiPass, and then he changes category passwords and their life times.

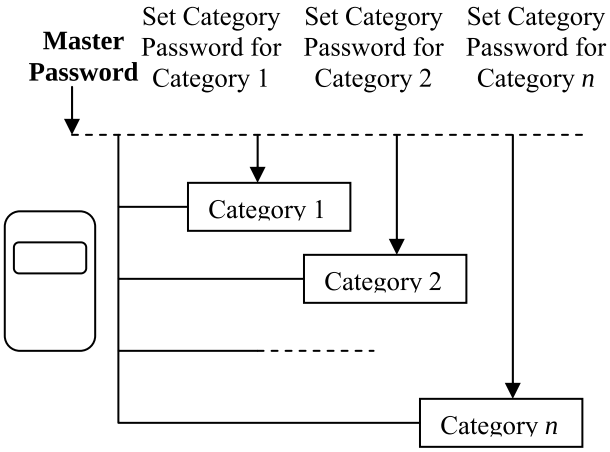

Master Password
Set Category Password for Category 1
Set Category Password for Category 2
Set Category Password for Category *n*
Category 1
Category 2
Category *n*

Figure 3: The owner uses his master password to log into the DigiPass and sets category passwords for transaction categories.

Using the DigiPass

After setting up category passwords, the device becomes ready to be used by either the owner or a legitimate holder nominated by the owner. If in the case of a legitimate user, the owner must define the lifetime of the category password(s) or number of valid attempts before giving it to the holder. By restricting the life time, the owner can assure that the holder cannot use the device for any unauthorized category or exceed the limits set by the owner in the case of authorized category.

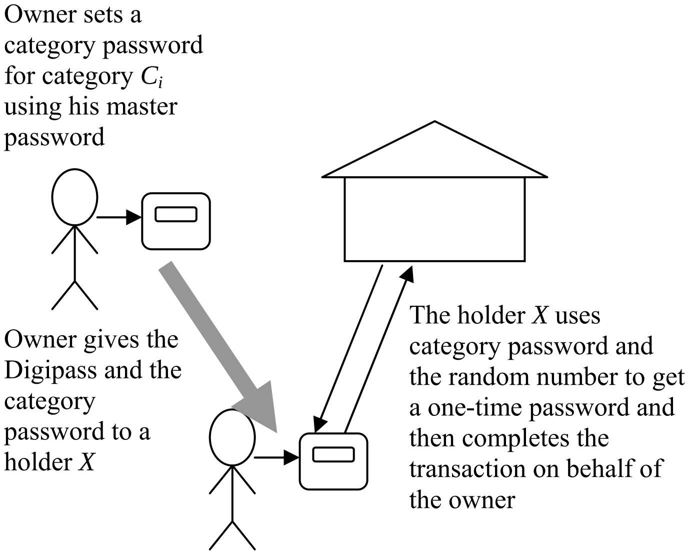

Owner sets a category password for category C_i using his master password
Owner gives the Digipass and the category password to a holder *X*
The holder *X* uses category password and the random number to get a one-time password and then completes the transaction on behalf of the owner

Figure 4: A holder completes a transaction on behalf of the owner.

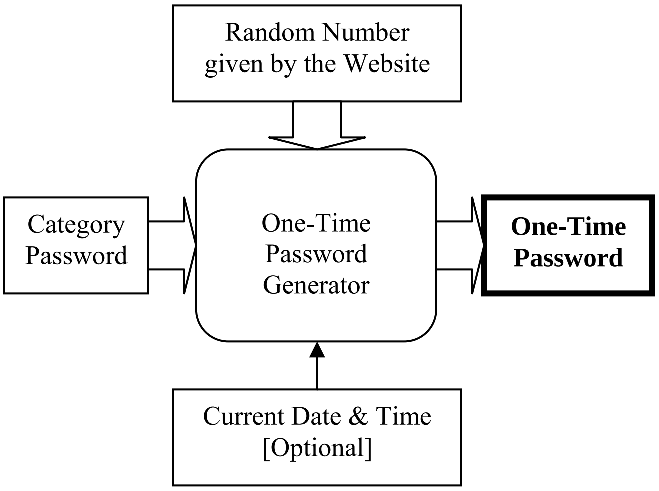

Random Number given by the Website
Category Password
One-Time Password Generator
One-Time Password
Current Date & Time [Optional]

Figure 5: One-time password generation process.

Figure 4 depicts how the holder completes a transaction on behalf of the owner and Figure 5 depicts the one-time password generation process. Once the holder informs his intention to make a particular transaction with an agency, for instance, paying a utility bill, the agency presents a random number associated with the particular transaction on the web page. The holder enters the category password, given by the owner, and the random number displayed on the web site to the DigiPass device. The DigiPass then generates a 'one time password'. This one-time password permits the holder to proceed and complete the requested transaction.

The following steps summarize the process of doing a web-based transaction using the enhanced DigiPass.

1. The customer gets the login page of the web site.
2. The customer selects the transaction category given in the web page, e.g.
   a. View the account balance
   b. Pay utility bills
   c. Transfer funds
3. The web server generates a random number and displays it to the customer.
4. The customer enters the category password, followed by the random number appearing on the web page to the DigiPass device. Consequently the DigiPass device generates a one time password to do the requested transaction.
5. The customer feeds the user name and the generated password to the system through the web page.
6. The web server responds to the customer by displaying the status information of the transaction.

4. Digital Exposure through DigiPass: Advantages

The proposed system has a number of advantages for people in developing countries. Researches have been conducted worldwide, especially in developing countries, with a view to identify better ways of using the ICT in their countries, since technologies in developed countries cannot be applied in developing countries "as is" basis [8]. These modified technologies and policies have to be amended to suite with economical, political and social structures of developing countries. Our approach is an attempt to modify and improve one of the existing technologies in developed countries to make it suitable for developing countries.

There is a struggle between security and usability of a system [3]. If the security is high, then the system becomes less user friendly and vice versa. In conventional systems, it is highly advised not to share one's password with someone else. In addition, there are guidelines for a good password, such as, a password should have a minimum length, contain numeric and special characters etc.. These measures are good precautions to keep a system in a secure state; however it drastically reduces the system's usability, since it creates a lot of hardships to users. Many people tend to forget their password and as a result they request new passwords [3]. This incurs cost to individuals, since their transactions get delayed for some period and on the other hand the issuing agency has to generate a new password and inform it to the user. Further researches should be conducted to identify the correct balance of security and usability.

Societies where people trust each other gain certain advantages over the rest. People in trusted societies do not paranoid about their security or privacy very much [7]. As a result people do share their belongings with others, since that would make their life easier. Another advantage is, they can save money otherwise would have been spent on providing additional security and privacy. Trust also varies with cultural, social and economical aspects.

Lack of access to the Internet is a major problem in developing countries. Over the past few years many cyber cafés, which provide access to the Internet, have been established to bride the digital divide. Since, most agencies tend to provide their services on line, it enables people to get things done by visiting a cyber café without going to agencies. However there is another problem to be considered, which is the cost and the time taken to visit a cyber café. For instance, if a family of four members want to do a transaction, each one has to visit a cyber café by himself/herself. In paper based system, if someone wants to get his account balance checked, he could give his pass book to his close associates and ask him to visit the bank to get the balance of his account. However, the same method cannot be generalized for electronic transactions. In the digital world, everything is in digital format and people are given passwords to access different services. In some situations, a single password is given to access all services offered by an agency. For instance, one password may give access to get one's account balance, pay his utility bills or transfer money to a third party account. To avoid a single sign on password which is vulnerable to many attacks, some financial institutions have introduced two passwords, one for sign-in and one for other transactions.

It is not wise to give one's single sign on password to another person since that may cause a number of problems, such as the second person may use it without the knowledge of the owner, or the second person may use it for unauthorized transactions. In a manual system, one could give his passbook to another and may ask him to get his balance since he knows that the agent cannot do anything other than getting the account balance. In a single sign on system giving that password to someone else is not a good idea, since that person might use that opportunity to transfer money to his own account.

The proposed system is more advantageous than the existing single sign on system or dual password system where it generates one time passwords for a single transaction. It requires the owner to have a single master password, and the owner can get independent passwords for every transaction. The owner sets a password for each transaction category and this category password combined with the on-line random number generates a 'one time password' for a given transaction. Each time the user wants to do an electronic transaction, a different random number is given by the server, which in combination with the category password generates a one-time password.

Since the DigiPass uses 4 digit passwords it is not that difficult to keep them in mind, and furthermore risk is very less since the owner has a hardware token. Consider an ATM account which has four digit numbers. Even though the password is very short, it provides enough security since the ATM card must be presented before doing any transaction.

This solution enables one to share his passwords with others while keeping control over his passwords. If something happens to his property, he knows the person who had the key and can identify the wrongdoer. The holder of the key also knows that if something goes wrong, he is responsible for that. The proposed mechanism uses this concept in sharing passwords. In this case, if a DigiPass is given with a password for a particular category, if something goes wrong in that category, the holder is responsible for it. For instance, if one is given a DigiPass to pay utility bills with the category password and if he has used it to pay more than the outstanding amount or paid twice, the owner can easily identify the person who was having the device when the particular transaction took place.

IT literacy is low in developing countries, therefore, it is difficult to educate them how to use a password and how to keep it secure. They might not realize its value and as a result they may write it on a piece of paper. They may reproduce it or write it down elsewhere. On the other hand, it would be easy for them to understand something they are familiar with. People are familiar with tangible keys and they have already realized the value of tangible assets, especially reproduction cost. Therefore, they would easily realize the value of the tangible DigiPass device.

There is no risk in showing a category password to a third person who loves to assist, if the password is set only for a single session. The third party cannot use again even though he has the device since the device is locked after the initial transaction. It also protects the people who use computers at cyber cafes. There might be key loggers or spy-ware programmers installed in the machines. Not like the single sign-on password, this is not vulnerable to spy-ware or key logger attacks since the system uses one time passwords.

While this facilitates sharing passwords it also prevents sharing of low sensitive passwords. The low sensitive passwords are used for accessing electronic journals, magazines, newspapers etc.. These passwords are used to protect commercial interest of the content providing agencies rather than protecting subscribers. Therefore, the subscribers are not reluctant to share these passwords unless they respect the intellectual rights of the content providers. This reduces cost for subscribers since one subscription is enough for many readers. But this practice makes huge losses to the content providers. The proposed mechanism does not solve this issue completely but it makes difficult to share passwords. Since, every time a unique and distinct password is required to access a web site and it is not convenient to contact the holder every time to get the right password for a particular session.

5. Conclusion and Future Work

This password generation process can be implemented as a separate unit or embedded in mobile phones. Implementing it as a separate unit adds an extra cost to the users and is also an extra burden for him to take it every where he goes. It gives a number of advantages if this mechanism is embedded in a mobile phone. In this case, it costs a very minimal amount and no extra burden is involved in carrying it since the mobile phone is always with the user. All mobile phones do not facilitate but fortunately upcoming mobiles do support third party programmes.

References:

[1] “Web Site Privacy with P3P” Helena Lindskog, Stefan Lindskog. Wiley publishing Inc 2003. p 17

[2] “Web Site Privacy with P3P” Helena Lindskog, Stefan Lindskog. Wiley publishing Inc 2003. p53

[3] “Password memorability and security: Empirical Results,” J. yan et al. Sept./Oct. 2004 IEEE Security and Privacy.

[4] “Customers, Passwords and Web Sites” B.Schneier, July/ Aug. 2004 IEEE Security and Privacy.

[5] http://www.sida.se/Sida/jsp/polopoly.jsp?d=321
“Sida- ICT in Developing countries” last accessed on 15/08/2005

[6] “Contracting on the Internet – Trends and Challenges for Law**”** Christina Ramberg
Law and Information Technology. Swedish Views 2002 *Information and communication technology commission. p 109*

[7] “Privacy In India : Attitudes and Awareness” Pnnurangam Kumaraguru and Lorrie Cranor

[8] eMoney Order System: The Smart way to Pay Online, IICT 2003

[9] ICTs, e-Governance and Rural Development, ICEG 2004

Dynamic Server Access by Mobile Agents Using Attribute Certificates

R.D.G Jayarathne
MSc. In Advanced Computing, University of Colombo School Of Computing
35, Reid Avenue, Colombo 07,
Sri Lanka
gaminij@atlinkcom.com

Abstract

Mobile agents are autonomous software entities that are able to migrate across different execution environments. Mobility and autonomy make permanent connections unnecessary and thus mobile agents have the advantage of providing low-bandwidth connections, asynchronous communication and many more. The characteristics of mobile agents make them ideal for electronic commerce applications in open networks. Such as Electronic Commerce web sites, distributed processing for high performance, data security and more. A mobile agent can search for special products or services and negotiate on behalf of its owner with other entities. Furthermore, mobile agents can be used as selling agents. Since the mobile agents are operating on the open network, it is not a good idea to trust and allow access to services provide. That can be restricted in many ways with the available technologies of authentication and any other mechanism. Scope of this paper is to find out how to secure dynamic server access by mobile agents using attribute certificates.

Keywords: Mobile Agents,Mobile Agent Platforms,Public Key Infrastructure,Certificate Authority,Public Key Certificates,Attribute Certificates,Attribute Authority,Role Based Access Control,Extensible Markup Language

1. Introduction

Use of mobile agents is an extreme form of distributed processing. This gives a huge flexibility in many different ways such as speed, use of external CPU power and resources, less network load, reduced network latency, etc. Hence this advance technology can be used for many systems such as transaction processing on distributed databases [SGE], sales agent systems search for quotations, clients accessing many other services over the internet and many more. The task of the system developer of near future will be driving through the network services that are available on the net and get the job done. But with all these advantages there are so many drawbacks that can be affected with the design of the integrated mobile agent systems. One such critical drawback is the security as many of the systems operating on the open networks find today. The security issues that; affects the mobile agent systems by the design are both ways from the service provider to the mobile agent and from the mobile agent to the service provider. Each party cannot trust what other party does and the origin of the agent while the agent travels over network. This paper presents an authorization solution for mobile agents to move over the network and access to services by producing special form authentication mechanism.

In brief the server authenticate the mobile agent who want to access a service in the server by a client certificate issued by a trusted Certificate Authority (CA) by both the client and the server. The public Key Infrastructure available today provides signed certificates for a considerably long period of time. But theses certificates have many drawbacks such as misuse of purpose issued, maintaining revocation lists, inefficient in usage and many others. But in this case we propose use of a special certificate with the signature for authentication and many other attributes for the validity of the request called attribute certificates. In here attribute certificate is valid for a small period of time and each time/day the mobile agent has to get a new certificate in order to present at the server for the identity and authorization of the request service.

1.1 Agents

An agent acts on behalf of people and does or helps in particular task or a subject area in the real world. The agent is limited to particular scope and it is operating in a particular environment that it is assigned to. The best example we can take from the real world is sales agents [AAFA] in many aspects. They act between their customers and the trader that they are working with. In this case if the agent is capable of providing better service if it is intelligent on acting upon a task.

Most of the situations computer science borrow the ideas in the real world and model them in the computing environment so that it is applied to solve a problem. The software agents that we talk about is also modeled in computer so that it is acting on behalf of a person or particular host, and assisting them by various means. When we consider the today's available agent systems there can be various characteristics that we can categories in to many forms. Mainly they are autonomous, acts on the various execution environment events, assigned to achieve particular goal acting on a particular execution environment. It is optional that these agents are capable of moving/mobile, able to communicate, being intelligent and able to learn from the environment that they act.

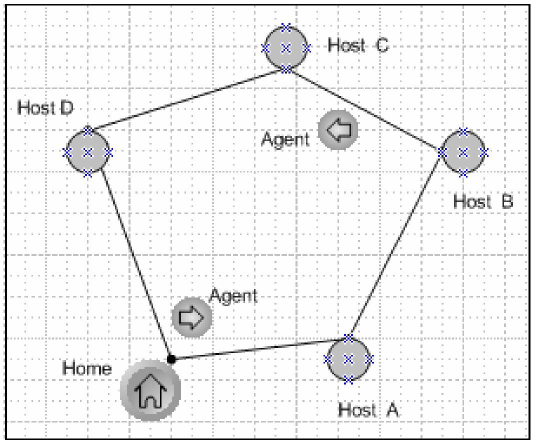


Figure 1.1 Sample Mobile Agent Platforms

1.2 Problem

Mobile agents begin their execution on a host that it owns and launch. On the origin host it self it will execute a part of its execution plan/code and move the interrupted execution code to a different host with current state of information. From the next host that it reaches needs to resume the execution in order to perform it's assigned task. This process itself will lead to lot of security risk on both the agent and the hosted environment of the services. In the research and practical community dealing with mobile agents today uses many sophisticated solutions to get rid of the high risk of this security aspect.

The current used methods protecting agents can be listed in following main areas:

Partial result encapsulation – The results collected in each host is encrypted, or signed.
Mutual itinerary recording – The itinerary for the previous, current and next host is recorded.
Itinerary recording with replication and voting – Execute few instances of the agent and result is selected by a voting scheme.
Execution tracing – Each step is recorded while the agent executes on the platform.
Environmental key generation – Set environmental condition as a key to decrypt a section of code to be executed.
Computing with encrypted functions – Functions are encrypted to execute after self extraction.
Obfuscated code (Black box) – The code is scrambled such that no one can read and understand [WJTK1].

Also the research community identifies protecting platforms from malicious agents, and major ones can be listed as follows:

Software-based fault isolation – Execution environment is limited depending on the development environment.
Safe code interpretation – The code interpreter analyze the harmful code segments.
Signed code – The code is signed with a digital certificate for identification.
Authorization and attribute certificates – The privileges are defined as attributes for actions.
State appraisal – State changes are appraised to detect harmful levels.
Path histories – Check the agent's path that was visited before arriving at current host.
Proof carrying code – The code segments provided by the authors is analyzed and checked against the agent code. [WJTK1]

1.3 Access services over network

The users of a geographical distributed organizational environment will want many interactions among departments in the process of digging in to the information that needs in the day today operations and even more. This is a general need of today's organizations where it may allow many inter organizational relationship maintenance over the their IT systems in the real world. When the interactions are in place it always happened to consider the restrictions and the limitations that applies for the departmental policies and the organizations. If these interactions are implemented and used, with a sophisticated mobile agent platform expecting a high level of the service in many senses that automatically bring up many other concerns. Mainly the level of the access, authentication mechanisms, departmental or organizational policies are in the priority.

1.4 Related Work

"Smart Certificates: Extending X.509 for Secure Attribute Services on the Web" [JPRS] is a paper published to extend the X.509 public key certificates with attribute services on the web interactions in security aspects. Smart Certificates defines the secured attributes of an entity as the key to the authentication where they are unique to a particular individual in an environment or organization. The smart certificates are such an entity where attributes describe the authentication information for a purpose that it will be used. In a particular scenario smart certificate can be used as ticket to bid for an e-commerce bid system, where it identifies and authenticate the particular bid and the transaction. In this approach it has been used for a web authentication mechanism for individual users. This lacks the generalization mechanism where it can be used for a group categorization. Hence it is not a general model where it can be applied for many such individuals.

The M & M approach: The M & M approach" [PRPLJ] the paper published to find out the feasibility of using mobile agents for communication among existing web servers. Even though this approach is not directly connected with the scope of this work, it has considered the security aspects of the agents and agent platform, which talk about the access policies and the privileges.
As the abstract architecture of the work they say, "In the M&M framework there are no agent platforms. Instead applications become agent-enabled by using simple Java Beans Components." The paper discussing the security aspects of the architecture, it has used secure socket layer to transfer the mobile agents over the network, so that third parties does not eavesdrop on them. And more than this the authors have considered the resource access levels by the agents through the applications they interact with using JVM security manager. It has mentioned that the difficulty of instantiating more than one security manger in the JVM.

2. Infrastructure

The concept of the secure services access using mobile agents needs to make use of many other technologies that is out and become an industry standards. Integration of important features that are with in those technologies are chained together to come up with a security mechanism over the mobile agents accessing many services provided over the network. Following sections discuss the main features of the technologies that are being integrated to model the new mechanism.

2.1 Role based access control

When individual user is considered in case of access control and the administration over a large network it becomes more complex and hence difficult to manage. Role based access control (RBAC) (also called role based security), as introduced in 1992 by Ferraiolo and Kuhn [RS1], has become a highly advance and useful model for advanced access control because it reduces the complexity and cost of security administration in large networked applications.

Access to resources in a system such as files, applications or devices by means of use, view, change and many other actions becomes critical resources in some situations. Therefore access control is restricting the access to those resources by means of various levels. In role based access control, accesses to those system resources are permitted by means of the roles that each individual is playing in a particular organization. The access is controlled on behalf of groups/roles, but not by each individual. The concept itself reduces the number of users that it needs to manage and administer because when it is grouped numbers become smaller.

Figure 1.2 shows the overall idea of the Role Based Access Control relations among user, Roles and permissions [RS1].

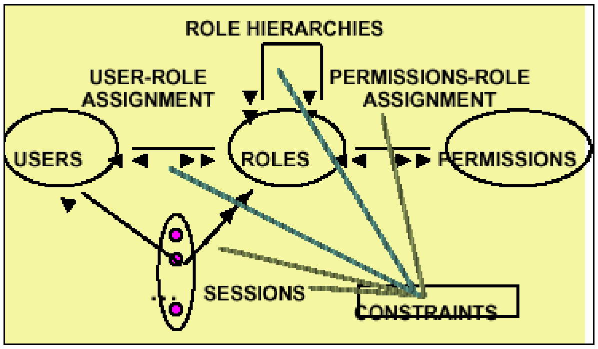


Figure 1.2 Role Based Access Control relations.

Need of Policies

When an organization defines there roles on the system it needs a comprehensive set of system/organization level policies and privileges explaining what actions and activates need to perform, in order to achieve a direct or indirect business requirement. Each role is a sub set of the organizational policies and privileges, which defines the rules, and some of them will overlap among roles which will take the priority to avoid such situations.

2.2 Public key certificates

The certificates become a need in the public key cryptography when the large-scale networks are used to exchange secret keys securely. Public key cryptography provides a way to evade this problem.

The public Key Infrastructure defined by X 509 defines public key certificates (PKC). A certificate authority (CA) a person or entity that checks to confirm that the identity or other information in the certificate belongs to the holder digitally signs this. In large scale, all the parties operating on a network might not trust a particular certificate authority. In this situation certificate authorities form a hierarchy, so that signature of certificate authority on each level is confirmed signing by the upper level certificate authority. This allows to trust any other certificate authority in the highrachy, which are signed in the certificate. When two such higherachies become separate, the two top certificate authorities sign each other to bridge the two chains of certificate authorities.

The certificate authority can revoke an issued certificate if the private key is lost or stolen, or if it is expired as well as becomes invalid due to some other reason. This will lead to maintain a revocation list in each certificate authority so that any one who needs to validate a certificate against the list can do so. But this becomes a burden when the lists become longer.

A certificate typically includes:

1. The public key being signed.
2. A name, which can refer to a person, entity or an Organization.
3. A validity period.
4. A location to maintain revocation list.

2.3 Attribute certificates

The public key infrastructure X 509 defines a certificate [SFRH1], which contains set of attributes that are belong to a particular entity. Attribute Authority certifies this certificate by digitally signing on the set of attribute and value pairs. If a particular person or entity trust the Attribute Authority it can trust the attributes specified in the certificate. An attribute certificate (AC) is a structure similar to a PKC. The main difference between attribute certificate and a pubic key certificate is there is no public key in the attribute certificate [SFRH1]. An AC may contain attributes that specify group membership, role, security clearance, or other authorization information associated with the certificate holder. The public key certificates are issued for a significantly longer period of time and hence it needs a proper management mechanism in order to keep them valid such as maintain revocations lists etc. The attribute certificate contains the set of attributes which are dynamic with the time, and hence it is not issued for a long period of time since the contents itself invalidates. The attribute certificate is also structurally similar to the public key certificates. The following list shows the generic structure defined in RFC3281 [SFRH1] for an attribute certificate.

```
AttributeCertificate ::= SEQUENCE {
         acinfo              AttributeCertificateInfo,
         signatureAlgorithm  AlgorithmIdentifier,
         signatureValue      BIT STRING
        }
 AttributeCertificateInfo ::= SEQUENCE {
      version               AttCertVersion -- version is v2,
      holder                Holder,
      issuer                AttCertIssuer,
      signature             AlgorithmIdentifier,
      serialNumber          CertificateSerialNumber,
      attrCertValidityPeriod  AttCertValidityPeriod,
      attributes            SEQUENCE OF Attribute,
      issuerUniqueID        UniqueIdentifier OPTIONAL,
      extensions            Extensions OPTIONAL
        }
```

In the above structure it shows that an attribute certificate is similar to a public key certificate except the “attributes” field, which can be, extend to set attributes that defines an entity. With this definition someone might find that this can be extended to any kind of properties that we need to promote for a system.

3. Problem specification and aims

The mobile agents it self is highly complex as we discussed in the above sections that explains the subject area of the research. Since it is mobile that improves the loopholes that can support the malicious agents or hosts as well as third parties do misuse the environment. When we look at today’s available technology for protecting each other in the mobile agent platforms it has actively looked in to the issues and there are many techniques that are proposed by various parties.

When it is looked in to the area where protecting agent host platforms from malicious agents, there are well known technologies available as listed in the above sections. But all of those techniques that are out there explain how to predict/ guess who is malicious and who is real in the sense of what they are supposed to do, and are they doing the right thing. But when it happened to be a large organization with so many users using the network, it always need an access control mechanism which defines who are supposed to do what and when. As we discussed earlier this is well defined by the RBAC with the rules defined by policies and privileges. The same principle will apply if the agent platform is distributed over the organizational network and it will not be easy to keep track of each individual agent and implement an access control mechanism for protecting the agent platform and the resource access.

The idea of this research is to propose a method to protect agent host platforms from agents with an assigned role by the organizations AA defining the roles for each agent. In here the use of attribute certificates to define the role and the validity period provide the flexibility to keep them dynamic so that the issued certificate/ granted permission is useful only for a small period of time, probably a single use.

Aim of this research is to prove that the concepts discussed above can be applicable in to a mobile agent platform in order to protect the agent host access by agents can be restricted by RBAC defined by an attribute certificate. Here the server can be stationary or mobile.

3.1 Prototype user and system requirements

Three personal computers on a network environment with Windows 2000 Server operating system installed
Java Runtime Environment (JRE) installed on two hosts
Other Machine hosting the Attribute Authority
Grasshopper 2.0 installed on the two hosts

3.2 Basic findings

The concept of "Dynamic Server Access by Mobile Agents using Attribute Certificates" itself list out the infrastructure need to prove the concept that it can be implemented as a real case study. This is an integration of many technologies that are out there today in many areas. Therefore in order to prove feasibility we need to find out the usable implementations that are out there today.

When it is considered the mobile agent platforms that are available today to get the environment setup, we looked at two of the mobile agent platforms that are available for evaluation. The two are IBM Aglets [MOGKKO] and the Grasshopper [GHC].

IBM Aglets: The authors of this system says "it mirror the applet model in Java. The goal was to bring the flavor of mobility to the applet. The term *aglet* is indeed a portmanteau word combining *agent* and *apple*t. We attempted to make Aglets an exercise in "clean design," [DBL] and it is our hope that applet programmers will appreciate the many ways in which the aglet model reflects the applet model". This is also a kind of suitable mobile agent platform for the proposed model in this research, if it was not more towards a web environment. Also the Interfaces provided by the platform are not well documented to plug in own code and display the output on each host.

Grasshopper: The Grasshopper is a platform it self uses the basic concepts in the mobile agents. Each host can be configured as a node in an open network. The grasshopper user guides says, "Every fundamental implementation aspect is handled, such as mobility, local/remote communication, agent localization, CORBA support, and security." [GPG] As it says the platform it self support the rich communication mechanisms provided by java, and tools that are capable of monitoring the host agent platform, agent status as and detailed log viewer to help the solution development.

The implementations for the attribute certificate defined in the RFC3281 [SFRH1] are available from many API vendors. The Akenti is available on the web as an evaluation version, and it is part of Job and resource management tool called "Fine-Grained Authorization for Job and resource Management Using Akenti and the Globus Toolkit®". [MAKVSB] When the organizations are distributed it needs a standard way of sharing remote resources across organizational domains, the need for fine-grained access control to these resources as it increases. The Akenti Attribute certificate [WSM] is designed to support for the above requirement and it matches with the exact requirement of the proposed model. It was implemented as a proprietary XML format, which can be extended with a Role based access control mechanism very easily.

4. Overall architecture

The model that we consider here is a general one which can be applied to any network structure moving a mobile agent to access server method which is stationary or mobile server running on a different platform. Consider the server is in the network and it is available for servicing. A host in the network needs to access a server method. Hence the host creates a new mobile agent and requests an attribute certificate from the AA in order to present as the host identity and the assigned rights on the access policy of the organization. When this request comes through a secured socket connection, the attribute certificate Authority (AA) verifies with its database of all registered IP's that the origin is not a malicious host. According to the registered IP's assigned rights are matched against organizational policy database and an attribute certificate is generated and issued accordingly.

When the certificate is received by the mobile agent, it starts own processing on the origin host and move over the network with the current status of execution. When the mobile agent moves over the network, in some node it will find the required server that is providing the service. Then the mobile agent presents his attribute certificate to the server for the authentication. The server processes the attribute certificate presented and first checks the signature for authenticity. If a trusted AA of the server/organization has issued the certificate, it processes the attributes and their values assigned by the certificate authority to find out the purpose of the certificate and the rights assigned by the organizational policy database. This is done by the Policy Engine, which will be discussed in a section below. If the policy engine is satisfied with the attributes contained, it will allow the mobile agent to access the service from the server and continue the processing. When the mobile agent finishes his processing it will return to the origin host.

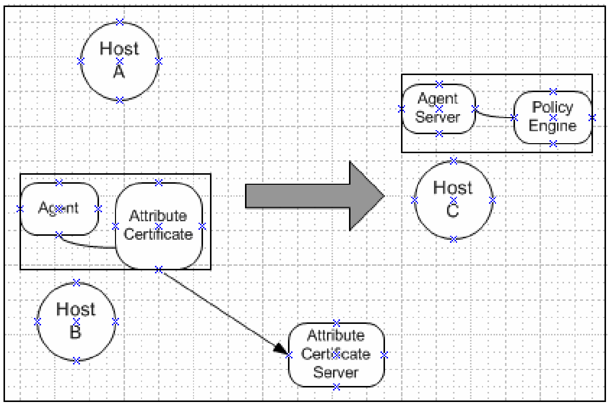


Figure 1.3 Sample Environment Model

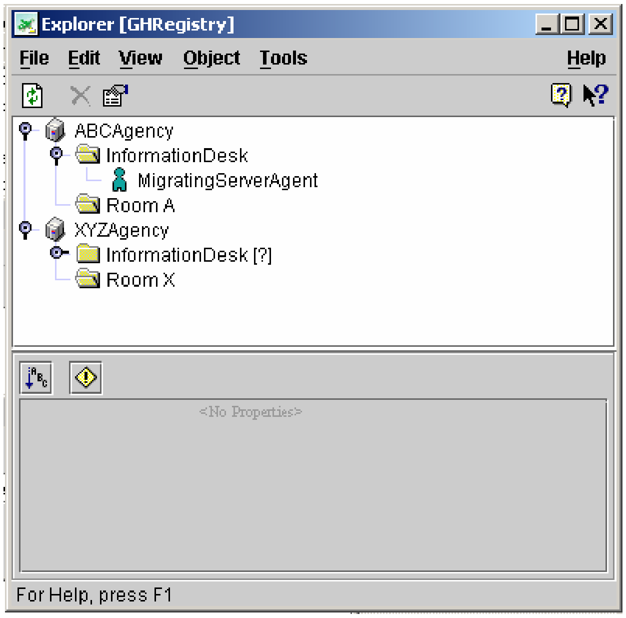
4.1 The Policy Engine for validation

The Policy Engine is a concept that we introduce here and implemented as a security manager for the mobile agent platform. The security manager in the java provides a way to check for the access rights for different resources that is accessible by different applications. This is implemented in to the code and it takes the advantage of checking the accessibility of the each resource before the execution starts. In here we propose a plug in where it connect to the policy database, which is defined by the organization on the enterprise wide network. The policy database will contain the policies defined with the roles that will interact with the system. A role will have a set of restricted privileges as defined by the policies.

The idea here is when the attribute certificate is read which defines the attributes, will be passed in to the Policy Engine so that it will validate against the policy database.

4.2 The agent platform setup

The proposed model contains two hosts, one as an agent owner host and the other as an intermediate host that runs stationary or mobile agent server. Each host is configured in a Region Registry in the Grasshopper platform so that it becomes a network node which can do the inter communication among regions. With in the Region Registry an Agency is created. An Agency can hold Places that a mobile agent can move and continue to process. In each agency there is a default place called an Information Desk, which is available for any mobile agent to arrive.



ABCAgency
InformationDesk
MigratingServerAgent
Room A
XYZAgency
InformationDesk [?]
Room X

<No Properties>

For Help, press F1

Figure 1.4 Grasshopper Agency with in a Region

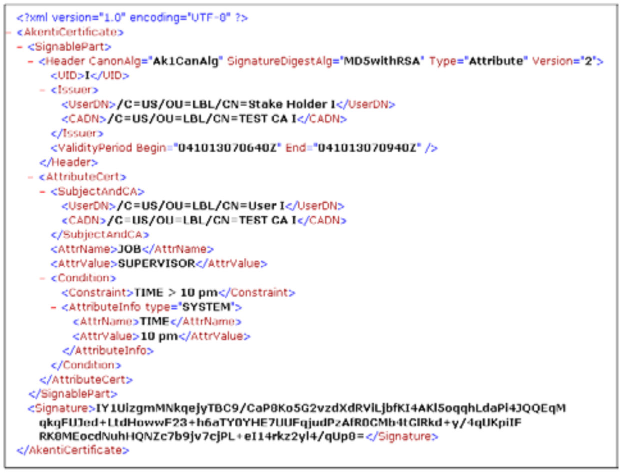
In the proposed model Host ABC and XYZ hosts two regions and one agency with in each region. Host ABC is the origination of the mobile agent and get the attribute certificate from the certificate authority and move the next hop specified by the Region URL in the following format.

socket://172.21.0.41:7000/ABCAgency/RoomA
<Communication Mechanism>:<Destination IP>/<Agency>/<Place>

When the agent moves hop by hop it will find the required server on an intermediate agency. At this node the agent get access to the server by presenting the certificate and does the processing.

4.3. Akenti certificate as attribute certificate [MAKVSB]

The implementations for the attribute certificate defined in the RFC3281 are available from many API vendors. The Akenti is available on the web as an evaluation version, and it is part of Job and resource management tool called "Fine-Grained Authorization for Job and resource Management Using Akenti and the Globus Toolkit®".[MAKVSB] The Akenti Attribute certificate is designed to support for the above requirement and it matches with the exact requirement of the proposed model.



```
<?xml version="1.0" encoding="UTF-8" ?>
- <AkentiCertificate>
 - <SignablePart>
  - <Header CanonAlg="Ak1CanAlg" SignatureDigestAlg="MD5withRSA" Type="Attribute" Version="2">
     <UID>I</UID>
   - <Issuer>
      <UserDN>/C=US/OU=LBL/CN=Stake Holder I</UserDN>
      <CADN>/C=US/OU=LBL/CN=TEST CA I</CADN>
     </Issuer>
     <ValidityPeriod Begin="041013070640Z" End="041013070940Z" />
    </Header>
  - <AttributeCert>
   - <SubjectAndCA>
      <UserDN>/C=US/OU=LBL/CN=User I</UserDN>
      <CADN>/C=US/OU=LBL/CN=TEST CA I</CADN>
     </SubjectAndCA>
     <AttrName>JOB</AttrName>
     <AttrValue>SUPERVISOR</AttrValue>
   - <Condition>
      <Constraint>TIME > 10 pm</Constraint>
    - <AttributeInfo type="SYSTEM">
       <AttrName>TIME</AttrName>
       <AttrValue>10 pm</AttrValue>
      </AttributeInfo>
     </Condition>
    </AttributeCert>
   </SignablePart>
   <Signature>IY1UlzgmMNkqejyTBC9/CaP8Ko5G2vzdXdRVlLjbfKI4AKl5oqqhLdaPl4JQQEqM
    qkqFUJed+LtdHowwF23+h6aTY0YHE7UUFqjudPzAfR0CMb4tGlRkd+y/4qUKpiIF
    RK8MEocdNuhHQNZc7b9jv7cjPL+eI14rkz2yl4/qUp0=</Signature>
</AkentiCertificate>
```

Figure 1.5 Sample Attribute Certificate viewed in Internet Explorer

The certificate that above Figure 1.5 lists has only one Attribute assigned, but with the flexibility of Akenti certificate structure [WSM] it can go on adding more attributes as needed. For example the Attribute value of the certificate is a SUPERVISOR, which is a role in the organizational policy database, and it has assigned a constraint that he is able to use the certificate only after 10 PM. When this certificate is evaluated in the policy engine it will grant or deny the access to the service validating against the system time. In this model the basic policy engine is implemented so that it can validate a role and its time constraint so that the server makes the decision weather the certificate is valid for grant or deny accessing the service. For example following logical expression is formed from the attributes assigned for the role as a supervisor and the supervisor only allowed to access the system only after 10 PM.

Eg. **((role==SUPERVISOR) && (time>10pm))**

4.4 The implementation of the system

The whole architecture of the implemented model can be vied by the Figure 1.6. The numbers listed on the figure with a small description will be contrasted from here.

Server starts

The Server that is providing the predefined resource access using public methods will be created on one of the nodes on the agent host platforms. After the creation it will be initialized to provide the services by loading the modules as the services provided. After the initialization as the traditional servers do, it will wait for the agent requests that have arrived on the same agent host.

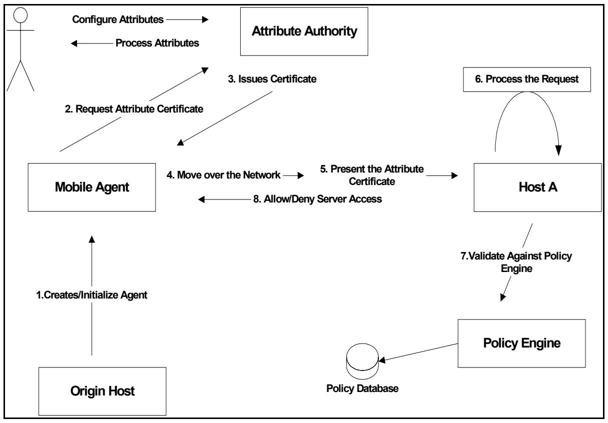


Figure 1.6 Overall Architecture & implementation

Attribute Authority starts

The Attribute Authority starts and initializes with the set of pre-registered IP address and a profile pairs that is available on the network management system, which will only be allowed to request an Attribute certificate. When the registration happens the profiles are checked against the network management system for the existence. After loading the register set of profiles, the Attribute Authority will wait for the certificate requests.

Creates/Initializes the mobile agent

The origin host will create the mobile agent as the need arises to assign a task. As the mobile agent's assigned task starts it will connect to an Attribute Authority, which is residing on a network node. Then the agent will send the impersonated profile to the Attribute Authority with a secure connection.

As the Attribute Authority receives the profile over the secure connection it will check the profile and the IP address which the request came from against the registered set of IP address and profile pairs. If this step passes it will check the profile for the roles assigned and their valid policies against the Organizations policy database in order to assign the attributes for the certificate. After certificate is generated it will be signed by the Attribute Authority and sent the agent who requested the certificate.

As the agent receives the attribute certificate it will save that as a state variable in the execution code and move to the next hop as it is instructed to do so.

Move over the network

While agent moves over the network on each agent host it will search for the particular server that is searching for in order to access the required method.

Present the attribute certificate

When the required server is found the agent will present the Attribute Certificate it carries in order to gain access to the server method for the resources that it needed.

Process the request

The sever which allow the agent to connect by presenting the Attribute certificate that it present will check for the validity of the certificate by using the signature on the certificate by the attribute Authority which is trusted by the server as well. If this check is passed then it will check for the valid period that is specified in the certificate against the current time of the agent host. If both the tests are passed server will allow it to proceed, otherwise the access is denied and the agent is sent back.

Validate against the policy engine

If the certificate is valid after checking the signature and the valid time period, the server will proceed to check for the next level. It will read the attributes that are specified on the certificate with the constraint value pairs. In here in order to implement a Role Based Access Control mechanism we use an attribute called "Role" and value as various roles that are valid for the organization in this scope. Then the role is validated against the policy database and finds the privileges that it is using the Policy Engine, which will be described next. Then the constraints specified in the certificate are validated individually. After that it can form a logical expression and get the end result weather a particular method can be accessible on the server which is predefined set of constraints that need to be valid. When this step is passed the agent will be able to access the resources on the host using the server method, which can be well restricted with the time and the privileges.

5. Discussion

Attribute certificate is a powerful tool for implementing a good security mechanism for protecting agent platforms. It was well proved that, it could be used to implement fine-grained access control mechanisms with attributes and their constraints. The scope of this paper is to propose a method to secure server access using mobile agents. Attribute Certificate is ideal for this scenario since it is issued for a small period of time and many other constraints which will expire soon, no revocation is needed and allowed to have flexibility for defining the policies to access servers. Hence it is a matter of putting down the server access policies in to an attributes having all other standard headers and the attributes providing the security for the certificate it self. The proposed model was to secure one side of the mobile agent security, which is protecting the server access against malicious mobile agents. It can also be a possibility that we can use a certificate or any other means to protect mobile agents against the malicious hosts that can be mobile or stationary.

It is a necessity that there exists a powerful standard for defining policy databases for an organization where it make the use of attribute certificates to protect the illegal server access. When the policy database defined simple it will be a very much easier for an attribute authority to interpret them in an attribute certificate as well as a policy engine evaluate them when an access request comes in with a presence of an attribute certificate.

It is a possibility without carrying the certificate from the origin host, request it on the agent host, or received after the migration. This might make the agent light weight while moving over the network. It may be an issue in certain cases where it is not possible to access the local resource to keep a socket connected.

5.1 Future Work

As mentioned in the discussion it is a requirement that it need a simple policy definition standard where a defined policy or set of policies can be interpreted as a logical expression. If such standard exists it will make the attribute authority's work simpler when interpreting them in to attributes of an attribute certificate. Together with this requirement it needs an implementation policy engine APIs where they can evaluate the policies defined in an attribute certificate come up with the final decision weather it allows or deny the access. These come in to one area where it needs a standard for policy definitions and a related evaluation mechanism, which is simple. The scope of this paper was to protect only the agent hosts from malicious agents with attribute certificates. But also it can check weather the same concept can be applicable to protect agents from malicious hosts.

References

[SGE] S. Papastavrou, G. Samaras, E. Pitoura "Mobile Agents for WWW Distributed Database Access", *University of Cyprus, CY-1678 Nicosia, Cyprus*

[AAFA] Alberto Pan, Antonio López, Fidel Cacheda, Angel Viña, "Mobile Agents based Architecture for building Virtual Markets", Departament of Electronics Systems, University of Coruña.

[WJTK1] Wayne Jansen, Tom Karygiannis "Privilege Management of Mobile Agents" National Institute of Standards and Technology

[WJTK2] Wayne Jansen, Tom Karygiannis “NIST Special Publication 800-19 – Mobile Agent Security”, National Institute of Standards and Technology, Computer Security Division,Gaithersburg, MD 20899

[RS1] Prof. Ravi Sandhu “RBAC Architectures and Mechanisms” INFS 767 Fall 2003

[SFRH1] S. Farrell, R. Housley “An Internet Attribute Certificate Profile for Authorization – RFC3281”, Baltimore Technologies, RSA Laboratories, April 2002

[MOGKKO] Mitsuru Oshima, Guenter Karjoth, Kouichi Ono, “Agglet Specification”

[GHC] Grasshopper Basics And Concepts, Release 2.2 IKV++ GmbH Bernburger Strasse 24-25 10963 Berlin, Germany

[DBL] Danny B. Lange “Mobile Objects and Mobile Agents: The Future of Distributed Computing?” General Magic, Inc. Sunnyvale, California, U.S.A.

[GPG] Grasshopper Programmer’s Guide, Release 2.2, IKV++ GmbH, Bernburger Strasse 24-25, 10963 Berlin, Germany.

[MAKVSB] M. R. Thompson, A. Essiari, K. Keahey, V, Welch, S. Lang, B. Liu “Fine-Grained Authorization for Job and Resource Management Using Akenti and the Globus Toolkit®” ,University of Houston, Houston, TX 77204.

[WSM] William E. Johnston , Srilekha Mudumbai, Mary Thompson, “The Akenti Access Control System: Attribute Certificate Generation”,Information and Computing Sciences Division, Ernest Orlando Lawrence Berkeley National Laboratory, University of California

[JPRS] Joon S. Park, Ravi Sandhu. Smart Certificates: Extending X.509 for Secure Attribute Services on the Web, 1999, The Laboratory for Information Security Technology, Information and Software Engineering Department, George Mason University.

[PRPLJ] Paulo Marques, Raul Fonseca, Paulo Simões, Luís Silva, João Silva,” Integrating Mobile Agents into Off-the-Shelf Web Servers: The M&M approach”, CISUC, University of Coimbra, Portugal.

FraMaS: Fraud Management System for Online Credit Card Transactions

K.D.C. Anuradha and K. De Zoysa
Department of Communication and Media Technologies,
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7,
Sri Lanka.
Phone: (94–1-) 2589123, Fax: (94-1-) 2587239

Abstract

Cyber Frauds are becoming major factor in financial losses. Hence the importance of the fraud prevention and detection are also becoming a hot topic in the industry worldwide. Therefore it is obvious that Internet transactions should be intelligently watched by a fraud detection system to minimize the risks. Although there are such systems available in the foreign market those solutions might not fit into our local scenarios where purchasing patterns and buying frequencies of local Internet customers vary from foreign customers. Hence, in this paper we present a ***Fra****ud* ***Ma****nagement* ***S****ystem called FraMaS, which is suitable for local purchasing patterns and buying frequencies.*

Keywords: Credit Card, Online Payment, Cyber Fraud, Electronic Transaction

1. Introduction

The amount of Internet transactions are being increased in *Sri Lanka* with the introduction of Internet Payment Gateways (IPG) in *local banks and financial institutions.* This will also increase the risk of frauds causing grater losses for the participating merchants. Therefore, it is obvious that Internet transactions should be intelligently watched by a fraud detection system to minimize these risks. The major challenge and the constraint in such system are to identify and differentiate the legal and genuine transactions from that of fraudulent. The system should not hamper the performance of the existing transactions while capturing the suspicious users.

Even though there are such fraud management systems developed world wide, they cannot be applied in the Sri Lankan context since the transaction patterns of the local users are not similar to their foreign counterparts. Thus, we developed **Fra**ud **Ma**nagement **S**ystem (FraMaS) suitable for Local Payment Gateways.

FraMaS calculates the estimated risk/fraud score for the transactions originating from existing Payment Gateways. It uses rule base mechanism for the risk calculation based on pre-identified fraud cases likely to occur in the local context. The likely hood of transaction being a fraudulent transaction is determined by using a naïve Bayesian classifier and thus FraMaS predicts the likely probable lost of the transaction.

2. Credit Card Frauds

It was evident that with advent of the Internet and its infrastructure in early 90's the revenue of online shopping and cyber transactions was increased. Business community realized this trend and started expanding their business over the web. However, then came the intruders who start exploiting the loop holes in these systems and started doing illegal transactions causing greater losses and reducing the consumer confident and faith on online transactions. The losses were on the rise and this make the experts to think of fraud management systems. The statistic from below indicates the seriousness of this issue [1].

With the analysis of existing frauds, customer complains and surveys it has been found that the criminal use of credit cards can be divided into the following categories:

- **Counterfeit credit card use:** This the major form of credit card fraud, involving losses with virtually all issued cards, with 47% of all dollar losses. Very well organized criminals with the help of hi-tech equipments have acquired the technology that allows them to "skim" the data contained on magnetic stripes, manufacture phony cards, and overcome such protective features as holograms.

Credit Card Losses 1999-2001			
	1999	**2000**	**2001**
Lost	20,967,361	18,815,619	17,614,803
Stolen	24,914,922	24,702,625	25,361,473
Non receipt	19,560,715	6,233,718	5,238,058
Fraudulent applications	9,002,216	9,612,764	23,858,324
Counterfeit	123,640,477	81,142,438	66,312,727
Fraudulent use of account	25,250,660	23,440,085	29,770,630
Misc	3,402,781	8,561,527	14,549,549
TOTAL	$226,739,133	$172,508,776	$182,705,564

Table 1: Credit Card Fraud Activity
Source: Payment Card Partners Committee

- **Cards lost by or stolen from the cardholder:** Lost and stolen cards represent 25% of all card fraud losses. Typically the cards are stolen from the workplace, vehicles, health clubs, golf clubs, etc.
- **Fraud committed without the actual use of a card (no-card fraud):** No-Card Fraud accounts for 13% of the all losses. Deceptive Telemarketers and now fraudulent Internet Web Sites obtain specific card details from their victims, while promoting the sale of exaggerated or non-existent goods and services. This in turn results in fraudulent charges against victims' accounts.
- **Fraud committed on cards not received by the legitimate cardholder (non-receipt fraud):** Non-Receipt Fraud where cards are intercepted prior to delivery to the cardholder account for 4% of all losses. Losses attributable to mail theft have declined as a result of "card activation" programs, where cardholders must call their financial institution to confirm their identity before the card is activated. In 1992, this category accounted for 16 % of the losses.
- **Cards fraudulently obtained by criminals who have made false applications:** Fraudulent Applications involve the criminal impersonation of creditworthy persons in order to acquire credit cards. Although false application losses represent only 6% of all losses, the numbers are increasing.
- **Other:** There are several other criminal uses of credit card which account for 5% of all card fraud losses.

3. Fraud Detection Mechanisms

As the fraud rates increase people start review the transactions before completing it. However, with the increase of transaction volumes automated fraud detection systems were integrated to the existing transaction systems. The techniques used by such systems are,

- Real-Time Authorizations
- Negative/Positive Files
- Rules/Exceptions
- Pattern Detectors
- Statistical Models
- Hybrid Solutions

Each of these methods has their pros and cons as described in the table 2 [2].

Fraud detection involves complicated analysis and processing of large amount of historical data, thus causing considerable delay in transaction processing. This in turn affects the speed and efficiency of genuine transactions causing greater inconvenience to customers. By using statistical approach this can reduce to certain extent due the nature of theory of statistics [3].

Most of the companies who have expose their business to the Internet, are awash in data from the digital revolution. Whenever a customer buy groceries, clothes, gasoline, electronics, pay a bill, surf the web or make a telephone call, a record is generated and stored in a database somewhere. Companies engaged in fraud detection, need to use this data to develop and fit mathematical models of consumer behaviour.

Detection Technique	**Pros**	**Cons**
Real-Time Authorizations	– Immediate Response – Reliable If configure properly	– Additional processing overhead on transactions – Not always reliable
Negative/Positive Files	– Defense against repeat offenders – Protects orders from good customers	– Good customers cards can become compromised

Detection Technique	**Pros**	**Cons**
Rules/Exceptions	– Allows merchant to apply expert knowledge – Highly customizable – Easy to determine what flagged the transaction	– Requires constant monitoring to ensure that the rules are effective – Rules are only as good as the people building them
Pattern Detectors	– Provide protection against card number generators – Can help identify fraud when each individual transaction seems normal – Could be highly automated	– Good customers could be locked-out – Good transactions could be trapped in suspicious sequences
Statistical Models	– Leverages historical fraud data to catch new fraud attempts – Risk score is determined by evaluating numerous factors simultaneously – Catches subtle patterns that would typically be overlooked by a human	– Requires ample, accurate, cleansed historical data (which most merchants don't have) – Since multiple factors contribute to the risk score, it is sometimes difficult to interpret the score

Table 2: Comparison of Detection Techniques

Considering the above facts it was decided to use a Hybrid solution where Rule Based and statistical approach (Bayesian) were combined to detect and minimize frauds in our approach.

Bayesian statistics simplifies the science of fitting models to data, which makes it easier to deal with realistically complicated models. Bayesian methods are revolutionizing marketing, finance and other business applications including credit card scoring (whether to grant you credit and, if so, what limit), detecting telephone call fraud and computer network security. The Bayesian rule says that one should update the probability of some event in light of new evidence.

Since the Bayesian uses subjective probability which says that probability can be used to measure one's subjective uncertainty about unknown quantity or event, it best fits for fraud detection where subjective uncertainty of transaction could be calculated given the evidence of failure of certain parameters [3].

Bayesian methods have also changed the face of computer network security and the detection of credit card fraud. Companies would monitor the spending on the card and anomalous behavior can be seen as a possible sign of fraud. Bays' Theorem is crucial in assessing the likelihood of fraud given the spending patterns.

4. System Architecture

The FraMaS communicates with existing payment gateway system and facilitate monitoring of transactions.

Real time monitoring screen for all transactions: This screen shows information on all transactions as they happen. Details of the merchant, the transaction code, date/time, and transaction amount are shown to an authorized user. There is no facility to block or suspend any transactions from this view.

Real time fraud screening mechanisms: This function allows the bank to create sensitivity levels and parameters to screen for what might be fraudulent transactions. Then FraMaS have an option of taking action if the parameters are triggered. For example, FraMaS can flag a transaction and alert the audit staff for further monitoring or suspension of the card, or merchant, auto suspends a transaction based on sensitivity levels.

Real time monitoring screen for card base: This function shows the available cards in the system and its details with emphasis given to the fraud score.

Real time monitoring screen for Merchants: This function shows the available merchants in the system and their details with emphasis given to the fraud score.

Real time fraud screening is based on a fraud scoring mechanism. This fraud scoring mechanism makes use of transaction patterns studied over a period of time to determine the fraudulent transaction patterns that deviate from the norm. These transactions are assigned different fraud scores depending on the different types of traces and the weight of such scores is dependent on the acquiring environment. The system built with intelligence to self-learn the statistical norms for trace parameters and correlation techniques to determine the deviation of such parameters present in transactions.

4.1. Screening Mechanisms

The following fraud screening mechanisms are used in an acquiring environment as the basis for tracing fraudulent transactions.

Risky Merchants Trace: Certain merchants have higher fraud risk than others. Those that sell downloadable software or expensive items like jewelry and electronics get hit the hardest. Merchants with goods that are easily transferable for cash are more prone to fraud. The risk values for each merchant are incorporated to their profiles and the merchants can be flagged in the system. If very high transactions occur or the same credit card is used to purchase from the merchant in rapid fashion the transactions will be highlighted. The bank will be able to set a rapidity value so that the system can begin to auto suspend a high fraud risk transaction.

High Value Transactions Trace: Fraudulent transactions tend to have higher transaction values than the usual transactions. The fraud scoring engine will compare the learned mean of the transaction values in the system and assign a higher fraud score for such transactions that deviate significantly from the mean. These transactions will be highlighted on an alert screen. The bank can decide on the upper bounds of a transaction and the system will suspend transactions above that upper bond.

Transaction Frequency Trace: The transaction frequency can be used as a valid parameter in assigning fraud scores for transactions. If several transactions are taking place in a small time span with the same card number, those transactions need to be assigned a higher fraud score. The transaction rate needs be learned by the system over the time and the system should alter this figure automatically to suit the seasonal high transaction rates.

Time of Day Transaction Pattern Trace: Each merchant has a statistical transaction pattern depending on the time of day. As an example for local merchants, who do not deliver goods overseas a transaction taking place at 2.00 am in the morning will carry a higher score than a day time transaction, if no previous transactions had taken place at that merchant site between 2am to 3am for the last two months. Based on these patterns, the fraud-scoring engine assigns higher scores for the transaction patterns that deviate from the usual patterns to detect any fraudulent behaviour.

Transaction Origination Trace: Transaction origination IP trace is used as a basis for fraud detection in this category. As an example if transactions pertaining to many different card numbers originate from the same IP address those transactions need to be assigned higher fraud scores as such scenarios usually depict fraudulent behavior.

Foreign BIN Trace: The BIN of the card number in the transaction can be traced to detect fraudulent transactions. As an example if a transaction takes place at a local merchant site who does not deliver goods overseas and if that transaction carries a foreign BIN, that transaction is more likely to be a fraudulent transaction and thus needs to be assigned a higher fraud score.

Declined Authorization Trace: The declined transactions due to invalid, unavailable or expired card numbers should be traced and logged in to the system with origination address, merchant ID, time and card number. The preceding transactions from the same origination address or merchant ID should be assigned higher fraud scores, as this would indicate an attempted fraud.

Address and Name Verification Trace: The address of the cardholder is traced against the card number and preceding transactions with same credit card with different address need to be assign higher fraud value.

The system built with the intelligence to identify different transactions take place in a fraud attempt and depending on their correlation and group them in to fraud cases.

The system has a facility to define different types of corrective actions in case of a suspected fraudulent transaction. These include passive alerts in the form of a report, active alerts in the form of a popup dialog box.

4.2. Security Settings and Parameters

System allows the bank to set the security settings and threshold values for each fraud cases. The following fraud cases were identified and are used in our system:

1. Transaction Amount is greater than merchant's Maximum Transaction Limit
2. Transaction Velocity exceeded
3. Identical Transactions with high velocity
4. Transactions originating from black listed IP
5. Local customer with foreign IP
6. Foreign customer with Local IP
7. Foreign customer with foreign IP
8. Transaction from blacklisted card numbers
9. Invalid card holder name
10. Transactions via high risk merchants
11. Transactions originating from high risk merchants

The system allows define the weights for each of the above fraud cases , the administrator can also set weights for the individual card numbers observing the past transactions done using that card. The system also automatically increases the fraud weight values of each credit card number by examining the number of fraud failure cases against such card numbers.

When a transaction is done the system checks whether the fraudulent cases are violated if so get the fraud weight values for the card number involved in the transaction. If fraud weight values are not set for the card number the default values are taken. Finally, system calculates the fraud score of the transaction and save the failure cases of the transaction.

Frauad Value of the transaction=∑(Violated Fraud Case weight X Fraud Value for the Card)

The transaction is then analysed using the naïve bays classifier to determine the probability of it being a fraud.

Naïve Bayes Classifier

Suppose that each instance

$x \in X$ is a sequence of attribute values $(a1,a2,a3 ..an)$ and that the discrete valued target function we are trying to learn is $f:X \rightarrow V$ where $V=\{v1,v2, \ldots vn\}$.

A set of *m* training instances of the form $<xi,vi>$ is presented followed by a new instance *x*. What is the most likely value of $f(x)$?

The system also allows setting up and maintains following system specific parameters:

- **Email Address** to be notified for high risk transactions
- **Thresh hold value** for alerts
- **Bin numbers** of local bins
- **Black listed** country codes

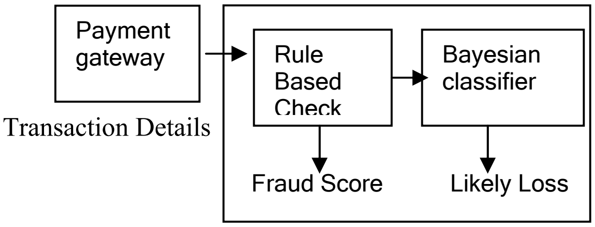


Figure 1: Overall System Architecture

Figure 1 shows the overall system architecture. The system has been designed as web based application and can be accessed via the Internet or cooperate Intranet. Since the fraud management system should be able to integrate with any payment system the web based solution is the ideal option compared to the traditional standard along systems.

5. Evaluation

Evaluation and the success of the system measured against by simulating the series of test transactions with different attributes and seeing how the system response to such transactions and also with the feed back from the client.

In whatever forms it comes, the subsequent consequences of any fraud are similar. It has social and economical impacts on the industry. Thus, the success of fraud management systems also can evaluate in terms of its financial and social gains.

The interview held with the customer (The hosting bank of a payment gateway) revealed that on average 100-200 transactions are done online using credit cards, and it has been observed that nearly 2 percent of those transactions are reported as frauds. If it is assumed that average transaction amount is Rupees 1000 the expected loss would be,

Loss=1000*200*0.02=Rupees 4000
=>4000*365=Rupees 1.46 Million per year

The above figure shows that within year the system has a return on an initial investment. Since the operational cost is minimum, the return gain on long run is unbelievable.

Here are some of the transaction types that have identified as frauds by the bank. Since there are not any fraud detection system the losses had to be born by the poor merchant.

Case 1: A fraudulent user has acquired the information of a credit card and he has promised to pay the mobile phone bill of others by only getting the half amount of the billed value and then paying the whole bill value with a stolen credit card. He has continued this fraud until the real cardholder had made a complain. The unfortunate merchant, in this case the mobile phone operator has lost nearly Rupees ten thousand due to this act.

Had the FraMaS software been installed, this fraud would have been identified at very early stage since the Fraud case 2 (Transaction velocity exceeded) would have been constantly violated causing alerts and high fraud score values for transaction.

Case 2: This is a case where very well organised group has used an embossing machine at a super market and embossed credit card numbers in fake cards which looks similar to original ones. These cards have then been given to foreigners who have used to pay services bills from other countries.

Had the FraMaS software been installed this fraud would have been identified at very early stage since the Fraud case 5 (Local customer with Foreign IP) would have been constantly violated causing alerts and high fraud score values for transaction.

Case 3: This was case where a person possessing an illegal card had tried to purchase an expensive ornament from a merchant store. He has been so wise to inform the merchant to collect from the merchant stall without giving his shipping address in the online transaction.

Had the FraMaS software been installed this fraud would have been identified at very early stage since the Fraud case 1 & 11 (Merchant Transaction Amount Exceeded, High risk merchant) would have been constantly violated causing alerts and high fraud score values for transaction.

These were some of the identified fraudulent intrusions but there are lot of other cases the bank was reluctant to reveal due to the security reasons. In additions, there would be some fraud cases unnoticed by the bank and the merchant until actual card holder deny the transaction.

Despite the FraMaS helps client and the hosting parties to track and minimise the risk and losses of frauds, following precautionary measures should be considered for better protection [5],[6]. These measures should be followed by all the parties involved in on line transactions. i.e. Merchant ,card holder's acquiring bank ,issuing bank etc.

- Always insist on a mailing address, zip or postal code and phone number of the buyer and then check them out to ensure they aren't fake.
- If possible try to get the customer signature and a faxed photocopy of the credit card (from a photocopy is fine).
- In case if you can't contact the buyer or he is unreachable, then don't process the order.
- Always try to use Address Verification services where they're available.
- Be extremely wary of shipping overseas - it can be hard to pursue claims abroad. Eastern Europe is seen by many as a high risk area.
- Check the email address against the name on the credit card. If the real name doesn't match the email name then you definitely want more reassurance before processing the order.
- Refuse to process orders from free email domains unless you have incontrovertible proof of the buyer's identity.
- Never ship products to postal box numbers. Always insist on a physical shipping address.
- Check the DNS table of the remote IP of the customer. Find out the remote server's geographic area and check it against the address of the customer. Few people connect to the Net using a long distance call. You also need to check the mailing address, phone number and email address of the server, though thieves can also set up servers too.
- Be especially careful of those wanting higher priced fast delivery or otherwise being price insensitive. Thieves don't care how much it costs as they don't plan to pay.

6. Conclusions

It has been the tradition of Sri Lankan customers to have a second thought before using an online payment system to pay their bills and services. On top that if they happen to know and experience a fraud against their transactions erodes the confidence and move them away from using such online systems. Customers are distraught when fraudulent charges show up on their statement.

Therefore, implementing the proposed Fraud Management System (FraMaS) at local IPGs will boost the confidence and minimise the losses to the all parties involved in the online business in Sri Lanka.

FraMaS has some draw backs and short comings which could be considered as the improvements for the future releases.

The major area of concern was given to the selecting of predefined rules to be checked against the transaction attributes, and definitions of these rule sets are always debatable since a dump software system could only check for these rules blindly so a smart hacker can easily by pass these rules and complete transaction as if it were a genuine transaction. This can be improved by introducing self learning and deducing system which will lean from the patterns of the transactions. For example, a neural network or data mining system could be incorporated with the FraMaS to analyse and learn the previous buying patterns of a user [7].

References

[1] "5th annual online fraud report (Credit card fraud tends and merchant response)", Midwave Research, Cybersouce Cooperation, 2004

[2] Julie Fergerson, "Fighting for online frauds", Clear Commerce, [Online] Available at http://www.ClearCommerce.com

[3] Steven L. Scott, "How Theory of Probability Can Help your business", Information & Operations Management University of southern California, USA, 1999

[4] James painter, "Use Of Bayesian statistics in Fraud detection", Tessella support services plc, April 2003

[5] Steve Patient, "Reducing online frauds", [Online] Available at http://www.webdevelopersjournal.com

[6] "Minimizing credit card Frauds", University of Utah [Online] Available at http://www.scambusters.org/ccfraud.html

[7] R Brause, M Hepp, "Credit card fraud detection by adaptive neural data mining", J.W.Goethe-University, Frankfurt, 1999

Strategic eSecurity for eSri Lanka:
Design and Development of Information Security Governance (ISG) Model

T J Pathirage
MBA(Sri J.), B.Com(Sri J.), CISA(USA), CISM(USA), CCSE(USA), AIB (SL), BS7799 Lead
DIR/CEO of CISS, Assistant General Manager – Seylan Bank Ltd

Abstract

Information Security is of critical importance in achieving e-Sri Lanka vision. It is a key to extending the enterprise to enable deep integration of organizations either public or private and to deliver services to government agencies, partners, suppliers and customer or citizens. The eGovernance infrastructure is woven with the Internet based technologies in order to grasp the opportunities and to leverage cost of ICT investment. The rewards are not free but attached with inherent risks, threats, and vulnerabilities, which may have dramatic impact of core business activities of the organizations.

A credible security architecture, reliable technologies and a good ICT (Information and Communication Technologies) Security governance system are vital in constructing security infrastructure support system for successful implementation of eSri Lanka Projects. A building of a trust mechanism using PKI technology is an essential component of the security infrastructure. Unfortunately, there is no one best method or the world class standard to manage ICT Security risk in emerging technological environment. Given the risks and rewards the ICT Security is best viewed, not solely as technologies but as a corporate governance issue in the information economy. Investment on ICT and Internet technology should be tight to actual business risk in order to achieve maximum value and rational allocation of resources both by government agencies and private sector organizations.

This paper focuses the framework that is necessary to systematically integrated ICT security into corporate governance system of the country and manages the risk at strategic level. The eSri Lanka demands sound framework for information security risk management framework..

1. Introduction

The **e**Sri Lanka vision is driven by the ICT and Information Systems that enable sharing of information and wider access to information repositories and services both in public and private sector organizations. ICT is a foundation medium for the equitable distribution of opportunity and knowledge within societies and communities and a key determinant of the competitive advantage of the Nation. Also in the context of global business environment, the significance of information is widely accepted, and information systems are truly pervasive throughout business and governmental organizations. The growing dependence of most organizations on their information systems, coupled with the risks, benefits and opportunities ICT carries with it, have made ICT governance an increasingly critical facet of overall governance. Boards and management alike need to ensure that ICT is aligned with enterprise strategies, and enterprise strategies take proper advantage of ICT. The new reality is that businesses and governments are in a constant balancing act, extending deeper access to their information assets and services while also complying with a myriad government regulations around information privacy and corporate governance. Information security and risk management is therefore, critical in providing such integration of information systems at national level.

The public sector, private sector and citizens frequently experience the consequences of inadequate information security. Information Security breaches are an increasingly common occurrence. Carnegie Mellon's, CERT, noted that there were 137,529 security incident reports in 2004, a 68% increase over the 82,094 reported in 2003. And, according to the 2004 CSI/ FBI Computer Crime and Security Survey the theft of proprietary information caused the greatest financial loss amongst respondents at $201.3 million. Although there seems to be a decline of financial loss due to computer security incidents to $ 141.4 in Years 2004, the attack methods have become complex and sophisticated (FBI/CSI, 2004). The denial of Service category emerged for the first time as the incident type generating the largest total loss (Replacing the theft of proprietary information for consecutive five years). Many national governments have recognized the importance of security, establishing initiatives to reinforce such measures as safeguarding infrastructures according to their sensitivity, investing in better authentication methods and making users of the infrastructure accountable for their actions.

Also in recent years, the Cyber security has emerged as a critical security issue in national information systems and infrastructures. As commonly used, *cybersecurity* refers to three things: measures to protect information technology; the information it contains, processes, and transmits, and associated physical and virtual elements (which together comprise *cyberspace*); the degree of protection resulting from application of those measures; and the associated field of professional endeavor.

2. IS Information Security Governance as Issue?

The industries have been experiencing major business scandals during the past two decade and as such the corporate governance has become an important issue. Defined, corporate governance is the set of policies and internal controls by which organizations are directed and managed. Information Security Governance (ISG) is a subset of corporate governance that relates to the security of information systems. Information security is all too often perceived as a wholly technical issue. For companies, educational institutions, and non-profit organizations to make progress in securing their information assets, however, executives must make information security an integral part of core business operations. The best way to accomplish this goal is to highlight ISG as part of the internal controls and policies that constitute corporate governance. This paper examine the strategic information security frameworks and models that help the formulation of Information Security Governance (ISG) model for both government and private sector organizations, which would ensure securer infrastructure and socio technical environment for sustainable competitiveness of eSri Lanka

IT Governance Defined

While numerous definitions exist for IT governance, the following two definitions will be used in this article IT governance is the responsibility of the Board of Directors and executive management. It is an integral part of enterprise governance and consists of the leadership and organizational structures and processes that ensure that the organization's IT sustains and extends the organization's strategy and objectives. value through investments in IT (*Board Briefing on IT Governance*, IT Governance Institute, 2001, *www.itgi.org)*

> IT governance is the organizational capacity exercised by the Board, executive management and IT management to control the formulation and implementation of IT strategy and in this way ensure the fusion of business and IT (Van Grembergen, 2002)

The definition from the IT Governance Institute states that IT governance is an integral part of enterprise or corporate governance. Indeed, to make sure that corporate governance matters are covered, IT first needs to be properly governed. IT governance should be an integral part of enterprise governance and, in this respect, a primary concern of the board of directors that is responsible for governing the enterprises. Boards may carry out their governance duties through committees and by considering the criticality of IT through an IT strategy committee.

3. Definitions of Information Security and Risk

Security relates to the protection of valuable assets against loss, misuse, disclosure or damage. In this context, "valuable assets" are the information recorded on, processed by, stored in, shared by, transmitted or retrieved from an electronic medium. The information must be protected against harm from treats leading to different types of vulnerabilities such as loss, inaccessibility, alteration or wrongful disclosure. Threats include errors and omissions, fraud, accidents and intentional damage. Protection arises from a layered series of technological and non-technological safeguard such as physical security measures, background checks, user identifiers, passwords, smart cards, biometrics and firewalls. These safeguards should address both treats and vulnerabilities in a balanced manner.

The objective of information security is ***"protecting the interests of those relying on information, and the systems and communications that deliver the information, from harm resulting from failures of availability, confidentiality and integrity"*** (IFA, 1998). While emerging definitions are adding concepts like information usefulness and possession-the latter to cope with theft, deception and fraud-the networked economy certainly has added the need for trust and accountability in electronic transactions such that for most organizations, the security objective is met when:

- Information is available and usable when required, and the systems that provide it can appropriately resist attacks and recover from failures **(*availability*).**

- Information is observed by or disclosed to only those who have a right to know (***confidentiality***)
- Information is protected against unauthorised modification ***(integrity).***
- Business transactions as well as information exchanges between enterprise locations or with partners can be trusted ***(authenticity and non-repudiation).***

Anything that harm above principle is considered as the risks to information Security.

4. The Importance of Information Security for E-Sri Lanka

Information systems can generate many direct and indirect benefits, and as many direct and indirect risks. These have led to a gap between the need to protect systems and the degree of protection applied. The gap is caused by: Widespread use of technology, Interconnectivity of systems, Elimination of distance, time and space as constraints, Unevenness of technological change, Devolution of management and control, Attractiveness of conducting unconventional electronic attacks against organizations, External factors such as legislative, legal and regulatory requirements or technological developments. This means that there are new risk areas that could have a significant impact on critical business operations, such as:

- Increasing requirements for availability and robustness
- Growing potential for misuse and abuse of information systems affecting privacy and ethical values
- External dangers from hackers, leading to denial-of-service and virus stacks, extortion and leakage of corporate information
- Privacy of personal information

4.1 Security Strategy

The Information security strategy shall plays a key role in driving the eSri Lanka imitative (Policy on e-Government ,Preliminary draft, 2003

- Secure electronic environment will encompass all aspects of security and trust pertaining to e-systems, e-transactions, e-mail and information/data and personnel (users, providers and developers) to enable government institutions to work together with each and with businesses and citizen and governments securely

- Relevant access controls and mechanisms will be built for secure e-transactions.

- Data exchange standards requiring encryption and digital key technology (Public Key Infrastructures (PKIs) will be specified for adaptation by institutions for secure transaction environment together with the infrastructure for implementation.

- Measures will be taken to Endeavour success of projects in building the trust of the users on e-technology.
- Establish the environment for information/data privacy enabling the participants to respect the legitimate interests of others so that the citizen could participate in e-g with security and dignity.

- Steps will be taken to build up an ethical conduct as security systems alone cannot bring about the required security without the support of the participants. Ethical conduct is therefore crucial and participants should strive to develop and adopt best practices and to promote conduct that recognize security needs and respect the legitimate interests of others supported by the higher authorities.

- The security of information systems and networks will be made compatible with essential values of a democratic society. Security will be implemented in a manner consistent with the values recognized by democratic societies including the freedom to exchange thoughts and ideas, the free flow of information, the confidentiality of information and communication, the appropriate protection of personal information, openness and transparency.

- Security measures will be built into the process reengineering phase to facilitate reverting to manual systems in the event of failure/disaster depending on the nature of the transactions.

- Collaborative programmes will be setup with International bodies dealing with security (CERT/CMU) to protect infrastructure from cyber threats of misuse and hacking and to minimize risk from security vulnerabilities.

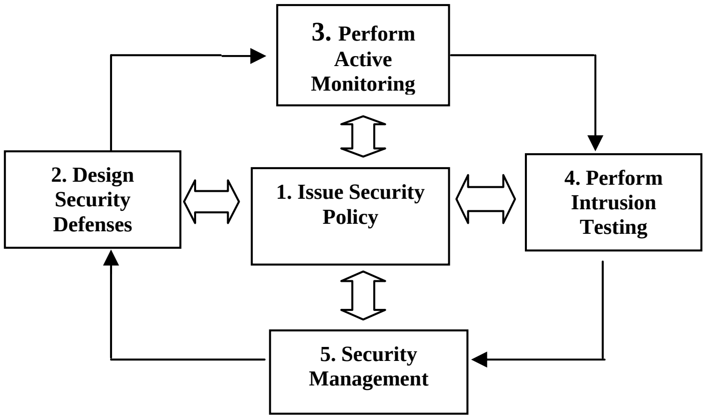
At the strategic level, ICTA is working on the following security initiatives to achieve the above security strategies.

- E - Security policy, strategy, standards and Guidelines
- Secure E-mail messaging strategy, standards and Guidelines.
- Internet Security Policy, strategy, standards and Guidelines.
- Information Security Policy, strategy, standards and Guidelines.

5. Literature Review

5.1 ISG Management Models

Information Security needs to be managed similar to the other business function of the organizations. Over the years industry has been developing many such frameworks and the most recognized IS management formwork is published by the IT Governance institute depicted below


3. Perform Active Monitoring
2. Design Security Defenses
1. Issue Security Policy
4. Perform Intrusion Testing
5. Security Management

Source: IT Governance Institute, Information Security Governance, 2001

Figure 1: ISG Process Framework

The relative priority and significance of availability, confidentiality, integrity, authenticity and non-repudiation vary according to the data within the information system and the business context in which they are used. For example, integrity is especially important relative to management information due to the impact that information has on critical strategy-related decisions. (refer Figure 1).

5. 2 Legislative and Regularity Demand for ISG Framework

In response to threats, governments around the world have enacted legislation at the regional and government levels. Although no local legislations are available in this area, the adaptation of international best practices is critically important. However, organizations are faced with a maze of regulations with IT security elements, including:

- US Sarbanes-Oxley Act
- US Health Insurance Portability and Accountability Act (HIPAA)
- US Gramm-Leach-Bliley Act (GLBA)
- US Federal Information Security Management Act (FISMA)
- Canadian Personal Information Protection and Electronic Documents Act (PIPEDA)
- EU Directive on Data Protection (EU Data Directive)
- California Security Breach Information Act (SB1386)

More regulation, however, is unlikely to improve the situation. Nor is relying solely on organizations' CIOs and IT departments to fix the problem. If we are to systematically strengthen information security, organizations must elevate the issue to a corporate governance priority. Specifically, CEOs, CIOs and Boards of Directors must integrate information security into their overall governance program at all levels of the organization. To accomplish this, organizations need a framework.

5.3 What Do Regulatory and Standards Bodies say

Financial regulators are instructing the banking industry to focus on operational and IT risk within which security and ICT are very significant (Basel II). All major past risk issues-they claim-have been caused by breakdowns in internal control, oversight in ICT. Organization may study the following best practice formworks in formulating their own ISG framework.

Table 1: Information Security Best Practices

1 **Guidance**	2 **Goals**
Control objectives for Information and Related Technology (CoBIT, 2004)	IT control objectives for day-to-day use
Organization for Economic Co-operation and Development (OECD)	Guidelines for the Security of Information Systems (1992)
The Systems Security Engineering Capability Maturity Model (SSE-CMM)	Software Security Quality
ITIL- IT Infrastructure	Vendor-independent approach

Library	for service management
ISO/IEC 17799:2005	Guidance for implementing Information Security
ISO/IEC TR 13335	Guidance on aspects of IT security management
ISO/IEC 15408	Definition of criteria for evaluation of IT security
TickIT	QMS for software development and certification criteria
NIST 800-14	Baseline for establishing and reviewing IT security programs
National Cyber Security Partnership (NCSP) recommendations	Corporate Governance- Task Force of the National Cyber Security Partnership (NCSP) in USA.

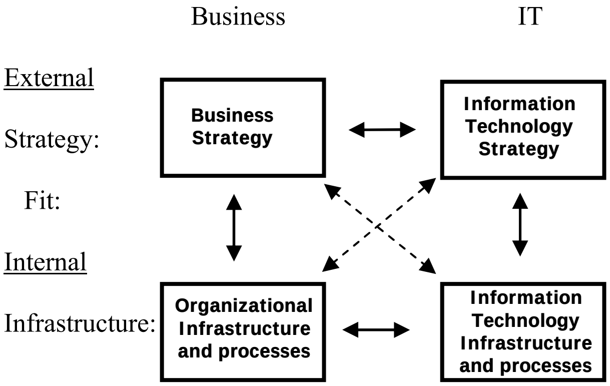
6. What Should Information Security Governance Deliver

Information security governance, when properly implemented, should provide the following four basic outcomes, which would lead the organization for sustainable competitive advantage;

i. Strategic Alignment
ii. Value Delivery
iii. Risk Management
iv. Performance Measurement

6.1 Strategic Alignment of ICT with IS Governance Framework

An important element of IT governance is the alignment of IT with the business. J. Henderson and N. Venkatraman developed their strategic alignment model (SAM) to conceptualize and direct the area of strategic management of IT. They were the first to describe in a clear way the interrelationship between business strategies and IT. The model is based on two building blocks: strategic fit and functional integration (figure 2). Strategic fit recognizes that the IT strategy should be articulated in terms of an external domain (how the firm is positioned in the IT marketplace) and an internal domain (how the IT infrastructure should be configured and managed). Strategic fit is equally relevant in the business domain, with similar attributes but focused to the business. Two types of functional integration exist: strategic and operational. Strategic integration is the link between business strategy and IT strategy, reflecting the external components, which is important because, for many companies, IT has emerged as a source of strategic advantage. Operational integration covers the internal domain and deals with the link between organizational infrastructure and processes and IT infrastructure and processes.


Business
IT
External
Strategy:
Fit:
Internal
Infrastructure:
Business Strategy
Information Technology Strategy
Organizational Infrastructure and processes
Information Technology Infrastructure and processes

Source: Adopted from Henderson, J. C., and Venkatraman, N., "Strategic Alignment: Leveraging Information Technology for Transforming Organizations," 1993.

Figure 2- Strategic Alignment Model

The model explores the interrelationship between business and IS where effecting a change in any single domain requires the use of three of the four domains to ensure that both strategy fit and functional integration is properly addressed

Although the SAM model clearly recognizes the need for continual alignment, it does not provide a practical framework to implement this. However, over the years, many alignment mechanisms have been developed and are used in organizations to achieve the business/IT fusion: business systems planning, critical success factors, the competitive forces model and the value chain of M.E. Porter, and business

process reengineering. Recently, Porter adapted his models to the e-business (e-commerce) phenomenon concluding that "the Internet *per se* will rarely be a competitive advantage" and "many of the companies that succeed will be ones that use the Internet as a complement to traditional ways of competing, not

Those that set their Internet initiatives apart from their established operations."

7. Information Security Risk Management

The risk management of ICT investment is widely accepted with emerging threats and vulnerabilities of IS systems. As news of break-ins and losses related to hackers, computer viruses and other Internet-based threats grows more frequent, enterprise stakeholders are becoming concerned about the risks, regulatory requirements and information security. Their need for assurance is putting the issue firmly in the lap of executive management and enterprise boards. The operational risk management is an emerging issue in the financial services industry in Basel 11 context.

Effective security is not just a technology problem, it is a business issue. Related risk management must address the corporate culture, management's security consciousness and actions. Sharing of information with those responsible for governance is critical to success. An information security programme is a risk mitigation method like other control and governance actions and should therefore clearly fit into overall enterprise governance. ICT governance itself is emerging as a subject matter and integral part of enterprise governance, with the goal of ascertaining that;

- ICT is aligned with the business, enables the achievement of business goals and maximises benefits
- ICT resources are used responsibly
- ICT related risks are managed appropriately.

The literature review reveled that there is no single global information Security Governance standard or framework acceptable across the countries or industries. The information security industry is growing fast and evolving with industry dynamics. The selection of the correct framework is a choice that best suit the delivery of business mission in the given business and organizational context.

8. Benefits of Implementing the ISG Framework

Our research revealed that benefits derived by organizations that implement the ISG framework go beyond facilitating compliance with applicable legislative, regulatory and contractual requirements. ISG and its associated information security program also result in tangible business benefits, including:

- Improved internal processes and controls:
- Potential for lower audit and insurance costs:
- Market differentiation through a continuous improvement process:
- Self-governance as a better alternative than regulation:

9. Stakeholders of Information Security Governance

Generally, information security has been dealt with as a technology issue only, with little consideration given to enterprise priorities and requirements. Responsibility for governing and managing the improvement of security has consequently been limited to operational and technical managers. However, for information security to be properly addressed greater involvement of boards of directors, executive management and business process owners is required. For information security to be properly implemented, skilled resources such as information system auditors, security professionals and technology providers need to be utilised. All interested parties should be involved in the process.

10. Role of the Board and Management in Developing ISG

Boards and management have several very fundamental responsibilities to ensure that information security governance is in force. They should understand why Information security needs to be governed (IT Governance Institute, "Information Security Governance", USA, 2001)

- Risks and threats are real and could have significant impact on the enterprise.
- Effective information security requires co-ordinated and integrated action from the top down.
- ICT investments can be very substantial and easily misdirected.
- Cultural and organisational factors are equally important.
- Rules and priorities need to be established and enforced.
- Trust needs to be demonstrated toward trading partners while exchanging electronic transactions.
  - Trust in reliability in system security needs to be demonstrated to all stakeholders.
  - Security incidents are likely to be exposed to the public.
  - Reputation damage can be considerable

10.1 Board Level Action

This includes: Become informed about information security; Set direction, ie. Drive policy and strategy and define a global risk profile; Provide resources to information security efforts; Assign responsibilities to management; Set priorities; Support change; Define cultural values related to risk awareness; Obtain assurance from internal or external auditors; Insist management makes security investments and security improvements measurable, and monitors and reports on programme effectiveness

10.2 Management Level Action

The six broader categories of actions that need to be implemented for effective ISG formwork (refer Figure) are :

- **Policy Development**-Using the security objective and core principles as a framework around which to develop the security policy.
- **Roles and responsibilities**- Ensuring that individual roles, responsibilities and authority are clearly communicated and understood by all.
- **Design**-Developing a security and control framework that consists of standards, measures, practices and procedures.
- **Implementation**-Implementing the solution on a timely basis, then maintaining it.
- **Monitoring**-establishing monitoring measures to detect an ensure correction of security breaches, such that all actual and suspected breaches are promptly identified, investigated and acted upon, and to ensure ongoing compliance with policy, standards and minimum acceptable security practices.
- **Awareness, training and Education**-creating awareness of the need to protect information, providing training in the skills needed to operate information systems securely, and offering education in security measures and practices.

11. Recommendations

There are some fundamental steps boards and management can take to ensure effective information security governance is implemented in their enterprise;

11.1 At the Board of Director Level

- Establish ownership for security and continuity with enterprise managers.
- Create an audit committee that clearly understands its role in information security and how it will work with management and auditors.
- Ensure that internal and external auditors agree with the audit committee and management how information security should be covered in the audit
- Require that the head of security report progress and issues to the audit committee
- Develop crisis management practices, involved executive management and the board of directors from pre-agreed thresholds onward.

11.2 At the Executive Management Level

- Establish a security function that assists management in the development of policies, Procedures, standards and guidelines and assists the enterprise in carrying them out
- Create a measurable and management-transparent security strategy based on benchmarking, maturity models, gap analysis and continuous performance reporting.
- Conduct an annual executive risk brainstorming session, prepared by security and audit professionals (internal and external), resulting in actionable conclusions and followed up until closure.
- Develop what-if scenarios on information security and risk, leveraging the knowledge of the specialists.
- Establish clear, pragmatic enterprise and technology continuity programmes, which are continually tested and kept up-to-date.
- Conduct information security audits based on a clear process and accountabilities with management tracking closure of recommendations.
- Develop clear policies and details guideline, supported by a respective and assertive communications plan that reaches every employee.
- Constantly assess vulnerabilities through monitoring system weaknesses (CERT), intrusion and stress testing, and testing of contingency plans.
- Make business processes and supporting infrastructures resilient to failure, especially targeting single points of failure.
- Establish security baselines and rigorously monitor compliance
- Run security responsiveness programmes and conduct frequent penetration tests.
- Harden all security and critical server and communications platforms by applying a high level of control.
- Base authorization on business rules and match the authentication method to the business risk.
- Include security in job performance appraisals and apply appropriate rewards and disciplinary measures.

11. 3 National Cyber Security Partnership (NCSP) Proposed Core Set of Principles for ISG :

Task Force recommendations are targeted for both industry and government adoption and champion better ways of providing; measuring and maintaining cyber security Recommendations focus on:

- Broadening recognition and adoption of existing standards and best practices;
- Furthering the use of existing capabilities through common software security configurations;
- Investing in federal research toward the development of better vulnerability analysis or "code scanning" tools that can identify software defects;
- Developing guidelines for secure equipment deployment and network architectures; and,
- Improving the "Common Criteria" process, used by vendors and customers to develop security specifications and conduct security evaluations.

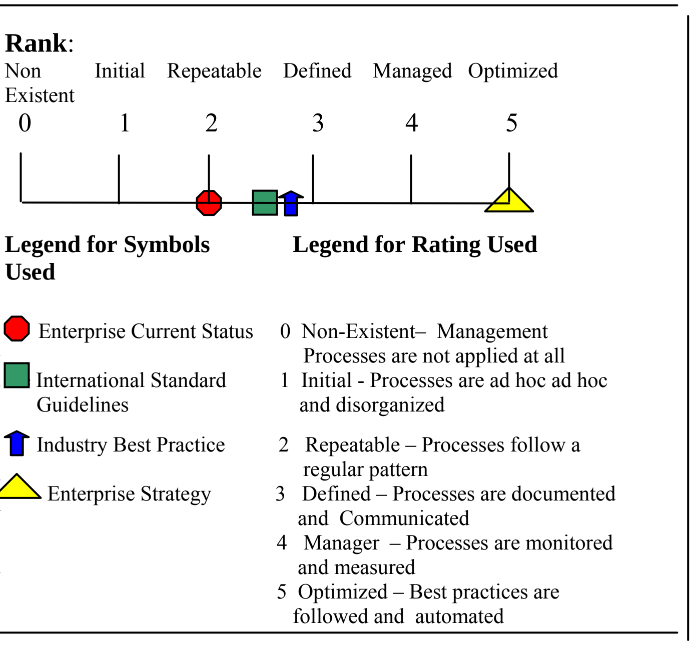
Also the following best practices were recommended by the NCSP which would provide sound practice for Sri Lankan context.

- CEOs/CIOs should have an annual information security evaluation conducted, review the evaluation results with staff, and report on performance to the board of directors.
- Organizations should conduct periodic risk assessments of information assets as part of a risk management program.
- Organizations should implement policies and procedures based on risk assessments to secure information assets.
- Organizations should establish a security management structure to assign explicit individual roles, responsibilities, authority, and accountability.
- Organizations should develop plans and initiate actions to provide adequate information security for networks, facilities, systems and information.
- Organizations should treat information security as an integral part of the system lifecycle.
- Organizations should provide information security awareness, training and education to personnel.
- Organizations should conduct periodic testing and evaluation of the effectiveness of information security policies and procedures.
- Organizations should create and execute a plan for remedial action to address any information security deficiencies.
- Organizations should develop and implement incident response procedures.
- Organizations should establish plans, procedures and tests to provide continuity of operations.

A well plan ISG programme must identify the critical success factors for monitoring a and subsequent performance evaluation purposes

12. Comparing the Maturity Level of the Organization with ISG

Boards of directors and executive management can use an information security governance maturity model (table 2) to establish rankings and benchmarking of security strategy in an organization. This model can be progressively applied as:


Rank:
Non Existent
Initial
Repeatable
Defined
Managed
Optimized
0
1
2
3
4
5
Legend for Symbols Used
Legend for Rating Used
Enterprise Current Status
International Standard Guidelines
Industry Best Practice
Enterprise Strategy
0 Non-Existent– Management Processes are not applied at all
1 Initial - Processes are ad hoc ad hoc and disorganized
2 Repeatable – Processes follow a regular pattern
3 Defined – Processes are documented and Communicated
4 Manager – Processes are monitored and measured
5 Optimized – Best practices are followed and automated

Source: IT Governance Institute, Information Security Governance 2001

Table 2: Information Security Maturity Model

- A method for self-assessment against the scales, deciding where the organization is
- A method for using the results of the self-assessment to set targets for future development, based on where the organization wants to be on the scale, which is not necessarily at the top level
- A method for planning projects to reach the targets, based on an analysis of the gaps between those targets and the present status
- A method for prioritizing project work based on project classification and an analysis of its beneficial impact against its cost.

13. Conclusion

ISG is no more a technical issue and it has to be treated as a governance level of both public and private sector organizations. Security ICT assets and providing secure access are critical in successful implementation of eGovernment initiatives. The effective ISG framework proves the strategic alignment for the organizations. However, the opportunities and the challengers are to be managed strategically in harnessing opportunities while managing the risk at acceptable level. eSri LANKA There is no world-class standard or one best way of doing this and the approach needs to be the organizational specific which take into consideration the global and country specific business ethics and good governance. The guidelines and the frameworks discussed in this paper provide best practice to public and private sector organizations in formulating eSecurity governance model in achieving and sustaining the eSri Lanka vision.

References

1. American Institute of Certified Public Accountants/Canadian Institute of Chartered Accountants, ***SysTrust Principles and Criteria for Systems Reliability*** V2.0.2001

2. Claudo, C., ***IT Governance why a Guidelines,*** Information Systems Control Journal Volume 3, 2003

3. Conner F., W., ***"Implementing Information Security Governance (ISG) Entrust Securing Digital Identities & Information,*** July 2004.

4. Corporate Governance Report, National Cyber Security Summit Task force, ***"Information Security Governance- A Call to Action"*** National Cyber Security Partnership www.cyberpartnership.org, 2004

5. Gordon, L., A., Loeb, M., P., Lucyshyn, W., and Richardson, R., ***"2004 CSI/FBI Computer Crime and Security Survey",*** Computer Security Institute, 2004-08-24

6. Information Security Survey 2002, ***Global Information Security Survey,*** Ernest & Young , 2002

7. Information Technology Association of America www.itaa.org

8. International Federation of Accountants, The International guidelines for Managing Risk of Information and Communications statement #1: ***Managing Security of Information***, 1998

9. International Standard Organization, ISO 17799, BS 7799-2, 2000

10. IT Governance Institute, ***"Information Security Governance",*** USA, 2001.

11. IT Governance Institute, CoBICT (Control Objectives for Information and related Technology) 3rd Edition, 2000, www.Itgivernance.org and www.isaca.org.

12. TechNet www.technet.org

13. White Paper, "***Information Security Governance (ISG), An Essential Element of Corporate Governance"***, Entrust, 2003- 2004.

14. Ernst and Young, **Global Security Survey**, 2003.

15. Henderson, J.; N. Venkatraman; ***"Strategic Alignment: Leveraging Information Technology for Transforming Organizations,"*** IBM Systems Journal, 1993,

16. Board Briefing on IT Governance, IT Governance Institute, 2001, www.itgi.org

17. Van Grembergen, W.; ***Introduction to the Minitrack IT Governance and Its Mechanisms,*** Proceedings of the 35thHawaii International Conference on System Sciences (HICSS), 2002

18. Government of Sri Lanka, Preliminary Draft ,***Policy on E-Government. Innovative e-government for empowered citizen.,*** , 2003

A Comprehensive Attachment Repository for E-Mails

H. Ekanayake
Department of Computation and Intelligent Systems
University of Colombo School of Computing
35, Reid Avenue, Colombo 7, Sri Lanka.

E-mail: hbe@ucsc.cmb.ac.lk

Abstract

Email has become one of the crucial medium of communication for today's information society. As a result email system has improved in various ways to provide more privacy, capacity and accessibility. However the attention that has been gathered for improving the manageability of the content is very less. This has become evident when someone wanted to search for an attachment, sent by the person x, six months ago. The effort becomes worst if that person has already deleted the message to preserve his mail box capacity in a favorable limit. This paper examines the current email framework and proposes a global solution to enhance the easy retrieval of email attachments without introducing any improvements to email client programs. The solution uses the descriptive power of email messages to retrieve keywords to annotate email attachments, and later, one could easily locate a document recalling the memory he has had when the document was an attachment to an email message.

Keywords: Electronic Mail, Email, Attachment, Repository, Search

1. Introduction and Motivation

Today a major part of communication happening over the world has grasped by electronic mail, or email. Lot of factors has determined its demand, such as, its low cost and its speed of communication. Email systems are available mainly in two forms, POP mail and web mail. The latter has now becoming very much popular since it does not require additional client software to be configured in a local computer other than an Internet browser. In addition, email service is provided in many forms, such as, free email service from Yahoo or Gmail, corporate mail, or packages from email service providers.

Earlier email was suffered as an unreliable communication medium due to its absence of providing a guarantee for delivery or security while in transit. However, over the past few years most of these shortcomings were fixed after introducing lot of features into the email system [1, 2]. A recently addressed drawback of email communication is the guarantee for delivery or the repudiation issue. This problem has been successfully resolved after introducing a notarization service for email system [3, 4, 7]. As a result today emails are used by government agencies and other organizations for their formal communication needs. Another advantage of using emails for such purposes is its demanding accountability. The header part of an email message provides a full description of the message, such as, to whom it has been sent and by whom, the constructed date of the message and a short summary of the subject. Emails also provide a facility to attach files and as a result important documents are exchanged over inter-organizations or intra (within)-organization through emails. This trend has created a situation where people do not tend to delete old messages from their mailboxes, since they think that these messages and their content might be useful in the future. The result has created a big problem; people are running out of their mail box capacities!

2. Potential Solutions

Literature was unable to show any similar system that could provide an easy attachment retrieval facility. However, the web disk facility [5], provided by today's web mail systems, has some potential of providing lot of related services. Web disk is a separate storage of hosting email attachments and documents temporary. Since its storage counts against the capacity of one's web mail capacity, the documents cannot be saved for longer periods. Furthermore, there is no comprehensive search facility provided for these documents.

Yahoo! Desktop Search [6] is a recent attempt that has been launched by Yahoo! Inc to enable online and/or offline search facility to find items, such as, files, email messages, email attachments, instant messages and contacts related to some user. While it perform the searching, it looks for items in Microsoft Outlook, Outlook Express, Yahoo! Messenger or other personal folders for some pre-defined item types. Then it presents the results in a much organized interface, so that one can easily identify the required item(s). This can be seen as an improvement to traditional search facilities provided by operating systems, where in addition to searching files in the local computer the new search looks into email programs. However none of the solutions are using the power of email messages discussed in this paper.

3. Aims and Objectives

The following aims and objectives have been identified for implementation:

- [O1] Provide a transparent service to the users: This is to prevent additional software installations on user's local computers.
- [O2] Enable the service for wide spectrum of email systems: Anyone who is using POP mail, web mail or any other email system should be able to receive the service.
- [O3] Multiple email accounts capability: A user can has more than one email accounts; however there should be only one attachment repository instance for that user.
- [O4] Capture as much as keywords from an email message to provide a much more elaborative searching facility.
- [O5] Provide a comprehensive results sheet as a response to a search.
- [O6] Provide an alternative way to attach documents to an email message, so that the sender has some control over the attachment even after the message has been sent.
- [O7] Reliable and secure service to the users.

4. Global Attachment Repository Approach

4.1. Advantages of Email Attachments

The following are some of the advantages that could be gained by preserving the documents as attachments to email messages:

- **It enables availability of the document in global scope**: Since most email services are globally available, the same accessibility applies for the attachments since the attachments are also part of an email message. Therefore if someone sends a document to some other person, by attaching it to an email message, the second person gains the ability to download the document to his computer despite of his location (at his working place or at his home) using the same steps.
- **Attachments are sufficiently described**: Attachments come bounded to an email message. The header part of that message contains information to describe who has been sent the message, who are the recipients, the date the message has been sent, subject line containing a summary of the message and other optional information (priority, path). The body of the message might contain information describing the attachments.
- **Attachments are touched! Therefore one can easily remember them**: Normally no one ignores an incoming email message, unless it is a spam. Since an email message is observed, the observer sees the content of that message including any attachments. This attempt triggers the observer's memory to remember the email message. Later, when he wants to find a document, which has been arrived to him as an email attachment, immediately he could recall the memory to trace the sender, date, subject line … and would surf his mail box to get the email message and then the document.
- **Reliability**: Locally stored documents are very much intensive to loose, because of unorganized habits, unintentional deletions, modifications and other misconducts. However, the potential to loose an email attachment is very less.

4.2. Interacting with the Global Attachment Repository

This section mainly focuses on the techniques which have been used to fulfill the aims and objectives identified at the beginning of the work.

Figure 1 depicts the technique which has been used to preserve the objectives O1, O2 and O3 (provide a transparent service to the users, enable the service for wide spectrum of email systems and multiple mail account capability).

If a user wants to send an email message to some other party, thinking that the attachment(s) would be useful for him in the future, all he has to do is to mention the GAR server email address (sec@ucsc.cmb.ac.lk) in the BCC address field of that email.

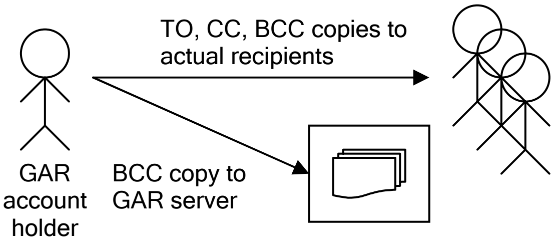


Figure 1: Posting an email

Here BCC address field is chosen, since email addresses mentioned under this field are hidden to the recipients (however there is no restriction to mention the email address under TO or CC address fields). This technique avoids the requirement of additional software to be installed on client computers in advance and as a result users who are using any type of email system can receive this service.

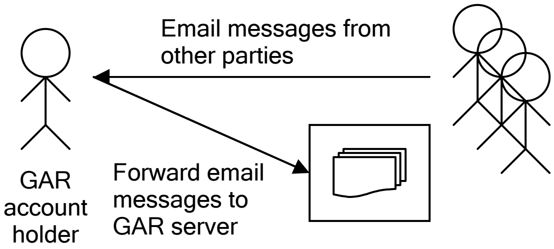

Email messages from other parties
GAR account holder
Forward email messages to GAR server

Figure 2: Forward email messages received from other parties

Figure 2 depicts the steps needed on a situation where the user receives email messages from other parties. However the user wishes to store them under the GAR server. The steps in this situation are also simple as forwarding the required email messages by mentioning the GAP server's email address as the recipient.

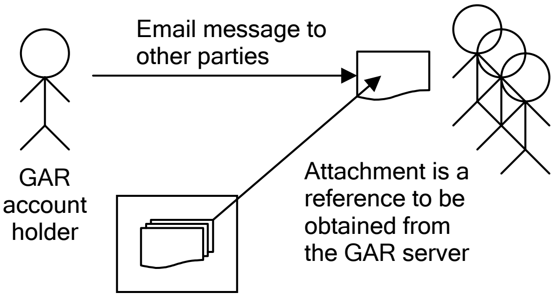

Email message to other parties
GAR account holder
Attachment is a reference to be obtained from the GAR server

Figure 3: Inline URLs to download attachments

Figure 3 depicts how the objective O6 (Provide an alternative way to include attachments) is fulfilled. The expectation of this objective is to reduce the size of an email message by offloading the attachments, but provide an alternative way to get those attachments when needed. The solution is to provide inline links (URLs) on the body of the message to download the attachments from a remote location. In this method the sender first sends the documents he wishes to attach to an email message to the GAR server in advance to constructing the actual email. Then the GAR server stores the documents in its repository and responds the sender with an email message containing links (URLs) to those documents. Later, without modifying those links (URLs) the sender fills out the body with his message and posts it to the actual recipient(s). The recipient(s) then follow the provided URLs to download the attachments. This method also adds an extra control over the attachments to the sender, for instance, the sender can modify or change the attached documents without the knowledge of recipient(s).

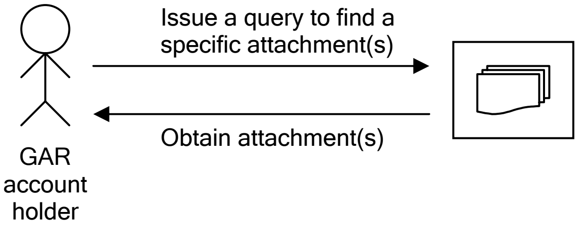


Figure 4: Searching for saved attachments

The most important feature of this architecture is the ability to retrieve a document that was an attachment to an email message, by tracing ones memory associated with it, which is depicted in Figure 4. The user issues a query using the options provided by the GAR server web interface. The server then responds with a nicely organized results-set where the user can easily identify the expected documents and download.

The GAR server captures as much as keywords from the email messages arriving to its mailbox to provide an elaborative search service to its users. The provided information would include the date, recipients, subject, a short summary of the content, details of the attachments and a hierarchical conversation like history if possible.

A SSL or similar secure channel could be used to implement a transport level security during the session of user interacting with the web interface. The GAR server also drops any unsolicited emails that come from unregistered users.

5. Prototype Implementation

The system is designed based on an architecture comprised of modules which is depicted in Figure 5.

POP Connector: This module is responsible for connecting regularly to a preconfigured POP mail server

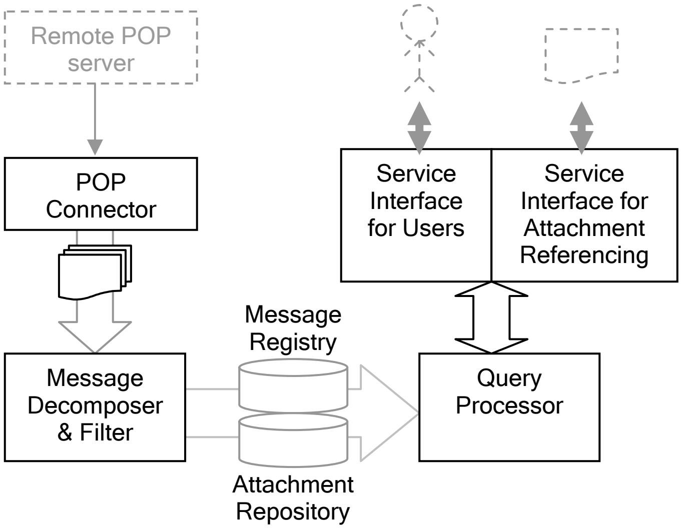

Remote POP server
POP Connector
Message Decomposer & Filter
Message Registry
Attachment Repository
Query Processor
Service Interface for Users
Service Interface for Attachment Referencing

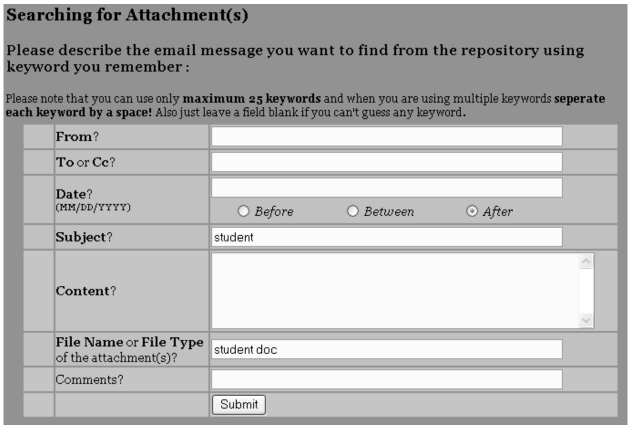
Figure 5: System architecture

to retrieve newly arrived email messages (as a result of copying or forwarding). The downloaded messages are then forwarded to the *Message Decomposer & Filter* module to process.

Message Decomposer & Filter: This module first extract the header part of a message to retrieve the details associated with *From*, *To*, *Cc*, *Date* and *Subject* fields. The address associated with the *From* field is validated against the registered users of the system. The verified messages are further processed to see whether it has any attachments. If the message has attachments those attachments are extracted from the message and placed under a preconfigured attachment folder (*Attachment Repository*). The details of successfully processed messages are saved in the database (*Message Registry*). All the unsuccessful messages are discarded.

Service Interface for Users: Registered users are interacting with the system through the web-based interface provided by this module. When a user wants to search for an attachment(s) he has to describe the email message as much as he can, recalling his memory, in the search window depicted in Figure 6. As a result the user is presented the matches found in the results window depicted in Figure 7. The matches are ordered by its rank. The user can then download the attachment(s) simply by clicking the URL(s) of the document(s).

Service Interface for Attachment Referencing: This module responds to attachments whose URLs are placed in the body of an email message as depicted in Figure 8.



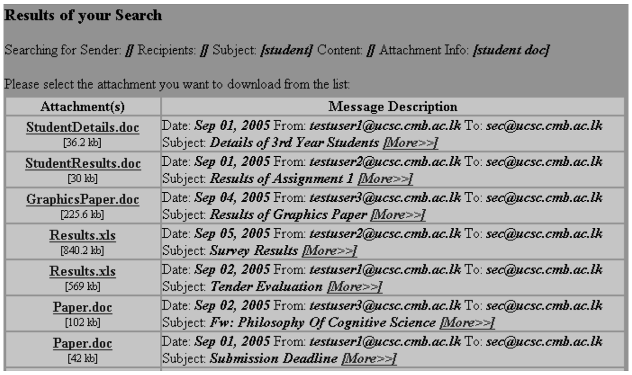
Searching for Attachment(s)

Please describe the email message you want to find from the repository using keyword you remember :

Please note that you can use only **maximum 25 keywords** and when you are using multiple keywords **seperate each keyword by a space!** Also just leave a field blank if you can't guess any keyword.

From?	
To or **Cc**?	
Date? (MM/DD/YYYY)	○ *Before* ○ *Between* ● *After*
Subject?	student
Content?	
File Name or **File Type** of the attachment(s)?	student doc
Comments?	
	Submit

Figure 6: Searching window



Results of your Search

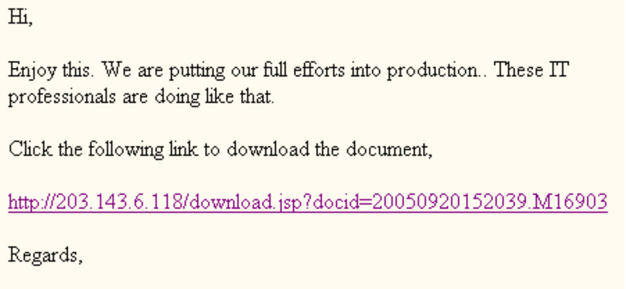
Searching for Sender: *[]* Recipients: *[]* Subject: ***[student]*** Content: *[]* Attachment Info: ***[student doc]***

Please select the attachment you want to download from the list:

Attachment(s)	Message Description
StudentDetails.doc [36.2 kb]	Date: ***Sep 01, 2005*** From: ***testuser1@ucsc.cmb.ac.lk*** To: ***sec@ucsc.cmb.ac.lk*** Subject: ***Details of 3rd Year Students*** *[More>>]*
StudentResults.doc [30 kb]	Date: ***Sep 01, 2005*** From: ***testuser2@ucsc.cmb.ac.lk*** To: ***sec@ucsc.cmb.ac.lk*** Subject: ***Results of Assignment 1*** *[More>>]*
GraphicsPaper.doc [225.6 kb]	Date: ***Sep 04, 2005*** From: ***testuser3@ucsc.cmb.ac.lk*** To: ***sec@ucsc.cmb.ac.lk*** Subject: ***Results of Graphics Paper*** *[More>>]*
Results.xls [840.2 kb]	Date: ***Sep 05, 2005*** From: ***testuser2@ucsc.cmb.ac.lk*** To: ***sec@ucsc.cmb.ac.lk*** Subject: ***Survey Results*** *[More>>]*
Results.xls [569 kb]	Date: ***Sep 02, 2005*** From: ***testuser1@ucsc.cmb.ac.lk*** To: ***sec@ucsc.cmb.ac.lk*** Subject: ***Tender Evaluation*** *[More>>]*
Paper.doc [102 kb]	Date: ***Sep 02, 2005*** From: ***testuser3@ucsc.cmb.ac.lk*** To: ***sec@ucsc.cmb.ac.lk*** Subject: ***Fw: Philosophy Of Cognitive Science*** *[More>>]*
Paper.doc [42 kb]	Date: ***Sep 01, 2005*** From: ***testuser1@ucsc.cmb.ac.lk*** To: ***sec@ucsc.cmb.ac.lk*** Subject: ***Submission Deadline*** *[More>>]*

Figure 7: Results window

The downloading of an attachment is a click on the relevant URL.



Hi,

Enjoy this. We are putting our full efforts into production.. These IT professionals are doing like that.

Click the following link to download the document,

http://203.143.6.118/download.jsp?docid=20050920152039.M16903

Regards,

Figure 8: A message with inline URL to attachment

Query Processor: The queries issued by the modules *Service Interface for Users* and *Service Interface for Attachment Referencing* are processed by this module. This module is responsible for analyzing the keywords provided by the user to identify relative importance of them, connecting to both *Message Registry* and *Attachment Repository* to retrieve required information, filtering them to obtain the most appropriate results set and presenting them in an orderly manner.

6. Conclusion

This paper has presented a simple solution to increase the manageability of attachments by developing a globally accessible attachment repository for emails with searching, organizing and other features. Moreover it has presented an implementation architecture for a prototype system. It can be anticipated that in the future these kinds of supplementary systems would be useful to increase the usability of current email framework.

Some improvements to this proposed architecture is discussed under the future work.

7. Future Works

This attempt is only a first step to a more comprehensive solution. The framework could be improved by adding lot of additional features; some of them are identified as below:

- The current system does not search inside the documents which are stored in the *Attachment Repository*. By enabling a full-text searching functionality this could be incorporated into the system. In addition, the searching functionality could be enhanced by incorporating a good probabilistic searching method to get the most appropriate matches and order them accordingly.
- The alternative attachment referencing functionality could be enhanced to make the documents downloadable only to predefined locations by detecting the IP addresses of the locations. This way one could make a document downloadable only to a corporate network, so no outsider can see the document even the email message has arrived to one’s mailbox accidentally.
- When the results of a search are displayed, they could be organized thematically, for instance, group the documents containing only jokes.
- There are situations where an email message represents an official letter. Then it is necessary to treat the body of that email message as a document and store it under the document repository.
- The security by means of encryption, and virus protection to the documents in the *Attachment Repository* could be incorporated to provide a more trusted service to the users.

References:

[1] "S/MIME Version 3 Message Specification", Request for Comments, [Online] Available at ftp://ftp.ietf.org/rfc/rfc2633.txt

[2] Karin Becker, Simone Nunes Ferreira, “Virtual Folders: Database Support for Electronic Messages Classification” Proceedings of the International Symposium on Cooperative Database Systems for Advanced Applications (CODAS), Kyoto, Japan December, 1996

[3] Hiran Ekanayake, Kasun De Zoysa, Rasika Dayarathna, “A Notarization Authority for the Next Generation of E-Mail Systems”, Proceedings of the International Information Technology Conference, December 2003

[4] “ReadNotify Email Notary and Timestamping Service”, [Online] Available at http://www.readnotify.com

[5] “Webmail Tutorial”, [Online] Available at https://ucsc.cmb.ac.lk/data/openwebmail/help/en/index.html

[6] Yahoo! Desktop Search, [Online] Available at http://desktop.yahoo.com

[7] H.E.M.H.B. Ekanayake, “Notarization Authority for Emails”, B.Sc. Dissertation, University of Colombo School of Computing, May 2003.

WS*o*JS – A Space based Web Services Architecture

A S Chandima and D N Ranasinghe
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7,
Sri Lanka.

E-mail: chandima_senevirathne@yahoo.com , dnr@ucsc.cmb.ac.lk

Abstract

This paper presents the design and implementation of a new architecture for Web Services, based on JavaSpaces. This architecture, known as a Space-based Web Service Architecture has superior performance qualities in load balancing, failure recovery and scalability. Evaluation proves the rationale for this type of Space-based Web Service Architecture for commercial applications.

1 Introduction

Web Services are developer-friendly services since it uses the well-known transport protocol (HTTP) and the base language is XML. Web services are based on set of standardized rules and specifications. Similarly a Web Service is an interface that describes a collection of operations that are network accessible through XML messaging. Further, web services are modular, self-describing, self-contained applications that are accessible over the Internet. It allows developers to build web-based application via any platform, object model, and programming language that developer wishes.

In Web Services, loose coupling is possible only if it does not use strict type models. But that does not imply that existing Web Services Model is completely loosely coupled. However it is the major fact, which drives more efficient Web Service Models. Another big challenge is Web Service Asynchrony. Though Microsoft .NET provides some kind of Asynchronous Web Services, it is not assured that it is completely asynchronous.

In our work, we introduce an alternative Web Service model with more benefits for the business world, than the existing Web Service architecture based on Javaspaces.

JavaSpaces technology is a high level coordination tool for gluing processes together into a distributed application which provides a fundamentally different programming model that views an application as a collection of processes cooperating via the flow of objects into and out of one or more spaces. A Space is a shared, network accessible repository for objects. Processes use the particular repository as persistent object storage and an exchange mechanism, instead of communicating directly and those processes coordinate by exchanging objects through space [4]. These are carried out with operations such as write(), read(), readifexist(), take(), takeifexist() and notify(). JavaSpaces has properties such as being shared, persistent, and associative and the ability to provide transactions, etc. As such it is one of the most effective space-based computing models in distributed computing and the use of its features in Web Services are likely to yield the following advantages for web services. Since space is shared, many web service providers can interact with many clients through spaces. This has relevance for p2p applications. Due to persistency, Web service requestor (client) need not be time coupled with the service provider. As such Service requestor and service provider need not be space coupled. Since the spaces are associative, third party assistance is not needed to search the web services in the network (i.e. via UDDI registry). JavaSpaces itself can do that mechanism.

2 Space–Based Web Service Architecture

The space based web services architecture attempts to exploit two features offered by the space namely, asynchronous access and loosely coupled communication. As a result, it can be shown that following properties such as absence of third party searching services, secure Web service access than existing conventional Web Service model through high security SOAP messages, dynamic client access management for same web service given by the different service providers and fault recovery for service provider’s failures will hold good for the space based webservices model.

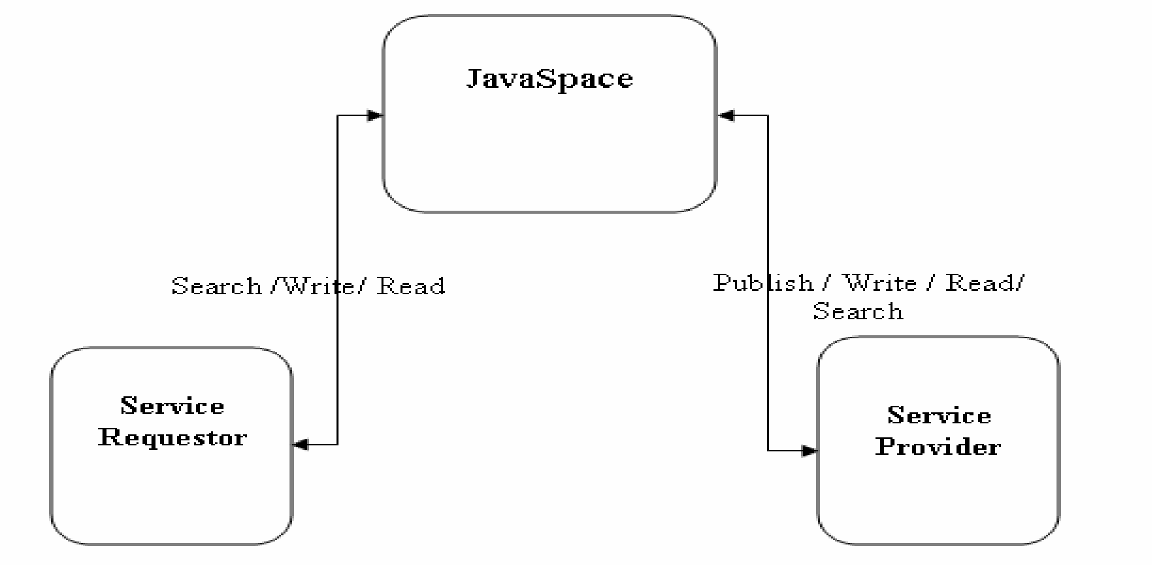


Figure 1: Components Interaction of Space-Based Web Service Model

The Space-based Web Services architecture is based on the interactions between three actors: Service Provider, JavaSpaces and Service Requestor. The interactions involve the publish, search, and write operations. Together, these roles and operations act upon the Web Services artifacts, the new Web service software module and its description. Typically, a service provider hosts a network-accessible (through JavaSpaces) software module (an implementation of a Web service). The service provider defines a service description for the Web service and publishes it in JavaSpaces as specific objects. The service requestor uses a search operation to find the service description from the JavaSpaces which verify the availability of particular service and also it uses the service description to know about the service and the service provider's details. JavaSpaces acts as a mediator, as well as a registry for services.

As defined in standard Web Service Architecture, service provider is the owner of the service from a business perspective. From an architectural perspective, this is the platform that hosts access to the service. Unlike the conventional architecture, here service provider publishes its services in JavaSpaces as objects (with its Service descriptions) Here Service provider is continuously searching for Service Request Objects from the JavaSpaces.

From an architectural perspective, service requestor is the application that is looking for and invoking or initiating an interaction with a service. In this system Service requestor search for Services in JavaSpaces. If it can find service object from space it writes the Service Request Object to JavaSpaces with soap request message and it can make read operation for retrieve Response Object through JavaSpaces.

The space is a searchable XML based object repository of service descriptions where service providers publish their service descriptions. Service requestors search services and obtain service information (as service descriptions) of services. As in conventional way, binding information given by service description is not needed for service requestor since there is no any connection between service provider and the service requestor. That description only used to generate related classes of Web Service. Similar to conventional architecture, to perform the operations of publish, search, and write in an interoperable manner, there must be a Web Services stack or high level architecture that embrace standards at each level.

3 Related work – XMLSpaces (Ruple)

Rogue Wave Software have recently introduced a new communication infrastructure that builds upon Web Services to provide for secure, loosely coupled, many-to-many, and document-centric communications over the Internet called Ruple. [16].

Ruple is an Internet shared memory space. It is based on concepts originally presented in Linda, a tuple spaces language developed at Yale University. Other implementations of tuple spaces include Sun's JavaSpaces and IBM's TSpace. Rogue Wave's Ruple is unique in that it stores XML documents rather than Java objects and that it is accessible over the Internet using standard Internet protocols such as HTTP and SOAP. It also offers a security model based on X.509 digital certificates. Applications can place documents in a Ruple Space and then retrieve them using an XML query expression.

A space can be located anywhere on the Internet, so spaces are easy to find using standard Internet protocols such as DNS. A space is also accessible as a Web Service. Spaces offer a flexible alternative for applications to communicate with each other, in contrast to the present connection-oriented technologies like CORBA or DCOM. These RPC-based approaches require that both applications be online and connected to be used; such a connection is analogous to a telephone call.

A key feature of Ruple is a completely simple programming model of only four methods and an equally simple state model. All entries are self-contained. Together these two properties make it extremely simple to replicate a Ruple Space, making it straight-forward to offer redundancy and high availability. While the Internet is a remarkably robust infrastructure, it also makes very few guarantees in the way of security and delivery. Ruple answers the question of how to build flexible, reliable systems on an infrastructure that makes such few promises.

Because of its ability to build asynchronous, loosely-coupled, multilayer, applications, Ruple revitalizes distributed computing in several ways by providing support for intermittently connected devices, making it possible to implement complex collaborative applications in a simple manner, and by providing extreme loose-coupling. The many possible uses for Ruple include workflow applications, simplified firewalls to support SOAP, transformation engines, portals, B2B applications, document routing and many others. [16]

The basis of Ruple is an XML Tuple Space implementation called the *Forum*. The Forum provides a foundation for applications that collaborate transparently and independently by selectively and securely exchanging a set of business documents. Following figure illustrate how the Ruple exchange its document through space.

So Ruple has more benefits for distributed computing such as simplicity, scalability, document centric, Asynchronous, loose coupling and Interoperability etc. Since Ruple is a general XML document exchange technology, it can be used to create a wide variety of collaborative applications across the Internet or within an organization. Simply, it can be used for application areas such as B2B Applications, Mobile Applications, XML Document Routing, Secure Document Exchange, Web Services Intermediary, etc. We contrast our design with Ruple in the next Sections.

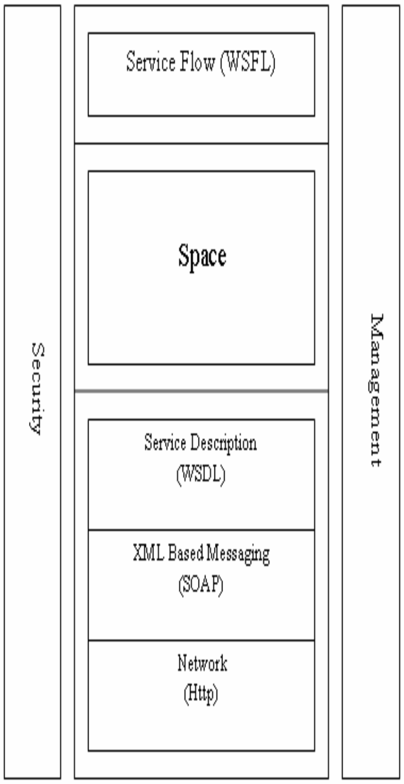


Figure 2: High Level Architecture

4 Design and Implementation

The foundation of the Web Services stack is the network. Web Services must be network accessible to be invoked by a service requestor. Web Services that are publicly available on the Internet use commonly deployed network protocols. The next layer, XML-based messaging, represents the use of XML as the basis for the messaging protocol. SOAP is the chosen XML messaging protocol for many reasons. The service description layer is actually a stack of description documents. First, WSDL is the standard description type for XML-based service description. This is the minimum standard service description necessary to support interoperable Web Services. The topmost layer, service flow, describes how service-to-service communications, collaborations, and flows are performed. WSFL is used to describe these interactions. Other than above conventional layers, this architecture has another layer called space. It has the capability to act as Service registry instead of registry layer. That is the major difference of this system with the conventional architecture.

4.1 Client

This System client consists of a Client program and Specific SOAP engine to interact with the JavaSpaces using objects with SOAP Messages. This specific SOAP engine is somewhat like the AXIS Soap engine but the only difference is that it has some mechanism to write SOAP request messages to objects which are compatible with the JavaSpaces. Then those objects are written to the Space using the same mechanism. Instead of direct XML serialized form of request which is sent to the network in the conventional web service access, here the engine can get the appropriate serialized SOAP request of the service call and write it as a SOAP object to the space. The basic view of the client is illustrated as follows.

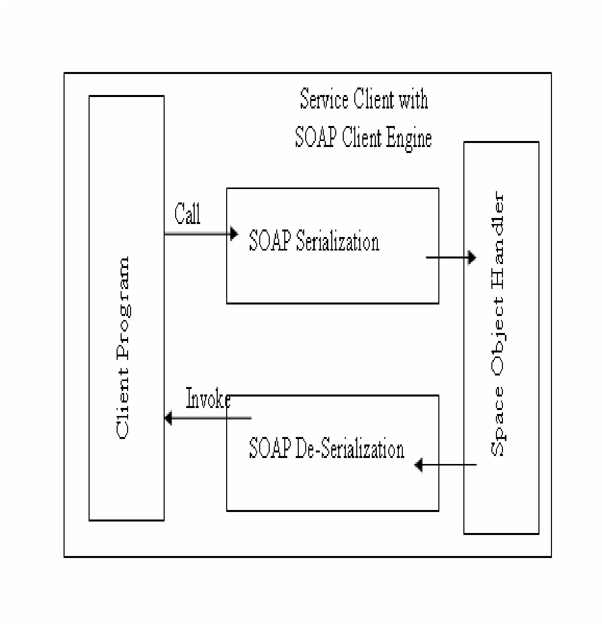


Figure 3: Client Side Design

Since the implementation of the SOAP engine is a complex task, Axis Client engine is used to manipulate the operations of SOAP serialization and de-serialization.

The methods of the Service Client represent main functionalities of the designed system for some extent. They include *findService()* which is used to search all Web Services, which are available in the Space. This search is done using the number of ServiceObjects in the Space. The method *requestService()* is used to write service request message to Space via RequestObject instance. At the time, client writes request to the Space ServiceObject is taken from the space for a moment. (Until the RequestObject writes to the Space). The method *invokeService()* is used to get a response message from the Space due to the corresponding request done using this method. The method *gerSeviceDetail()* receives WSDL definition and related details are done using this function. This WSDL document is used to generate classes related with Web service. ServiceObject pass this definition as its parameter.

4.2 Service Provider

As defined in the client's system design, Provider also has same sort of architecture, but with some slight differences. The basic view of the Server side design is illustrated via following figure.

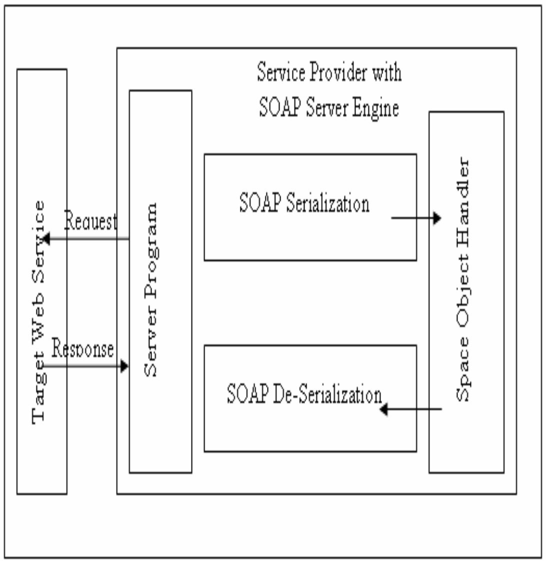


Figure 4: System Design - Service Provider

Service Provider is mainly used to publish the services and send Service Response. Also it includes continuous search process from Space to retrieve the service requests sent by the Service Requestors. The methods associated with the service provider are *publishService()* which is implemented as an alternative to the publish service functionality. Appropriate ServiceObject with WSDL is written to the Space, using this functionality. The method *executeServie()c* represents the execution of Web Service due to the corresponding request. Then the corresponding ResponseObject is written to the Space, after execution process is finished.

The system is implemented in a networked PC environment based on Windows XP operating system, jini 2_0_1 middleware, Apache Tomcat Web server, Apache Axis SOAP engine and the JDK 1.4.2 runtime environment. The objects used to transfer the data through JavaSpaces are implemented as serialized objects, using interface called *net.jini.core.Entry* which is given by Jini2_0_1 libraries.

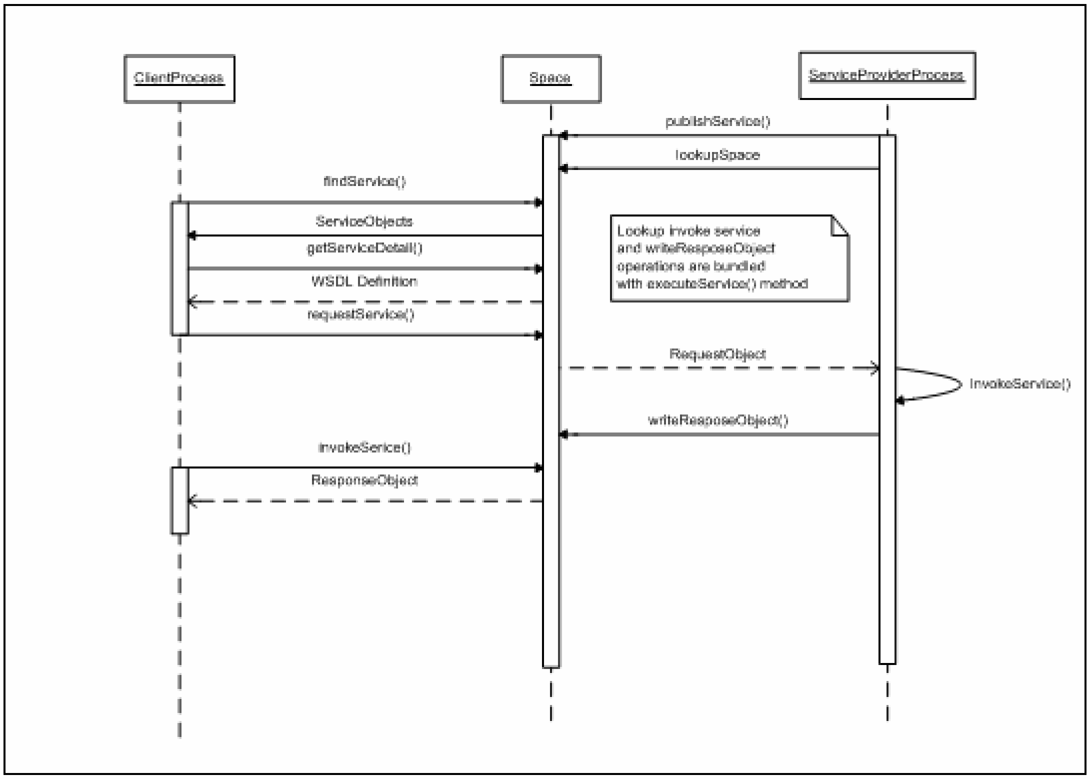


Figure 5: Sequence Diagram - Overall System

5 Performance

The following special features of the space based web services model can be identified.

5.1 Scalability

The system allows growth of the number of requestors and providers by centralizing the objects in the Space. When the number of clients and providers grow, this system has the capability of handling the increasing volumes of transactions whilst fulfilling its high performance requirements.

Although, JavaSpace acts as a centralized object coordination component, multiple parties can access space simultaneously. Because of this facility of the JavaSpace, multiple clients can request service from the Space at the same time and providers can publish and respond to clients concurrently through Space. Each party who access the Space, has connection with the space only if they execute one of few methods in JavaSpaces. Otherwise they are connectionless components. But there is a connection oriented situation only if they use notification service form the Space. (e.g., notify())

The Space handles the ServiceObject with WSDL. This architecture manipulates those published services, better than the UDDI registry activity in the conventional architecture. So, large numbers of Web Services can be published in the space with out having any doubt about their availability.

5.2 Load Balancing

When two or more service providers publish the same Web Service in Space, Service Requestors or Clients can access any of these ServiceObjects which consist of the WSDL of the service, provided by any of those service providers. Here, it assumes that ServiceObjects which has the same service name does not consist of the different type of WSDL definitions.

At the time the Service Requestor request some service from space, it takes the ServiceObject a moment and after reading relevant information from that object, rewrites it to the Space. So at the time one client read the ServiceObject others are unable to access that object (only for that moment).

Since JavaSpace uses its own Web Server, this prototype has the capability to be accessed over the intranet as well as Internet (Since it uses HTTP protocol). Unlike in the conventional architecture, this system is facilitated to find the same type of services for different clients, when they call for the same services concurrently. i.e The system is able to balance the load of accessing the same service by getting response from different providers who provide the same service.

5.3 Fault Tolerance

When a problem occurs in the Web Service in the Service Provider's side, it automatically takes out the published ServiceObject from the Space. Then, that Service is no longer available for clients.

In this implemented prototype, the Client can manipulate the fault recovery process by requesting that service. i.e. Suppose a Client sends some RequestObject due to some ServiceObject for the Space. And when the Service Provider finds some problem with the Web Service invocation, it takes the corresponding ServiceObject from Space and rewrites that RequestObject to space for a response by any other Service Provider which provide that requested service. So Client has guarantee that answer will definitely come from somewhere.

The Conventional Web Service Architecture does not support this type of mechanism for fault recovery in a Service Provider.

5.4 Decoupled communications

To find a Web Service through the Space, clients need only the name of the Web Service. A client does not need to know about the location of the Service provider as well as availability of the Service provider. At any time a Service Provider starts, its search operation about the requests in Space, responses are written to the Space. Associative lookup facility given by the Java Spaces, has given vital assistance to do this decouple communication

5.5 Portability

In this implemented prototype, Service Provider acts as a temporary client (virtual) and executes the Web Service in conventional ways. Because of that it has been proved, how easy it is to merge this kind of architecture with conventional model.

Finally, in the conventional architecture, SOAP Engine is a vital aspect for both client and the provider (such as AXIS Client and Server Engines). Unlike the conventional way, here in our implementation, the responsibilities of the XML engine are a little different. The following requirements could be implemented to come up with the more complete implementation of the system. When the Requestor gets a ServiceObject with the WSDL file there should be a mechanism to process this WSDL file and convert it to an actual language implementation. When the Requestor requests any service, the client side engine should generate the SOAP request message and write it to the space after serialize it. Also when requestor gets the response SOAP message it should de-serialized by the client engine. Similarly, the provider's Server engine should have the ability to de-serialize the request messages, which are found from the Space and rewrite the responses SOAP message after completing the Web Service invocation process inside the server engine.

6 Conclusion

In this work, we have proposed, designed and implemented a new web services architecture. The space based webservices model utilizes the decoupling and persistent properties of Javaspaces to achieve a 'asynchronous' form of Web Services, the elimination of the UDDI registry, and other properties such as load balancing, fault recovery and scalability.

References

[1] Armstrong E, Ball J, Bodoff S, Evans I, Green D,*The J2EE^{TM} 1.4 Tutorial, 2004*

[2] Chandima A. S. *Webservices on top of Javaspaces*, BSc in Computer Science dissertation, University of Colombo School of Computing, May 2005

[3] Colgrave J, *UDDI & OGSI* Research paper, April 2003

[4] Freeman E, Hupfer S, Arnold K *JavaSpaces Principles Patterns, and Practice*, Sun publications 1999

[5] Foster I., Kessalmen C, Nick J. The Philosophy of Grid; *An Open Grid Services Architecture for Distributed System Integration*. Globus Project 2002; http://www.globus.org/research/papers/ogsa.pdf

[6] Gulathi A, Murray P, *The Evolving Ecosystem of Web services, Application integration and E-commerce framework*- 2002

[7] Jong I; *Web Services/SOAP & CORBA*, April 2002

[8] Kreger H. *Web Services Conceptual Architecture* (WSCA 1.0) IBM Software Group, May 2001

[9] Maximilum E. M, Singh M.P, *Reputation And Endorsement for Web Services*, 2003

[10] Misra J, *Loosely-Coupled Processes;* University of Texas, Department of Computer Science 2000

[11] Radage K. *Web Services And Mobile Interoperability*; Oracle research Paper January 2004

[12] Sun Microsystems Ins, 2004, *Web Services Description Language* (WSDL) note;

[13] Sun Microsystems Ins, 2004, *Web Services Performance comparing J2EE & .NET framework;*

[14] Tartanoglu F, Issarny V, Romanovsky A, Coordinated Forward Error Recovery for Web Services, 2003

[15] Tartanoglu F, Issarny V, Romanovsky A, Dependability in the Web Services Architecture, 2003

[16] Thompson P, Rogue Wave Software, *Ruple: XML Space Implementation*, 2002

[17] Thompson P, Rogue Wave Software, *XML Spaces Beyond Web Services*, 2001

[18] Tuecke S, Czajkowski K. *Open Grid Service Specification*; July 2002 http://www/gridforum.org/ogsi-wg

[19] Tuecke S., Foster I, *Open Grid Services Infrastructure* (OGSI) version 1.0; June 2003; http://www.ggf.org/ogsi-wg

[20] UDDI publication, *UDDI Specification*, Open Draft Recommendation Document 2003; version 3.0; http://www.UDDI.org/pubs/uddi_v3.html

[21] World Wide Web Consortium, Simple Object Access Protocol (SOAP 1.1), W3C Note, 2000. http://www.w3.org/TR/soap

[22] World Wide Web Consortium, Web Services Description Languages (WSDL 1.1), W3C Note, 2001. http://www.w3.org/TR/soap

Secure Dynamic Group Establishment Protocol for Web Services

C. Atupelage[1] and K. De Zoysa[2]

Department of Communication and Media Technologies,
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7, Sri Lanka.
Phone: (94–1-) 2589123, Fax: (94-1-) 2587239
E-mail: [1] chamidu.atupelage@mazarin.lk
[2] kasun@cmb.ac.lk

Abstract

As the rate of new technology introduced to the world increases, the type of disruption growth is also high. There are several researches and technologies based on web services recently, however there is no secure protocol framework definition to configure a dynamic collaborative group of web services. Therefore this paper proposes very new approach to make shared security in between members of the web services group. The approach is presented as a new protocol capable of stabilizing a Dynamic collaborative Web Service group. This protocol is named as the Web Service Key Management Protocol (WS-GKMP).

The paper will discuss mainly the definition of the WS-GKMP, implementation of the protocol as an API, a sample application developed using the WS-GKMP API, and evaluation of the protocol.

Keywords: Web Services, Group Communication, Group Security, Key Distribution, GSAKMP

1. Introduction

Web services offer certain benefits over the other technologies making them suitable for e-business. They are faster and cheaper to develop, easier to deploy and be discovered, and offer more flexibility and interoperability [1].

The core concept behind is that of a group communication, which integrates several web services as one business communication group. However these architectures still don't define a clear proper security mechanism to handle threats behind faced by the web service.

The security architecture designed for the web is limited when used for the web service architecture, and the need for new standards is apparent. Since solutions to this problem are still emerging, creating new standards will enable the Internet world to quickly adapt to the new security architecture of web service groups.

Security in this paper is represented by five main factors when it comes to quality of the web service. It represents the following aspects of the development of web service. [1]

- Integrity
- Assurance
- Verification
- Confidentiality
- Availability

When defining this protocol we needed to ensure security for our product so that the result of any transaction is in unaltered state. This is a great challenge, with the large amount of attacks taking place on the internet, and we need to be ready for these and thus consider these factors. Since the protocol is for group communication we need to consider group communication issues as well.

A secure group communication system should provide confidentiality and integrity of data being exchanged between the group members, integrity and possibly confidentiality of server control data, client authentication, message source authentication and access control of system resource and services.

In the design stage of this protocol, we mainly considered the three available technologies as follows

- XML Signature[6]
- XML Encryption [7]
- GSAKMP [3]

The protocol is mainly designed by following the GSAKMP message formats. XML Security technologies are used to archive the security requirements.

The implementation of this protocol provides a new secure API which can be easily integrated in to the Apache Axis engines.

2. Background

2.1 Web Services

Web services mean different things to different people. For some, a Web service is simply a Web application that performs a service through the Web. To others, a Web service must involve a monetary transaction. Some people think that a Web service application is one that uses the SOAP protocol. In general, a Web service is an application that accepts XML-formatted requests from other systems across a network (Internet or Intranet) via lightweight, vendor-neutral communications protocols. [8]

The technologies and protocols that Web services rely upon are designed to be relatively lightweight, leaving many of the more complicated features such as security, session-handling, and transaction management to be handled by extensions to the Web services specifications.

Since Web services are based on standard open protocols, Web service systems offer interoperability across all platforms. This interoperability is one of the key features that make Web services so attractive for Enterprise Application Integration (EAI).

2.2 Business Scenarios for Web services

Most recent communication systems and business to business applications have emerged around the web services architecture. Huge business applications have to integrate several web services at once to improve their performance and their availability, as dynamic co-operative web services.

Such collaborative applications often require secure message dissemination to a group and an efficient synchronization mechanism. Secure group communication systems provide these services and simplify application development.

2.3 Related Work

Group key agreement is a fundamental building block for secure group communication systems. Several group key agreement protocols were proposed in the last decade, all of them assuming the existence of an underlying group communication infrastructure.

As indicated above, the focus of this work is on the performance of group key management protocols for collaborative, peer groups. Therefore, here we have considered only key distribution and key agreement protocols.

In general, group key agreement protocols were designed to accommodate different membership changes: joining of a new member, leaving of a member, network partition, and network merge.

This section briefly describes currently available group key management protocols and network authentication protocols, which we are concerned.

2.3.1 Kerberos

Kerberos is a network authentication protocol. It is designed for the purpose of making strong authentication between client and server applications using secret-key cryptography [9]. Initially the client has to show its identity to the Kerberos server using the username/password mechanism. Kerberos issues the ticket to the TGS (Ticket granting server) with other secure credentials and sends it to the client in encrypted format using client password. The client can use that ticket to get access to the TGS. This request includes some other parameters such as the ultimate destination of the server and TGS responsible to authenticate the client by using relevant credentials and time stamp. Once the TGS authenticated the client it issues a TGT (Ticket granting ticket) and secure credentials needed to access the required server. In this process each party synchronizes with time [10].

2.3.1 GDH Protocol

Cliques GDH IKA.3 is a contributory key agreement protocol which is essentially an extension of the two-party Diffie-Hellman protocol. The basic idea is that the shared key is never transmitted over the network. Instead, a list of partial keys (that can be used by individual members to compute the group secret) is sent. One member of the group – group controller– is charged with the task of building and distributing this list. The controller is not fixed and has no special security privileges [4].

2.3.2 CKD Protocol

The CKD protocol is a simple centralized group key distribution scheme. The group key is not contributed, but it is always generated by one member, namely, the current group controller. The group controller establishes a separate secure channel with each current group member by using authenticated two-party Diffie-Hellman key exchange. Each such key stays unchanged as long as both parties (controller and the member) remain in the group. The controller is always the oldest member of the group. A special case occurs when the group controller itself leaves the group. In this situation, the oldest remaining member becomes the new group controller. Before distributing the new key, the new group controller must first establish secure channels with all of remaining group members. [4]

2.3.3 GSAKMP

The Group Secure Association Key Management Protocol (GSAKMP) is a general protocol for creating and managing cryptographic groups on a public network. A cryptographic group is a logical association of users or hosts that share cryptographic keying material.

While GSAKMP provides mechanisms for cryptographic group creation, other protocols may be used in conjunction with GSAKMP to allow various applications to create groups according to their application-specific requirements [3].

2.3.4 GSAKMP Light

The specification of a GSAKMP-Light (GL) profile is a way to shorten the number of messages exchanged during secure group establishment. The GSAKMP protocol assumes that group members joining a secure group have no information about the specific security mechanisms used by the group (for example, the key length, encryption protocol, etc). GSAKMP-Light provides a profile for the case where group members have been previously notified of these security mechanisms used for joining a group, during the group announcement or invitation. This profile does not sacrifice any of the security properties of the full protocol.

Like GSAKMP, the GL profile is provably secure, supports distributed architectures, allows multiple data sources within a single cryptographic group, and provides group management mechanisms [3].

3. Web Service Group Key Management Protocol (WS-GKMP)

3.1 Introduction

We used the architecture of GSAKMP [3] in our new Web Service Group Key Management Protocol (WS-GKMP). In design, we were mainly concerned with the security needed to avoid eavesdropping, tampering and impersonating threats. In our protocol, we used existing security standard in PKI and XML, mainly XML Signature [6] and the XML Encryption [7].

3.2 WS-GKMP Design

The main architecture of the WS-GKMP is illustrated in the figure-2. One centralized Group Controller (GC) and Group Members (GM) can establish a secure group using a group key as shown in the figure. GM in the system directly connects to GC and exchanges some basic messages to obtain a group key. In general, WS-GKMP has one centralized group controller and any number of group members. All parities in this protocol should trust the message processing on GC and GM.

According to the web service architecture GM has to play the web service role as well as web service client role. In addition, GM maintains a secure location protected from unauthorized accesses. (This place is used to store group keys and the other confidential information).

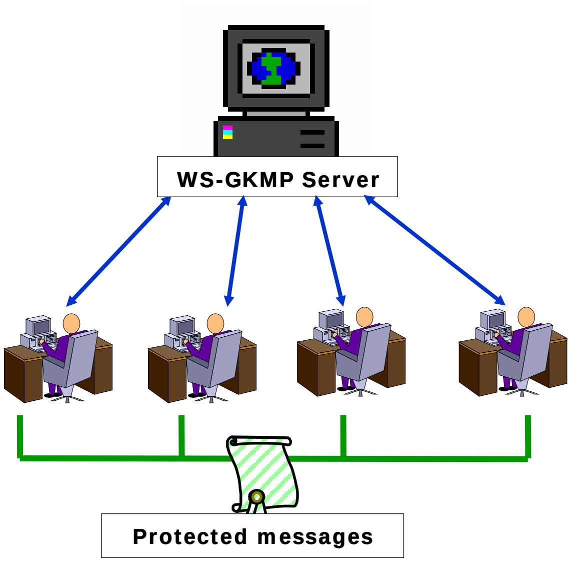


Figure 1 (Main system architecture)

3.2.1 Group Controller Web Service

In web service infrastructure, the group controller is a web service which provides several set of operations. In this protocol, we have assumed each and every member trust the GC. Thus, GC should be operated by a trusted third party.

The GC should provide following services to their members.

- Request To Joint (WSGKMP_RTJ)
- Request To Leave (WSGKMP _RTL)
- Request To Alive (WSGKMP _RTA)

In order to provide these services, GC performs following operations:

1. Accepts *"request to join"* messages coming from new members.
2. Authenticates the key requesting messages.
3. Maintains a secure key store.
4. Maintains a log file for connected members.
5. Stores member information.
6. Sends group key to its members.
7. Accepts, *"request to alive"* message and removes members who don't responding for this message within the given time period.
8. Accepts *"request to leave"* message.
9. Performs a rekey operation.

3.2.2 Group Member Web Service

As mentioned above, the group member plays two roles. The member can be any business integration service, business partner for other system or an individual client. GM type depends on the application type which uses WS-GKMP. The GM should provide following services to the other members.

- Rekey (WSGKMP _REKEY)
- Message Pass (WSGKMP _MSG_PASS)

In order to provide these services, GM has to perform following operations.

1. Requests a group key from the GC.
2. Downloads and saves the key in a secure location.
3. Sends alive message to GC and maintains the existence.
4. Encrypts every message passed to other members using the group key.
5. Decrypts every message received from the other member.
6. Accepts *rekey* message from GC and updates the group key.

3.3 WS-GKMP Message Processing

Every message passed between the GC and GM is a SOAP message. They include WS-GKMP message type WS-GKMP version and Group Identification number in the SOAP header by default. The identification number and the version are embedded in the SOAP message as default constants for a specific group. The WS-GKMP message type element has several sets of values (Ex: wsgkmp_rtj, wsgkmp_rtl, wsgkmp_alive, etc). The sender of a message should be responsible for including the required parameters to the SOAP message before sending, and the recipient should be responsible to verify it as soon as a message is received.

3.3.1 Member Joining to the Group (WSGKMP_RTL)

This process is the initial step of the group establishment. It is illustrated in figure-3. A GM calls the service of the GC to join the group and the GC performs the verification process and sends the group key to GM.

3.3.1.1 WSGKMP_RTJ (Request to joint message)

As the initial step, the member has to connect to the group with own certificate and request the group key.

Input parameters:

WSGKMP Version.
WSGKMP Group Identification number.
WSGKMP message type("wsGKMP_rtj")
Member name
Alias URL
Certificate
XML Signature and digest value.

The SOAP body part of the message should consist of version of the WS-GKMP and the group identification number. Here the GC checks over the compatible version and group identification number.

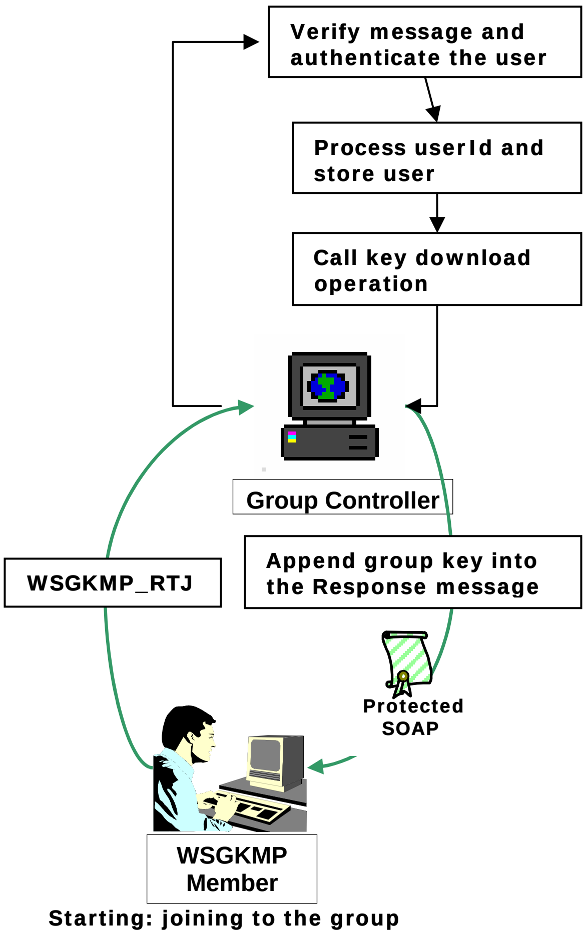

Verify message and authenticate the user
Process userId and store user
Call key download operation
Group Controller
WSGKMP_RTJ
Append group key into the Response message
Protected SOAP
WSGKMP Member
Starting: joining to the group

Figure 2 (WSGKMP_RTJ)

Other parameters are embedded in the SOAP Header Element. The message type needs to recognize the WS-GKMP message; each message has each unique message id.

Member name and member URL is needed because initially GC needs the member information. Especially the URL should be directed to the own (member) web service and the rekey service should hosted along with this URL by a member.

The WS-GKMP signature element consists of other certificate details, digest values, key information, relevant algorithms and signature details. After processing the SOAP message, it passes to the GC as an initial request.

3.3.1.2 Verify WSGKMP_RTJ at a GC

After SOAP message reaches the GC it has to verify and validate the member.

Input parameters:

Complete WSGKMP_LRT SOAP message

Initially the group controller analyses the SOAP message structure with message type. It checks over the version compatibility and group identification number. Then it gets the certificate and verifies it up to the trusted certificate. If that verification is successful it again verifies the message over the **<WSGKMP:Signature>** element and especially the digital signature upon SOAP body. If there are no errors, it adds some unique part to the name of the member. (Ex: if the received name is "john" then new name is "dixbbykk_john"). If there are no errors, the member details are stored by the group controller at some secure location where it should protect them from unauthorized accesses.

3.3.1.3 WSGKMP_KEY_DOWNLOAD (key down load message)

After verification the request to join message the GC passes the group key and newly generated name to the member.

Input parameters:

WSGKMP Version.
WSGKMP Group Identification number.
WSGKMP message type ("wsGKMP_key_download")
Group key
Member alias name (generated by the GC)

As the first step GC encrypts the SOAP body by the group key (symmetric key) where the body is consists of member name (generated by the GC). Thus, the name here is blinded for interceptors. Then the group key is encrypted by the requested (newly joined) member's public key and appended it to the SOAP Header.

After processing all the steps, the SOAP message is passed to the requested member as a response message.

3.3.1.4 Verify WSGKMP_KEY_DOWNLOAD at a GM

This is the response message to the corresponding request (WSGKMP_RTJ).

Input parameter

Complete WSGKMP_KEY_DOWNLOAD message

Initially, a member analyses the SOAP header and checks the message type. If it is a wsGKMP_key_download type, the group key is decrypted using the private key. Then the group key is stored in some pre defined secure location.

Finally, the SOAP body is decrypted by using this group key and the alias assigned by the group controller is read. The alias is also stored in the same secure location.

After that, the member is an authorized member in the group. Thus, he can participate in further activities in the group.

3.3.2 Member Leaving from the Group

This process is started by a member before it going to leave from a group. After the message passed to a GC, it removes the information of that specific member from the current list and generates a new group key. The new group key is distributed to the group members. Figure-4 illustrates this operation graphically.

3.3.2.1 Member Leaving from the Group WSGKMP_RTL

If some connected member needs to leave from the group it should send the request to leave service the GC with a "wsGKMP_rtl" message.

Input parameter

WSGKMP Version
Group Identification number
WSGKMP Message type ("waGKMP_rtl")
Member alias name

The message type is needed to identify the message by GC. The alias is included into the SOAP body and that part is encrypted by the group key. After processing this operation, the GM passes this message to the GC.

3.3.2.2 Verifying WSGKMP_RTL at a GC

After receiving the SOAP message, GC performs verification operation. Then the GC checks the message type and decrypts the SOAP body. After that, the GC extracts the alias name and verifies it with the current data store. Finally, the GC removes the user detail form the store and performs rekey operation.

Input parameters:

Complete WSGKMP_RTL message

3.3.2.3 Rekey Operation (WSGKMP_REKEY)

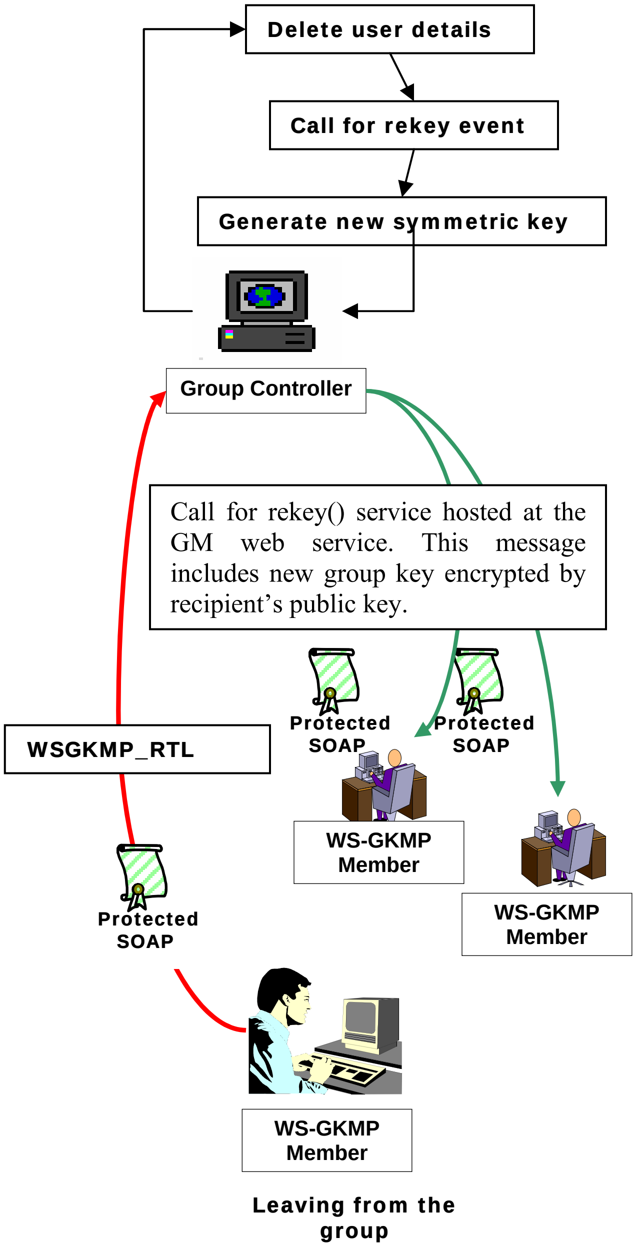
This is very important operation process by GC that distributes new group key to group members.

Input parameters:

WSGKMP Version
Group Identification number
WSGKMP Message type

New Group Key
Member alias name

As the first step, this process generates a new group key. Then it calls individual member's rekey services over the entire group (with the exception of the member who just left) with a new group key. This is very similar to the key download process, except for the message type and the mode. The message type here is wsGKMP_rekey and the mode is request. In the key download (see section 3.3.1.3), message type is wsGKMP_key_download and mode is response.



Call for rekey() service hosted at the GM web service. This message includes new group key encrypted by recipient's public key.

Figure 3 (WSGKMP_RTL)

3.3.2.4 Verify Rekey Message at a GM

After calling the rekey service of the member, it verifies WSGKMP messages and updates their group key.

Input parameters:

Complete WSGKMP_REKEY message

This verification operation is similar for the key down load process (refer 3.3.1.4).

3.3.3 Secure Message Exchanging

After connected to the group, the members can exchange secure messages. One member can request the message pass service of the other member. The body part of the message passing is encrypted by the group key. The recipient of this message can decrypt it by using the same key.

The message type here is "wsGKMP_msg_passs" and the process is illustrated in figure-5.

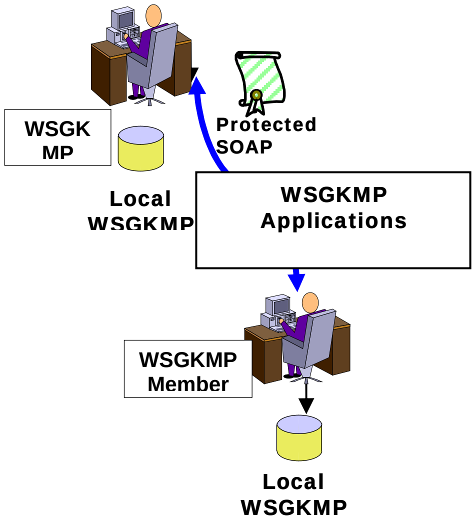


Figure 4: (WSGKMP_MSG_PASS)

3.3.3.1 Message Passing (WSGKMP_MSG_PASS)

If the message type is WSGKMP_MSG_PASS, every message should be encrypted before passing it to the members.

Input parameters:

WSGKMP Version
Group Identification number
WSGKMP Message type "wsGKMP_msg_send"
Alias of member
Secured message

The message sender is known as a requester. The body part of this SOAP message is encrypted by the group key.

The alias name also embeds in the body in the encrypted format, which is useful to know the owner of the message.

3.3.3.2 Verify WSGKMP_MSG_PASS at a GM

After the message reaches the destination, the service verifies the header of the message. When the encrypted message is received by a recipient, it decrypts the message and read it.

Input Parameters:

Complete WSGKMP_MSG_PASS message

3.3.4 Alive message

After a member joins a group, there should be a way of knowing whether all members are alive or not. Sometimes the members may be crashed or shutdown. Thus, in this situation, the group key must be updated. In order to do that work, WS-GKMP contains an "alive" message.

All members in the group should call "alive" service hosted at GC to inform about the existence within equal time gaps defined by protocol policy. This service is very useful to keep track on GMs. Figure-6 illustrates the process.

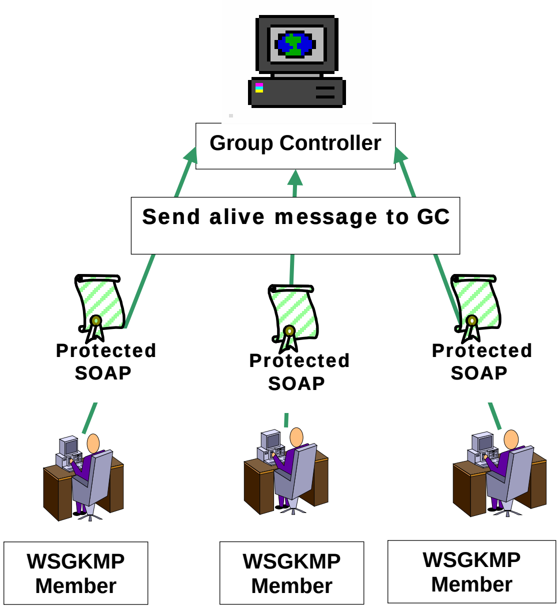


Figure 5: (WSGKMP_RTA)

3.3.4.1 Member sending alive message WSGKMP_RTA

In the group communication session, the GM sends the "alive" message to the GC.

Input Parameters:

WSGKMP Version
Group Identification number
WSGKMP Message type "wsGKMP_msg_send"
Alias of member

The alias name is included in the SOAP body as encrypted format. Other parameters are embedded in the header part of the SOAP.

3.3.4.2 Verify WSGKMP_RTA at a GC

When the "alive" message reaches to a GC, it has to perform several tasks. Initially, it decrypts the body, reads alias name and stores it in temporally location. Then GC checks over the requested alias names and current member list and identifies the existing and non-existing users.

Input Parameters:
Complete WSGKMP_RTA message

When GC removes a member, it should perform rekey operation (refer the section 3.3.2.3).

4.0 WS-GKMP Implementation

For the demonstration and evaluation, we have implemented the WS-GKMP API and very simple chat application that uses the API.

The API was implemented using Java and it uses the Apache Axis Engine. The developer only needs to install the API and build their application on top of the WS-GKMP. The protocol runs inside the Axis engine and all security integrations are handled inside the Axis engine.

The following section describes a very simple chat application implemented using WS-GKMP. The architecture of the system is shown in the figure-7.

Here the chat member is also the WS-GKMP client. All members initially have to connect to the group using request to join message. Then the messages are passed to the chat server and others will be able to download the messages from the chat server. In the chat server, all receiving messages are stored against the member id. The chat server uses a timestamp to recognize the sequence of the messages. In addition, messages such as "alive" works under the user application. We are not going to describe details of the chat server implementation, because it is out of the scope of this paper.

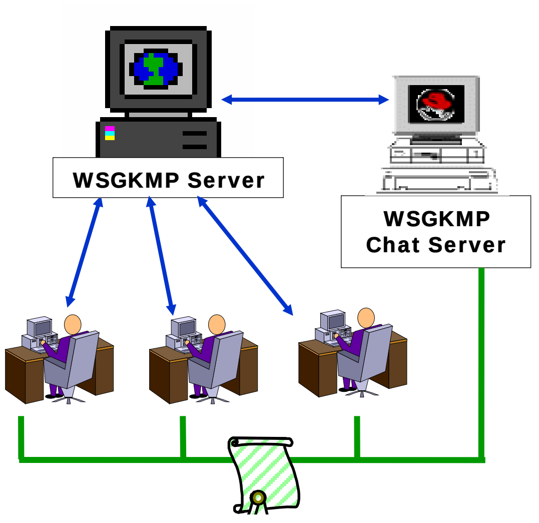


Figure 6 (WSGKMP Secure Chat Application)

5. Evaluation

As mentioned in previous the two major events we should evaluate in WS-GKMP are member-join and member-leave events. Thus, we evaluated the time intervals against to the group size of the WS-GKMP protocol for the above events. The matrix below describes the status of the environment, where we used to carry out this evaluation work.

Venue :	MSC Computer laboratory, University of Colombo School of Computing.
Cluster Size :	11 machines 40 applications.
Configuration :	2.7 GHz Pentium IV Single-processor Servers Machine
OS :	Windows XP SP1
Web Server :	Apache Tomcat 5.0
Web Service Server :	Apache Axis
Java Version :	JDK 1.4.0
Certificate Authority:	UCSC Certificate Authority. (http://ca.cmb.ac.lk)

5.1 Key Establishment

As the key establishment time, we measured the time needed by a member to be completely stabilized within in the group. This measurement is taken by differencing the time of request (GM calls to GC for membership) and the time that the GM completely receives the session key.

GMs request constantly for membership from the GC. The test is carried out by incrementing the group size in steps of five. Figure-8 shows the average time needed for the key establishment process against the group size.

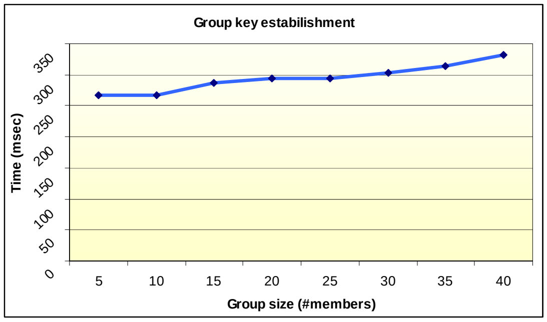


Figure 7 Group key Establishment Time

Since the time scale is in milliseconds, the graph is close to straight line. It is an evident that the GC does not have huge work load whe the numbers of the group are increased. It implies that when we increase the size of the group up to forty members, the GC can handle the group easily.

Among the time consuming processes used by GC the secret key generation process takes higher rank. But in the initialization process GC has to perform this task only once (If no extraordinary circumstances occur). Therefore the GC only has to perform the member request verification of the member request and key distribution.

Therefore the graph (Figure 8) implies the size of the group would not make considerable effect for the group establishment process.

5.2 Rekey Process

Rekey is the one of the most important events in the secure group establishment protocol. In WS-GKMP we have considered two cases in testing this event. For reference, the time calculation periods against to the sub events occurring in the Rekey process are described in figure 8.

Case 1:

In case 1 it calculate the time interval between the processes of any member calls for RTL message and the last member updated from the new session key. This can be calculated from the point a member start a RTL (t1) to the time of the last member updating with the new Group Key (t6).

Rekey process time= $t_6 - t_1$

The graph (Figure 10) plots group size against the the time intervals calculated according to the above formula. The graph shows, when the group size increased utilization time need to update with the Session Key in the group also increased. This is to be expected, as according to the architecture of WSGSP the GC has to individually call for the Rekey services and distribute the new key through the group

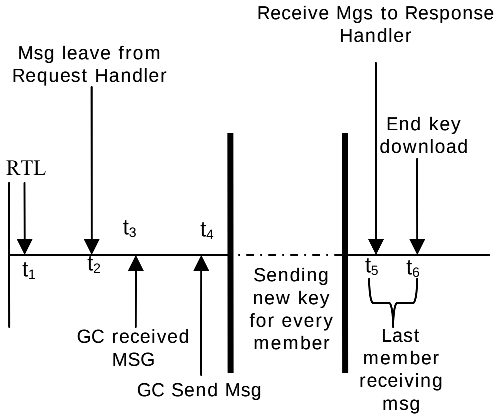

Receive Mgs to Response Handler
Msg leave from Request Handler
End key download
RTL
t3
t4
t1
t2
Sending new key for every member
t5
t6
Last member receiving msg
GC received MSG
GC Send Msg

Figure 8 Rekey event time line

Therefore large groups need considerable resources for a key establishment processes to occur while in the middle of the communication.

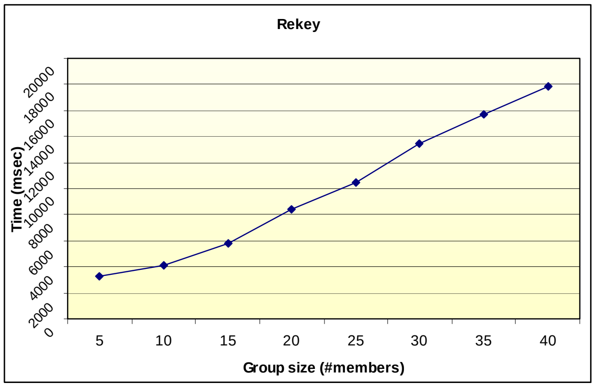


Figure 9 Rekey Time vs Group Size

Case 2:

In case2 we consider the processing time consumed by the GC when group updating with a new key. Here the total time taken by GC to complete the rekey event is calculated and plotted against to the group size. The formula used here is derived from Figure 10 as follows.

Time taken to complete the rekey process by GC= t_4 - t_3

Figure 10 is shown the plotted graph representing above formula. When the group size is increased the curve reaches a near exponential shape. Evidently, when the group size became larger and larger, the processing power needed by the group controller is high.

When we make a comparison between figure 7 and 9 it seem confusion, due to the time consumption calculated from the GC. As a major reason for this case is, when the rekey event occurs in the group, the GC has to call rekey services hosted at each member point. According to the architecture of the web services (in practically), invoking a service from the other party would take much more time resources than responding to an invoked service. This may vary according to the network traffic at the destination end

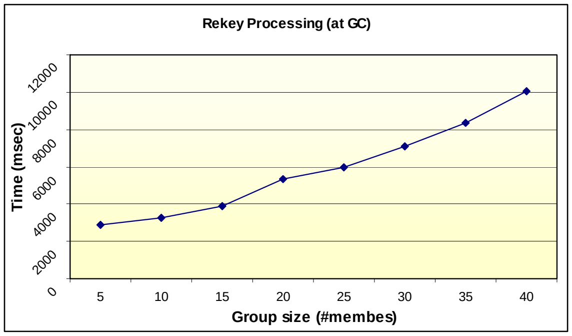


Figure 10 Rekey Processing at GC

6. Future work

This is the initial version of WS-GKMP and can be mentioned as WS-GKMP Version 1. A lot of enhancement of this protocol is required before it can be embedded with more sophisticated secure mechanisms. The following tips depict some enhancements which should be embedded in later versions of the WS-GKMP.

- The current version of the WS-GKMP doesn't have any feature or mechanism to recover the group form GC crashes. Basically it only keeps secure data (Member's certificates and session key) in some secure location. Therefore the only possibility this version is to replace or restart the server which should make it possible to access secure data.
- In this current version the GC directly communicates with all members in the group. This is an extra overhead for the GC when it increases the size of the group.
- The API developed including WS-GKMP functionalities only works with Java applications. Since the concept of the web service is for language independent applications the WS-GKMP should have deferent implementations for different languages.

7. Conclusion

The WS-GKMP (Web Service Group Key Management Protocol) provides a strong framework to fulfill security needs in web service group communication.

This application is mainly designed for high level Web application, while applications such as Kerberos are designed for low level standalone applications. Further such applications use symmetric cryptography while WS-GKPM uses strong public key cryptography. The latest version of the Kerberos is capable of authenticating client and server existing in deferent realms. WS-GKMP is capable of carrying this out by using two or more Web domains.

WS-GKMP can be used to establish secure dynamic multi-party web service application sessions. At present, banking systems, business applications, and airline systems use integrated web service applications and WS-GKMP will be very useful protocol for them. However, we noticed that when the number of GM's is increased the resource consumption at the GC is also increases. Thus, WS-GKMP might require some optimization such as introducing the logical key hierarchy (LKH) [5] before it can be applied to large (>1000 users) groups.

References

[1] "Professional Web Services Security", Ben Galbraith, Whitney Hankison, Andre Hiotis, Murali Janakiraman, D. V. Prasad, Ravi Trivedi, Wrox Press, 1st edition December, 2002

[2] "Introduction to Web services and the WSDK V5.1", Carlos Valcarcel, Jacob Weintraub, David Fraser, Kyle Gabhart and Rick Hightower, IBM Technical Library, http://www-130.ibm.com/developerworks/webservices/

[3] "A Survivable Group Security Architecture", Sead Muftic, Gernot Schmölzer, D. Sierra, Kasun De Zoysa, NSA/LUCITE Project Report, CSPRI/GWU, December, 2002

[4] "On the performance of group key agreement protocols", Y. Amir, Y. Kim,C. Nita-Rotaru and G. Tsudik, Proceedings of the 22nd IEEE International Conference on Distributed Computing Systems, Viena, Austria June 2002.

[5] "Secure Group Communications Using Key Graphs", Chung Kei Wong, Mohamed Gouda, and Simon S. Lam, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 8, NO. 1, February 2000

[6] "An Introduction to XML Digital Signatures" http://www.xml.com/lpt/a/2001/08/08/xmldsig.html

[7] "XML Encryption Syntax and processing" http://www.w3.org/TR/xmlenc-core

[8] "Web Services Journal", http://webservices.sys-con.com/

[9] "Kerberos: The Network Authentication Protocol" http://web.mit.edu/kerberos/www/#what_is

[10] "Applied Cryptography", Bruce Schneier 2nd edition

Can a Mobile Phones Save Lives?: Towards a Mobile ECG Monitoring System

P Wimalaratne[1] and D.P.S. Kularathna[2].

University of Colombo School of Computing
35, Reid Avenue, Colombo 7, Sri Lanka
IFS Pvt Ltd.
501, Galle Road, Colombo 6, Sri Lanka.
[1]spw@ucsc.cmb.ac.lk and [2]prabha.kularathna@ifs.lk

Abstract

This work presents a prototype developed for detecting the arrhythmic heart beats of a patient and to deliver the information including the ECG signal to a central Control Centre for appropriate action to be taken by the mediacal staff.

A prototype was designed and implemented including a wearable ECG sensor module which was designed and developed as a part of this work. The sensor is capable of acquiring ECG signals and feeding the signal to a neural network based software capable of detecting cardiac arrhythmias in real time. The arythmia detection software runs in a mobile phone and is capable of communicating with a personal computer based application where the medical staff such as mecical experts or paramdics can take appropriate action at the Control Centre.

The focus was on designing a mobile phone based Arithmia Detector which detects the health condition through a ECG sensor which communicates with a the mobile phone and the detection is carried out using the software on the phone and any abnormalities detected will be communicated with a control centre by transmitting a report which includes the ECG signal. Further Analysis will be carried out at the control centre and appropriate further action will be actaken place. A neural network based prototype running on a mobile phone in real time is usesd to detect the arithmia. The software is capable of analyzing each and every beat from the input ECG signal in real time. The system provided an accuracy rate of 81.0% in detecting arrhythmic beats, against 45 records of 30 minute ECG signals from MIT-BIH Arrhythmia Database [13]. *The program was tested on a Nokia 6630 handset and provided sufficient level of performance to obtain a real-time .*

1. Introduction

The interest in automatic arrhythmia detection dates back to 1960's - the age the concept of *artificial intelligence* began to evolve [5]. Attempts have been there to make this process remote, where the patient and the consultant are in geographically separate locations [14]. Recently, some have proposed the mobile phones [15] and other modern wireless techniques such as Bluetooth ([16], [17], [20]) to transmit the ECG signals. However, having an arrhythmia detecting process operating in a mobile - wearable device is challenging. The need of heavy computational power and memory makes such processes difficult to be implemented in wearable devices.

However, during the last couple of years, mobile phone technologies have achieved a significant amount of gain in both processing power and memory, making the automatic arrhythmia detection a possibility with the mobile phones. At the same time, mobile phone networks with technologoies such as GPRS are being established all over the world, covering even the remotest areas.

The contemporary remote ECG monitoring systems use dedicated-proprietary mobile networks to provide their service. The establishment and maintenance of such wireless networks are extremely expensive. Nevertheless, arrhythmia detection is not typically performed in the mobile devices.

Existing GPRS mobile network coverage can be used so that the service provider does not have to implement and maintain dedicated mobile networks eliminating a need for costly special purpose networks. Moreover, GPRS networks are being used all around the world, so that, such a system can be implemented wherever the necessary quality of service (QoS) standards are guaranteed by the network operator. Mobile phones are becoming increasingly powerful in processing, more and more in memory capacity, and at the same time, with decreasing prices. The vision of the authors is to develop a prototype that makes a a Mobile ECG Monitoring System a reality with the use of mobile phones.

However, the key challenge in this scenario is developing a program which can detect the arrhythmic beats, being able to be executed in a mobile phone, operating at real time, which also meets necessary levels of accuracy as a critical medical application. The prototype implementation attempts to address these issues.

The schematic diagram in *Figure 1.1* illustrates the mobile ECG monitoring system The patient wears a *mobile unit* which consists of a small electronic circuit (ECG sensor) which picks ECG signals via the electrodes connected to patient's body, and a mobile phone which analyzes the ECG signals, beat by beat, in real time, which is capable of distinguishing cardiac arrhythmias. If any arrhythmia is detected, all necessary information (including the ECG trace) is sent to the *control center* over the air, where medical team such as a panel of medical experts continuously accepts incoming requests from the patients and provides medical care. Another potential application of the same system would be to activate the repose of an emergency medical response team such as 999 calls in the United Kigdom. The system can alert the paramedics and transmit the location details. The accuracy of the detection could be double checkd by analyzing the ECG signal by a medical expert at the control centre. A request contains the electrocardiograph containing the affected beats, which is the main source of information to the medical expert.

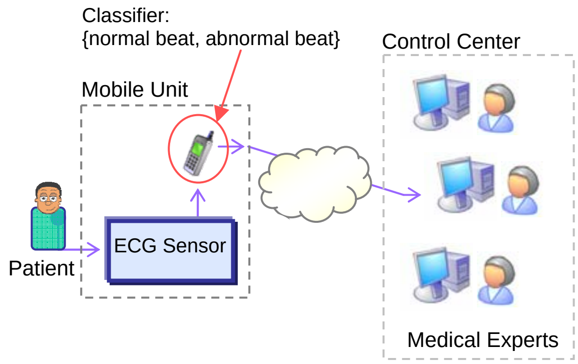


Figure 1.1 The Mobile ECG Monitoring System

The mobile phone contacts the control center only when it has detected arrhythmic heart beats and thus sending the resultant electrocardiograph to the control center via GPRS. The objective was to minimize the traffic generated by the system while altering the control centre when necessary.

The arrhythmia detection process is based on an artificial neural network. Neural networks are being widely used for automatic arrhythmia detection and they are quite stable on noisy and erroneous inputs[1]. Power spectrum values of Fourier transform are used as inputs to the neural network since variation of the duration of heart beats prohibits the use of time domain data as input.

The design and development of the prototype system is described in the following Mobile ECG Monitoring System and its evaluation is discussed in the following sections.

2. The Prototype Implementation

The prototype includes an electronic component capable of sensing ECG signals (*the mobile unit*), a program which runs in the mobile phone detecting the cardiac arrhythmias and a server-side program that accepts all incoming requests from the mobile phones and presents to the medical experts.

2.1 The Mobile Unit

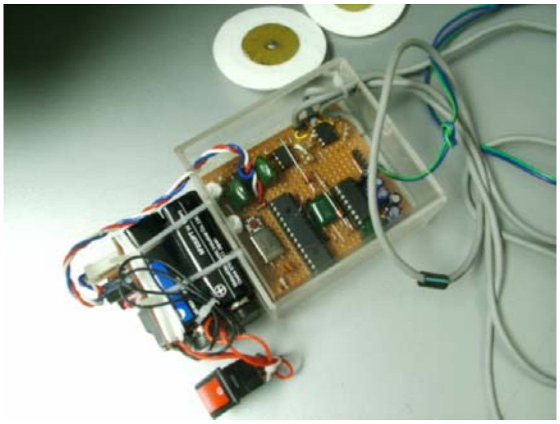


Figure 2.1 The ECG Sensor

The ECG signal is acquired from the patient's body surface using ML-II (modified lead-II). The signal is converted with an ADC at a 200 samples/sec rate, with 10-bit resolution and can be sent to the mobile phone via Bluetooth.Due to the contartints, instead of a Bluetooth chip, a USB-Bluetooth dongle was used through a PC. A microcontroller (PIC 16F876A) was used both as the ADC and the USART interface to the COM port of the PC.

2.2 Mobile Phone Program

The program that runs in the mobile program is the core of the mobile ECG monitoring system, which is capable of analyzing the incoming ECG signal in order to detect

[1] Neural Network based approach was chosen solely because of the success the neural networks have achieved in the field of arrhythmia detection, as a starting point. However, other methods such as Petri nets, hidden Markov models, etc. remain as candidates.

arrhythmic beats. Following sections describe the process of arrhythmia detection step by step.

2.2.1 Preprocessing

The ECG amplifier provides a low pass filter, eliminating all high frequency components of the ECG signal, which are essentially noise. A *rolling average algorithm* was used to eliminate the *baseline wander*[2] (fluctuation of the baseline with respect to time) of the ECG signal (equantion 2.1).

$$y(nT) = y(nT) - \frac{1}{2NT}\sum_{i=-N}^{N} y(nT + iT) \qquad (2.1)$$

2.2.2 Segmentation

The conventional approach is to break the ECG signal in the time domain in such a way that one segment contains a single heart beat. Several techniques hasve been proposed from the use of simple threshold based R-peak detection [1] to the use of derivatives ([1], [2]) to techniques such as hidden Markov models [3]. Frequency domain techniques such as wavelet transforms [4] also have been investigared.

A slope based R-peak detection scheme was used based on the first derivative of the ECG signal, using the *least square polynomial derivative approximation* [5] (equation 2.1). The peaks that cut a certain threshold were identified to be the *R-peaks* (see Figure 2.2).

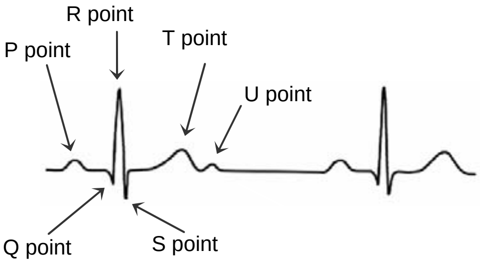


Figure 2.2 ECG of normal sinus rhythm

$$y(nT) = \frac{1}{10T}\left[2x(nT) + x(nT - T) - x(nT - 3T) - 2x(nT - 4T)\right] \qquad (2.1)$$

The absolute value of the amplitude was used when searching for the extreme-point in order to make the segmentation algorithm capable of handling situations where the axis of the QRS wave is inverted. The absolute values of the negative peaks are divided by two to avoid them becoming superior than the positive peaks which would lead to erroneous results. Figure 2.3 depicts a section of ECG signals, its first derivatives and the threshold.

The first-derivative based approach provided a sufficient level of accuracy over the test data record set in general. However, while the algorithm performed extremely well for a large portion of the records, it poorly performed on some smaller set of records. The reason was that, the height of the R-peaks of some records was lower than the height of P-peaks of some other records. It was seen that, this problem can be remedied by making the threshold adaptive. However, computational complexity and the memory requirements of the algorithm should be considered since the target platform is a mobile phone.

2.2.2 Feature Extraction and Selection

We use Fourier Transformation as a function that maps a time-domain representation of a signal into a frequency domain representation. The Fourier Transformation provides a consistent mapping between time and frequency domains, given by:

$$H(f) = \int_{-\infty}^{\infty} h(t)e^{2\pi i f t}dt \qquad (2.2)$$

and

$$h(t) = \int_{-\infty}^{\infty} H(f)e^{-2\pi i f t}dt \qquad (2.3)$$

where $h(t)$ is the amplitude of the signal as a real-valued function of time and $H(f)$ is the amplitude of the sine and cosine components as a complex-valued function of frequency.

Since the output frequency amplitudes of Fourier transform are in complex form, the power spectrum of the output is obtained by multiplying each amplitude with its complex conjugate:

$$P(f_n) = H(f_n).H^*(f_n)$$

In the implementation, a *fast Fourier transform* algorithm was be used, which is capable of applying the Fourier transform in $\boldsymbol{O}(N \log_2 N)$ time, extracted from [6]. (2.1) Each segment was placed in a buffer size of 512 elements, capable of storing 2.56 seconds of data of the 200 samples/sec signal. In the output of the Fourier analysis, 0th element of the buffer represents the 0Hz component and 256th element carries the 100Hz component.

The frequency of the heart beats lie within the band 0.5Hz – 50Hz [5]. Only the frequencies that lie within this interval is used. Consequently, the first $256/2 = 128$ elements of the output vector of the Fourier transform were taken. However, the 0th element is not used in order

[2] Baseline wander occurs when the strength of contact between the skin and the sensor electrolyte changes due to motion of the patient

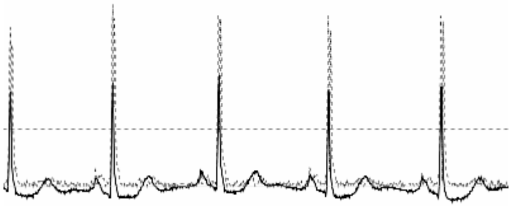


Figure 2.3 ECG signal (solid), First derivatives (dotted)

to remove the DC component of the ECG signal. Hence, this selection also acts as an effective noise filter.

Duration of a heart beat (i.e. RR-interval) is a vital fact used in ECG analysis and it is intuitive to think that, using this measurement in conjunction with the frequency spectrum of the heart beat may lead to a good increase in accuracy of the classifier providing a time dimension to the feature set. Thus, the RR interval is also used as an element in the feature vector. Advantages of the RR-intervals are that it represents the *time scale* of the overall heart beat and it is easier to measure.

The neural network classifier used is *stateless*. i.e., it processes each segment (heart beat), one at a time, without considering any information of the past segments. Duration of the previous heart beat provide the information to the classifier to identify the *variations* of the heart beats over time. Therefore we also use the difference ($RR_{current}$ – $RR_{previous}$) as an element in the feature vector.

Collectively there are 129 elements in the feature vector. The two time-domain measurements were scaled by 10^5 to get them in the order of power spectrum values so that the classifier would not neglect those values.

Fourier transform implicitly works as a noise filter as we consider only the frequency band: 0.5Hz to 50 Hz, dropping all signals having the frequencies outside of this range as noise. This particularly helps eliminating the interference of power lines near 60Hz[3].

Each wavelet (i.e. P-wave, R-wave, etc.) has its own range of frequencies which are localized in different areas in the frequency spectrum. Consequently, Fourier transform clusters these different wavelets in distinct regions in the frequency spectrum. It is intuitive to see that it allows the neural network to grasp the *information* of each wavelet in a consistent manner in specific regions of the inputs.

2.2.3 Arrhythmia Detection

Artificial neural networks have been widely used in arrhythmia detection and classification. Various approaches have been taken in applying the neural network classifier to the process. For example, discrete wavelet transform has been used to decompose the ECG signal into a series of coefficients which forms the input feature set for the neural network [4]. Another approach is the use of a set of physical measurements of the heart beat as the neural network input, such as QRS duration, PR interval, QT interval, etc. [10], [18]. A set of fixed polar coordinates of the poles of the z-plane has been used with the neural networks considering the fact that the frequency changes are reflected in the angular variation of the poles and damping is reflected in the magnitude of the variations [19].

We use a feed forward back propagation neural network [11] for arrhythmia detection. A neural network is a network structure of *units* (nodes) connected by a collection of directed *links*. A unit is a mathematical model based upon the functionality of a neuron of the brain. An *activation function* takes the weighted sum of the input values and produces the output as the output of the unit. Links connect the units and have associated weights. A *feed forward* network is an acyclic graph where there is a layer of inputs, a layer of outputs and zero or more intermediate layers called *hidden layers*. Nodes in each layer are connected with those of the adjacent layer by the links.

A variation of back-propagation algorithm was used, with momentum and an adaptive learning rate. The algorithm is capable of adjusting the learning rate depending on the rate at which the error is reduced. The neural network discriminates the segments (heart beats) into two classes {*arrhythmia*, *normal*} given the inputs.

The neural network has 129 inputs, one hidden layer with 2 units and the output layer with 2 units. Typical *sigmoid* functions [12] were used as activation functions in the hidden and output layers. A feed-forward - back-propagation neural network was chosen because it has shown very successful accuracies in problems of many domains and also, is very fast in execution with affordable memory requirements for our implementation [11].

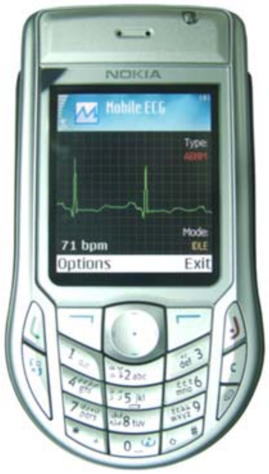
When testing with various numbers of units in the hidden layer and with different activation functions in hidden and output layers, it was possible to obtain 96.5% accuracy (see 3.2) over the test data set with just 2 units in the hidden layer. An accuracy of 89.5% was obtained with 5 layers. Other combinations did not produce comparable accuracies. The 2-unit configuration was selected under the assumption of, giving the neural network more freedom to represent complex functions raises the risk of *overfitting* where the network is optimized to represent the training set very well but fail to represent the function for the general case.

Though it is possible to represent the binary output {*arrhythmia*, *normal*} with a single output unit (say, [output > 0.5] ⇒ *arrhythmia*, otherwise ⇒ *normal*), two

[3] We assume the effect of the noise of some frequency bands above 100Hz that wrap-around onto the 0Hz-50Hz range due to aliasing effect [7] are negligible.

output units were used following the strategy often called *1-of-n output encoding* [11]. This provides more degrees of freedom to represent the target function (with *n* times as many weights available in the output layer). In addition, the difference between the two outputs can be used as a measure of confidence of the output. If both outputs fall below 0.5, which means the result is ambiguous, it is considered to be arrhythmic.

When training the network, 0.1 and 0.9 is used as the output values instead of the common Boolean values 0 and 1, respectively, since the sigmoid functions cannot produce these values for given finite weights. If attempted to train the network to fit 0 and 1, the gradient descent will force the weights to quickly grow out of bounds which results in an unusable neural network [11].



Cross validation is used when training the network where at the end of each training round with the training data set; the performance is validated against another validation data set. Whenever the error with respect to validation data set increases, the training process is immediately stopped. If the training is continued, the error with respect to the training data set may decrease but the network will lose the ability of classifying unseen data. In other words, this is the point where the neural network starts *memorizing* instead of *learning*. The final accuracy measure was taken by testing the neural network with an entirely different test data set which is completely independent of the training process.

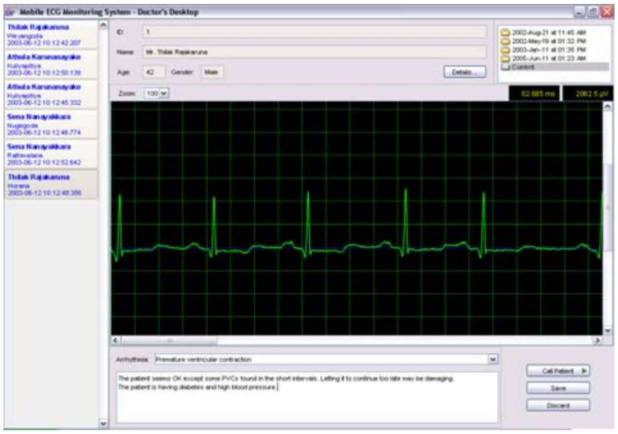
2.2.4 Implementation

In the prototype implementation of the Mobile ECG Monitoring System, the program was implemented as a J2ME (Java™ 2 *micro edition*) application on a Nokia 6630 handset. In addition to the arrhythmia detection process, the program provides a means of sending an electrocardiograph to the *control center* over GPRS, whenever an arrhythmia is detected.

The GUI displays the ECG signal that is currently being processed. As well as sending to the Control center, the program is capable of recording the ECG as a backup for later retrieval.

The combination of the ECG sensor and the mobile phone communicating via Bluetooth and linked with the control center via GPRS provides a great degree of mobility, allowing the patient to go anywhere as long as the patient is under the coverage of the mobile network. However, the quality of service of the network requires a very high level of reliability since the life of the patient depends on the reliability of the mobile network, as well as that of other components of the system.

2.3 Control Center


Mobile ECG Monitoring System - Doctor's Desktop

Thilak Rajakaruna
Weyangoda
2003-06-12 10:12:42.207

Athula Karunanayake
Kuliyapitiya
2003-06-12 10:12:50.138

Athula Karunanayake
Kuliyapitiya
2003-06-12 10:12:45.332

Sena Nanayakkara
Nugegoda
2003-06-12 10:12:48.774

Sena Nanayakkara
Rathmalana
2003-06-12 10:12:52.642

Thilak Rajakaruna
Horana
2003-06-12 10:12:48.356

ID: 1
Name: Mr. Thilak Rajakaruna
Age: 42 Gender: Male

2002-Aug-21 at 11:45 AM
2002-May-19 at 01:32 PM
2003-Jan-11 at 01:35 PM
2005-Jun-11 at 01:23 AM
Current

Zoom: 100
62.685 ms
2062.5 µV

Arrhythmia: Premature ventricular contraction

The patient seems OK except some PVCs found in the short intervals. Letting it to continue too late may be damaging.
The patient is having diabetes and high blood pressure.

Figure 2.6 The terminal for medical experts

A PC based application was developed as the back end of the Mobile ECG Monitoring System where the medical experts carry out their consultation procedures. All the requests arriving from the mobile units of the patients are accepted by a central access point called *interface server* and sent to the instances of terminals, each of which is used by a medical expert.

Figure 2.6 shows the terminal application. The stack of tokens at the left hand side displays the pending requests to be served. Elements at the right hand side to it display various information of the request currently under consideration, including the electrocardiograph. Patient's medical history can be accessed by the list at the top-right corner.

4. Conclusions

A prototype of a Mobile ECG Monitoring System was implemented along with a program capable of detecting arrhythmias in a mobile phone, in real time. A 81.0% of success was achieved with the 45 MIT-BIH ECG records with the arrhythmia detector, with mean time taken to process a single beat being 13.81 ms in a Nokia 6630 handset. A successful performance level was achieved in the arrhythmia detection operation in the mobile phone.

Neural networks are not commonly deployed in mobile phone based application. One potential reason would be the insufficient performance and memory capacity of standard phones. However, the prototype developed in this work demonstrates that neural networks can be deployed in the mobile phone based applications.

Acknowledgements

The authors would like to express their gratitude to Mobitel Pvt Ltd for their support and specifically for providing the mobile phone required for this research. The authors would like to acknowledge Dr D.N. Ranasinghe of UCSC for his assistance.

References

1. Touch, J. D. (1986), "A statistical Method for detecting peaks in Electrocardiogram Signals", December 1986 http://www.isi.edu/touch/pubs/tri.pdf

2. Kunzmann, U., Wagner, G., Schöchlin, J. & Bolz, A. (2002), "Parameter Extraction Of Ecg Signals In Real Time", *FZI Forschungszentrum Informatik Karlsruh*e, Germany

3. Hughes, N. P., Tarassenko, L. & Roberts, S. J. (1998), "Markov Models for Automated ECG Interval Analysis", *Lecture Notes in Computer Science State-of-the-Art Surveys*, Springer Verlag, p. 299-314 http://www.robots.ox.ac.uk/~nph/Pubs/nips03.pdf

4. Prasad, G. K., Sahambi, J. S. (2003), "Classification of ECG Arrhythmias using Multi Resolution", http:// www.ewh.ieee.org/ecc/r10/Tencon2003/Articles/739.pdf

5. Tompkins, W. J. (2000), *Biomedical Digital Signal Processing*, Prentice Hall of India Private Limited, New Delhi

6. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2002), *Numerical Recipes in C,* Second edition, Cambridge University Press, Cambridge

7. Smith, S.W. (1999), *The Scientist and Engineer's Guide to* Digital *Signal Processing,* 2nd edition, California Technical Publishing, San Diego, California

8. Watrous, R. & Towell, G. (1995), "A patient adaptive neural network", *Computer in Cardiology*, Vienna, Austria.

9. Pektatli, R., Özbay, Y., Ceylan, M. & Karlik, B. (2003), "Classification of ECG Signals using FUZZY clustering", *International XII. Turkish Symposium on Artificial Intelligence and Neural Networks* - TAINN.

10. Gao, D., Madden, M., Schukat, M., Chambers, D. & Lyon, G. (2004), "Arrhythmia Identification from ECG Signals with a Neural Network Classifier based on a Bayesian Framework", www.it.nuigalway.ie/m_madden/profile/pubs/ai2004b.pdf

11. Mitchell, T. (1997), *Machine Learning*, McGraw-Hill Companies, Inc., New York

12. Skapura, D. M. (1996), *Building Neural Networks*, Addison-Wesley Publishing Company, New York

13. MIT-BIH Arrhythmia Database (2003), http://www.physionet.org/physiobank/database/html/mitdbdir/mitdbdir.htm

14. Whipple, H., Dickson, F., Horibe, H. & Stark, L. (1965), "Remote, Online, Real Time Computer Diagnosis of the Clinical Electrocardiogram", *Communications of the ACM*, vol. 8, no. 1, pp. 49-52

15. Đaja, N., Reljin, I. & Relgin, B. (2001), "Telemonitoring in Cardiology – ECG Transmission by Mobile Phone", *Annals of the Academy of Studenica*, Institute of Oncology, Sremska Kamenica, Yugoslavia.

16. Jie, C. S. (2004), "A Pocket PC Based ECG Monitor" (2004 June 15) http://www.nexxusscotland.com/_cms/files/Carl_Hudson.s_presentation_20.01.05.63.ppt

17. Lee, R. G. (2004), "A Portable Tele-Emergent System with ECG Discrimination in SCAN Devices" (2004 June 24) http://pads1.cs.nthu.edu.tw/~scan/ppt/scan.ppt

18. Silipo, R. & Marchesi, C. (1998), "Artificial neural networks for automatic ECG analysis", *IEEE Trans. Signal Processing*, vol. 46, no. 5, pp. 1417-1425.

19. Lawrence, S., Burns, I., Back, A., Tsoi, A. & Giles, L. (1998) "Neural Network Classification and Prior Class Probabilities", *Tricks of the Trade, Lecture*

Notes in Computer Science State-of-the-Art Surveys 1998, Springer Verlag, pp. 299-314.

20. Healey, J. & Logan, B. (2005) “Wearable Wellness Monitoring Using ECG and Accelerometer Data”, to appear in *IEEE International Symposium on Wearable Computing*, Osaka, Japan. http://www.hpl.hp.com/techreports/2005/HPL-2005-134.pdf

Performance Evaluation of Application Aware Transport Services for High Bandwidth Sensor Actuator Networks

T. Banka, P.Lee, A.P. Jayasumana* and V. Chandrasekar
Department of Electrical and Computer Engineering, Colorado State University
Fort Collins, CO 80523
Email :{tarunb, leepanho, anura, chandra}@engr.colostate.edu

Abstract

Advances in networking have led to the emergence of high-bandwidth sensor actuator network applications. In many of these applications, it is required to transmit high bandwidth data to multiple end users while meeting their heterogeneous QoS requirements. An application aware Deterministic Overlay One-to-Many (DOOM) protocol is proposed, that concurrently meets the heterogeneous real-time rate and data framing requirements of multiple end users under dynamic network conditions. DOOM protocol supports application aware congestion control by performing dynamic selection of data for transmission at a given rate as per the end user requirements. Moreover, DOOM protocol performs efficient scheduling of high bandwidth data for the transmission to the multiple end users. Effectiveness of this protocol in concurrently meeting heterogeneous QoS requirements of multiple end users under dynamic network conditions is evaluated using Planetlab and a network emulation test-bed.

Keywords: Transport protocol, Multicast, Sensor-actuator networks, QoS, Overlay networks

1. Introduction

Modern society is increasingly dependent on sensor-based systems to preserve and improve the quality of life. Sensing, the process of sampling an environment or a phenomenon, helps us measure, process, interpret, predict and react based on the sensed phenomena. Sensor networks rely on networking technology to sense phenomena over large areas or at high granularities using multiple sensors. Desire to understand and monitor our environment for reasons that include structure integrity, weather, natural disasters such as tsunamis and seismic activity, and homeland security, will continue to push sensor network based applications. Pacific Tsunami Warning Center (PTWC) [26] is a sensor based real-time monitoring network to detect and locate major earthquakes and resulting tsunamis in the Pacific region. Costal-Marine Network (C-MAN) and Moored Buoy [25] are networks of sensors deployed for measuring parameters such as relative humidity, precipitation, barometric pressure, wind direction, speed, and sea temperature. International Monitoring Network (IMN) for seismic activity [24] is a network of more than 300 monitoring points all over the globe to observe seismic activity due to nuclear tests or natural reasons.

Associated benefits of near real-time sensing and ability to process and take actions based on the sensed information have led to the emergence of sensor-actuator networks [2]. A sensor-actuator system is one where sensing is accompanied with the actuation of various devices, for controlling a phenomenon or the measurement process itself. Note that we use the term sensor network and sensor actuator network interchangeably in this paper as our results are applicable to both.

There is a broad spectrum of sensor-actuator network applications. Sensor actuator networks can vary from resource-constrained mote based, low-bandwidth wireless sensor networks such as habitat monitoring [1] to medium bandwidth systems such as collaborative video surveillance [20,22], to high-bandwidth systems such as CASA (Collaborative Adaptive Sensing of the Atmosphere) [16,23]. In high-bandwidth systems, the data rate of a sensor may be of the order of tens of Mbps to hundreds of Mbps. Many such systems do not have energy or computation constraints that the traditional mote-based wireless sensor networks do. These networks [5,11,22] have the potential to revolutionize the way we sense our environment. Collaborative Adaptive Sensing of the Atmosphere (CASA) [16, 23], based on a network of radars for monitoring and prediction of weather phenomena is a prime example of this

*Corresponding Author: Prof. Anura P. Jayasumana
Email: anura@engr.colostate.edu
Phone: 970-491-7855 Fax: 970-491-2249

This work is supported by the Engineering Research Center program of the National Science Foundation under NSF award number 0313747.

emerging class of high-bandwidth systems. The radar senses the atmospheric conditions to monitor hazardous events such as tornadoes and hail storms, while detection algorithms and end users control the operation of radars to cover regions in the lower troposphere that are of interest or to meet requirements of prediction algorithms.

Many of these sensor network applications perform mission critical functions and have distinct application-specific QoS requirements. Key QoS requirements include bounded delay, critical high bandwidth requirement, application-specific data framing for acceptable data accuracy, and maximum acceptable loss threshold [5,7]. There is a need to meet such application-specific QoS requirements with available network resources and dynamic network conditions for their proper operation.

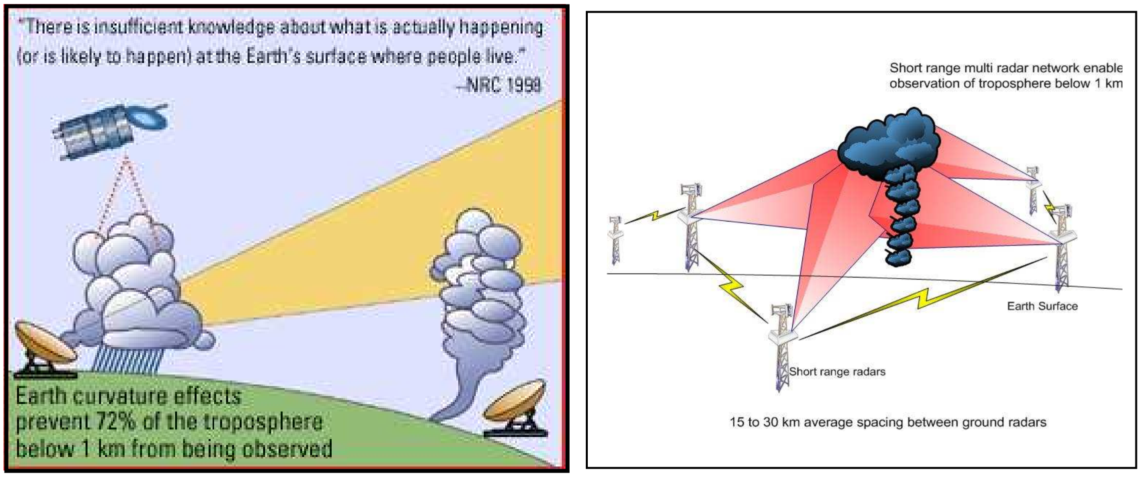
Ability of a network to meet heterogeneous high-bandwidth and application specific data framing requirements depends on the transport protocol and available network infrastructure. In this paper, we demonstrate the effectiveness of DOOM one-to-many protocol based transport services in meeting heterogeneous high bandwidth and data framing requirements of the CASA sensor actuator network application.

Late 90's have seen the emergence of overlay network concept that provides a quick and easy deployment path for new protocols and services over the already existing networks without the need for changing underlying network infrastructure [21]. Moreover, overlay networks provide a scalable solution for supporting application specific QoS requirements as it is not always efficient to support such requirements for every application at the lower layers [18]. Beside that overlay nodes are special nodes in the overlay network with significantly more resources in terms of computation, memory and storage. It enables complex application specific operations as well as transport oriented operations to be performed at the overlay node level rather than performing them at the router level in case of Internet. Moreover, under normal conditions, an ISP can be relied upon to meet end user specific critical QoS requirements using a scheme such as DiffServ. However, a CASA like system often has to operate under adverse conditions, such as severe weather or tornados, that can potentially disrupt services provided by some of the links or ISP's. These systems need to be designed to operate even under adverse conditions, adapting to degradation of service in parts of the network. Deployment over overlay networks enables selection of alternate bandwidth rich paths, i.e., provide the ability to adapt to available bandwidth in an application specific manner [3,9]. Thus an overlay-network based solution is an attractive proposition for the design of these systems.

Network dynamics such as packet drops and delay can degrade the perceptual quality of the applications [15]. Ubiquitous transport protocol like TCP and UDP are not sufficient for meeting real-time rate and data framing requirements of multiple end users under dynamic network conditions [7]. Lack of universal support for IP multicast has led to the emergence of end-system approaches [13]. Most of the current research in multimedia streaming is focused on developing multicast solution specific to properties and requirements of video streaming. Many of the protocols use adaptive layering algorithms for congestion control [14] and such layering may not be always applicable for high-bandwidth applications such as radar data streaming. Difference in characteristics of sensor data due to fine granularity of the data demands the need for development of new one-to-many data dissemination protocol that are cognizant of the characteristics of the sensor data. Moreover, most of the present multicast solutions are receiver driven for reasons of scalability [13,14]. In case of high-bandwidth sensor networks, scalability requirements are not as stringent as the requirements of large scale video streams distribution applications, i.e., the number of end users involved may be much smaller. Sender-driven one-to-many data dissemination protocols provide an attractive option for the high-bandwidth sensor network applications.

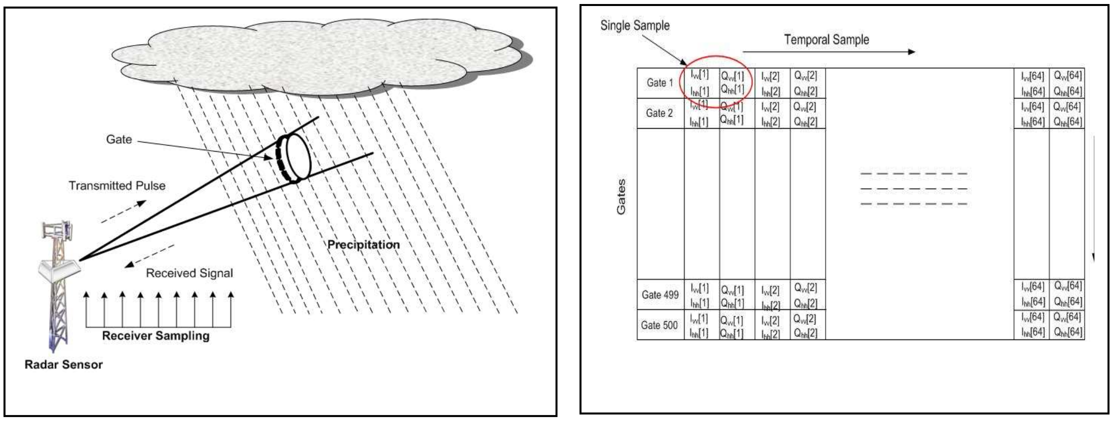
In this paper we propose an application aware **D**eterministic **O**verlay **O**ne-to-**M**any (DOOM) data dissemination protocol. DOOM is a sender driven high-bandwidth, one-to-many data dissemination transport protocol with following key goals: (i) Given the rate determined by the congestion control algorithm, sub-sample and frame the data based on the intended use of the data by the end user (ii) Efficient scheduling of different sub-sampled sets of high-bandwidth data for transmission to multiple end users within bounded time. Additional details and performance results for DOOM are provided in [6].

Section 2 describes CASA in detail. Section 3 explains DOOM protocol. Section 4 shows the performance results of DOOM protocol based transport services in meeting heterogeneous QOS requirements of the end users/applications. Section 5 presents the summary of the paper.


"There is insufficient knowledge about what is actually happening (or is likely to happen) at the Earth's surface where people live."
–NRC 1998
Earth curvature effects prevent 72% of the troposphere below 1 km from being observed
Short range multi radar network enable observation of troposphere below 1 km
Earth Surface
Short range radars
15 to 30 km average spacing between ground radars

(a) **(b)**

Figure 1. (a) Limitation of current state of the art in observing atmospheric phenomena (b) Network of short-range radars for observing lower troposphere



Gate 1	$I_{vv}[1]$ $I_{hh}[1]$	$Q_{vv}[1]$ $Q_{hh}[1]$	$I_{vv}[2]$ $I_{hh}[2]$	$Q_{vv}[2]$ $Q_{hh}[2]$	$I_{vv}[64]$ $I_{hh}[64]$	$Q_{vv}[64]$ $Q_{hh}[64]$
Gate 2	$I_{vv}[1]$ $I_{hh}[1]$	$Q_{vv}[1]$ $Q_{hh}[1]$	$I_{vv}[2]$ $I_{hh}[2]$	$Q_{vv}[2]$ $Q_{hh}[2]$	$I_{vv}[64]$ $I_{hh}[64]$	$Q_{vv}[64]$ $Q_{hh}[64]$
Gate 499	$I_{vv}[1]$ $I_{hh}[1]$	$Q_{vv}[1]$ $Q_{hh}[1]$	$I_{vv}[2]$ $I_{hh}[2]$	$Q_{vv}[2]$ $Q_{hh}[2]$	$I_{vv}[64]$ $I_{hh}[64]$	$Q_{vv}[64]$ $Q_{hh}[64]$
Gate 500	$I_{vv}[1]$ $I_{hh}[1]$	$Q_{vv}[1]$ $Q_{hh}[1]$	$I_{vv}[2]$ $I_{hh}[2]$	$Q_{vv}[2]$ $Q_{hh}[2]$	$I_{vv}[64]$ $I_{hh}[64]$	$Q_{vv}[64]$ $Q_{hh}[64]$

(a) **(b)**

Figure 2. (a) Radar operations (b) Digitized Radar Signal (DRS) block periodically generated by radar

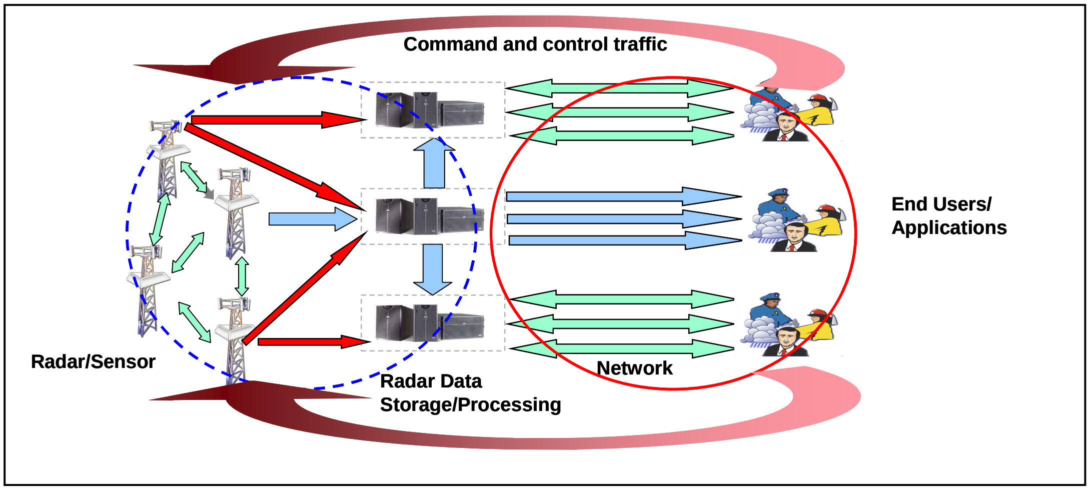


Figure 3. Different data transfer scenarios in a CASA network

2. High Bandwidth Sensor-Actuator Network Applications

This section explains communication network and the associated QoS requirements of CASA system.

2.1. CASA Sensor-Actuator Network

The vision of the CASA is to revolutionize our ability to observe lower troposphere through Distributed Collaborative Adaptive Sensing (DCAS), vastly improving our ability to detect, understand, and predict severe storms, tornados, floods, and other atmospheric and airborne hazards. Fig. 1(a) [5,16] shows the limitations of current state of the art for observing atmospheric phenomena using long-range autonomous radars. Current technology is unable to monitor the lower troposphere due to earth curvature limitations. In CASA, a network of short range radars is used instead to sample the previously unobserved region of the atmosphere as shown in Fig. 1(b). CASA network forms a tightly coupled network of radar and processing nodes. Some of the key features of CASA network include presence of heterogeneous network infrastructure.

Historically radars have been designed and operated in a “central unit” environment where the radar transmitter/receiver and information processing were all carried out at the radar node. Due to advances in high-speed networking, there is no need to do computation at the radar node itself any more. Moreover, multiple smaller, cheaper radars can be networked to sample the atmosphere in an efficient manner and can be deployed over rooftops or cell towers as shown in Fig. 1(b) [5,16]. A network of small radars provide more flexibility by enabling re-tasking of the radar for effective sampling of the atmosphere based on the existing atmospheric conditions. Moreover, different radars can now operate in different scanning modes/bands unlike static operations of autonomous long-range radars. This system is a sensor-actuator network in that the radars sense the atmosphere, yet the scanning strategies of radars are controlled dynamically in real-time depending on the features being sensed and the requirements of the end-users.

In a typical operating scenario, radar transmits short pulses of energy, which are scattered back by the target, received by the receiver, and digitized for further signal processing. Figure 2(a) shows the radar operation and Figure 2(b) shows a block of data (Digitized Radar Signal) periodically generated by the radar node [8]. Depending on the radar operating parameters, it can generate data at rates of tens of Mbps to hundreds of Mbps. Under certain scenarios it is required to transmit high bandwidth data to the remote end users. In certain cases, this high bandwidth data can be processed at the radar node and only low bandwidth processed data transmitted over the network.

2.2. Networking Challenges in CASA

In a CASA network shown in Fig. 3, there are heterogeneous QoS requirements that need to be satisfied by transport services for the proper operation of the system [5]. The solid red and dotted blue circles in Fig. 3 show different regions of the CASA network with different QoS requirements. The blue dotted circle highlights the part of the network that supports communication between radar nodes and the distributed processing/storage nodes. Depending on the resources available at the radar node, data can be locally processed or transmitted over the network (which may consist of both wired and wireless links) in real time for remote processing. In this case sustained data rates may be in order of tens of Mbps to hundreds of Mbps. At the same time, low bandwidth streams like command and control signals, monitoring probes, and radar health traffic streams share the bottleneck links with high bandwidth radar data streams. Thus the protocols need to be friendly to the cross traffic streams in the network. Moreover it is necessary to support one-to-many, many-to-one and many-to-many data transfer scenarios for both high and low bandwidth data. For example, many-to-one communication will be required to gather data from multiple radars for integration at a central server. Radar data streaming from the radars to the points of computation also has unique QoS requirements: a minimum acceptable rate based on the end application (CASA needs to support multiple applications with different requirements simultaneously), better accuracy of end results with higher bandwidth, a bound on delivery time beyond which data is not useful, a bound on bursty losses, and more.

The solid red circle in Fig. 3 highlights the network that provides communication between processing/storage nodes and the end users. End users may send request for real-time processed or unprocessed data; similarly non real-time access of archived data may be requested. In certain cases, depending on the mode of operation and end user requests, reliability of data may be important. In

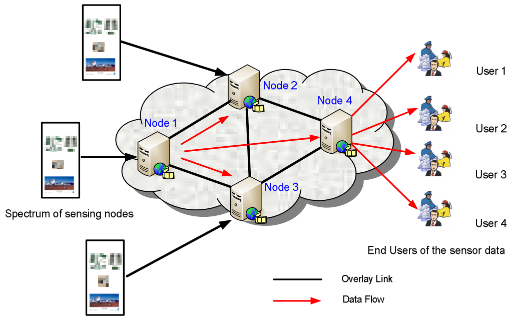


Figure 4. One-to-Many data dissemination scenario in networked sensing systems

other cases, real-time transmission may be critical, and certain types of losses tolerated. End users would use wired networks for data access from the storage or processing nodes. Moreover, each end user can have heterogeneous real-time and data framing requirements for acceptable data accuracy due to variable resources like bandwidth and computation resources availability at receiver end.

CASA networks with a large number of radar sensors may need multiple DOOM servers for one-to-many data dissemination. Since end users are geographically distributed at different locations, it is desired to deploy DOOM servers at strategic locations in the network for efficient dissemination of the sensor data in terms of bandwidth conservation. As seen in Fig. 4, a one-to-many DOOM server can distribute data to multiple other DOOM servers for distribution of data to a larger region of the network. In Fig. 4, node 1 performs one-to-many data dissemination, where nodes 2, 3, and 4 are the recipient nodes. Similarly node 4 further performs one-to-many data dissemination where end users 1-4 are the recipients of the data.

3. Deterministic Overlay One-to-Many (DOOM) Protocol

In this section, we describe design of the application aware overlay one-to-many DOOM protocol for data dissemination in CASA like sensor-actuator networks.

3.1. Design Goals of DOOM Protocol

DOOM protocol has the following goals (i) concurrently satisfy heterogeneous real-time rate requirements, (ii) satisfy distinct data framing needs of multiple heterogeneous end users and (iii) efficiently use resources of the DOOM server for servicing high bandwidth end users.

DOOM performs the congestion control for each end user using TRABOL [4], which is a UDP based application layer congestion control protocol. The key feature of TRABOL is that during network congestion, it performs transmission rate adaptation while considering end user specific minimum and target rate requirements [7]. Unlike with TCP, the transmission rate for a particular end user does not fall below its critical minimum rate requirement. Similarly, TRABOL does not exceed the target rate for any end user, which is determined by the network and computation resources available at the end user. Spatiotemporal dependency of radar data enables DOOM to support different data framing requirements by selecting most relevant data for transmission.

DOOM uses a time-multiplexed deterministic data scheduling scheme which is integrated with TRABOL to support multiple data framing requirements in real time under dynamic network conditions. Current implementation of DOOM supports two CASA specific data framing types, Type 1 and Type 2. But it is no way limited to this particular application, and can be extended to support other similar applications where data has spatiotemporal dependencies.

Existence of spatiotemporal dependency of the sensor data is the key factor that enables application specific data framing by the DOOM overlay server. Next, we describe the difference in importance of different parts of a block of data generated by a radar for different end users due to spatial and temporal dependencies. Fig. 2(b) shows a block of data [8], periodically generated by a radar sensor every *heart-beat* interval, while scanning particular direction in the atmosphere. The data is also known as Digitized Radar Signal (DRS), all samples in a row of the DRS block correspond to a finite volume of atmosphere at certain physical distance from the radar. Multiple samples are available for a volume of atmosphere at a particular distance; DRS block shown in Fig. 2(b) consists of 64 samples each for 500 different volumes of the atmosphere, referred to as gates, at different distance from the radar.

In sensor actuator networks, different end users may use data generated by same sensing node in different ways for meeting different system specific goals. In case of radar sensors, two common end user algorithms are reflectivity computation and Doppler velocity computation [8]. Random data loss during network congestion can have an adverse impact on the performance of these two algorithms. In order for reflectivity computation algorithm to have low error in the end results, it is desired to receive uniformly spaced samples in time for each gate of a DRS block, i.e., Type 1 data framing. Alternatively, Doppler velocity algorithm requires a pair of adjacent samples for their reliable computation [8], i.e., Type 2 data framing. End user applications can specify their unique data framing needs to the DOOM protocol that helps them to have an acceptable performance with a subset of the data.

3.2. DOOM Protocol Details

Each end user contacts the DOOM server to initiate a data transfer session. End users independently specify their critical minimum rate (MR) requirement below which data is not useful. Depending on the available computation and network resources, end user also specifies the maximum rate (TR) above which data cannot be received by the user. Moreover, each user also specifies their preference for different data framing types, e.g., Type 1 or Type 2, based on the specific use of the data by the end user. End user specific rate and data framing requirements are stored in the user-*list* as shown in Fig. 5. A static *rate-table* of supported transmission rates is defined, starting with lowest rate to the maximum possible transmission rate as shown in Fig. 5. Data samples to be transmitted for a given transmission rate and framing requirement of data are determined at the initialization time and stored in data schedule tables.

Maximum possible transmission rate (*Rate n* in Fig. 5) can be determined by the data generation rate of a single sensor node. In case of CASA, each radar sensor generate data at 100Mbps. Minimum rate can be determined by the lowest rate overlay server want to support for any end users (Rate 1 in Fig. 5). In the current implementation we have considered minimum supported rate as 1 Mbps and maximum rate 100Mbps. Number of rates supported in rate table is determined by the granularity requirement of the end user applications. In the current implementation, 1Mbps granularity is supported, i.e., two adjacent rates in rate table differ by 1Mbps.

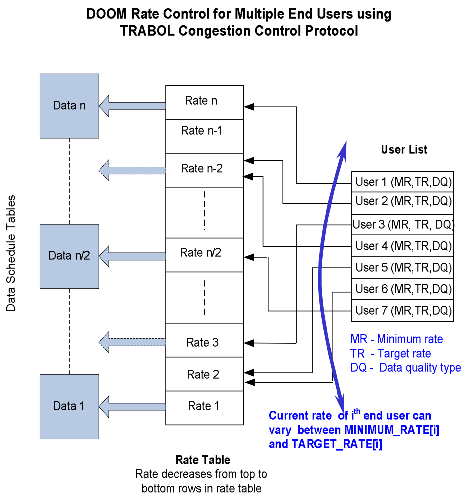


Figure 5. DOOM Protocol Implementation

```
                         DOOM Server
heart-beat:  Time for one block data generation
time_slot:   Time window for scheduling (10ms)
data      :  Data scheduled for transmission
user      :  End user scheduled to get data
user-list:   List of all users getting data
USER_COUNT: Number of user requests
ACK       :  Acknowledgments received from a
             user (received packet count)
WHILE (1)
 {
  // Repeat following every heart-beat
  // interval
  IF (USER_COUNT >0)
   {
    // Determine new transmission rate
    // of each client every heart-beat interval
   FOR (EACH USER IN client-list)
     {
      // Use TRABOL congestion control to
      // determine next transmission rate
       IF (ACK RECEIVED)
          {
          determine_TRABOL_rate(user, ACK)
          }
        else
          {
          determine_TRABOL_rate(user,
                                        NO_ACK)
          }
       // User next rate information is updated
      update_rate_table(user)
     }
    // Use Time multiplexing to transmit data
     FOR (EACH TIME SLOT )
       {
        FOR (EACH USER)
           {
            // Get data schedule table for the client
            data_schedule_table =
                              get_reference(user)
           // Determine data to transmit for a given
           // client in the current time slot
           data = lookup(time_slot,
                          data_schedule_table)
          send_data(data, user)
          }
       }
     }
  }
```

Figure 6. DOOM algorithm for one-to-many data dissemination

Fig. 6 shows the flowchart of the DOOM protocol. TRABOL congestion control algorithm independently determines the transmission rate of each end user served by the DOOM server based on the loss feedback for the previously transmitted block of data. Once current transmission rate is determined by TRABOL for all end users, a time-multiplexed scheduling algorithm is used to select samples for transmission in a given time slot at current transmission rate. Data is selected while considering data framing requirements of a particular end user. Total number of scheduling slots used is determined by the *heart-beat* interval of the sensor (periodic time interval after which sensor generates block of data). In the current implementation, time slot of 10ms is used to select data for transmission. When *heart-beat* interval is 100ms, for example, it is partitioned into 10 slots of 10ms each for scheduling block of data for the transmission. Multiple end users can be scheduled to receive data within the same time slot. This scheduling scheme makes sure that different end users are scheduled to receive data as per their rate and data framing requirements within *heart-beat* interval.

4. Performance Evaluation

Planetlab [17] and Emulation based test-bed shown in Fig. 7 are used for the performance analysis of DOOM protocol. Planetlab is an overlay network deployed over the Internet. It provides unprecedented ability to deploy and test new applications and protocols under realistic network conditions over the Internet. Further, it allows one to implement applications and underlying overlay networking services. Alternatively, NISTNET [10] network emulator is used to emulate different network dynamics. NISTNET is used here to emulate different bottleneck bandwidth scenarios varying between 105Mbps to 215Mbps. Moreover, NISTNET is used to introduce different ACK delays, which vary between 5ms and 600ms. Radar data generation is emulated using archived radar data. We consider the case when radar node generates data at constant rate of 100Mbps. Multiple end users send requests to the DOOM server, specifying their target and minimum rates along with their data framing requirements. These requests can vary from one end user to another.

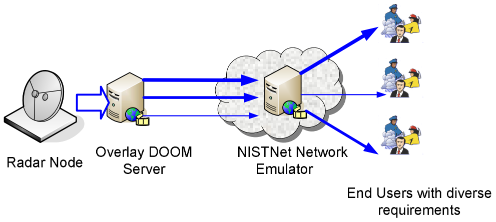


Figure 7. Network emulation test bed

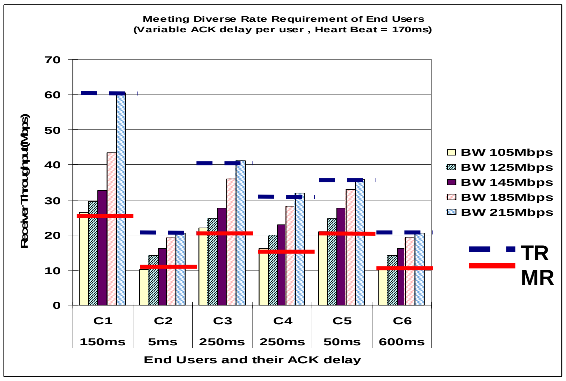


Figure 8. Throughput variation of DOOM at different clients for different bottleneck bandwidths (each client has a different ACK delay)

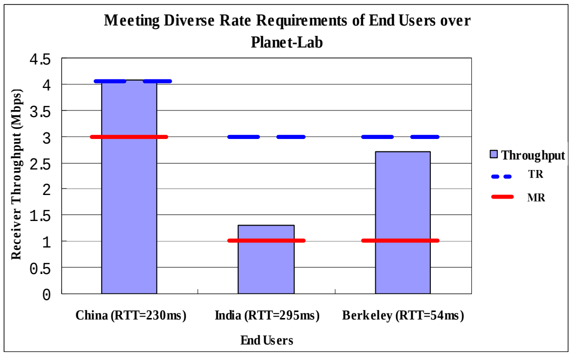


Figure 9. Effectiveness of DOOM in meeting rate requirements of different end users over Planetlab (data generation rate is 10Mbps and sensor heart-beat is 220 ms)

Fig. 8 shows emulation based results for evaluating DOOM effectiveness in meeting heterogeneous rate requirements of the multiple end users with different target rate (TR) and minimum rate (MR) requirement under varying bottleneck bandwidth conditions. For different end users the ACK delay, i.e., the time before loss feedback is received is variable. As the figure indicates, when bottleneck bandwidth exceeds the sum of minimum requirements of all end users, i.e., 100Mbps, then each user is able to meet its minimum rate requirement. As bottleneck bandwidth increases, different users share the incremental bandwidth equitably.

When bottleneck bandwidth exceeds cumulative target rate requirements of all end users, i.e., 200Mbps, all end users are able to

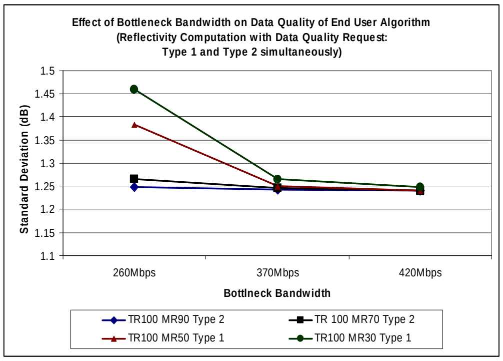


Figure 10. Quality of received data vs. bottleneck bandwidth for multiple clients using emulation based test-bed.

receive data at their target rate requirements. Fig. 9, demonstrates the performance of DOOM in meeting heterogeneous rate requirements over Planetlab test-bed. Overlay DOOM server at USA (Colorado) is used to stream CASA radar data to three Planetlab sites at China(Beijing), India(Bangalore), and USA(Berkeley). RTT of the three end users, measured from Colorado to China, India, and Berkeley was 230ms, 295ms and 54ms respectively. Case is considered when radar node generates data at 10Mbps and three end users have target rates (4Mbps, 3Mbps, and 3Mbps) and minimum rate (3Mbps, 1Mbps, 1Mbps) respectively, indicated by dotted blue and solid red line in Fig. 9. Each of the end users show different network congestion characteristics. Losses are observed by end users in India and Berkeley (USA), activating TRABOL based congestion control used by the DOOM protocol. Alternatively, the end user in China didn't suffer any losses thus receiver throughput is equal to its target rate, 4Mbps. Note that for all the end users, receiver throughput is above their minimum rate requirements. This experiment using Planetlab is thus able to demonstrate that DOOM protocol can meet different rate requirements of multiple end users in realistic network conditions over the Internet.

An end user can have a distinct data framing requirement, determined by the acceptable accuracy of the end results when received data is used for the computation of a certain characteristic. The quality of data received at the end user thus has to be based on the particular end user application. In case of radar applications, standard deviation in the end results is used to measure the quality of received data. A lower standard deviation is an indication of better quality of the received data for some end user applications. Fig. 10 therefore illustrates the performance of DOOM protocol in concurrently meeting radar application specific data framing requirements of different end users. In this case, two end users request for Type 1 data, and other two end users request Type 2 data. As seen in Fig 10, standard deviations for users with similar data type requests are very close to each other, indicating similar data quality of the received data.

Results in Fig. 8-10 demonstrate the effectiveness of application aware DOOM protocol in meeting heterogeneous real-time and data quality requirements of end users under different emulated network conditions and over Internet using Planetlab.

5. Summary

In this paper we underscore the emergence of broadband sensor actuator networks and highlight different QoS requirements of end users and applications, taking CASA as an example. Application aware transport services are the key to meeting heterogeneous real-time and data framing needs of multiple end users. Effectiveness of one-to-many data dissemination protocol, DOOM, in meeting QoS requirements of end users is demonstrated using an emulation test-bed as well as Planetlab overlay test-bed. Emergence of CASA like broadband sensor-actuator networks has the potential to hasten the development of other mission critical broadband sensor actuator networks. Transport services are an important area of research that has the potential to play a definitive role in the success of such mission critical systems. Future work includes large-scale performance evaluation of DOOM protocol.

References

[1] I.F. Akyildiz, Su Weilian, Y. Sankarasubramaniam, and E. Cayirci, "A Survey on Sensor Networks," IEEE Comm. Magazine, Vol 40, Issue 8, Aug. 2002

[2] I.F. Akyildiz, and I. Kasimoglu, ``Wireless Sensor and Actor Networks: Research Challenges," Ad Hoc Networks Jour. (Elsevier), Vol. 2, No. 4, pp. 351-367, Oct. 2004

[3] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris "Resilient Overlay Networks," Proc. ACM SOSP, Banff, AB, Canada, Oct. 2001

[4] S. Bangolae, A. P. Jayasumana and V. Chandrasekar, "Gigabit Networking: Digitized Radar Data Transfer and Beyond," Proc. IEEE International Conf. on Communications (ICC'03), Vol. 1, pp. 684-688, Anchorage, 2003

[5] T. Banka, B. Donavan, V. Chandrasekar, A. P. Jayasumana, and J. F. Kurose, ``Data Transport Challenges in Emerging High-Bandwidth Real-Time Collaborative Adaptive Sensing Systems, "Poster/Demo Session, IEEE INFOCOM 2005, Miami, FL, March 2005

[6] T. Banka, P. Lee, A. P. Jayasumana, and V. Chandrasekar, "Application Aware Overlay One-to-Many Data Dissemination Protocol for A Class of High Bandwidth Sensing Systems," Proc. IEEE/ACM 1st Int. Conf. on Communication System Software and Middleware (COMSWARE 2006), New Delhi, India, Jan. 2006

[7] T. Banka, A. Maroo, A.P. Jayasumana, V. Chandrasekar, N. Bharadawaj, and S.K. Chittababu, "Radar Networking: Considerations for Data transfer Protocols and Network Characteristics," Proc. 21st Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society (AMS), 19.11. Jan. 2005

[8] V. N. Bringi, V. Chandrasekar, "Polarimetric Doppler Weather Radar: Principles and Operations," Cambridge University Press, Aug. 2001

[9] J.W. Byers, J. Considine, M. Mitzenmatcher, and S. Rost "Informed Content Delivery Across Adaptive Overlay Networks," IEEE/ACM Trans. on Networking, Vol. 12, Issue 5, pp. 767-780, Oct. 2004

[10] M. Carson, D. Santay, "NIST Net: A Linux-based Network Emulation Tool," ACM SIGCOMM Computer Communications Review, 33(3): pp. 111–126, 2003

[11] V. Chandramohan, K. Christensen, "A First Look at Wired Sensor Networks for Video Surveillance Systems," Proc. 27th IEEE Conference on Local Computer Networks, pp.728-729, Nov. 2002

[12] D. Chen, and P. K. Varshney, "QoS Support in Wireless Sensor Networks: A Survey," Proc. of the 2004 Intl Conf. on Wireless Networks (ICWN 2004), Las Vegas, NV, June 21-24, 2004

[13] Y. H. Chu, S. Rao, S. Seshan, and H. Zhang, "A case for end system multicast," IEEE J. Select. Areas Comm., vol. 20, pp. 1456–1471, Oct. 2002

[14] A. Ganjam, H. Zhang, "Internet Multicast Video Delivery," Proc. of the IEEE, Vol. 93, Issue 1, pp. 159-170, Jan. 2005

[15] W. Jiang, and H. Schulzrinne, "Modeling of Packet Loss and Delay and their Effect on Real-time Multimedia Service Quality", Proc. 10th Intl. Workshop Network and Operations System Support for Digital Audio and Video, June 2000

[16] D.J. McLaughlin, V. Chandrasekar, K. Droegemeier, S. Frasier, J. Kurose, F. Junyent, B. Philips, S. Cruz-Pol, and J. Colom, "Distributed Collaborative Adaptive Sensing (DCAS) for Improved Detection, Understanding, and Prediction of Atmospheric Hazards," Proc. 21st Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society (AMS), 11.3, Jan. 2005

[17] L. Peterson, T. Anderson, D. Culler, and T. Roscoe, "A Blueprint for Introducing Disruptive Technology into the Internet," Proc. 1st ACM Workshop on Hot Topics in Networks, HotNets-I, Oct. 2002

[18] L. Subramanian, I. Stoica, H. Balakrishnan, R. Katz, "OverQoS: An Overlay Based Architecture for Enhancing Internet QoS," Proc. 2nd Symp. on Networked Systems Design and Implementation, San Francisco., CA, May 2005

[19] Y. Sankarasubramaniam, B. Akan and I. F. Akyildiz, "ESRT: EventtoSink Reliable Transport in Wireless Sensor networks," Proc. *MobiHoc 2003*, Annapolis, Maryland, June 2003

[20] D. Talbot, "Seamless Surveillance," Technology Review, Feb. 2004

[21] J. Touch, Y. Wang, L. Eggert , "Virtual Internets," ISI Technical Report ISI-TR-2002-558, July, 2002

[22] L. Yuan, C. Gui, C. Chuah, and P. Mohapatra, "Applications and Design of Heterogeneous and/or Broadband Sensor Networks," Proc. Int. Conf. on Broadband Networks, (BROADNETS 2004), Oct. 2004

[23] "CASA: Collaborative Adaptive Sensing of Atmosphere," Website: http://www.casa.umass.edu

[24] "International Monitoring Network for CTBT," http://www.ctbto.org/

[25] National Data Buoy Center, http://www.ndbc.noaa.gov

[26] "Pacific Tsunami Warning Center," http://www.prh.noaa.gov/ptwc

Sharing Excess Bandwidth among Cooperative Organizations for Web Access

S. Gamage* and G. Dias**
Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
Tel: +94 11 2650301 Ext: 3100, Fax: +94 11 2650912
{ *sumith , **gihan }@cse.mrt.ac.lk

Abstract

Internet is a very critical infrastructure today. However developing countries still rely on low bandwidth links, due to the high cost of International bandwidth. On the other hand these connections are heavily used, since people access international content, due to scarceness of local content. This makes out that the international bandwidth is a scarce resource which we should utilize efficiently.

Demand for International bandwidth is very high at peak hours; but lower at off-peak time. Further, the peak hours are not the same for all organizations. Many organizations are happy to share their excess bandwidth, provided that they are able to receive excess bandwidth of others when required.

In this paper we propose a solution for the sharing of the bandwidth among cooperative organizations for web access. It is based on Squid proxy. This system utilizes total bandwidth as a whole without noticeable performance degradation at individual organizations.

1 Background

Internet is a very critical infrastructure today, irrespective of whether it is in a developed or developing country. Most of the developed countries have a high percentage of population with high bandwidth internet connection. Many of the modern researches are focusing this era of the technology enhancement.

However in developing countries such as Sri Lanka and most of the south Asian countries, they still rely on low bandwidth links, because the cost of International bandwidth is restricting them going for high bandwidth links. Even for the entire University system we have only around 5~10 Mbps in Sri Lanka. Many of the organizations rely on their dialup links or ADSL connections. Only a very small number of organizations have dedicated 128kbps ~ 2Mbps links.

Most of these Internet connections are used at educational institutes. The students, researchers and academic staff are heavily using the Internet for their academic purposes. Further most of the knowledge base is distributed around the world. We have only a very little amount of *localized content* in developing countries. Therefore most of our traffic is *outbound traffic* where the International bandwidth is heavily used.

Apart from the above problem we have realized the following interesting issues, as well.

- Even the demand is very *high at peak-hours*; it is significantly very *low at the off-peak* hours.
- Peak and off-peak hours of one organization is not the same as that of another organization.
- Particular organization might have a substantial amount of excess bandwidth available even at a peak time (Examination period of the University) as well as an unexpected high bandwidth requirement at general off-peak times (An IT exhibition at a weekend).
- Many of the organizations that have their own links (departments in the same University / Universities in the Lanka academic network) are *not reluctant to share* their *spare bandwidth* with their cooperative partners, *provided that* they will receive the excess bandwidth of others at their peak time, if it is available.
- By sharing bandwidth among cooperative organizations like this, we get the maximum out of *existing* bandwidth *as a whole*.

Our research addressed the above issues and we came up with a system which *dynamically shares bandwidth* among cooperative organizations. The system is developed as a *plug-in* for the Open source Web proxy software called *Squid* [1]. It consists of four major components:

- Self bandwidth monitoring mechanism
- Inter-Proxy Bandwidth Negotiation Protocol
- Module for bandwidth granting on demand
- Module for Optimal user redirection via multiple uplinks

The next section of this paper discusses the other approaches in the world which are addressing the above specific problems. Then it explains our approach of solving the unaddressed issues in bandwidth sharing. The final section will conclude the paper with the present status of the system.

2 Existing Systems

The researchers working with low bandwidth links strongly believe that we have to live with it for quite a long period. Therefore the internal users of organizations should manage, assuring proper priority schemes giving higher opportunities to the most essential personnel and groups.

Delay pools [2] in Squid caching proxy server is based in this principle. Delay pool is the bandwidth management module in the Squid proxy server. It allows allocation of users in different priority pools depending on various user defined criteria.

Then the concept of *Dynamic Delay Pools* [3] extended the bandwidth management of Squid. It allows the excess bandwidth of unused delay pools (user groups) to be shared among the heavily used delay pools. This solution facilitated the transferring of bandwidth among internal users utilizing bandwidth at a single organization.

Another approach for utilizing the bandwidth is to make the users tend to use it more at the off-peak time. *Offline downloader* [4] is one of the good solutions for this. It allows users to put their requests of downloading to the system at peak times. Then the system downloads that huge file at the off-peak time. When the download is completed the system sends a notification to the user with a link to the local copy of the downloaded file.

Introducing a *pricing scheme* [5] (not necessarily required to be monetary) for the Internal bandwidth make users tend to shift their usage to off-peak. One simple way of doing this is by varying the bandwidth unit price to be inversely proportional with the demand [6].

There are various ways of addressing this problem at different levels. One way is to compress the data at the sending end of the limited bandwidth link and to decompress it at the receiving end [7]. Traffic Control (*tc*) [8] in Linux uses different queuing disciplines at the Kernel level prioritizing the packet appropriately when sending through a gateway.

All above mentioned methods are known to be good for optimizing the bandwidth usage at a single organization. Here in our paper we are targeting loosely coupled cooperative organizations on which there is not that much of central control (work as individual entities), but are willing to share with each other. They should be well connected internally. The best example is the *University system in Sri Lanka*. Every University have their own policies. Some of them have their own International links. But all of them are well connected via *LEARN* (Lanka Educational And Research Network). At the same time Universities are willing to work cooperatively to utilize intemperance resources they have, provided that their own University also gets the benefit out of it.

Internet Cache Protocol (ICP) [9] is one of the developments gratifying above objectives. ICP is primarily used in a cache mesh to locate specific web objects in neighboring caches. One cache sends an ICP query to its neighbors. The neighbors send back ICP replies indicating a "HIT" or a "MISS."

Here the ICP is used only for the communication about cache objects. There are also some other versions of communication protocols, like *Cache Digest* (CD), *Caching Neighborhood Protocol* (CNP), for the same purpose [10]. Any of these protocols does not consider the bandwidth measurements of their neighbor proxies.

Intention of this paper is to negotiate bandwidth among the cooperative proxies based on the availability of information in addition to cache object information.

3 Overview of the System

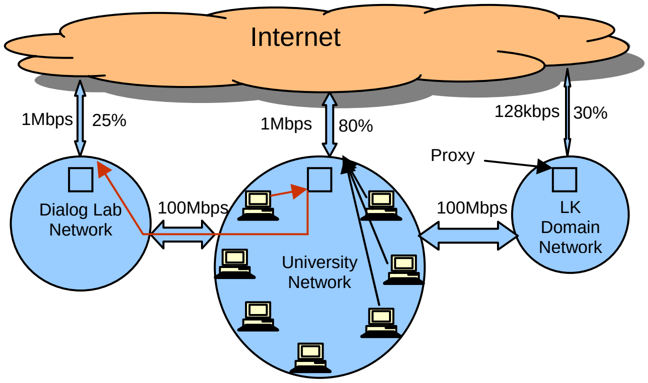


Fig 1: Overview of the System

Fig 1 shows the current model we are using at the University of Moratuwa. We have one big network at the *University of Moratuwa* connecting all the students and staff. Then there are two other dedicated networks for the *Dialog Research Lab* and *LK Domain Registry*. All these three entities have their own Internet links. Additionally three networks are connected to each other via University LAN.

Each of the above networks has a proxy for the local users who are accessing the web through their own bandwidth link. All those were *Squid proxies* [1] running *dynamic delay pools* [3] for local bandwidth management and *ICP* [9] for inter-proxy communication.

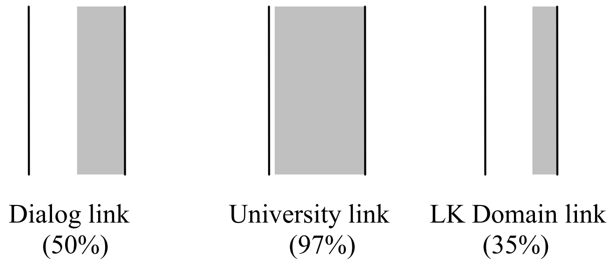


Fig 2: Unutilized Bandwidth

Under these conditions most of the time our observation was that the links of Dialog Lab and LK Domain were under utilized while University link was fully overloaded as illustrated in Fig 2. This is fairly true for most of the cooperative organizations.

Our assumptions are as follows in a general case:

1. *Cooperative organizations* (departments in the same organization, branches of one big group of companies, University network) have two or more direct Internet links.
2. Those organizations are well connected via a LAN/WAN and there is not any problem of bandwidth among the cooperative organizations.
3. Individual organizations are mutually agree on sharing the excess bandwidth of the local link for the betterment of other cooperative organizations.
4. All cooperatives are using our improved version of Squid proxy server for web accesses.

4 System's Logic

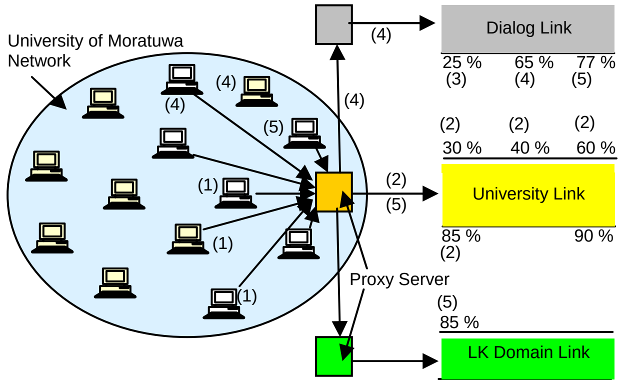


Fig 3: System's Logic

Fig 3 shows the abstract scenario of the system in action. The steps of the scenario as are follows:

1. All the users of a network (say users of University of Moratuwa network) should place all their web requests through their local proxy (University of Moratuwa proxy, in this case).
2. If the local link is not congested the request is forwarded to the destination directly (through University of Moratuwa link). This will continue for all the requests while enough local bandwidth is available.
3. When the local link is about to congest, the system checks the bandwidth status of the neighbor proxies (Dialog Lab and LK Domain Registry).
4. If any of them have enough excess bandwidth to donate, the request is redirected via first selected proxy.
5. When a user is redirected via a neighbor, that user is added to a list, bound to that proxy. Therefore the system can assure that the users selected to send through a certain link will keep on sending through that link for the entire period, the link is available.
6. Finally if all the neighbors are also not free enough, the requests are sent through the local link, even if it is congested.

5 Main Implementation Modules

We came up with the modular architecture shown in Fig 4 that facilitates simple, rapid implementation of the system while providing much greater flexibility for expansion. We have dedicated the next couple of paragraphs to brief each of those modules. Fig 4 has a dash-line around each module named with the corresponding paragraph number.

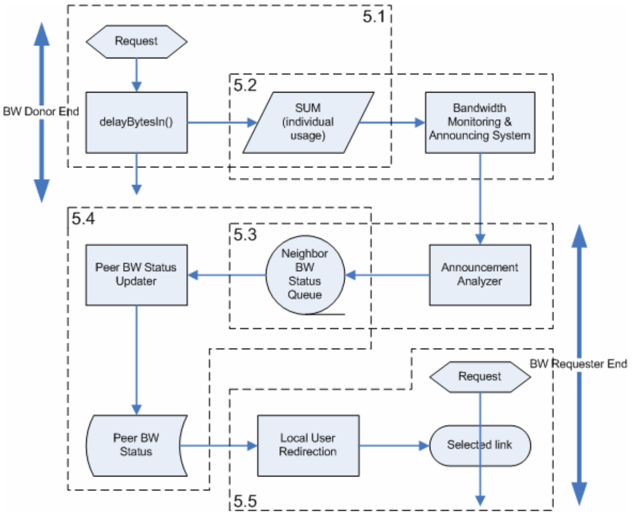


Fig 4: Flowchart of the Entire System

5.1 Squid Usage Reporting System

The objective of this system is to request bandwidth from (or offer bandwidth to) neighbors depending on the local usage. We should have a proper way of tracking local bandwidth usage for this purpose. We tried different alternatives for this:

1. Using the *round trip time* of a ping command
2. Getting the link status by directly communicating with the router via *SNMP*
3. Using *Excess bytes* variation of dynamic delay pools in Squid
4. Using Squid *access log* analysis

Even though the first two options give more accurate and reliable answers, they require additional application involvements. Since our system is an extension to the Squid proxy system, we realized that it is better to use Squid application data, if possible, to make the system compact and consistent.

The third option was less accurate. Our basic observations showed that *excess_bytes* of delay pools changes proportionally to the additional excess bytes availability. However later we realized that even when the link is highly congested, *excess_bytes* increases since the users can not use their allocated bandwidth. Fourth option does not give real time status, because of the Squid proxy server waits until the completion of file downloading to report in the log file.

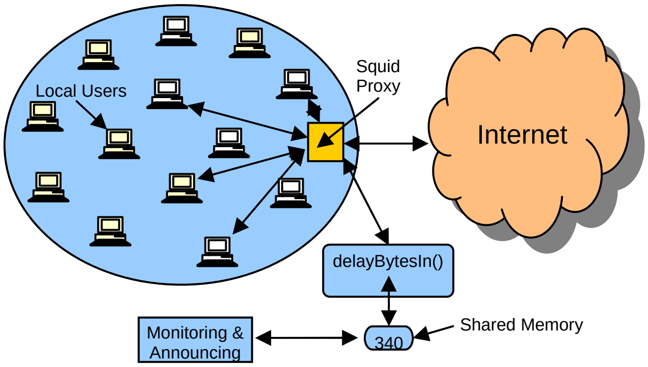


Fig 5: Squid and Bandwidth Monitoring

The final solution we came up with, was based on *delayBytesIn()* function of Squid application. All the delay data sent to the client on their request passes through this function. Therefore this is the best reference point at which we can collect web usage. At the same time it excludes all the local web traffic that we do not want to account for, because it bypasses the delay pools.

Squid Usage Reporting System simply sums-up the amount of data transferred to the local clients, to a shared memory location, as shown in Fig 5. This puts only an insignificant overhead to the Squid process.

5.2 Bandwidth Monitoring & Announcing

Then the Bandwidth Monitoring Process accesses the shared memory in predetermined time periods (P) and reads the bandwidth usage value (U) and reset. Then the average rate of bandwidth usage (R) is simply obtained by equation (1):

$$R = U / P \quad \text{--} \quad (1)$$

Next this value is announced to all the neighbors via *Inter-Proxy Bandwidth Negotiation Protocol* (IPBNP) [11].

5.3 Announcement Analyzer

Announcement Analyzers on all the neighbors keep on listening for incoming Bandwidth Announcement. It puts all the relevant announcements based on Squid *Cache Peer* Configurations [12] to a *shared queue*.

5.4 Squid Peer Status Updater

Then Squid continuously keeps on checking for any new entry in the shared queue. If there are any entries in the queue, *Squid Peer Status Updater* reads them one by one and updates the neighbor bandwidth availability information based on *moving average algorithm* [13]. The algorithm works as follows.

Assume the calculation of bandwidth usage at a particular instance (after ith bandwidth report from the neighbor). If we denote ith bandwidth update by $\boldsymbol{E_i}$ and the instantaneous average by $\boldsymbol{B_i}$, moving average can be calculated by the equation (2):

$$E_i = (1 - \alpha)E_{i-1} + \alpha B_i \quad \text{----------------------------------} \quad (2)$$

This ensures suppressing sudden peaks, while gradually changing average with the real variation of bandwidth usage depending on the value of ‘α’. If the neighbor bandwidth usage is increasing that should adapted by the bandwidth requester quickly. Therefore we are higher value for α when increasing the bandwidth usage, while having a lower value for ‘α’ when decreasing the bandwidth usage.

5.5 Local User Redirector

All the modules discussed above make sure proper update of bandwidth availability status of each neighbor proxy. The final module we discuss here selects corresponding forwarding proxy (or obviously direct link) appropriately to optimize the total bandwidth of all the organizations as a group. Fig 6 illustrates the algorithm used for this purpose.

1. When any request arrives to the web proxy it checks the *Local Bandwidth Usage* first. If the usage is less than the considering threshold, the request directly sends through the local link.

2. This threshold value has an upper (75~80%) and a lower (60~70%) bound. When we start and keep-on using the local link the users are not redirected to neighbor links until the local link reaches the upper bandwidth threshold. On the other hand, if any set of users were redirected through a neighbor link, they will not get back to the local link, until the local link usage drops below the lower threshold or the neighbor reports that its link is getting congested. This minimizes unnecessary frequent swapping of users among local and neighbor links.
3. When the local link starts getting congested, the systems poles through neighbors. If it does not have neighbors or if it could not find any neighbor with enough excess bandwidth, there is no other solution than sending them through the local link, even under congestion.
4. If there is one or more neighbor proxies with enough excess bandwidth, that user is added to the first neighbor's *access list*. All the requests coming from that user will continue to be sent through the newly selected neighbor link, until it satisfies one of the conditions mentions under point 2 or 5.
5. When the neighbor link usage reaches the *central neighbor boundary* (~65%), system stops adding new users. It will continue serving the existing users in the list through the selected neighbor link. If the neighbor link usage rises (This may be due to local user access or requests coming from neighbors) to *upper neighbor boundary* (~75%) users are removed from the list when the request arrived. This process guarantees the prioritize use of local users of the neighbors even when the local or remote uses makes the link congested. If the usage drops below *lower neighbor boundary* (~40%), new users are added to the neighbor list.
6. At any time if the requesting user is in the list of a particular list, that user is redirected through that proxy. This is done via the newly introduced function *getLinkOptimalProxy()* for peer selection. The way this works is very similar to the other standard peer selection functions in Squid, but uses the algorithm given above.

Combine all above modules together, makes the complete system that shares the excess bandwidth among the specified neighbor proxies efficiently.

6 Conclusion

This paper discussed an efficient mechanism for effective bandwidth sharing among cooperative organizations as an extension to *Squid*, a widely used web proxy server. The main advantage of this proposal is its *dynamic adaptability* to the changes of bandwidth usage pattern of individual organizations.

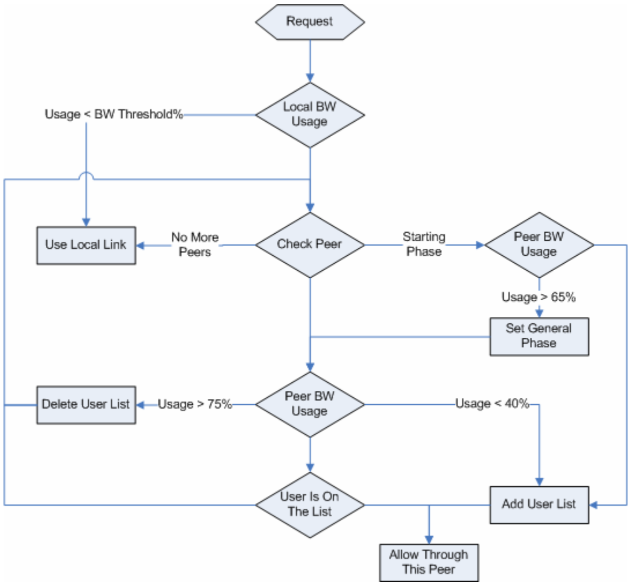


Fig 6: Local User Redirection Algorithm

System assures sharing *only* the excess bandwidth which is wasting otherwise. Therefore this proposal will increase the productivity as a whole within the group without sacrificing performance of individual organizations.

Currently we have deployed the system at the University of Moratuwa for a selected user group.

References

1. Home Page, "Squid Web proxy Cache", http://www.Squid-cache.org, Referred: January 2003
2. Squid: A User's Guide, "Delay Classes", http://Squid-docs.sourceforge.net, Referred: June, 2005
3. Gihan Dias & Chamara Gunaratne, "Using Dynamic Delay Pools for Bandwidth Management", 7th International Workshop on Web Content Caching and Distribution, Boulder, Colorado, August 2002.
4. Home Page, "Offline Browser" http://www.surfoffline.com, Referred: June, 2005
5. Austin Poulton, Peter Clayton and F F Jacot-Guillarmod, "A Bandwidth Management and Pricing Proxy", citeseer.ist.psu.edu/548382.html, Referred: December, 2004
6. Chamara Disanayake, Gihan Dias and C. R. de Silva, "A Market-Based approach to control Web bandwidth Usage", APAN, Cairns, Australia, July 2004.
7. Pradeepa Gurusinghe, Gihan Dias, Vishaka Nanayakkara, "On-the-fly Inter-proxy Data Compression for Web Access", SANOG 4, Kathmandu, Nepal
8. Milan P. Stanic, "tc – traffic control" webpage, http://www.rns-nis.co.yu/~mps/, Referred: July, 2005
9. D. Wessels and K. Claffy, "Internet Cache Protocol" and "Application of ICP": RFC 2186 and RFC 2187.

10. Cho-Yu Chiang, Mikihiro Ueno, Ming T. Liu, Mervin E. Muller “Modelling Web Caching Hierarchy Schemes”: Department of Computer and Information Science, The Ohio State University, Columbus, Ohio pp. 20–p28.
11. Sumith Gamage & Gihan Dias, "Dynamic Bandwidth Negotiation among Web Proxies", ERU (Engineering Research Unit) Symposium 2004, Sri Lanka.
12. Duane Wessels, “Configuring Hierarchical Squid Caches”, http://www.Squid-cache.org/Doc/Hierarchy-Tutorial/, Referred: January, 2005
13. Lars Burgstahler, Martin Neubauer, “New Modifications of the Exponential Moving Average Algorithm for Bandwidth Estimation”, Institute of Communication Networks and Computer Engineering, University of Stuttgart.

SIP over “Fast-Track” TLS for Mobile VoIP

C. Ling
Department of Computer and Systems Sciences,
Royal Institute of Technology (KTH),
Forum 100 SE-164 40 Kista,
Sweden.
Email: lingc@kth.se

Abstract

In this paper, a new SIP secure schema based on TLS Fast-Track Session Establishment technique is proposed. This schema consists of three phases: In the Fast-Track handshake-enabling phase, all the Fast-Track determining parameters have been stored in user agent (UA) side through a successful TLS handshake; in the registration phase, a Fast-Track handshake between UA and its register server is used to establish a TLS connection and all the SIP register messages are sent through this TLS tunnel; In the Interdomain request phase, all the TLS connections have been established hop-by-hop based on the proposed SIP over “Fast-Track” TLS schema. The TLS secure channel secures all the SIP messages for a call setup attempt. Security analysis proves that this schema achieves the necessary secure requirements for SIP used in a mobile VoIP system. Experimental results show that this schema minimizes the network bandwidth consumption and the processing cost at the same time. Because of these characteristics of the SIP over “Fast-Track” TLS schema, it is especially suitable for mobile VoIP systems.

Keywords: SIP, SIP over “Fast-Track” TLS, mobile VoIP

1. Introduction

Voice over Internet protocol (VoIP)[1] refers to the transmission of the speech across a data network. This form of transmission is conceptually superior conventional circuit switched communication in many ways. By using it, people can make economical long distance calls from PC to phone, from phone to phone, or from PC to PC.

In recent years, Wireless Local Area Network (WLAN) has been developed rapidly. Years of experience with WLAN have made this technique mature. The introduction of VoIP to a WLAN (mobile VoIP) seems to be natural. It can give users both mobility and convenience. Users can enjoy VoIP services, wherever they want. However, VoIP over WLAN (VoWLAN)[2] technique brings new challenges. The VoWLAN system transmits their packets through air. An attacker can easily intercept and modify them. Therefore, it is necessary to implement some security mechanisms [3], such as SIP over TLS (Transport Layer Security (TLS) protocol is an Internet Engineering Task Force (IETF) stander. The primary goal of the TLS protocol is to provide privacy and data integrity between tow communicating applications. This protocol is composed of two components : the TLS Record Protocol responsible for data transfer and the TLS Handshake Protocol responsible for the establishing TLS session states between the communicating peers.), SRTP, etc, in such a mobile VoIP system. However, almost all these security mechanisms will consume additional bandwidth of network and increase the computation cost of User agents (UAs). The limited bandwidth of WLAN and the limited computation ability of mobile devices seem to be a serious problem to implement these security mechanisms.

This paper presents a possible solution to minimize the bandwidth consumption and the computation cost for using TLS in SIP. First, we introduce a TLS Fast-Track Session Establishment technique, which uses client-side caching for TLS. Then, we adapt the existing SIP over TLS to SIP over “Fast-Track” TLS. According to experimental results, we conclude that our solution minimizes processing cost and network bandwidth consumption at the same time.

2. Introduction of “Fast Track” TLS [4][5]

In order to reduce the bandwidth consumption and computation on both ends, the TLS protocol incorporates a mechanism for session reuse. Unfortunately, session reuse relies on a server session cache. Heavily loaded servers, such as public VoIP registrar/proxy server server, have hundreds of TLS connections per minute. They cannot store so many cached sessions. However, clients rarely connect to numerous TLS servers, and can cache information for a longer time.

2.1 Cacheable parameters of Fast-Track

Fast-Track session establishment takes advantage of this observation to improve the TLS handshake efficiency. Because some servers' public parameters change infrequently, Fast-Track clients can maintain a cache of long-lived server parameters. These parameters include:

- The server certification chain;
- The server's Diffie-Hellman group (if any) for EDH key exchange;
- Whether the server requests client authentication; if so, what client certification types and certificate authorities the server is willing to accept;

Moreover, some parameters are negotiated between client and server, but also change infrequently. They also can be cached by the client. These include:

- Preferred client-server cipher suite;
- Preferred client-server compression method;

Thus all of these items need not to be retransmitted in a Fast-Track handshake. This mechanism saves significant bandwidth. Moreover, the handshake messages are rearranged, reducing the total number of roundtrips from four (there are four roundtrips in original TLS handshake) to three. These saving can be significant, especially for wireless networks.

2.2 Negotiation of Fast-Track

The Fast-Track session establishment requires that both client and server must support Fast-Track. Moreover, Client cannot establish a Fast-Track session, when the server is not aware of it or is not willing to participate.

In establishing a negotiation session with server, a client may include the fasttrack_capable extension in the ClientHello (or ClientHelloFT) message. The server may then include the fasttrack_capable extension in the ServerHello (or ServerHelloFT) message assenting to the use of Fast-Track in the further handshakes with that client.

The client must determine some parameters in the negotiation session. These include [6]:

```
●   struct {
            opaque dh_p<1..2^16-1>;
            opaque dh_g<1..2^16-1>;
            } ServerLongLivedDHParams;
●   struct {
                CipherSuite cipher_suite;
```

```
                CompressionMethod;
                compression_method;
                Certificate server_certs;
                select (KeyExchangeAlgorithm) {
case
  diffie_hellman: ServerLongLivedDHParams llparams;

 case rsa: struct { };

 };
 ClientCertificateType certificate_types<0..2^8-1>;
 DistinguishedName certificate_authorities<0..2^16-1>;
  } FTDetParams;
```

Then, the client must send a hash of these determining parameters to the server in the later Fast-Track handshake, using the fasttrack_hash extension.

2.3 Handshake of Fast-Track

To engage in a Fast-Track handshake, a client and a sever must agree on the Fast-Track determining parameters, which will be reused later. The client can obtain these parameters during a previous enabling handshake. To start a Fast-Track handshake, the client sends a ClientHelloFT message with a fasttrack_hash to server. If the server want to start a Fast-Track handshake and the fasttrack_hash matches the hash of determining parameters on server's side, the server can accept the Fast-Track handshake attempt by replying with a SeverHelloFT message. Finally, the client sends a finishing message to finish this handshake. The message flows for a successful Fast-Track handshake are summarized in Fig 1.

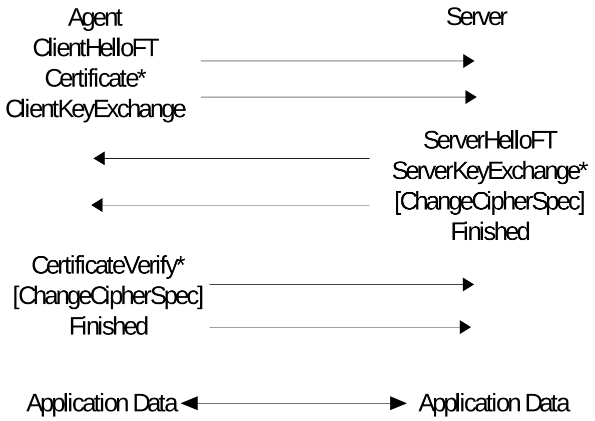


Fig 1. The message flows for a successful Fast-Track handshake

If a sever doest not want to accept a Fast-Track handshake, it can also deny a Fast-Track handshake attempt and start a ordinary TLS handshake or resume a previous TLS session.

3 Introduction of SIP over TLS for mobile VoIP

Mobile VoIP systems can use TLS [7] for communication with proxy, redirect, and register servers as well as user agents to protect SIP [8] signaling. TLS can be specified as the desired transport protocol within a Via header field value or a SIP-URI [9]. The identifier will be tls. To implement SIP over TLS [10], there are two different scenarios: For a SIP register message, the register server and the UA must agree on a root Certificate Authority beforehand. When the UA registers, it sends a TLS connection request. After the register server receives that request, the register server offers its certificate to the UA. The UA verifies server's certificate and establishes a TLS connection with that register sever, before it sends out a SIP register message. For other SIP messages, there are two different hops: On the hop between the user agent and its SIP proxy server, the TLS tunnel established in the register phase can be reused. On the hop between SIP proxy servers, a new TLS connection from one SIP proxy server to another proxy server should be established. Since both of the participants in this TLS connection are servers that possess site certificates, mutual TLS authentication should occur in this hop.

The advantage of SIP over TLS is that this model provides peer-to-peer mutual authentication between the two parties by using the TLS handshake, allowing hop-by-hop message confidentiality and message authentication, as well as a dynamic negotiation of cipher suites and secure key distribution. At the same time, TLS is able to protect SIP signaling messages against loss of integrity, confidentiality and against replay attack.

The drawback of SIP over TLS: TLS requires a reliable transport stack (TCP-based SIP signaling). TLS cannot be applied to UDP-based SIP signaling. The certificates and the negotiation information transmitted during TLS handshake consume lots of network bandwidth. In order to verify certificates and negotiate cipher suite and master key, the computation cost of UAs has increased dramatically. Because UAs use WLAN to connect to their registrar/proxy server servers and UAs are usually installed in a mobile device or handset in mobile VoIP systems, these drawbacks can seriously damage the performance of mobile VoIP systems.

4 Propose the SIP over "Fast-Track" TLS for mobile VoIP

In order to overcome some disadvantages of SIP over TLS above, a new secure SIP schema is needed. At the same time, we found the UA in a mobile VoIP system usually connects to the same registrar/proxy server server. Hence the certificate of the server and some parameters of the TLS connection between the UA and its server don't change frequently. Based on these findings, we propose a SIP over "Fast-Track" TLS schema for mobile VoIP. This schema can efficiently minimize network bandwidth consumption between registrar/proxy server servers and UAs as well as decrease the computation cost load on user agents. There are three phases in this schema: the Fast-Track handshake-enabling phase, the registration phase, and the Interdomain request phase. We will consider each in turn.

4.1 Fast-Track handshake-enabling phase

In this phase, the UA and the registrar/proxy server need to establish a successful TLS handshake as before. The only difference is that both the TLS ClientHello message and the TLS ServerHello message must include fasttrack_capable extension. In this way, the UA and the registrar/proxy server agree to the use of Fast-Track in future handshakes. The UA collects the determining parameters for Fast-Track handshake and stores these parameters in its cache during this phase.

4.2 Registration phase

When a UA registers (i.e. provide its registrar with its IP address), it should first establish a "Fast-Track" TLS connection with its registrar/proxy server. In this phase, the UA assumes that its handshake with the registrar/proxy server will reuse the Fast-Track determining parameters negotiated in the previous enabling phase. It sends out a

ClinetHelloFT to the registrar/proxy server. The registrar/proxy server can accept the Fast-Track handshake attempt by replying with a ServerHelloFT message. The UA sends back a finished message to establish TLS connection with registrar/proxy server. The UA then creates a register request that is addressed to a Request-URI corresponding to the site certificate received from the registrar/proxy server. Finally, the UA sends the register request over the existing TLS tunnel and the registrar/proxy server sends back a 200 OK message to finish the registration. The message flows for a successful UA register phase is illustrated in Fig2.

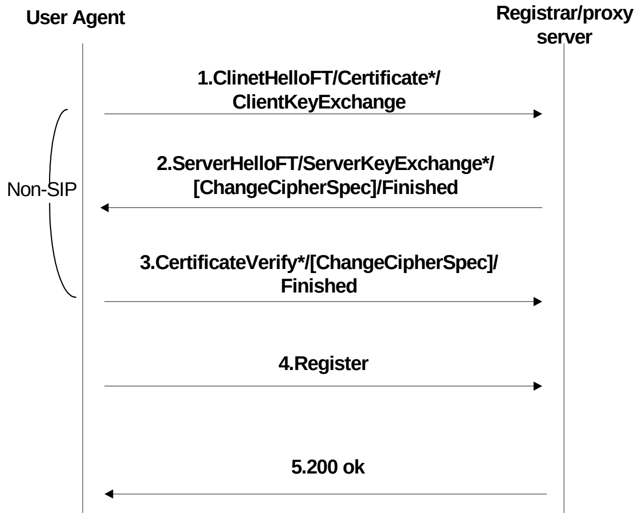

User Agent
Registrar/proxy server
1.ClinetHelloFT/Certificate*/ ClientKeyExchange
Non-SIP
2.ServerHelloFT/ServerKeyExchange*/ [ChangeCipherSpec]/Finished
3.CertificateVerify*/[ChangeCipherSpec]/ Finished
4.Register
5.200 ok

Fig.2 The message flows for a successful UA register phase

4.3 Interdomain request phase

After the registration, user agent A would like to initiate a session with user agent B in a remote administrative domain. For simplicity's sake, we assume that the register server also acts as a proxy server for UAs.

We distinguish two different hops in this interdomain request phase:

1.The hop between user agent A and its proxy server.

Because the user agent A has completed the registration process described in the preceding section, it can reuse the TLS connection to their local proxy server to send the INVITE request (otherwise the user agent A should initiate a new Fast-Track TLS connection to its proxy server at this point). When the INVITE message arrives and has been approved by user agent B's proxy server, the proxy server should identify the existing TLS channel, if any, associated with user agent B. The INVITE can then be proxied to user agent B through this secure channel.

2.The hop between user agent A's proxy server and user agent B's proxy server.

In order to provide hop-by-hop SIP security, user agent A's proxy server should establish a TLS connection with user agent B's proxy sever. Normally the ordinary TLS handshake process will be used to establish the TLS connection in this hop. There are two reasons: a) either user agent A's proxy server or user agent B's proxy server could be a client or a server in the TLS handshake (If user agent A initiates a session with user agent B, user agent B's proxy server would be a server in a TLS handshake. If user agent B initiates a session with user agent A, the user agent A's proxy server would be a server in a TLS handshake). It is very difficult to efficiently implement "Fast-Track" TLS handshake in this scenario; b) since both of the participants in this TLS connection are servers instead of mobile devices and the network between these two servers is likely to be a broadband connection rather than a wireless network, these nodes are likely to have more computation power and greater network bandwidth to establish TLS connection. Hence the cost of an ordinary TLS handshake won't have a significant impact on performance of communication on this hop.

Each server of this connection should inspect and verify the certificate of the other, noting the domain name that appeared in the certificate for comparison with the header fields of the SIP message. Mutual TLS authentication should occur there. The message flows between user agent A's proxy sever and user agent B's proxy sever are summarized in Fig. 3

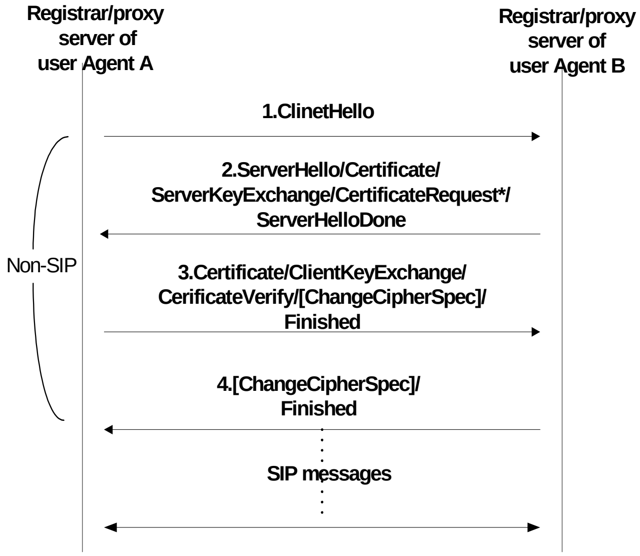

Registrar/proxy server of user Agent A
Registrar/proxy server of user Agent B
1.ClinetHello
Non-SIP
2.ServerHello/Certificate/ ServerKeyExchange/CertificateRequest*/ ServerHelloDone
3.Certificate/ClientKeyExchange/ CerificateVerify/[ChangeCipherSpec]/ Finished
4.[ChangeCipherSpec]/ Finished
SIP messages

Fig.3 The message flows between user agent A's proxy sever and user agent B's proxy sever

4.4 All the message flows of a SIP over "Fast-Track" TLS call setup attempt

Figure 4 shows all the message flows of a SIP over "Fast-Track" TLS call setup attempt. From the figure, it is not difficult to find out that the hop between user agent A and its registrar/proxy server, the hop between SIP registrar/proxy server A and SIP registrar/proxy server B, and the hop between user agent B and its registrar/proxy server are all protected by TLS secure tunnel. The TLS handshake flows between UAs and their registrar/proxy servers are three instead of four. It means that we can achieve security and efficiency at the same time.

5 Security analysis of SIP over "Fast-Track" TLS

Because our SIP over "Fast-Track" TLS schema provides hop-by-hop SIP security as showing in Figure 4, we are going to carry out security analysis hop-by-hop.

On the hop between user agent A and its registrar/proxy server—server A, we used Fast-Track to establish a secure TLS channel. Since the security analysis of TLS has been done by Mitchell et al. [11] and Paulson [12], we are not going to tall about it in this paper. We only present common arguments about security of Fast-Track and security of SIP over "Fast-Track" TLS.

Except for the fasttrack-capable extension, the ClientHelloFT message and ServerHelloFT message sent by user agent A and server A are almost the same as ClientHello message and ServerHello message in an ordinary TLS handshake. The extension_data field of the fasttrack_capable extension is empty. This means there is no sensitive information in fasttrack-capable extension. Hence, the Fast-Track hello messages don't have greater security risks than ordinary TLS hello messages.

After the first enabling-handshake phase, the user agent A verifies server A's certificate and collects some parameters from that handshake. Server A's certificate and some long-lived parameters are stored secretly in user agent A's side. They are not open to tamper. Even if these determining parameters have been tampered with, the mismatch of a SHA-1 hash of the determining parameters will be detected in the course of handshake. Then the Fast-Track TLS connection attempt will be rejected.

Furthermore, the finished message in Fast-Track has been protected by MAC. So both the user agent A and server A can verify the integrity of that message. Hence, a third party can not modify the finished message with it being detected. From the analysis above, we know that the Fast-Track won't cause any security problems. We can turn to analysis the security of SIP over "Fast-Track" TLS now.

If server A requires that user agent A provide a certificate in the Fast-Track session establishment phase, the mutual

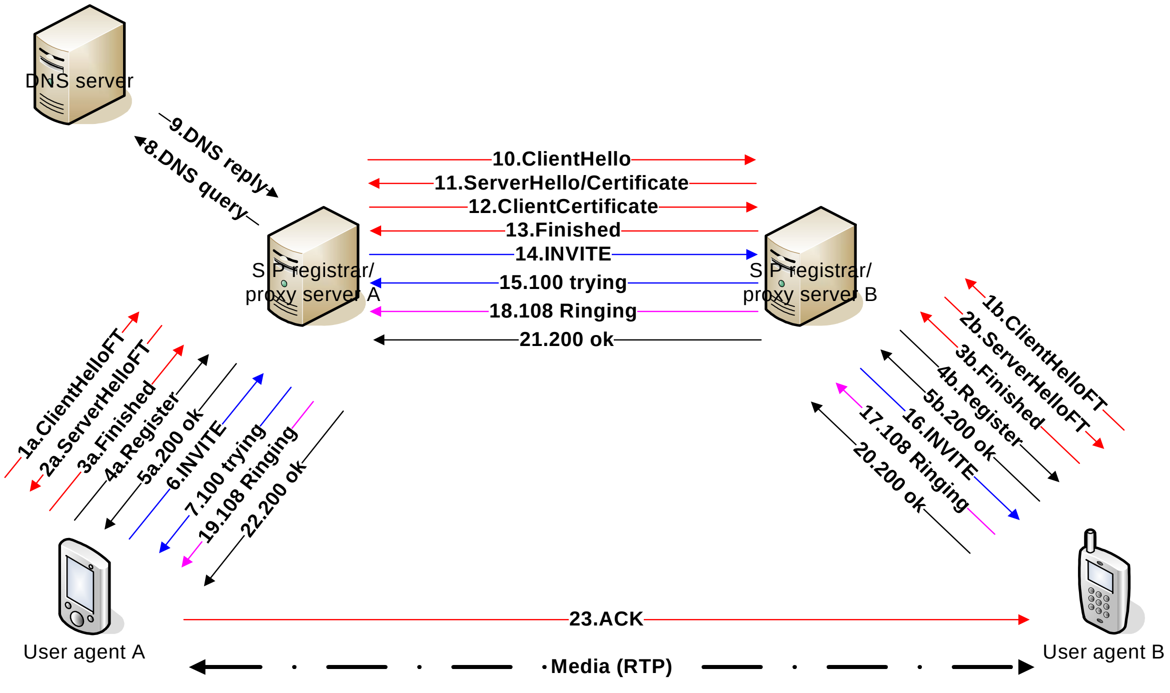


Fig.4 All the message flows of a SIP over "Fast-Track" TLS call setup attempt

authentication between both parties has been done. Server A knows user agent A is really who it claims. User agent A knows that the registrar/proxy server is not an attacker who might redirect the UA, steal passwords, or attempt an similar attack.

Since we use TLS-DHE-RSA-WITH-3DES-EDE-CBC-SHA cipher suite in Fast-Track, the SIP messages are encrypted by 3DES algorithm, which is a very strong cryptographical algorithm. Now, both the UA and server would send and receive SIP messages over this authenticated channel. They can be assured that those SIP messages have been authorized by the other party, and no one has modified and repudiated them.

The same situation goes for the hop between user agent B and its registrar/proxy server—server B.

On the hop between registrar/proxy server A and registrar/proxy server B, we use ordinary TLS handshake to establish the TSL connection. The registrar/proxy server A can verify that the certificate received from the remote side corresponds with registrar/proxy server B's domain, and registrar/proxy server B also can verify registrar/proxy server A in this way. Hence, a secure and mutual authenticated channel has been established between two servers. All the SIP messages transmitted on this hop can be protected.

We can use DNSSEC to protect the hop between the registrar/proxy server A and the DNS sever. However, this is beyond the scope of this paper. We are not going to describe it here.

6 Performance comparison between SIP over TLS and SIP over "Fast-Track" TLS

In order to prove the advantage of SIP over "Fast-Track" TLS for mobile VoIP systems, we carry out two tests:

1. We measure the CPU load of UA as well as the total traffic transmitted between UA and its registrar/proxy server, when we use SIP over TLS in a mobile VoIP system.
2. We measure the CPU load of UA as well as the total traffic transmitted between UA and its registrar/proxy server, when we use SIP over "Fast-Track " TLS in a mobile VoIP system.

6.1 Test resources

We use a MiniSIP as our user agent in the test. MiniSIP is a SIP user agent which was developed by Erik Eliasson at KTH, Stockholm, Sweden. It is a SIP based soft phone, which works under LINUX. MiniSIP can work both on the iPAQ and on a workstation/laptop running Linux. MiniSIP can be installed on a mobile device with Wi-Fi support and can be used as a mobile VoIP phone**.**

We use two laptops as client machines to install MinSIP. Each client machine has the following specifications:

IBM T40 Laptop with Pentium M 1500 MHz with 256 MB Ram

Operating System: Linux Federal Code 2.

MiniSIP Version 0.1 with LibMIKEY Support

WLAN Card: Intel® PRO/Wireless LAN 2100 3B Mini PCI Adapter

6.2 Testbed Environment

Fig. 5 shows our testbed setup. The wireless network in our testbed is an 802.11b WLAN in the Forum Building in Kista, Stockholm. The wired network in our testbed is an Ethernet in the same building. We install Ethereal 0.10.0 on both client machines to monitor traffic.

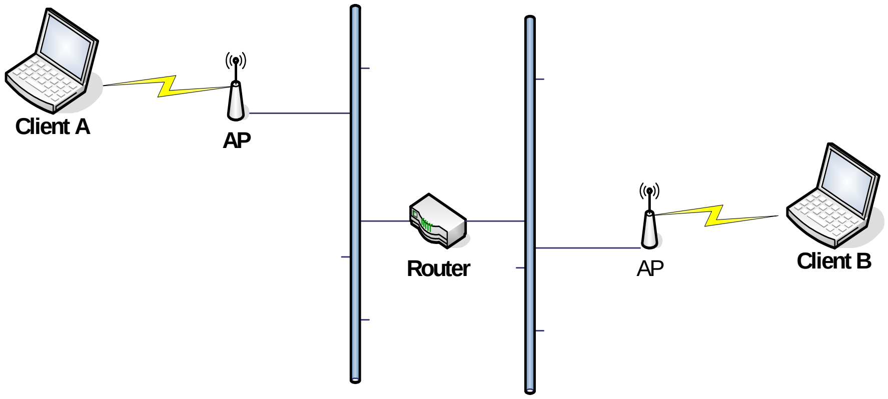


Fig.5 Testbed Environment

6.3 Test result

Because MiniSIP already has TLS support, we only need to configure MiniSIP to use SIP over TLS test. We use TLS-DHE-RSA-WITH-3DES-EDE-CBC-SHA cipher suite and none client certificate mode for TLS connection in this test.

In the SIP over "Fast-Track" TLS test, it was difficult for us to find a SIP user agent which had the SIP over "Fast-Track" TLS built in, and we did not have enough time to modify MiniSIP to support "Fast-Track" TLS. So we choose an alternate method to carry out this test. First, we use Fast-Track to establish a TLS connection, then, we use MiniSIP to send SIP messages over that TLS tunnel. We also use TLS-DHE-RSA-WITH-3DES-EDE-CBC-SHA cipher suite and no client certificate mode for TLS connection in this test.

In the test, we calculate all the traffic between the UA and its registrar/proxy server for SIP call setup attempt. We also record the client machine's CPU load during the TLS connection establishment. The test results are presented below:

Traffic (bytes)	SIP over TLS	SIP over "Fast-Track" TLS	Saving
The hop between user agent A and it's registrar/proxy server	4803	3519	1284
The hop between user agent B and it's registrar/proxy server	4203	2922	1281
Total	9006	6441	2565

Table 1 the traffic in bytes between the UA and its registrar/proxy server for SIP call setup attempt

CPU load (%)	SIP over TLS	SIP over "Fast-Track" TLS	Saving
Client machine A	48.1%	38.6%	9.5%
Client machine B	40.2%	28.8%	11.4%

Table 2 the client machine's CPU load

From the table above, we see that SIP over "Fast-Track" TLS schema saves a lot of bytes per connection between UA and its registrar/proxy server. Hence, we can save a large amount of traffic, when there are hundreds of connections between a UA and its registrar/proxy server, using SIP over "Fast-Track" TLS schema. The performance of network will be improved in this way.

7 Conclusion

The experimental results demonstrate that SIP over "Fast-Track" TLS schema indeed minimizes the traffic between UA and its registrar/proxy server and improves CPU load of UA side machine. Because of these two characteristics, The proposed SIP over "Fast-Track" TLS schema is very suitable as a secure SIP for mobile VoIP systems. It can provide SIP security and efficiency for a mobile system at the same time.

8 Further work

Because of limited time, some work remains to be done in the future. As we mentioned in section 6.3, we used an alternative way to test SIP over "Fast-Track" TLS. The improvement of UAs to support SIP over "Fast-Track" TLS should be done in the future. Additional, such as the comparison of the initial delay between SIP over TLS and SIP over "Fast-Track" TLS, might also be interested.

References

[1] Luan Dang, Cullen Jennings, and David Kelly, *Practical VoIP: Using VOCAL*, O'Reilly, 2002, ISBN 0-596-00078-2.

[2] Khurram Jahangir Khan, and Ming-Shuang Lang, Voice over Wireless LAN and analysis of MiniSIP as an 802.11 Phone, course report, KTH, June 2004. <http://www.it.kth.se/courses/2G1325/2g1325_Khurram_and_Ming-Shaung--20040629.pdf>

[3] Israel M. Abad Caballero, Secure Mobile Voice over IP, Master of Science Thesis, KTH, June 2003. <ftp://ftp.it.kth.se/Reports/DEGREE-PROJECT-REPORTS/030626-Israel_Abad_Caballero-final-report.pdf >

[4] Hovav Shacham, Dan Boneh, and Eric Rescorla, Client-Side

Caching for TLS, *ACM Transactions on Information and System Security*, Vol.7, No.4, Pages553–575, November 2004.

[5] Hovav Shacham and Dan Boneh, TLS Fast-Track Session Establishment, Internet Draft: draft-shacham-tls-fasttrack-00.txt, Work in progress, 2001.
<http://www.infres.enst.fr/~badra/draft-shacham-tls-fasttrack-00.tx t >

[6] S. Blake-Wilson, M. Nystrom, D. Hopwood, J. Mikkelsen and T. Wright, Transport Layer Security (TLS) Extensions, RFC 3546, June 2003.
< http://www.faqs.org/rfcs/rfc3546.html >

[7] T. Dierks, and C. Allen, "The TLS Protocol, Version 1.0", RFC 2246, January 1999.
< http://www.faqs.org/rfcs/rfc2246.html >

[8] Henry Sinnreich and Alan B. Johnston, *Internet Communications Using SIP: Delivering VoIP and Multimedia Services with Session Initiation Protocol*, Wiley, 2001, ISBN: 0-471-41399-2

[9] G. Q. Maguire Jr., ‘2G1325/2G5564 Practical Voice Over IP (VoIP): SIP and related protocols, Spring 2005, Period 3, Lecture notes.
<http://www.imit.kth.se/courses/2G1325/VoIP-2005.pdf >

[10] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley and E. Schooler, SIP: Session Initiation Protocol, RFC 3261, June 2002.
< http://www.faqs.org/rfcs/rfc3261.html >

[11] J. Mitchell, V.Shmatikov, and U. Stern, Finite-state analysis of SSL 3.0. *In Proceedings of USENIX Security*, A. Rubin, Ed. USENIX, 201–216, *1998.*

[12] L. C. PAULSON, Inductive analysis of the Internet protocol TLS, *ACM Transactions on Information and System Security,* Vol.2, No.3, 332–51, 1999.

Unifying Keywords and Visual Contents in Image Retrieval for Images Collected from Major Web Sites

L. Jayaratne*,A. Ginige** and Z.Jiang***
School of Computing and Information Technology
University of Western Sydney
Sydney, Australia
*k.jayaratne@uws.edu.au,**a.ginige@uws.edu.au and ***z.jiang@uws.edu.au

Abstract

The state-of-the-art image retrieval approach is to incorporate image semantics with low-level visual primitives to enhance the retrieval performance. In this paper we exploits the vast power of image semantics from both the text associated with the image in a web page and low-level visual features for abstraction of higher-level semantic contents to facilitate the semantic-based access to collection of images from major web sites. The contribution of our work includes conceptual level explanation of how the clustering of the image basic visual features to support browsing of images by semantic contents, as well as to improve the accuracy of semantic matching. We explore the unification of keywords and low-level visual contents for image retrieval, and propose a seamless joint approach using different evidence combination mechanisms to best of two worlds. These strategies were implemented in our prototype search engine named I-Search and we report the accuracy and effectiveness of our search engine with experimental results to improve precision and recall performance on collection of images from a major web site. The significance of this work includes a greatly improved understanding of content-based indexing of images and substantially enhanced image searching capabilities on the web.

1 Introduction

More than a decade ago some researchers were estimating that as many as an astounding one million digital images were being produced every day [1]. The growth of the World Wide Web (WWW) over this period has produced unprecedented opportunities for accessing and sharing digital images. However, the technologies for creating collections of digital images at present outstrip our capabilities for effective and efficient retrieval of these images. End-user access is a topic of vital concern to all those involved in building image databases.

There have been two distinct research communities driving research in image retrieval forward. The first has its roots in traditional information retrieval, using techniques and methods from text based information systems for indexing and retrieval. The second major area is the field of computer vision, which uses methods and algorithms to analyze and index images based on their visual content, such as color, shape, texture and layout [2].

Indexing and retrieval of images using techniques and methods from traditional information retrieval has *semantic understanding* as one of its main strengths. The most widely used technique relies on textual description of images, mostly using keywords or free text. This form for indexing has high expressive power, it can be used to describe almost any aspect of image content. In addition, a wide range of existing text retrieval software can automate the process of searching and retrieval. Keyword-based image retrieval is therefore more preferable, since it matches the query keywords against the image annotated directly with the higher-level semantic features. These systems [3] can be adopted for web images since the surrounding HTML tags typically describe the semantics of these images. However, this approach entails a tedious and labor-intensive process of manual annotation, which may also introduce errors due to the differences in human subjectivity. The combination of rich image content and differences in human perception makes it possible for two individuals to have very diverging interpretations of the same image. As a result, the description is prone to be subjective.

Another approach to retrieval from image databases is rooted in the field of computer vision. As a response to the difficulties large scale image collections posed to traditional techniques, content-based image retrieval (CBIR), was proposed. Rather than rely on annotation and textual descriptions, CBIR systems use automatic feature extraction for indexing and retrieval of images based on image content. The features used for retrieval can either be primitive or semantic, but the extraction process must be predominantly automated [4].

CBIR systems are classified into different levels of abstractions. At the lowest level, retrieval is based on *basic primitive features*, such as color, shape, and texture. Higher level describes higher level of *semantic attributes*, such as object types, individual objects, events, actions and so on. In this paper, we will highlight what is considered one of the greatest challenges to CBIR systems; *bridging the semantic gap*. Most current commercial solutions only operate on the lowest level, using the extracted basic primitive features as basis for a mathematical similarity search [4]. Recent research has shown that while primitive features work well for very specialized application areas, such as crime prevention (For fingerprints) and medicine [5], they do not provide adequate support for more general application areas. Therefore most of the concurrent research is aimed at bridging the gap between low-level primitive features and higher-level semantic content, the "*semantic gap*" [4]. For example, as mentioned by Santini and Jain [6], CBIR system solely depends on extraction and comparing of primitive features has no understanding of the images' semantic contents; fail to recognise similarities in semantic content at the higher-level.

In the recent years, it has been suggested that a combination of techniques from both areas of research might yield better results than either approach separately [6], thus bridging the semantic gap. Huang and Rui [7] suggest that an interdisciplinary research effort is required in order to build a successful Image Database System, is the *only* way satisfactory retrieval performance can be achieved. It is suggested that CBIR not is a replacement of text-based retrieval, but a complementary component. We therefore propose an alternate, modified architecture based on the same ideas, but extending text annotation with the aim of finding higher-level semantic contents in each image for low-level primitive features.

In this work, we select the textual information in a web page that can be used to construct a good semantic representation model for the web images. These include semantics from text associated with the image such as title of the image, ALT string, and its caption and semantics from text associated with the web page such as HTML meta data and title of the page. In addition, we evaluate mechanisms for classifying and finding images on the web by combining evidences due to different types of information and used an approach suitable mapping to relate the most salient pictorial features to higher level semantic attributes (in terms of labels) for the web images. For this we have developed a procedure to qualitatively measure the saliency of a feature towards this classification, which capture the visual characteristic of each of the images were computed and analyzes images' associated text with the aim of finding clues about semantic contents in each image. This builds on the *idea* that text based annotation can work *side-by-side* with a CBIR system. By extending text annotations with the aim of finding higher-level semantic contents in each image for low-level visual features, I hope to gain the benefits of a combinatory approach, in which data from visual content analysis of images is combined with annotation-based search within a unified framework in order to benefit from the effectiveness of the former and the efficiency of the later.

The rest of the paper is organized as follows. Section 2 briefly discusses some of the recent work in the literature on CBIR, and higher-level image classification using low-level visual primitives. We provide details of our proposed method in Section 3. The results of our prototype search engine ***I-Search***, evidence of combinatory approach, and quantitative evaluation of web image search engines are given in Section 4. We finally outline concluding remarks for the future directions with areas for further research in Section 5.

2 Related Works

Various systems have been proposed in the recent literature for CBIR, such as QBIC [8], Virage [9], SIMPLIcity [10], Netra [11], VisualSEEK [12], and MARS [13]. A number of web image search engines have been built in recent years, including both research prototypes and commercial products. Among the former category are WebSeer [14], WebSEEK [15], WebHunter [16], Diogenes [17], iFind [18], and 2M2Net [19]. Commercial web search engines such as Google^{TM}, Altavista^{TM}, Yahoo^{TM}, Lycos^{TM}, and Ditto^{TM} offer image search facilities. We will overview some of these systems in terms of the features they use to index images and how they resemble or differ from *I-Search*.

The commercial image search engines, Google^{TM}, Altavista^{TM}, Yahoo^{TM}, Lycos^{TM}, and Ditto^{TM}, apparently do not perform visual analysis and *rely* heavily on the images' associated text. Though most of these commercial image search engines are very popular, we note that they are not using any visual criteria to classify images as graphs, cartoons, photographs, clip-art, etc. Therefore we need to apply some criteria to the HTML documents over the WWW, which essentially provides a rich source of image collection, to especially *eradicate* the inclusion of *unwanted* images such as symbols, logos, etc from the query results. Secondly, the results show that using the textual content of the HTML document as part of the image's semantic content cannot provide best performance, even though such an approach is used by the systems WebHunter [16] and Diogenes [17]. Our proposed approach shows the identification of proper semantic content from various combinations of the textual contents in HTML document, can lead to very effective retrieval results. We thus expect our prototype search engine *I-Search* to outperform the existing web image search engines, maintaining a high level of recall of relevant images whilst optimizing the precision of the search.

A number of attempts have been made to understand higher-level semantics from images using low-level visual primitives. Szummer and Picard [20] proposed algorithms for indoor-outdoor image classification on a large database of 1,343 images. Vailaya and Jain [21] proposed a hierarchical indexing scheme for the outdoor images to be further dichotomized into city images or landscapes at the next level of classification hierarchy using the k-NN classifier based on edge direction coherence vectors using the leave-one-out method. However, the success of such clustering-based indexing scheme is often limited to attempting rather specific classification problem, largely due to the low-level visual feature based representation of image content for any trivial image collection. To achieve the goal of automatic categorization and indexing of web images in a large database, we need to develop robust schemes to identify salient features of images that capture a certain aspect of semantic content of these images. This is a well-known and important problem in pattern recognition and computer vision.

All the approaches described above differ not only in the classification schemes but also on the features used. In this paper, we show how a high level concept can be learned from images using relatively simple low-level visual features, and qualitatively measure the saliency of a feature towards a classification problem based on the discrimination power of hue component of HSV color histogram using the K-means classifier. The approach adopted here to relate salient visual features to meaningful labels. The experiment yielded mainly the following 4 categories: full faces, natural sceneries, events and city images.

3 Proposed Method

What do users typically want to search for in an image database? Users either have an a priori idea of what they are looking for (e.g., Figure 1 (a): Yasser Arafat's face) or they have a rather vague abstract notion of the pictures (e.g., Figure 1 (b): looking for scenery photographs for Taj Mahal).

3.1 Visual Features

In Figure 1 (a) and (b) images of the each figure comprise some elementary patterns, such as straight lines and curves, with some spatial relationships with each other. But, each of the images is quite distinct from the others in terms of shapes and relative positions because of differences in dimensions, focus and viewing position. And also similar differences are visible because of differences in viewing distances. In Figure 1 (b) the color of the structures, that can be used to recognize the building material, might also undergo transformations because of the differences in daylight or other lighting conditions. Therefore in order to identify the patterns with some spatial relationships in spite of such differences, there is a need for classification of database images based on the abstract concepts.

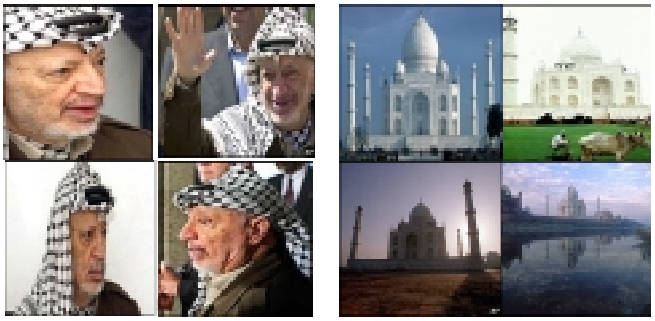

(a) **(b)**

Figure 1: (a) Typical images for full face of Yasser Arafat; (b) Typical scenery photographs for Taj Mahal.

3.1.1 Framework of Classification

There is currently no algorithm available to automatically determine the semantic content of arbitrary images accurately. The approach adopted here to relate salient visual features to meaningful labels. Understanding the entire content of an image may not be possible even with the state-of-the-art feature extraction and matching algorithms in computer vision. But we believe that it is feasible to attempt to classify images using low-level features geared towards semantically meaningful classes.

In order to generate a grouping based on the classification, we generated a 700 x 64 symmetric matrix. Based on the symmetric matrix, we performed a K-means clustering [22] of the 700 images. In our experiments, we used *K*=5. By setting *K*=5, we obtained highest natural demarcation with less significant mismatch for all the images. At the highest level of accuracy in terms of real life interpretation for all the images 4 prominent clusters, namely Full face, Events, Natural Scene and City Image were found. The resulting diagram of proper visual interpretation for cluster centers of 4 main categories is shown in Figure 2.

The images under the classes of buildings and street can be grouped into a single class, labeled city images (55 images). The classes of forests, farms, grounds, mountains, beaches/sea, and sunset/sunrise, can be grouped into the broader category of natural scenes (242 images). The Tanks, fighting, protests, and blasts can be grouped into the category of events (153 images). The images with a full face can be grouped into full-face (247 images) class, and images not belonging to any of the classes, into the miscellaneous (3 images) class. The details are listed in Table 1.

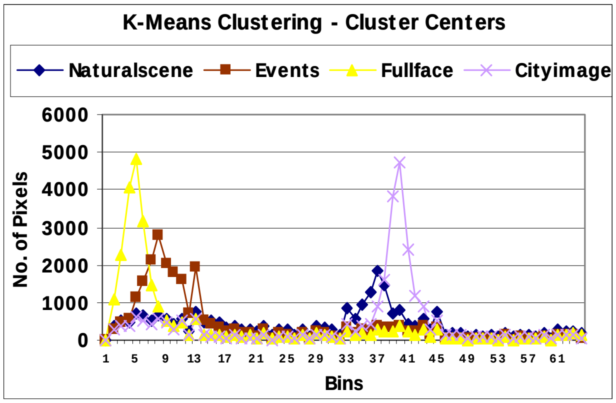


Figure 2: Cluster centers of 4 main categories generated for groupings.

Category	Classification
Full Face	247
Events	153
Natural Scenes	242
City Images	55
Miscellaneous	3
TOTAL	700

Table 1: Classification results using K-means classifier for color histogram feature

The image classification problem of interest here can be defined as follows: The input to the classification system is an image and the output is the label with which the system assigns one of the five image classes. We assume an existing set of images (training set of 700) to which the input image is compared. The labels for the training set of 700 images are assigned by human subjects. There obviously exist images that can be placed into multiple classes (e.g., a human face in an event, city skyline at natural scenery). Figures 3, 4, 5 and 6 show 20 images each from the full-face, events, natural scenes, and city image clusters generated by the classifier, respectively.

The proposed system does have the capability to reject some images that do not belonging to any of the classes, and these are called miscellaneous. While this experiment was useful in identifying meaningful semantic categories for retrieval process, we feel that in addition to generating a multi-class classification, it may also be feasible to attempt further classification for misclassified images based on the features that have high discriminability for a particular class (for example features of a human face in an image like eyes?). We call this process *semantic pattern matching*.

Category	Classification		
	True Class	Misclassified	Accuracy
Full Face	1036	29	97.3%
Events	602	90	87%
Natural Scenes	703	90	88.7%
City Images	133	52	71.4%
Miscellaneous	3	-	-
TOTAL	2477	261	90.5%

Table 2: Classification results using classifier for color histogram feature

3.1.2 Classification Results

Table 2 shows the classification results for the experiments. The overall accuracy of 90.5% (261 images were misclassified) was obtained using color histogram feature vectors [23]. Of the 261 misclassified images, 29 were full face, 90 were events, 90 were natural scenes, and 52 were city images. The details are listed in Table 2. A total of 2,477 images (90.5%) were classified with a confidence of over 75% (based on human observations) to their true class. These results demonstrate that the color histogram features have sufficient discriminatory power for the classification problem considered here.

3.2 Text/HTML Features of an Embedded Image

Based on the relationship between an image embedded in web document and its keywords, we identified some parts of the textual content that are well related to the embedded image. These are image title, image alternate text, image caption, page title and HTML meta data.

Extensions to the HTML specifications include new tags allowing the *context* summary of semantic attributes inside the HTML document. Obviously these tags have to be entered by the document author. However, when present, these keywords may reflect *better* and *real* content of the document. Thus, we extract the keywords from the meta tags appear for the image semantics. We exclude the body text (main text) that contains too much unrelated information not semantically related to the images in the HTML document. It adds only *noise* into the semantics of the images. Therefore, we use only above five parts to represent image semantic content.

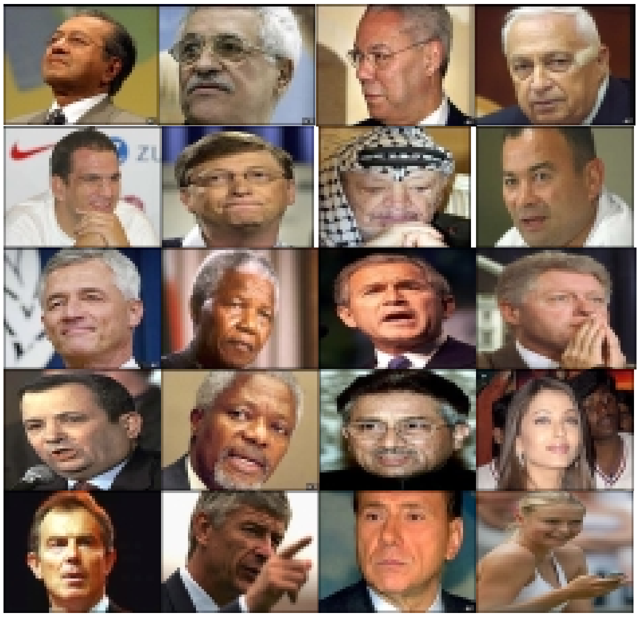


Figure 3: Representative images of full face class generated by the classifier.

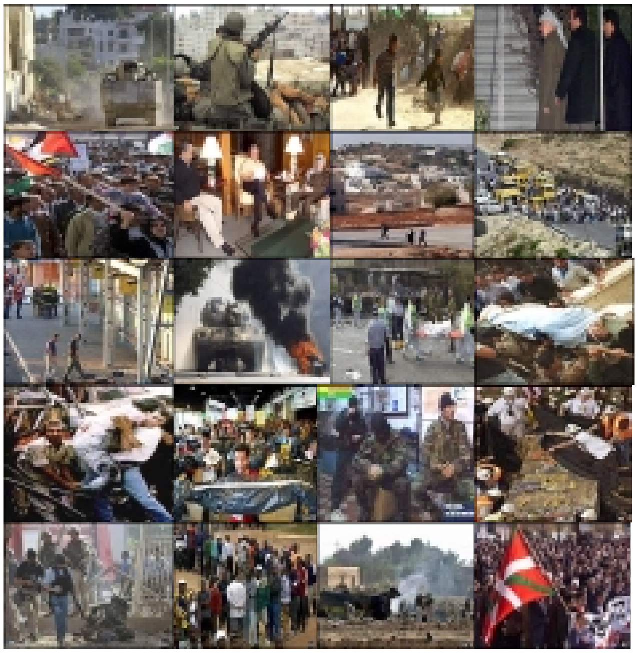


Figure 4: Representative images of event class generated by the classifier.

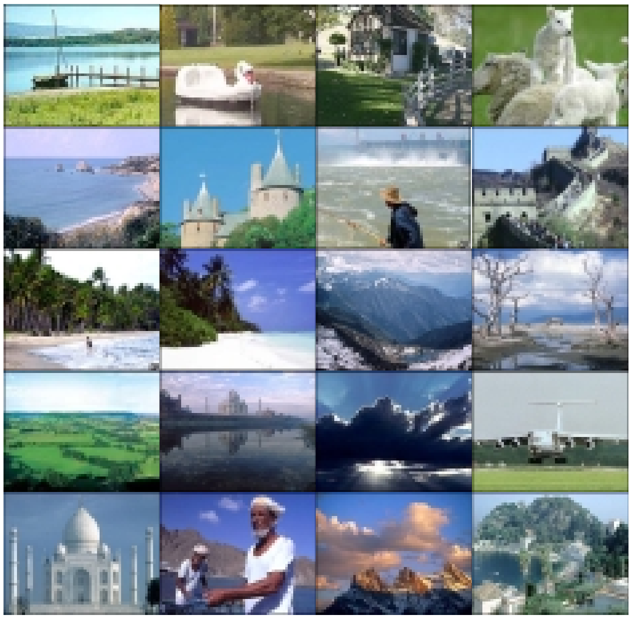


Figure 5: Representative images of natural scenes class generated by the classifier.

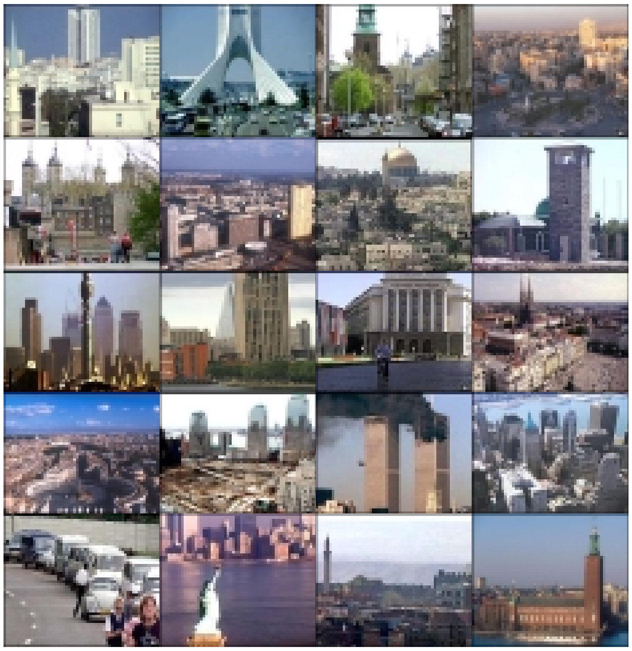


Figure 6: Representative images of city images class generated by the classifier.

The implementation of our text engine uses some elementary techniques from information retrieval to associate keywords to images and retrieve images based on the keywords. Keyword association occurs automatically whenever possible. In this case of web images, we parse the web page and collect the keywords for the corresponding images. For this purpose we wrote a *parser* it automatically captures the image's essential semantic content by text surrounding the page title, meta data, image title, ALT string and image caption tags.

We have presented a model for representing image/query semantics for text information. To calculate the semantic similarity between a query and an image, we start from determining the similarity between two basic components in an image to measure the *impact* of each keyword list (we will report more detail results on this in

Section 4). In our implementation, we store keywords of keyword list (i.e. image title, image Alt text, image caption, page title and HTML meta data) of the textual content separately. All the lists belonging to an image are connected to the image. Once the system has assigned the keywords to the images in the inverted file (Figure 7), we can calculate the text-based search to determine the match between the query and an image.

Let *I* be an image with associated a keyword *A* with a "*weight*" W. Similarly, the query contains a keyword *A*. The similarity between the query and image *I* is given by

$$\frac{W}{\sqrt{A.position}}$$

where ***A.position*** is the "*position*" associated with the matched keyword *A* in the list of keywords of the image *I* (see Figure 8).

	Img_1	Img_2	Img_3	....
Kw_1	W_{11}	W_{12}	W_{13}	....
Kw_2	W_{21}	W_{22}	W_{23}	....
Kw_3	W_{31}	W_{32}	W_{33}	....
	.	.		

Figure 7: Inverted File - Keywords against images with degree of relevance (Weights).

Therefore to compute the semantic similarity between multiple keywords in a query and an image, we propose the following formula:

$$\boldsymbol{Similarity}_{\text{image, query}} = \sum_{i=1}^{list.size()} \boldsymbol{t_i.weight\ X} \frac{\mathbf{1}}{\sqrt{\boldsymbol{t_i.position}}}$$

where $\boldsymbol{t_i}$ represents the $_i$th term in the keyword list of the image that has been matched with a term in the query. We also used $\boldsymbol{t_i.weight}$ and $\boldsymbol{t_i.position}$ to denote the "*weight*" and "*position*" associated with the matched term $\boldsymbol{t_i}$.

Keyword	Weight	Position
arafat	=> 2.8	1
yasser	=> 2.2	2
palestinian	=> 1.7	3
Leader	=> 1.7	3
peace	=> 1.7	3
process	=> 1.2	6
talk	=> 1.2	6
middle	=> 0.7	8
east	=> 0.7	8

Figure 8: The concept of "*position*" of a keyword for an image.

Various theories have been developed for an automatic assignment of term weights to the images and queries [3]. Since our proposed approach captures the image's essential semantic content by page title, meta data, image title, image ALT string and image caption, it is *well-structured* and *space-efficient* because it keep only few keywords for image representation. Therefore "*term position"* is defined as the *closeness* of the term from the view of relevance for that term in the list of an image.

Now consider the coefficient

$$\frac{1}{\sqrt{t.position}}$$

called *term discrimination value*, which measures the degree to which the use of a term position will help to distinguish some images to those to which the term is assigned from the remainder of the collection. For each potential index term as a *content identifier* to a collection of images, a *discrimination value* (DV) can then be computed as a function of the ranking of position. The greater the difference in term positions the more the images will become dissimilar, and therefore the better the particular term will function as a *discriminator* [24]. In Section 4, we will describe the *I-Search* image retrieval system that we have implemented based on the term position and provide experimental evaluations showing its effectiveness in image retrieval.

3.3 Integration of text/HTML Features with Visual Features

Once the visual and text/HTML features are computed for a particular web image, they can be combined in a number of different ways. But unfortunately, the textual descriptions are far from being structured and the quality of multimedia may be low. If the two approaches could be combined, it seems likely that the strength of one field could offset weaknesses of the other. Therefore we workout measurements that need to be made to validate the hypothesis and build the *idea* by extending text annotations work *side-by-side* with proposed classification knowledge to achieve this goal.

Since web queries tend to be very short and vague, the user's intent may be ambiguous. When the user issues a query like "Yasser Arafat" there is no way of knowing if the user is interested in images of only the face of Yasser Arafat or Yasser Arafat in white house for middle east peace talks without further feedback from the user. Hence some of the images returned may be very irrelevant. *I-Search* overcomes this problem by incorporating semantic attributes in terms of semantic labels (at the point user describes the picture he is interested in with a text query he can select the category) into its search mechanism. For example, query for a Yasser Arafat can now adjust the image ranking bringing images with dominant label full face to the front, instead of mixture of relevant and irrelevant images.

The textual descriptions of five lists connected to the image are not equally important. According to our experiments (will discuss in detail in Section 4.2) the importance order from high to low is like this: image Alt text > image title > meta data > page title > image caption. The reason we store keywords of each list separately, we want to differentiate the semantic importance of each list due to their positions and relationship with an image, and then later extend our text-based search to find higher-level semantic contents for visual features. In the experiments reported in this paper (in Section 4.4) we used the keywords in image Alt text as *actual* semantic contents for low-level features, because it derives that the keywords in the image Alt text are very highly semantically related higher-level semantic attributes of the images' visual contents.

To combine the visual features with keywords in image Alt text, first we convert keyword annotations in image Alt text for each image into a vector, use the technique of vector space similarity to determine the match between the query and an image. Let *I* be an image with associated a set *P* of keywords. Every word in *P* receives a "weight" 1/√|P|. Similarly, the query contains a set of Q of keywords, each one with a weight 1/√|Q|. The similarity between the query and image *I* is given by

$$\frac{|\mathbf{P} \cap \mathbf{Q}|}{\sqrt{|\mathbf{P}| * |\mathbf{Q}|}}$$

This measure has a simple geometric interpretation. Vectors in a (very high-dimensional) space represent the list of keywords with one axis for each word in the list. The vector corresponding to a list has component 1 along an axis if the corresponding keyword is part of the list, 0 otherwise.

To calculate their similarity for visual similarities, we propose the formula:

$$\boldsymbol{Sim}_{\text{visual}} = \delta * \frac{|\mathbf{P} \cap \mathbf{Q}|}{\sqrt{|\mathbf{P}| * |\mathbf{Q}|}}$$

where δ indicates the category of image. In the experiments reported in this paper we used value 1 if the corresponding image match with image semantic category user select, 0.25 otherwise. This clearly shows that $\boldsymbol{Sim}_{\text{visual}}$ between 0 and 1, the ideal case of this match is equal to 1.

Finally we propose a formula for integration to compute the similarity between a query and an image as follows:

$$\boldsymbol{Sim}_{combined} = \lambda * \boldsymbol{Sim}_{\text{text}} + \mu * \boldsymbol{Sim}_{\text{visual}}$$

Where $\boldsymbol{Sim}_{\text{text}}$ and $\boldsymbol{Sim}_{\text{visual}}$ are the numeric "*degree of relevance*" weights calculated to each image on a web page by the text/HTML analysis and visual analysis respectively. The other two parameters λ and μ are constants reflecting the contribution of the text/HTML and visual features. In the current implementation of our system, we have experimentally determined that setting λ to 0.2 and μ to 0.8 gives the best result, where λ + μ = 1. Obviously, this formula provides a comprehensive similarity metric that addresses both the semantic aspect and the visual feature aspect of images.

4 Experimental Results

To study the effectiveness of the proposed method, we implemented the proposed framework in a prototype search engine *I-Search* and conducted an extensive performance study. This section reports our study and findings.

4.1 The *I-Search* Retrieval System

When recall and precision are averaged over a number of user queries, experimental evidence can be obtained for the retrieval effectiveness of various processes by using image collection in several subject areas, together with various types of user queries. When the experimental results show similar trends in a number of unrelated environments, the experimental results are generally valid. Therefore, as shown in the following table (Table 3), we performed twelve text queries on our system for experiments for different types of user queries.

Q1	Tanks
Q2	Beach
Q3	Jerusalem
Q4	David Beckham
Q5	Taj Mahal
Q6	Deforestation of Amazon
Q7	World Trade Center
Q8	The Great Wall of China
Q9	Palestinian Leader Yasser Arafat
Q10	UN Secretary General Kofi Annan
Q11	Former South African Leader Nelson Mandela
Q12	Manchester United Football Manager Sir Alex Ferguson

Table 3: Queries for experiments.

For each query, we used each term to extract the list of images containing that term. The resultant from each term forms candidate images. We then manually examine the candidate images to eliminate those that are not semantically related to the query to get final set of relevant images. Then the precision-recall curve is computed by precisions obtained at the standard recall values for the query posed to the system.

4.2 Tuning the Weights

As mentioned earlier, there are 5 keyword lists of textual contents connected to an image and different lists may have different significance in identifying the image semantics. The reason we store keywords of each list separately, in the first experiment, we evaluate the performance of each list exclusively to study their impact on overall retrieval effectiveness. Figure 9 shows the results.

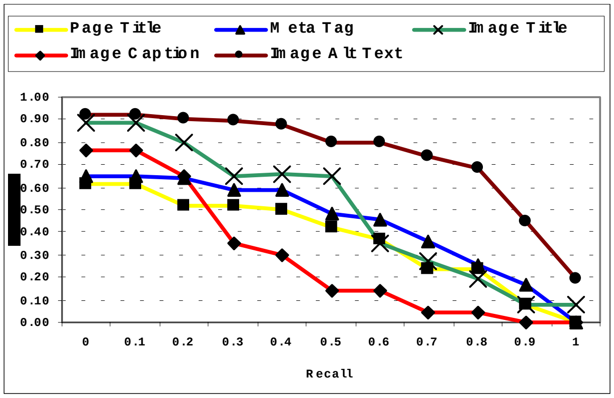


Figure 9: Each keyword list alone to represent image.

From Figure 9, we can see that image caption, it cannot maintain its precision over recall, although it has high at the beginning. This is mainly due to lack of information in image caption. Therefore it is not very effective. For both page title and meta data, since all the web pages have page title and meta tags, it can achieve high recall, but the precision is not satisfactory. Image title can be used to improve the precision a lot. On the other hand, page title and meta data can improve the recall substantially. Finally image Alt text, identical to the performance of overall system (we refer this scheme as IRS) can get 70% precision with recall of 80%. From this result, we have a clear picture of the relative importance of each keyword list.

To determine the weights to be assigned to the proposed model that combines all the keyword lists, we narrowed the search space based on the results from Figure 9 by adopting some simple heuristics. For example, since the image Alt text is the most important, we fixed its weight to 1.0. Moreover, the image Alt text is more important, the weights assigned to the other keyword lists cannot be more than its weight, relative to its weight, obtained the following weight assignments: page title(0.5), meta data(0.6), image title(0.7) , image caption(0.4).

As shown in Figure 9, the results show that except image Alt text, using a single keyword list exclusively cannot provide the best performance, even though such an approach can be viewed as a form of traditional text-based search without any *semantic understanding* involved. Though this experiment is meant to tune text-based search of our proposed method, we note that it is also a comparative study among the various systems, including both research prototype and commercial web image search engines. The existing image search engines do not possess a *good* semantic representation that can be used to improve both precision and recall of retrieved images; leading to therefore poor performance. The challenging task is to find the best way to construct a semantic representation that can be fully *utilized* the textual information to improve both precision and recall of retrieved images. Therefore our proposed model *exploits* the vast power of image semantics from the text associated with the image in a web page.

4.3 Tuning the System Performance

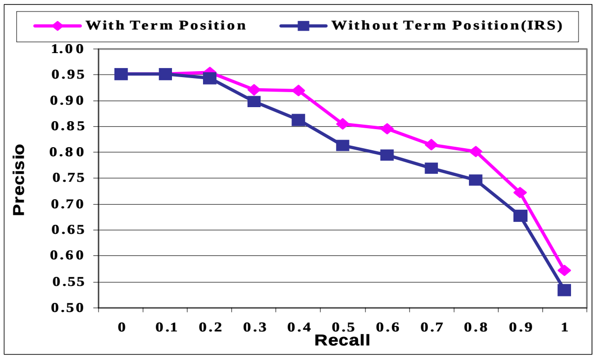


Figure 10: Performance comparison: with term Vs. without term position

There is another parameter that we have considered, the *term position* [24]. Recall that the term position is defined as the *closeness* of the term from the view of relevance for that term in the list of an image. In our proposed system a significant portion of the total weight of all the keyword associations with an image is contributed by a small number of keywords that are highly relevant to the semantic content of the image.

The term position explores the importance of match order for that term in the list of terms of an image. It determines as the position of the keyword in the list increases, and then derives that keyword more semantically related to the image. In other words, as the position of the keyword in the list increases, images with higher similarity measures will be returned to users ahead of images with lower similarity values. Therefore we can get more relevant images being displayed earlier, i.e., ranked higher. This increased proportion results in a much higher proportion of all possible relevant images being recalled. The effectiveness of this new *concept* [24] has shown in Figure 10, the resultant images with term position are more relevant than images without term position (IRS) and therefore term position has the impact in terms of image ranking during the presentation of the returned images.

4.4 Evidence of Integration

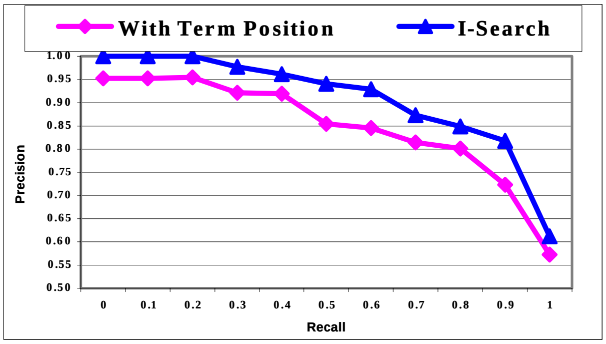


Figure 11: Performance Comparosion: I-Seach Vs. with term position.

Figure 11 shows the performance of our system *I-Search* in terms of precision and recall. In addition, to verifying the effectiveness our system through the performance measure, we have also compared it against with term position of the system. It is easily seen from the result that by combining higher level semantic with low-level visual features, the retrieval accuracy is improved substantially. In addition, more and more relevant images are being retrieved as the low-level visual features are combined. Therefore our proposed seamless joint approach implemented in prototype search engine *I-Search* build the *idea* by extending text annotations work *side-by-side* with attach semantic labels to achieve our final goal.

5 Conclusions and Future Work

A set of experimental retrievals was performed where search engines were compared in terms of average precision in answering for different types of user queries (Table 3). Two search engines were selected for this evaluation: Google^{TM} and Yahoo^{TM}. In reviewing the following results it should be noted that image retrieved by a search engine was regarded as relevant if the terms in the query clearly recognizable in the image. Although they index images they may also index other images of interest such as document images and inanimate object images. An evaluation accepting any image of interest as relevant in response to a query may reveal different results than what is reported here. Another important point is that the retrieval performance of search engines over time as they index more material. They may change their search algorithms as well. Therefore the numbers reported in this paper represent only a snapshot of their performance as of this writing.

Table 4 shows the search results of this experiment. We report the average precision of different search engines over top 20 images for 12 queries. In table 4 there are two rows for each search engine, the first row shows the average precision: the number of relevant images among the top 20 retrieves images (the number 20 was chosen based on the observation that users of web search engines typically do not browse beyond the top page of results contains 20 images). The second row in Table 4 shows the total number of images returned for the query. If the total recall number is less than 20, then the precision is computed over this total. The average precision of *I-Search* in Table 4 (0.92) is higher than the average precision of two of the other search engines.

Search Engine		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Avg.
Google	Prec.	.95	.90	.75	.95	.90	.60	.65	.90	.80	.85	.20	.50	.75
	Total	190,000	1,850,000	162,000	16,500	29,700	375	94,200	15,500	4,540	2,970	11	34	
Yahoo	Prec.	1.0	.60	.80	.95	1.0	.60	.90	1.0	.90	.90	.88	.50	.84
	Total	60,749	1,187,445	51,388	3,666	12,446	66	24,819	2,909	253	568	8	6	
I-Search	Prec.	1.0	1.0	.90	1.0	1.0	.85	.60	.92	1.0	1.0	.80	.94	.92
	Total	27	58	52	102	34	18	277	75	485	265	441	123	

Table 4: Performance comparison of different search engines over 12 queries. Legend: Avg.: Average precision over 12 queries, Pres.: Average precision for a single query over top 20 images.

We have presented the results of our experimental evaluation based on the combinatory approach for content-based image retrieval for the web. The web offers a rare opportunity for indexing multimedia: A significant amount of multimedia content is accompanied by textual content (Yahoo^{TM}, Google^{TM}, AltaVista^{TM}, Lycos^{TM} and Ditto^{TM}, apparently do not perform visual analysis and *rely* heavily on the image associated text). This enables search engines like *I-Search* to take advantage of both textual and visual clues in indexing this content. To address these challenges the strategies were implemented in our prototype search engine ***I-Search*,** and the results obtained suggest the significance of this work includes a greatly improved understanding of enhancing searching capabilities and efficient retrieval system for content-based indexing of images on the web.

References

[1] R. Jain, "NSF Workshop on Visual Information Management Systems: Workshop Report," In *Proceedings of SPIE: Storage and Retrieval for Image and Video Databases*, Pages 198-218, 1993.

[2] V. N. Gudivada, and G. S. Jung, "An Algorithm for Content-Based Retrieval in Multimedia Databases," In *Proceedings of Int'l Conf. Multimedia Computing and Systems*, Pages 193-200, 1996.

[3] H. T. Shen, B. C. Ooi, and K. Tan, "Giving Meanings to WWW Images," In *Proceedings of ACM Multimedia*, Los Angeles, CA, USA, Pages 39-47, 2000.

[4] J. P. Eakins, and M. E. Graham, "Content Based Image Retrieval: A report JISC Technology Applications Program," Newcastle, Institute for Image data Research, University of Northumbria.

[5] L. H. Tang, R. Hanka, H. H. S. Ip, and R. Lam, "Extraction of Semantic Features of Histological Images for Content- Based Retrieval of Images," In *Proceedings of SPIE Medical Imaging*, San Diego, USA, Vol. 3662, Pages 360-368, 1999.

[6] S. Santini, and R. Jain, "Beyond Query by Example," In *Proceedings of ACM Multimedia*, Bristol, UK, Pages 345-350, 1998.

[7] T. S. Huang, and Y. Rui, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues," *Journal of Visual Communication and Image Representation*, Vol. 10, No. 4, Pages 39-62, 1999.

[8] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Wquitz, "Efficient and effective querying by image content," *Journal of Intelligent Information Systems*, Vol. 3, Pages 231-262, Pages 1994.

[9] A. Hampapur, A. Gupta, B. Horowitz, C. F. Shu, C. Fuller, J. Bach, M. Gorkani, and R. Jain, "Virage: Video Engine," In *Proceedings of SPIE: Storage and Retrieval for Image and Video Databases V*, Pages 188-197, 1997.

[10] Li Jia, Z. Wang, and Wiederhold Gio, "IRM: Integrated Region Matching for Image Retrieval", In *Proceedings of ACM Multimedia*, Los Angeles, CA, Pages 147-156, 2000.

[11] W. Y. Ma, and B. S. Manjunath, "Netra: A toolbox for navigating large image databases," In *Proceedings of IEEE International Conference on Image Processing*, Vol. 1, Pages 568-571, October 1997.

[12] J. R. Smith and S. F. Chang, "VisualSEEK: A fully automated content-based image query system," In *Proceedings of ACM Multimedia*, Pages 87-98, November 1996.

[13] Y. Rui, T. S. Huang, S. Mehrota, M. Ortega, "A Relevance Feedback Architecture in Content-Based Multimedia Information Retrieval Systems". In *Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1997.

[14] J. Swain, F. Charles, and A. Vassilis, "WebSeer: An Image Search Engine for the World Wide Web", Technical Report TR-96-14, University of Chicago, Department of Computer Science, July 1996.

[15] J. R. Smith, S.F. Chang, "Visually Searching the Web for Content", In *Proceedings of IEEE Multimedia*, Vol. 4, No. 3, Pages 12-30, September 1997.

[16] O. Munkelt, O. Kaufmann, and E. Wolfgang, "Content-Based Image Retrieval in the World Wide Web: A Web Agent for Fetching Portraits", In *Proceedings of SPIE,* Vol. 3022, Pages 408-416, 1997.

[17] Y. A. Aalandogan, and C. T. Yu, "Evaluating Strategies and Systems for Content-Based Indexing of Person Images on the Web", In *Proceedings of ACM Multimedia*, Los Angeles, CA, Pages 313-321, 2000.

[18] Y. Lu, C. Hu, X. Zhu, H. J. Zhang, and Q. Yang, "A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems", In *Proceedings of ACM Multimedia*, Los Angeles, CA, Pages 31-38, 2000.

[19] J. Yang, Y. Zhuang, and Q. Li, "Search for Multi-Modality Data in Digital Libraries" In *Proceedings of 2nd IEEE Pacific-Rim Conference on Multimedia*, Beijing, China, Pages 482-489, 2001.

[20] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," In *Proceedings of IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*, Bombay, India, January 1998.

[21] A. Vailaya, A. Jain, and H. J. Zhang, "On Image Classification: City Images vs. Landscapes," Pattern Recognition, Vol. 31, No. 12, Pages 1921-1935, 1998.

[22] A. K. Jain, and R. C. Dubes, "Algorithms for Clustering Data," Englewood Cliffs, New Jersey, Prentice Hall, 1988.

[23] L. Jayaratne, A. Ginige, Z. Jiang, "On Image Classification: To Retrieve Images from major Web Sites," In Proceedings of the 6th International Information Technology Conference, Colombo, Sri Lanka, Pages 147-156, 2004.

[24] L. Jayaratne, A. Ginige, and Z. Jiang, "Effective indexing of web images with keyword positioning," In Proceedings of Sixth Asia Pacific Web Conference, Hangzhou, China, Pages 864-868, 2004.

E-Suvidha : Information And Facilitation Center.
A Success Story Of Nanded District: A Case Study.

S.D. Potekar
Scientist-C & Senior Systems Analyst,
Government Of India, Ministry of ICT,
National Informatics Centre, Nanded (MS) India,
Phone: +91-2462-235803, +91-9422874288
Email: potekarsun@yahoo.com

Abstract

Electronic public services have become a reality in Nanded District through the E-Suvidha initiative. The main focus has tended to be on delivering quicker and cost economic services through E-Governance to pull out a common man from the foe of middleman. An in-depth study of this application has been carried out to evaluate the technical issues, administrative issues and social issues and to identify the benefits to the common man. The paper subsequently brings out the critical success factors and concluded with the recommendations to progress towards perfection. The path breaking experiment have put impact in such a way that many district modules can be shared on these lines to achieve smart Government.

Keywords: E-Governance, Digital Government, Government to Citizen Relationship (G2C), District Administration, Information and Facilitation Center (IFC)

Disclaimer: The ideas and experiences put forth in this paper are my personal views and do not necessarily represent the views of the organization I work for.

1. Appcation under study:

In November 2001, Nanded District Administration has set up an Information and Facilitation Center, named “E-Suvidha”, for the citizens to provide better access to government information and services to achieve better Government to Citizen Relationship (G2C) through e-Governance.

2. Application Description:

E-Suvidha is an attempt of providing better G2C through E-Governance. The name E-Suvidha is given to Information and Facilitation Center (IFC), “Suvidha”, in Marathi language means facility and as various facilities are provided to citizens electronically, it is named as E-Suvidha.

This center is operational from November-2001. The Government of Maharshtra has formulated a policy to start such centers in the entire state later in August 2002 and then Nanded District Administration has bottled the project as per the policy.

E-Suvidha is managed by a society called ‘SETU Samiti’ registered under Maharashtra Societies Registration Act. The District Collector is President of the Samiti. The Resident Deputy Collector (RDC) assists the president as Secretary and the various departmental heads as members of the Samiti.

A state level body has been registered which guides and monitors the district level bodies. The State Level society has the Chief Secretary as chairman, Principal Secretary (Administrative Reforms), Principal Secretary (Finance), Principal Secretary (Planning), Principal Secretary (Revenue), Secretary (Information Technology) as permanent members and Director IT as a member secretary.

This center has adopted Quality Management System (QMS) and awarded with ISO9001: 2000 certificate with following scope in the year 2004:
“A friendly interface and service provider between the public and Government offices for issuance of various Government certificates/permissions/ licenses to citizen.” (Quality Manual of eSuvidha, 2003).

The services offered by E-Suvidha encompass a wide range of government departments (District Magistrate, Sub Divisional Magistrate, Taluka Magistrate, Land Records, Regional transport office, etc.) and can be accessed from one-stop center located in the premises of office of the District Magistrate, by any citizen, on payment of a nominal transaction fee.

The district Nanded, in which E-Suvidha is implemented, is a typical district of southern part of

Maharashtra State in India. The district spread over area of 10422 Sq. Kilo Meters. It is divided into three (3) Sub Divisions and Sixteen (16) Talukas having 1611 villages. The total population of the district is 28.7 Lacks.

3. Need of Study:

Utilizing the benefits of ICT for providing effective and transparent services to the citizens is one of the core focus areas of the IT policy of the Government and ICT application for masses is at the top priority in Nation Development Program. E-Suvidha is a pioneering approach in this direction.

E-Suvidha center is in service efficiently by the Nanded district administration since 2001 and is one of the foremost centers in Maharshtra State. Studying the critical success factors and evaluating the application in respect of ttechnical issues, administrative issues and social issues is essential.

Another reason why the case study of E-Suvidha is significant because this is the only center in maharashtra State getting ISO 9001: 2000 certification. This case study will definitely prove helpful for all District Administrators and many District models can be shared on these lines.

4. Objectives:

- In-depth study of working of E-Suvidha.
- Identifying and evaluating the benefits of E-Suvidha.
- Identifying gaps in working patterns and providing suggestions for improvement.

5. Methodology:

The study was carried out in two steps:

1. Study to understand the service delivery mechanism created by the E-Suvidha. The details were obtained by reviewing the literature like "Mahiti Pustika", website, http://setu.maharashtra.gov.in and from the dialogues with the officers controlling the functioning of E-Suvidha. (Author is one of the technical support providers to this center and also is manager in the Management Committee of E-Suvidha).
2. Study to understand the common citizen's perceptions on the system. The information was obtained from the citizens, who had visited E-Suvidha and benefited in terms of savings of time and cost of service delivery.

6. Detailed Study Report:

6.1 Background of eSuvidha:

District Administration works as main bridge between common man and Government. It is the nodal agency to implement various government policies and schemes at District level. The aims, mission, objectives, policy and design parameters are discussed in the web site of SETU. ("SETU- A bridge for facilitation between Citizen & Govt". Aims to achieve, Mission, Objectives, Strategy. http://setu.maharashtra.gov.in/.)

A common man comes to the District Administration for variety of certificates, permits and other important documents. A common man does not necessarily know the exact procedure for filing the application and need to run around different places for completing the formalities. After filing the application he does not get precise information about the status of his file and its time limit. So he keeps visiting the office and start meeting the staff resulting in loss of time and travel expenses. After several such visits he starts doubting about the government procedures. At one fine time he meets a gentleman (middle-man or agent) who promises him to complete his work in less time (many times give him home service) and demand for money. The case is even worst in rural area, the poor villagers believe in what ever they were told by the agents roaming outside the government offices. The agents mostly operate without any legal authority and misguide the public for their own benefits. The common public thus blames to the Government machinery and the end result is the loss of faith in the Government functioning.

6.2 Goal and Aims:

With a planned goal of improvement in G2C relationship the eSuvidha is established with the following aims. The idea is visualized in May 2001 and came in to reality in November 2001.

To pull out a common man from the foe of middleman and to provide the quicker services, an idea of establishing a one-spot IFC is conceived.

- Improve communication between citizens and Government.
- Provide transparency in the Government functioning.
- Elimination of corrupt middlemen.
- Prove quicker and cost effective services.
- Provide one-stop cost effective services.
- Enhance citizen empowerment.

6.3 Working Pattern:

The eSuvidha is established for providing one-stop services centre for the general citizen. For filing an application to any of the eight selected offices, for the selected 64 types of services, the common man is intended to come to this single stop where he will complete all necessary formalities for filing an application for the selected services and get the required certificates or permits in specific time period.

The applicant gets all information related to services, gets the application format with all necessary attachments and one can submits the application in the eSuvidha centre only. After submitting it gets verified for necessary attachment as per the rules and regulations and then accepted. After acceptance one gets a computerised token number and receipt of the application. The token number and the time period with the date are mentioned on the receipt. The applicant will get the certificate/licence/permission from the eSuvidha centre on the mentioned date.

The fulfilled applications then send to the concerned department for processing and on completion it comes back to E-Suvidha centre for distribution. In case of any

A help desk is provided in the centre for assisting in filling up the application format and photocopying is also possible in the centre if required by the applicant. Besides accepting the applications, one counter for printing computerised affidavits is available and one gazetted Government Officer is also posted in the centre for verifying and signing the affidavits. Thirty-five types of commonly used affidavits are identified and computer printout is made available.

6.4 Role of Information and Communication Technology (ICT):

As explained earlier, E-Suvidha is an application of ICT for masses. The server located in eSuvidha is Pentium-IV with six clients (Three P-III and three P-II). The counters are equipped with three Dot Matrix Printers (DMP) and two laser printers. Three hours backup is assured with 2 KVA UPS. Three web cameras are used for affidavit printing sections. The server system runs on Windows NT with IIS server; client PCs run Windows 98. MS SQL Server, Visual Basic, Java Development Kit and MS Access were used to develop the applications, plus an Indian language font set. One of the applications is developed under Linux 7.2 environment, DB2 as backend and Visual Basic as front end with Indian language font set.

6.5 Stakeholders:

Government officials in Nanded District (Eight selected offices) are major stakeholders, as are all the citizens in the District, and the operators. Other stakeholders include senior officials in the Maharashtra Government; IT professionals from National Informatics Centre (District level officers and State level Officers), the IT vendors, designers and implementers.

7. Financial Issues and Technical Issues:

7.1 Financial Issues:

There was no financial support requested by the District Magistrate from Government to establish this center. All expenditure was done by funds raised locally at the District level. MP/MLA local area development fund was made available as one of the sources of expenditure. Public Works Department (PWD) have helped to carry out necessary civil and electrical works.

National Informatics Centre (NIC) has provided the software at free of cost. NIC, District Centre has provided the training and necessary technical assistance for implementation of the software successfully.

7.2 Technical Issues:

Identification of the appropriate hardware platforms and software application packages for effective delivery of public services is an important issue in G2C applications. This is handled by NIC in this project. The technical support from NIC, District Centre is the strongest part in the successful implementation of this project. NIC has given the recommendations for the hardware configuration, verified the configuration on supply, designed and set up the Local Area Network (LAN). Installation of the system software and application software was carried out by NIC with day-to-day technical support.

8. Utilization at a Glance:

The success of any G2C project can be evaluated on the basis of its usage. Technically strong system may fails if the response of the common public is poor. The figures in the Table-1 give the details of utilization of the facilities provided by eSuvidha. In addition to these 37766 affidavits were printed through the centre.

Table-1: Utilization of Services

Office Name	No of applications		
	Processed	Completed	Pending
District Magistrate	12657	12595	62
Sub-Divisional-Magistrate	13102	11624	1478
Land Records	9181	9052	129
Tahsildar / Taluka Magistrate	35041	29509	5532
Regional Transport Officer	25841	25700	141
Total	95822	88480	7342

The utilization data collected till date is summarised as follows:

- The total number of citizens availed the service since November 2001: 1,91,704.
- Approximately 192 applications are getting processed daily.
- Number of citizens visiting the centre daily is more than 192.
- 92.28 % citizens have completed the work in scheduled time.
- 7.6 % citizens could not finish the work in scheduled time.

The figures above are obtained from the latest reports of the E-Suvidha. (September 20, 2004)

9. Project Evaluation:

The evaluation of this project is done with three dimensions: (1) Technical Evaluation, (2) Administrative Evaluation and (3) Social Evaluation. Evaluation factors and the remarks are as given in Table-2, Table-3 and Table-4 respectively.

Table-2: Technical Evaluation

Evaluation Factor	Remarks
Availability of Computers and peripherals	Sufficient number of computers and peripherals available.
Application Software development.	NIC has developed application software fulfilling the need of administrative requirement and giving local language support for better interaction.
Technical Support: Hardware and Software Maintenance	NIC gives the required technical support. Additionally a Program Manager is a technical person in the centre.
Design of Network	Network designed connecting eSuvidha with three other offices in the premises.
Power Management	UPS provided

10. Impact:

10.1 The impact observed from the Government office point of view:

- The project has improved the Government working culture in the office and brought discipline.

Table-3: Administrative Evaluation

Evaluation Factor	Remarks
Planning	Pre implementation and post implementation planning is perfect by District Magistrate.
Follow up and monitoring	RDC and DM regularly monitors and insures the smooth functioning
Involvement of Officer level staff.	Management committee meets regularly and Officers regularly visits the centre for getting feedback.
Involvement of staff.	Improved interest of staff.

Table-4: Social Evaluation

Evaluation Factor	Remarks
Usage.	192 applicants file the applications daily.
Satisfaction of common man	92.28% persons got the required certificates or permits as per the defined schedule.
Acceptance of new trend	Improvement in the faith in Government functioning is observed.

- Transformation of the working procedure to the new trends of e-Governance is observed.
- Increase in the efficiency of the Government machinery involved.
- The crowd has diverted to the eSuvidha results in increase of concentration of the staff working in Government.
- Increase in trust of common man towards Government functioning.

10.2 The impact observed from the citizen point of view:

- The Centre has facilitated the submission of applications in one location.
- The prompt service is assured with the time limits.
- Cost effectiveness.
- Time saving.
- Citizens getting attached with the center and confidence towards Administration increased.
- The center has created more job opportunities. More than 10 youth got employment.
- Reduction in the time delay or pending cases.

11. Assessment: Failure or Success?

E-Suvidha is an ISO 9001:2000 certified center in the year 2004 and continued the certification for another year, this is the only center in Maharashtra state getting such certification. This certification should be seen as justifiable recognition for design innovation, relevance and potential. It has demonstrated a model by which transparency of Government information and services could be increased, and corruption, favoritism and other costs reduced. However, it faces a number of practical difficulties that have constrained the benefits delivered to date.

In the past, citizens often had to pay middle-man in order to have services performed; some of them faced unfairness of being ignored or their applications rejected; some of them had to rush different places for routine works like photo copying and attesting and many of the citizens faced the prospect of losing a day's wages plus paying transport costs each time they were forced to visit offices for completing the necessary formalities for filing an application to the Government office. The E-Suvidha system has replaced this with a lower-cost, faster, and more transparent service for a limited range of services. Corruption and payment to the middleman have been replaced by transparent payments to SETU Samiti. Transport costs and wage losses have been reduced significantly; this is direct benefit for the poor.

A majority of E-Suvidha users seem positive about it, but many of the users are far from the impact as an e-Governance project. There is an understanding among many users that E-Suvidha has reduced corruption, and improved transparency. Yet, at the same time, few users continue to see a need for middleman to be paid. Majority of the users experienced the responsibility of E-Suvidha in delivering quicker services, but few of them has reported that the stipulated time has crossed to get the certificates or permits.

There are similar centers functioning at all Talukas in the District but all of them are isolated and working independently. Common man from Taluka place has to go to these centers for availing the services of Taluka office and for District level services he has to come to District head quarter. There is no association of these two levels.

Overall, the E-Suvidha can be considered, as moderately successful attempt and having scope for improvement to achieve the perfection in order to meet its goals and aims.

12. Critical Success Factors:

- **Infrastructure set-up:** The District Magistrate has set-up the centre by raising fund locally without any special allocation of funds.
- **Less willingness from the internal staff:** There was resistance to accept new change from internal staff; the challenge was taken by the District Magistrate to change the mindset with the new trends of e-Governance. This required the special efforts and administrative skills to improve the interest.
- **Technical support:** NIC, District Centre has provided strong technical support also imparted necessary training.
- **Monitoring:** SETU Samiti, which involves the senior officers in the district, governs the activities in E-Suvidha centre. The samiti meets at regular interval to understand and resolve the problems in the implementation.
- **Promotion of benefits to the common man:** Promoting the benefits to the citizens to develop faith of common in Government machinery is one the success factors of E-Suvidha.

13. Recommendations for improvement:

- **Reducing Pendency**: The E-Suvidha acts as a communication vehicle and it does not automate the core processes of creating documents. The working pattern of E-Suvidha is such that all the applications are actually processed by the Government Machinery. Thus achievements of the E-Suvidha are linked to this back end support. Accelerating and automating the working pattern of the back end can reduce the pendency. The regular follow-up and keen monitoring the application associated with E-Suvidha by the officer level staff is another factor for reducing the pendency.

- **Association with Taluka level IFC**: The IFC centers at Taluka level that are working independently right now can be linked with E-Suvidha so that the citizens at Taluka can file the applications at their local IFC center and need not come to district head quarter. This will save the time and reduce the cost.
- **Extension counters:** The access to the citizens can be made more comfortable if the couple of extension counter opened in the city. These counters can be set up as private enterprises, paying annual license fee to the SETU Samiti. Web based applications are the best substitute that can used for interlinking.
- **Expanding the scope:** The present scope of the E-Suvidha is limited to eight offices and 64 types of services. Some of the core service processes remain untouched, the services like paying property tax and paying electricity bill can be included which are of the most concern of the citizens.

* *

Reference

1. Definition of ICT given by another author. Retrieved September 30, 2004, from http://www.bambooweb.com/articles/i/n/Information_Technology.html.
2. Earl Mardle has given a definition of e-Governance. Retrieved September 30, 2004, from http://www.bytesforall.org/E-Governance/html/egov_lastchance.htm.
3. Dr. Thomas F. Gordon, e-Governance and its Value for Public Administration. Models and Software for e-Governance. Retrieved August 30, 2004, from http://www.tfgordon.de/publications/Gordon2004a.pdf.
4. e-Governance Conference Presentation, Bangalore November 1999. e-Governance in the Nation State: A Great Opportunity, or the Last Chance. Retrieved September 30, 2004, from http://www.kn.com.au/commentary/e-Governance_presentation.htm.
5. M J Xavier & M P Gupta. Government to citizen (G2C) SERVICE DELIVERY MODELS- An Indian Experience. Retrieved July 8, 2004, from http://www.the-south-asian.com/August2003/government_to_citizen_g2c_1.htm.
6. Rural Informatics in India - An Approach Paper. Retrieved July 8, 2004 from http://ruralinformatics.nic.in/files/4_9_0_273.doc.
7. Definition of e-Governance. Retrieved July 13, 2004 from http://link.wits.ac.za/research/research.html.
8. Definition of e-Governance. Retrieved July 13, 2004, from http://www.unescap.org/rural/ICTEGMNov2003/Malaysia-JamesGeorge.doc.
9. Quality Policy of SETU. Quality Manual of eSuvidha center.
10. SETU- A bridge for facilitation between Citizen & Govt. Aims to achieve, Mission, Objectives, Strategy. Retrieved June 8, 2003, from http://setu.maharashtra.gov.in/.
11. Application of I.T. in Government – MBT's Initiatives. Retrieved June 10, 2004, from http://egov.mit.gov.in/mbt.asp.
12. C-DAC, major initiatives in e-Governance. Retrieved July 10, 2004, from http://www.cdacindia.com/html/egovidx.asp.
13. eSeva - Citizen Service with a Difference. Retrieved September 10, 2004, from http://www.ap-it.com/eseva.html.
14. Alok Kumar Sanjay, Vivek Gupta, eGovernment for Development. eTransparency Case Study No.11. Gyandoot: Trying to Improve Government Services for Rural Citizens in India. Retrieved Sept. 10, 2004 from http://www.egov4dev.org/gyandoot.htm.
15. Subhash Bhatnagar and Nitesh Vyas, Gyandoot: Community-Owned Rural Internet Kiosks. Retrieved Sept. 1, 2004, from http://www1.worldbank.org/publicsector/egov/gyandootcs.htm.
16. CENSUS OF INDIA. Retrieved September 10, 2004, from http://www.censusindia.net.
17. Dr. Sada Shankar Saxena, (July 1, 2001). "Designing India's Development in the 21st Century: Some District Level Challenges" Retrieved July 11, 2004, from http://thinkcycle.media.mit.edu/thinkcycle/main/development_by_design_2001.

* * *

Qurixx – A Portable Operating System which could be shifted across any Heterogeneous Computer

W. A.A.N. De Silva and A. Perera
Department of Computing,
Informatics Institute of Technology,
Sri Lanka

Abstract

Modern computer operating systems are redefining the way humans interact with machines, literally adding a whole new facet to cyberspace. Operating systems developed in this arena are many. Nevertheless, they are all fabricated around the same concept and thus, materialize an unpleasant limitation. These are based only on the computers that they are installed on, and do not enable the user to use a single portable operating system across many computers. In consequence, the inconveniences faced are many. Unfamiliarity of the settings of a foreign computer and unavailability of users' favorite software on other computers are among few of them. This area has been a key field of research for more than 3 decades. However, these initial forays resulted in mostly failures. This paper discusses a concept of a portable operating system which can be executed on any heterogeneous computer instantaneously, devoid of any installation process. The first step of the solution involves creation of 'Self-bootable' operating system using 'Live-CD' technology. This operating system is run on a volatile virtual disk created on top of physical memory. This way it ensures that no data on the computer is harmed by executing this operating system. Notion of 'hardware auto-detection' is also bundled with this operating system to identify hardware devices uniquely on the running computer. The core OS content is placed in the above mentioned read only Live CD. In contrast to this, an external portable storage device which supports both read and write is employed to maintain User's software, customizations, etc… Analysis of the developed prototype has demonstrated its practical applicability thereby laying the first step in revolutionizing the entire concept of computer operating systems.

1. Introduction

Any adult or a teenager who has spent his or her quota of time on a computer or a mobile phone would have already experienced the awe-inspiring glimpses of Information and Communication Technology (ICT). From the hand held palm-pilot or the mobile phone, to the super computers in NASA, humans would forever be in the reliance of ICT. Playing a lead role in ICT, computers, in fact 1 billion of them across the globe are used numerously to assist human beings in areas ever imaginable. Simply, ICT has matured to a juncture where humanity is at the mercy of computers.

Bringing out the functionality of these computers are their operating systems which resides in their hard drives. A computer is nothing but a piece of metal without its operating system. Many vendors, over the years have produced many different operating systems. Microsoft Windows™, Linux™ and Mac OS™ can be taken as a few pioneer inventions. Nevertheless, they are all fabricated around the same concept and thus, materialize an unpleasant limitation. Present operating systems are based only on the computers that they are installed on, and do not enable the user to use a single portable operating system across many computers. Thus, it has lead to many setbacks. On the other hand, standard Live-CD [1] operating system such as 'knoppix'[5] and 'mepis' introduces a different concept. However, they contain a read only file structure and does not enable installing software or to carry out any OS customizations.

2. Problems in traditional Operating systems

Problem domain can be divided in to 3 major areas.

2.1 Inconsistency of operating systems

People use many different computers in their day to day lives at homes, offices, schools, or even at cybercafés. However, all these computers consist of their own operating system and thus, keeping them synchronized in terms of software, appearance or even in terms of data requires a great effort. For example, one's personal requirement to install a new software package turns out to be an agony of installing the same software many times on all the computers that he or she is using. This requires a

considerable amount of effort, knowledge and time as well. For example, installing and configuring ‘Visual Studio .Net™
’ or a similar product requires more than 3 hours on an average computer. On the other hand, this has led people to carry the documents and other files that they are working on, with them. The famous USB™ (Universal Serial Bus) storage device solves this crisis up to a certain extent. Nevertheless, this requires synchronization of all files in the USB storage device with the computer, before and after the usage. This again is a very tedious task people have to carry out recurrently.

2.2 Duplication of installed software

Typically, computer users use more or less the same bulk of software on all most all the computers that they are working. They generally prefer the same operating system, office package, internet browser and the media player on all the machines. Typically, this consumes nearly 3 GB of disk space on each computer. This is an unnecessary duplication of software across computers. Without even knowing, people waste disk space like this everyday.

2.3 Global unavailability of user’s operating system

Scores of people use computers which do not belong to them in their day to day lives. It can be at a cybercafé, friend’s computer or even at a public computer lab. Not only do such people find it uncomfortable without their favorite set of software, but also without their own operating system customizations such as the wallpaper, desktop theme, font, etc... The reason behind this is the static nature of the operating systems in this era. In other words current operating systems are based on the computers that they are installed on, and do not enable the user to use a single portable operating system across all his computers.

3. The Solution

The solution employees a unique combination of ‘portable storage device’, ‘RAM disk’[3] and ‘hardware auto-detection’[4] notions to form the required outcome. If the operating system needs to be moved from one location to another, then it indisputably needs to be stored on a medium which is portable. Thus, hard disks are out as a solution. Nevertheless, this set back can be overcome with the introduction of an optical storage device (CD/DVD).

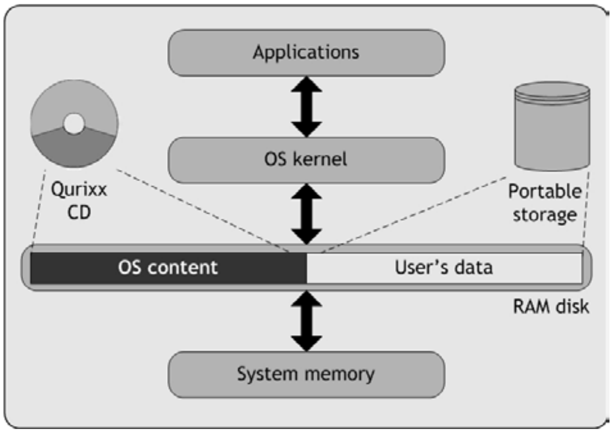


Figure 3.1 – Qurixx overview

A CD/DVD would be an ideal choice not only for the reason that they are commonly used, but also because they provide ample space to store the operating system. However, CD/DVD technology does not support writing back using standard CD/DVD ROMs. An operating system obviously necessitates both read and write access to its main file system. Thus, CD/DVD along is not sufficient as a storage solution. Nevertheless, it can be used to store static (unchanging) component of the OS. This obstacle can be overcome with an introduction of a RAM disk. However, the limitation of the RAM disk is its volatile nature of storage space. As a result, user’s data cannot be permanently recorded in a RAM disk. Therefore, an introduction of portable storage device is also necessary to store the data permanently in the system. With a use of RAM disk along with a CD/DVD and a portable storage device, the storage problems can be overcome. The RAM disk can act as the main storage space and the portable storage device and CDROM can be linked to this space. Figure 3.1 explains the proposed storage structure.

However, there exists another problem of working with unknown hardware devices of the host computer. Qurixx has no prior knowledge on the hardware devices as it is run on different computer at different times. Therefore, a mechanism to auto-detect and adopt hardware devices must be made available to achieve complete portability. As soon as the previously mentioned storage structure is in place, this mechanism should auto-detect the hardware configuration before the initiation of the operating system.

4. Design of Qurixx

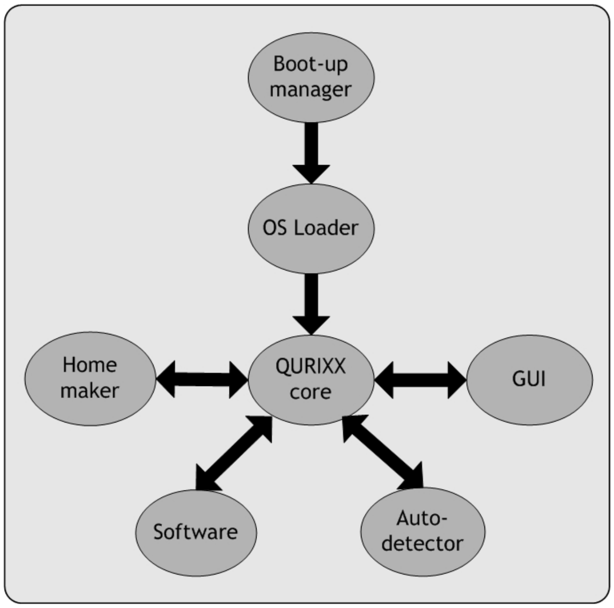


Figure 4.1 – High level design

Qurixx uses a novel design aimed at overcoming the weakness of current OSs. The entire system consists of 7 independent but correlated components considering their nature of operations. This component based approach permitted the developer to break down the big challenge into small manageable components. This also smoothened the progress of incremental development. However, a special attention to interfaces of the components was necessary as they are the only gateways to the components.

4.1 Boot-up manager

The primary goal of the boot-up manager is to ensure a hands free boot-up of the portable operating system. At the time of boot-up, this is the first program to be executed.

The first program to execute at the time of boot-up is specified in the 'Master Boot Record' (MBR) of the boot media. Therefore, the Qurixx CDROM is made in accordance to the bootable 'El-torito'[2] standard with such a MBR. This MBR contains the address of the 'boot catalogue'. Having a boot catalogue is typical Linux standard and it increases the flexibility of the system. This boot catalogue contains the addresses of operating system(s) and the address of the preferred 'boot loader' in a universal format. Therefore a new operating system or even new boot loaders can be added to the system by simply adding a new entry in the boot catalogue.

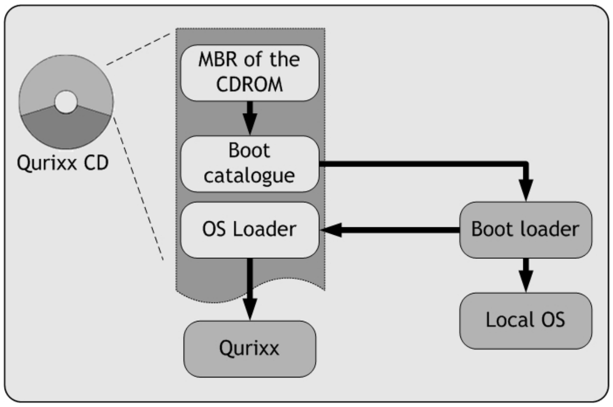


Figure 4.2 – Boot-up manager architecture

Subsequently, the preferred 'boot loader' is called with the information in the boot catalogue. The boot loader plays the key role of the boot-up manager. Boot loader initiated the boot process of the Qurixx by transferring the execution to the OS loader. Meanwhile, this also performs boot time vital transactions such as clearing RAM disk, capturing user boot parameters, etc…

4.2 OS Loader

Prime function of the OS loader is to craft the environment required for the full operating system to run on. This is a series of actions and cannot be done using a simple boot up program. Thus, a Micro Linux Kernel[6] which contains a chain of needed software is employed for this purpose.

A Micro Linux Kernel is a downsized version of a full Linux kernel which is typically used to craft the setting necessary for an OS to run. These contain a sub set of the features (only the needed few to perform its task) of full kernel and usually fall below 50 KB in size. The advantage of this type of a Micro Linux Kernel is that it does not require a secondary storage medium to run on. Therefore, a Micro Linux Kernel technology is ideal for the OS loader as it is run at a stage where a secondary storage device or an alternative is not available.

At the outset, this OS Loader is used to create a virtual disk space needed for the OS to run on. Thereafter, it initiates and links the portable storage devices and the data content of the CDROM with the virtual storage device. It creates necessary symbolic links to attach the portable storage device to the file system directories where user keeps his dynamic data. This way user's data are permanently recorded in the portable storage device.

4.3 Auto-detector

The main responsibility of this component is to explore and identify the system configuration of the host computer. However, this process is done using a 3rd party hardware auto-detector. Redhat Corporation's Kudzu[7] is used for this purpose. Firstly, the attached hardware devices are discovered with the help of Kudzu. Once the hardware devices are completely identified they are initiated in the system for them to be used by the operating system.

4.4 Qurixx Core

This is the operating system core which provides the base level OS functionalities. Linux kernel 2.4 is used for this purpose.

4.5 GUI

GUI is where users touch and feel the operating system. Thus, a better GUI obviously dispenses a better operating system. This component provides a graphical user interface to users. User friendliness and the consistency of the interface are considered as design goals of high importance. KDE Desktop manager[8] and a mixture of other 3rd party components were used for the development of this GUI. However, lots of customizations were done to fit it in to the needs of the portable operating system.

4.6 Home maker

The concept of portable operating system is based around the thought of portable storage device. Thus, the responsibility of this component is to create a portable home space for users. In another words this component associates portable storage devices with the system to be used to save user's own data. This process takes place only when the operating system is up and running (usually once) and not at the time of boot up.

4.7 Software

This layer consists of user's 3rd party software. Even though this doesn't influence any implementation decisions of this prototype, this layer was portrayed in the architecture to add more readability.

5. Implementation

An open-source approach was chosen for the development of the Qurixx prototype foreseen a heavy usage of open-source applications with in the Qurixx prototype. Qurixx prototype was implemented using a combination of ANSI C, PERL and Shell programming. Shell programming was selected because of its compatibility with Linux and open-source applications. Similarly, ANSI C and PERL were selected for speed and string manipulation respectively.

Linux kernel was used to provide the base OS level functionalities used in the portable operating system. Thus, Qurixx took the shape of typical Linux conventions. Both USB and FireWire technologies were enabled in the Linux kernel leaving the question of type of the portable storage device on the hands of the user. Boot-up manager was implemented using the 3rd part 'ISO Linux' boot loader. On the other hand, 'OS Loader' component was implemented as a Shell script. Similarly, Home maker component too was implemented as a shell script. In contrast to these components, GUI and Auto-detector components were implemented with the help of 3rd party components. They employed KDE desktop manager and Redhat Corporation's kudzu respectively.

6. Underlying Issues

Firstly, finding locations of operating systems reside in the host computer was a big challenge. This information was necessary to provide the user a choice between Qurixx and local OS at the boot up. A constant memory location could not be used as it varies completely from a computer to another. However, most of the modern operating systems maintain a boot entry in the MBR of the primary storage device. Author utilized this fact as a get away loophole to solve the problem. A new record which directs to this location was added in Qurixx's boot catalogue to provide an entry for the local OS.

On the other hand, one of the key requirements of the auto-detector is to be always up to date with the knowledge on hardware devices. Thus, another challenge faced was accommodating these hardware live updates. Nevertheless, author utilized the Kudzu's (3rd part tool which was employed for hardware detection) ability of accommodating hardware live updates to solve this problem.

Another challenged faced was accessing the CDROM for other CDs (other than the OS CD) while running the operating system. However, this problem was over come by introducing the 'Qurixx mass storage supporter' component to store the entire directory structure in a specious portable storage device. As a consequence, there is no need to access the CDROM at all as the needed data can be located in user's portable storage device.

Even though Linux kernel was employed to power up Qurixx, it needed configuring and adding new modules to fit the portable operating system. Thus, author went through another challenge of recompiling the Linux kernel. However, this problem became harder than it seemed as no proper literature was available on WWW. Therefore, expert's advice was seized to get over this challenge.

7. Related Products

The Live CD based operating systems such as Knoppix and Mepis, looks some what similar to Qurixx at a glance. However, fundamental concept it self goes far away from just a Live CD. The primary difference of the Live-CD based operating systems over Qurixx is that it restricts modification of the file structure. The entire operating system is available on a read only media (CD or DVD) and thus, cannot be written back using a normal CD-ROM. As a consequence, the ability to customize, install software or even saving new files is immobilized in Live-CD based operating systems [1]. As appose to this, Qurixx provides a fully customary operating system allowing users to install software or even change the file structure completely. Even though the underline technology seems bit similar to a Live-CD based operating system, the final outcome goes far beyond just being a demonstration OS to a real customary portable operating system.

8. Conclusion

This paper delineated a whole new concept of a portable operating system to overcome a general problem. It is to be noted that this is the first successful attempt ever of its kind. The proposed tool uses a distinctive combination of 'Live-CD', 'Hardware auto-detection', 'RAM disk' and 'portable storage device' notions to achieve true portability.

The results show that ICT can make solid turning points even in so called 'cutting edge' of all domains. The future holds more experiments for which this paper has made the first step.

9. References

[1] Livecd. (2004). LiveCD - Wikipedia. [Online]. Wikipedia. <http://en.wikipedia.org/wiki/livecd/> [2004, Nov, 18].

[2] El_Torito. (2005). El-Torotp - Wikipedia. [Online]. Wikipedia. <http:// en.wikipedia.org/wiki/ El_Torito_%28CD-ROM_standard%29> [2005, Jan, 19].

[3] RAM disk. (2005). RAM disk - Wikipedia. [Online]. Wikipedia. <http://en.wikipedia.org/wiki/ RAM_disk> [2005, Jan, 17].

[4] Autodetection. (2005). Autodetection - Wikipedia. [Online]. Wikipedia. <http://en.wikipedia.org/wiki/ Autodetection> [2005, Jan, 19].

[5] Knoppix – Live on CD. (2004). [Online]. Knoppix.org.<http://www.knoppix.org> [2004, Nov, 18].

[6] C2. (2005). Micro Kernel. [Online]. C2.com. < http://c2.com/cgi/wiki? MicroKernel> [2005, Jan, 19].

[7] Redhat. (2004). Fedora project - Kudzu. [Online] Red Hat, Inc.<http://fedora.redhat.com/projects/ additional-projects/kudzu/index.html> [2004, Oct 31].

[8] KDE. (2005). [Online]. kde.org. <http://www.kde.org> [2005 Mar 8].

Thermal Comfort for Passive dwellings via optimum Roof Architecture (RoofOpt)

M.S.R Perera and B. Modasia
Department of Computing
Informatics Institute of Technology, Sri Lanka

ABSTRACT

The occurrence of hot discomfort during the daytime is a serious problem for the citizens living in tropical regions. This drove the citizens to look intently on thermal comfort conditions. In tropical regions, the most prominent component that affects thermal comfort is the roof architecture as roofs are exposed to direct solar radiation. Conversely, in the modern world the houses are influenced by modern architecture where, the designer only concentrates on the aesthetic side of the dwelling. Therefore, to avoid thermal discomfort, designers use their experience, knowledge to determine a better dwelling structure through different passive methodologies. But professionals very rarely realize that the current passive techniques will result in a satisfactory solution to the dweller.

This paper presents a framework that will provide an intelligent artifact which will determine the optimum roof architecture according to thermal comfort conditions in a dwelling. The proposed framework consists of three layered architecture and consists of five main components. Each major component is further divided into sub modules. Data Extractor (DE),User preference component (UPC),Case based reasoner(CBR) ,Fuzzy decider (FD) component and Two dimensional designer components (2DDesign).This software tool promotes the concept of "Thermal Comfort", a novel, easy to use, intelligent which can be used to obtain the optimum roof architecture, insulation material and thickness in tropical climatic conditions.

1. INTRODUCTION

Warm humid climatic conditions prevail in many parts of the world. This climate is experienced in populous regions in South and Central America, South Asia, South- East Asia and Africa. In such regions experiencing warm humid climatic conditions, hot discomfort is one of the major problems in houses and buildings [1].

Generally, the countries with tropical climatic conditions have a high density of population. Most of these countries still remain as developing countries. With the economic development, the energy consumption for thermal comfort is also rising.

However, in recent times, a new trend has emerged in the housing sector due to the influence of modern architecture. This adverse situation is further deteriorated by adopting various features for houses and buildings that would be more suitable for temperate climates. Many designers in this region have ignored the climate in their designs, primarily because they are pre-occupied with fashionable building forms. They have tried to separate the building from nature rather than integrate it. Once separated, indoor thermal comfort should be achieved using air-conditioners, fans etc. These will require considerable amount of energy for operation resulting in high cost to the dweller and the country as a whole. This is not a desirable situation for most of these countries, which are still in the developing stage. Therefore, it is an essential fact that when designing dwellings, the designer should give due consideration to comply with nature to achieve thermal comfort by maximizing the natural resources.

Currently, architects/designers use their experience to design the appropriate roof architecture and insulation thickness/material for a particular building. Mostly, a great effort is spent on

- Deciding the optimum combination of the roof parameters
- Deciding the appropriate insulation thicknesses & materials for a particular dwelling to achieve thermal comfort.

Because of the unstructured, heuristic nature of this method, that is

- There is no proper calculation formulae available to determine the appropriate insulation thicknesses based on climatic factors.
- There is no proper algorithm to determine the thermal comfort for a combination of roof parameters

Optimum roof architecture, appropriate thickness, thermal performances, energy consumption and budget cannot be scientifically determined.

Thus, comes a need for a better software architecture, to “design thermal comfort dwellings”. There have been numerous attempts found to designing low energy dwellings and thermal comfort dwellings. RESFEN 2.1 [2], Building Design Advisor [3], ASHRAE Thermal Comfort Tool [4], DEROB-LTH [5] etc are some of the commercial applications found in the market today. Most of these concentrate on energy consumption, low energy or thermal comfort achieved to measure a hypothetical human at a point in space through mathematical calculations, algorithms/ simulations etc. These applications are specifically designed to cater to the cold climates like USA, Europe etc. In order to provide an optimum solution, provide alternative solutions, an intelligent tool has to be devised.

Continuous increase in computational power has encouraged the development of software tools in many different fields such as medical, automotive, finance etc. During the research carried out, it was revealed that the underlying concept has only been confined as a theoretical fact. It has not been practiced in real life scenarios in Sri Lanka. Dealing with changing climatic conditions, managing enormous amounts of data and performing mathematical calculations are impossible tasks to a normal human being. As a result, lack of practical exposure to the phenomenon “Thermal Comfort” is minimal. In order to approach the architectural community, an artifact should be devised to guide the architects, designers, to achieve thermal comfort in their architectural presentations and expose the professionals to take up this phenomenon to the architectural community. It is important to note that there is no proper tool designed to decide on optimum solutions, provide alternative solutions and weigh the pros and cons of solutions offered in this field of architecture.

2. PROBLEMS IN DESIGNING OPTIMUM ROOF ARCHITECUTRE AND DECIDING APPROPRIATE THERMAL INSULATION MATERIALS

In designing optimum roof architecture and appropriate reflective/ resistive insulation materials for a particular dwelling situated in a particular region or zone [6] is a problematic issue in many facets. Many problems associated in deigning can be overcome successfully by incorporating complex expert knowledge/ experience, experimental results and importantly natural resources such as sunlight, ventilation and humidity. Great emphasis was laid on acquiring proper and accurate expert knowledge and also on their experimental results. Not only expert information, climatic information was also obtained in order to analyze the thermal comfort values per region and to obtain its comfort zones [6]. The research gives an account of the strengths and weaknesses of software tools that have been utilized in the architectural world.

In order to deal with expert knowledge, climatic information and also to obtain optimum results, alternative results according to thermal comfort levels, budget etc are needed to comply with to devise an artifact which can emulate expert intelligence. We now summarize some of the more important issues highlighted here in which are also relevant to our discussion as well.

A. O*btaining proper Roof parameters.* When designing a proper, optimum roof according to thermal comfort levels, budget, land information etc, quite a large amount of parameters are affected. For an example, orientation, roof shape, roof angle, roof materials, ceiling materials, ceiling shape etc are affected [7]. Deciding different roof architectures are classified among the dwelling types, new dwellings or existing dwellings.

B. *Acquiring the thermal comfort information per region.* This is one of the most important and critical task. Because of its complexity in obtaining comfort regional climate details [6]. This is mainly achieved by capturing maximum and minimum temperatures, humidity and air velocity in a particular region. These details will be directly fed into the Graph [8] to produce comprehensive information on thermal comfort temperature, humidity and appropriate air velocity.

C. *Use of Artificial Intelligence techniques*
According to the research carried out, it revealed that most of the software tools use mathematical techniques to provide direct solutions for the user. It is not appropriate in determining the optimum solution or alternative solutions and reasoning each solution and most importantly dealing with climatic conditions. Designing roof architectures based on thermal comfort conditions, budget etc needs expert knowledge/ experience. Therefore, using artificial intelligence techniques would be an ideal method to deal with the problem.

D. *Tackling optimum roof architecture.* Mostly, a great effort is spent on deciding the optimum combination of the roof parameters for a particular dwelling to achieve thermal comfort. To decide on the optimum architecture, genetic algorithm [9] or case base reasoning [10] can be incorporated.

Choosing the appropriate AI technique was a challenging task.

E. *Tackling appropriate thermal insulation materials.* Before deciding the proper insulation [11] thickness, the professional has to be aware of the climatic parameters that are being inspected. There are many factors involved such as the room temperature, roof temperature, humidity, air velocity, number of levels in a dwelling, the existing roof material etc. Based on these factors a proper insulation thickness [12] has to be devised per room in a dwelling. Because comfort varies from one part to another in a dwelling. Therefore, proper insulation material layout has to be manufactured.

From the above discussion we infer the following conclusions.

A. Proper roof parameters have to be obtained in order to present the optimum results according to thermal comfort conditions. Along with the resulting heating loads, or any extra features added, have to be displayed. Assumption: Only four types of Roof shapes have been built-in, in the **RoofOpt** prototype.

B. In order to obtain the optimum results, alternative along with the reasons case base reasoning AI technique have to be used. because, past expert experience/knowledge is being inferred.

C. To decide on the proper thermal insulation thickness for a particular dwelling per room, including changing climatic information and professional experience, Fuzzy logic AI techniques have to be used.

3. DESIGN OF XML GENERTATION

This feature has been designed to retrieve and process climatic information through XML files, which have been made use of to extract updated climatic information from the meteorology department, if an EDI link exists.

This is mainly developed using "**System.XML**" inbuilt library in VISUAL .NET environment. The library will provide a mechanism to read information and write information in to the XML file where each climatic detail has to be written and read using **WriteXML** and **ReadXML** respectively. Details of these inbuilt types can be found in [13].

The construction of XML files will be explained in two different sections.

Write climatic details to the XML file:

The climatic information retrieved from the local database is fed in to a temporary Dataset through standard stored procedures.

Using the **System.Data** and **System.Data.SqlClient** libraries it was able to directly manipulate the database through prewritten stored procedures. The main concept of handling the climatic data stored procedure at the coding level was to gather climate information to the interface level which will be used to manipulate data retrieval, insert, delete and update at the database layer.

```
sqlCommand1.Parameters.Clear();

sqlCommand1.CommandType =
CommandType.StoredProcedure;

sqlCommand1.CommandText =
"SD_SP_GetClimateData";

SqlParameter ParaNum
=sqlCommand1.Parameters.Add("@RegNo",SqlDb
Type.Char,10);
```

Figure 3.1 - Retrieve Climatic information from database tables

Through a Data reader each dataset parameter is written to a XML file. Following will portray the creation of the XML file using **XMLTEXTWRITER.**

```
XmlTextWriter writer = new
XmlTextWriter(@"F:\Softwareproject\Myfinal
yearproject\cbr\CBR_Tester\WritingUserInpu
t.xml",null);
writer.Formatting = Formatting.Indented;
writer.WriteStartDocument();
writer.WriteStartElement("Query");
writer.WriteAttributeString("CID",id)
```

Figure 3.2 - Creating an XML file and start writing the document

Read Climatic details from XML file:

To retrieve data from an XML file, the method **XMLReader** [13] was used to read each node at a time.

To recognize each node **XMLNodeType.Element** and **XMLNodeType.Attributestring** is used to read the value specified under that Node.

```
switch (Reader.NodeType)
{
   case XmlNodeType.Element:
   name = (Reader.Name);
    if (Reader.Name == "Query")
```

```
{
  while
   Reader.MoveToNextAttribute()
   {
      SID = (Reader.Value);
}
```

Figure 3.3 - Reading the XML file

For brevity, lengthy discussions related to the XML is avoided, more information can be found in [13].

4. DESIGN OF TWO DIMENSIONAL DRAWINGS USING GDI+

To achieve user friendliness in presenting information to the user, 2D designs were incorporated. GDI+ is quite a complex, challenging design library in constructing changeable designs (E.g.: Whenever the user changes roof details, design changes accordingly).For more information on GDI+, refer [14].

Visualize different roof shapes and insulation thicknesses to the designer, namespaces **System.Drawing.Drawing2D** and **System.Drawing** was incorporated. These namespaces provide basic shapes such as lines, rectangles, circles, pies etc. In order to design roof shapes in different angles, many complex mathematical equations have to re apply on the mentioned basic shapes.

```
GraphicsPath myPath = new GraphicsPath();
Rectangle pathRect = new
Rectangle(181,198, 45,45);
myPath.AddRectangle(pathRect);
Pen myPen = new Pen(Color.Transparent, 2);
e.Graphics.FillPath(mySolidBrush, myPath);
e.Graphics.DrawPath(myPen, myPath);
```

Figure 4.1 - Designing roof shapes using rectangles and pies.

```
Y = (referen - sweepAngle);
Z = (Y/5);
startAngle = System.Convert.ToInt64((Z *
2.5)+ referen);
}
else if (sweepAngle >= referen)
{
   Y = (sweepAngle - referen);
   Z = (Y/5);
   startAngle =
System.Convert.ToInt64(referen -(Z*2.5));
}
```

Figure 4.2 - Drawing the roof shapes according to the specified angles.

5. DESIGN OF MICROMEDIA FLASH TWO DIMENSIONAL DESIGNS

This concept of designing in micro-media flash allows the user to portray the optimum results obtained in the form of a graphical representation of the overall roof architecture. Presentation of roof materials, insulation materials and ceiling materials along with the roof design is created using Macromedia flash MX and Active Server [15].

This was developed using *action script* [15] and designs are generated in movie files.

Manipulating values obtained from the program are stored in action variables. Images and shapes are stored as objects, which are manipulated according to the variable values specified and place the shapes/images on the X, Y plane. During the analysis stage it was found that Flash cannot be extended to create 3D drawings. This can be incorporated in to the program by using Directive X in designing three dimensional Drawings.

```
setProperty(line, _height, hypo);
setProperty(line_r, _height, hypo);
setProperty(line_rb, _height, hypo);

line_r_x = getProperty(line_r, _x);
line_r_x_tem = ((line_r_x)+Number(length));
setProperty(line_rb, _x, line_r_x_tem);
```

Figure 5.1 - Sample from action script. Movie files generation.

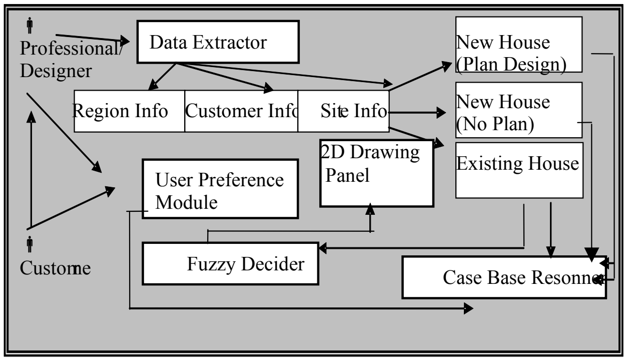
6. FUNCTIONAL BLOCK

In section 2.0 some of the key issues handled by designing the optimum roof architecture and insulations are highlighted. Investigations and surveys are carried out on existing applications and ongoing research discovered a new mechanism of solving the thermal discomforts using expert knowledge. The model is mainly categorized and described under five major components.

- Data Extractor
- User Preference Module
- Fuzzy Decider
- Case Base Reasoner
- 2D Drawing Designer

Each Major Component will be sub divided into sub components which will be described and discussed in the following sections.

Fuzzy decider and the case base reasoner are the key components in the system which presents with the optimum results according to thermal comfort condition as well as the climatic details obtained per room in a dwelling. Designer/professional will have to obtain climate data, customer requirements and site details through the **Data Extraction** component and then acquire preferences /problem from the customer by the **User Preference** component. Customer requirements, roof architecture preferences details also thermal comfort, budget etc are fed in to **the Case Base Reasoner**. This in turn will present with the optimum roof architecture details. By Moving into the **Fuzzy Decider** a customer will be able to obtain an appropriate thermal insulation thickness per room in a dwelling.**2Dimensional designer** will portray the optimum roof details as well as insulation thicknesses in graphical format in charts.



6.1 - Overall Functional Block Diagram

6.1 COMPONENT DESIGN

Each component is further divided into sub modules, where each module performs its task independently. Communication between each module will be handled through shared information.

6.1.1 Data Extraction

This particular module will be mainly targeted at gathering raw information from external sources such as regional & Climate Information from Meteorology Department or from professionals, Customer Information & Dwelling Designs and finally Site Details gathered during the site inspection. Site details will be then categorized into two sub modules, new dwelling designs and existing dwelling designs.

- Climate Information
- Customer Information
- Site Information
  - New Dwelling Design
    - Plan available
    - Plan not available
  - Existing Dwelling Design

6.1.1.1 Climate Information

This particular feature is added to obtain climate information to calculate the thermal comfort per region. This will allow the professional/designer to directly retrieve thermal comfort values for particular customer site location. Here, it is able to extract climate data from XML file formats (if a direct connection is available with the Meteorology Department) or a professional will be able to enter information by keying in data manually.

6.1.1.2 Customer Information

This will extract basic customer information like customer name, address etc while assigning a customer a unique identifier. Along with this information the professional/Designer should input, whether the roof is designed for a new house or an existing house. Depending on the type of category, he will be placed in the appropriate *site module*.

6.1.1.3 Site Information

This is one of the important sub modules, where most of the other major components will be accessed within it. Site Information is categorized into three sub components. For example, when a customer needs to obtain a solution for a existing house, the roof design solution. This sub module directly interacts with the *Case Base Reasoner* and the *Fuzzy Decider* for appropriate solutions. The following sub modules will be discussed in wider context.

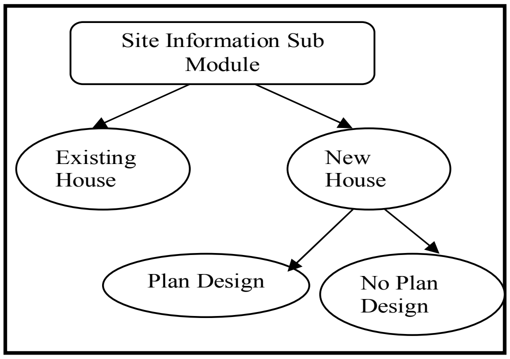


Figure 6.1.1.1 – Decomposition of Site Information Module

6.1.1.3. a Existing House Sub Module

This module will mainly target customers who need to find a comfortable roof through insulations or renovate his particular roof appropriately. When a professional visits to the particular site location he will be able to gather information such as the roof details of the existing house and the temperature, humidity. Air velocity levels of the existing house. These climatic conditions will be inspected in rooms that directly touch the roof.

6.1.1.3.b New House Sub Module (Plan available)

This module is one of the sub modules of the New House Module. This contains information about the customer who is willing to verify and design proper roof architecture according to the thermal comfort levels, budget and cooling, heating loads. This will offer the customer an opportunity to verify whether the drawn roof architecture provides adequate interior thermal comfort.

6.1.1.3.c New House Sub Module (Plan not available)

This module is mainly designed for the customer who has no plan drawn prior to the construction of a new house. It will allow the designer to choose proper roof architecture according to the site location, thermal comfort levels and customer preferences.

6.1.1.4 User Preference Module

This is one of the most important modules, which allows the user to specify roof architecture preferences and to verify whether it satisfies the thermal comfort levels, budget or the user requirements.

This module will be used in three specific areas.

1. In existing dwellings at the renovation stage - Will allow the user to key-in current roof architecture details to verify whether it meets the thermal comfort levels, user required budget and usage of insulations.
2. In New dwellings at the Construction stage - Will allow the user to key- in, roof architecture details specified in the plan to verify whether it meets the thermal comfort levels, user required budget and usage of insulations.

User desired stage (User preference) – Will allow the user to key-in, user preferred roof architecture details to verify whether it meets the thermal comfort levels, user required budget and usage of insulations

6.1.1.5 Fuzzy Decider

The fuzzy Decider will be designed only to the *existing site Information* sub module. This will extract climatic readings from a particular dwelling to decide the appropriate insulation thickness for each room in a dwelling. In order to decide on the appropriate insulation thickness, a *fuzzy module* has been designed, where it is governed using the **fuzzy logic principle** [16].

6.1.1.6 Case Base Reasoner

This particular module is directly linked to the *User preference module*. It will capture all the preference parameters into the *case base reasoning module*, where it will analyze the parameters with the case parameters. Each case in the database is weighted, based on the weights assigned at the user preference module. **Case base reasoner** is designed as an independent module. Based on the user criteria, the appropriate case base module will be overloaded. Weighted results will in turn transfer to the User *preference module*, giving out the optimum results and alternative results by weighing each solution.

6.1.1.7 2D Drawing Designer

This is the centre module, where the two dimensional drawing space is created. This panel will allow the designer to look and feel the roof shapes, ceiling shapes also the insulation thicknesses that should be installed per room. This is used to transform numeric or shapes into visualized form.

Integration of the Macromedia Flash MX was through “Text files”, where optimum roof architecture parameters captured at the *Case Base Reasoner module* is written on to the text file. These parameters are extracted by the Flash action scripts, which has been described in 5.0.

7. CASE BASE REASONNING MODULE

Case base Module is categorized into 7 main components in which each will be interacting through a number of items present and the their fields. Basic Case base architecture is depicted below [10].

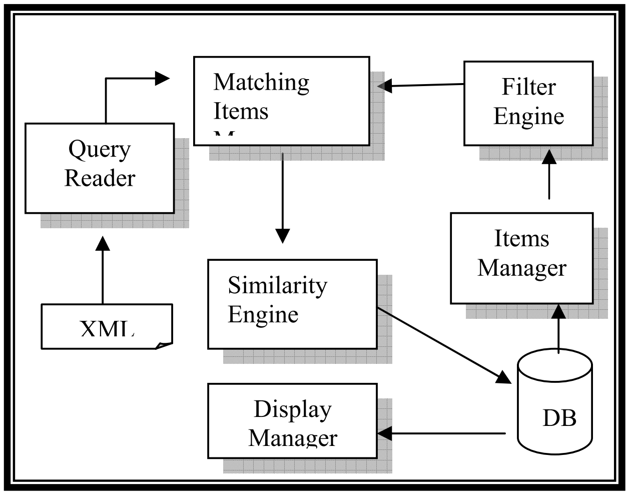


Figure 7.1 - Case base reasoning module architecture

7.1 Query Reader

This will extract the user preferences and the values specified from the XML file format. This information will be categorized into filter criteria and similar criteria .Filter criteria will be entered into the filter engine whereas similar criteria and the weights are entered in to the Matching Items Manager. Reading the XML files has been described in section 4.0.

```
<!--
Thermal Comfort Parameter -->
  <C fieldname="thermalcomfort"
Operator="<=" Value="20" />
  <C fieldname="thermalcomfort"
Operater="~" Value="28" />
  <W fieldname="thermalcomfort"
Value="7"/>
- <!--
Roof budget Filter Parameter -->
  <C fieldname="budget" Operator="<="
Value="20000" />
  <C fieldname="budget" Operator="~"
Value="65000" />
  <W fieldname="budget" Value="10" />
```

Figure 7.2 - User specified values and fields

7.2 Items Manager

This will extract past case values such as records and matches that with the user specified fields in the XML. This specific segment will hold all the case values in the database.

7.3 Filter Engine

According to the specified criteria (seen in Figure 7.2) the cases stored in the items manager are filtered by the filter engine. Along with the filtered values, filtered field names are sent to Matching items Manager.

7.4 Matching Items Manager

The filtered items obtained from the filter engine and the similar items obtained from the Query reader capture and passed to the Similarity Engine.

7.5 Similarity Engine

This segment will capture the filter items, similar items as well as their weight from the matching item manager. Using **K-nearest algorithm** [10], similarity between the filter items from the database is matched with the similar items obtained from the query reader. This will in turn calculate the distance/ similarity as a percentage. Each solution is ranked according to the percentage found.

7.6 Display Manager

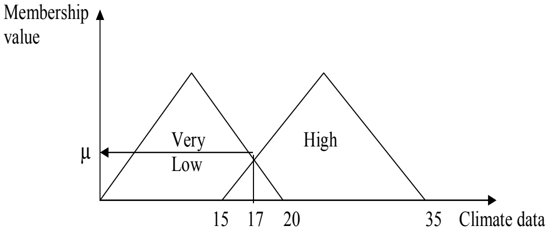
Similar items gathered from the similarity engine are directly transferred to the database for easy retrieval using stored procedures created by **System.Data** and **System.Data.SQLClient** libraries.

8. FUZZY LOGIC MODULE

The Fuzzy decider component will extract climatic data to decide on the appropriate insulation thickness for a particular room. Please note that each room will be evaluated separately in the fuzzy module. Centre of Gravity defuzzification method was incorporated to acquire the optimum insulation thicknesses. The particular algorithm is one of the promising algorithms found among the other fuzzy logic algorithms. [16]

8.1 Fuzzifying Climatic data

When the climatic data enters into the fuzzy module, it will check the appropriate membership function for each climatic parameter (Room Temperature, Humidity, air velocity and Roof Temperature). Each membership function, a membership value is calculated (µ value).This value will be passed onto the inference engine with the membership name.



```
string[] theTemp =
(GetMFTempName(TempValue).Split ('|'));
// membership name
string[] theHumid=
(GetMFHumidName(HumidValue).Split('|'));
string[]
theAirVel=(GetMFAirVelName(AirVelValue).Sp
lit('|'));
if (theTemp.Length == 2)
   {
      TempCheck1 = theTemp[1];
      MembershipFuncnew(TempCheck1,TempValue
);  // get membership value
```

Figure 8.1 - Obtaining membership value and Membership name

8.2 Checking for the appropriate rules

The passed membership names and the values for each climatic parameter will be evaluated by running through each rule specified. Please note that each rule is generated by identifying the relationship between the climate and insulation thicknesses provided by the professionals. (E.g If Temp = High and Humidity =

High and Air Velocity = Low, then the interior will be uncomfortable. Therefore, thickness should be thick.) These rules will be evaluated and the appropriate rules will be fired [16].

8.3 Getting the appropriate thickness based on the climate data

Appropriate thickness, membership names and the calculated membership values (μ value) will be extracted by the Inference engine procedure. Applying these values in the output membership function the appropriate thickness per room can be obtained. Using **Centre of Gravity** method, each membership function's area will be calculated and divided by the upper and lower boundaries of the total thickness membership functions to obtain the centre value. This will be deduced as the appropriate thickness value for the specified climatic data.

9. UNDERLYING ISSUES

In this section, some of the more important issues handled by the **"RoofOpt"** is discussed. As mentioned in 1.0, there is an argument between "designing optimum roof architecture based on thermal comfort conditions" using mathematical models and using artificial intelligence. There are a few software applications developed using AI techniques to the architectural community. This is possible only if there are predefined algorithms in designing roof architectures. To capture optimum results based on thermal comfort, budget etc can be obtained through professional knowledge, experience they have gained. Also it is relevant to use research/experimental results in determining the optimum solution.

In deciding the appropriate thermal insulation thickness per room according to climatic conditions to is never possible using mathematical equation. Most of the manufactures have their own R-values labeled per region [11]. But, these values are not always true when a particular site location is concerned, also the house architecture. Therefore, using professional knowledge experience in deciding the appropriate thicknesses per room will result in a proper solution suited for a particular dwelling.

Currently, the **RoofOpt** presents with the optimum solution by searching the highest weightage assigned by the *Case based reasoner* module. In order to automate the searching process of the optimum result among the solutions, Genetic algorithm [16] can be incorporated. We again stress the importance of obtaining the optimum results among the solutions, which is not completely automated.

10. IMPLEMENTATION

The [*]prototype was developed using .NET Framework in C# programming language due to its ability to create graphical designs, platform independence and powerful graphical user interfaces. SQL Server 2000 was chosen as Database Management System because of its capability to create stored procedures to perform the transaction automatically, Optimizing application performance during data retrieval, Recovery, security etc.

"RoofOpt" is implemented in three distinct layers to support the development of three tier architecture. Because of this nature, it was able to achieve independence in each layer.

For example, changes carried out at the middle layer or at the GUI layer will not be affected on the Data Layer. Here, the GUI Layer is consists of Data Extractor, User preference Module and displaying the optimum results in three different format such as graphical, tabular and 2D designs whereas the middle layer will extract information from the GUI layer and the Data layer to perform the particular operation. The *fuzzy decider*, outputs will be directly displayed on the GUI and saved in the Database. But in the case of *Case base Reasoner*, the outputs will be directly saved in the database and retrieved by the GUI layer.

In order to decide on the appropriate insulation thickness based on the climate information captured per room, it will be computed using fuzzy logic algorithm.

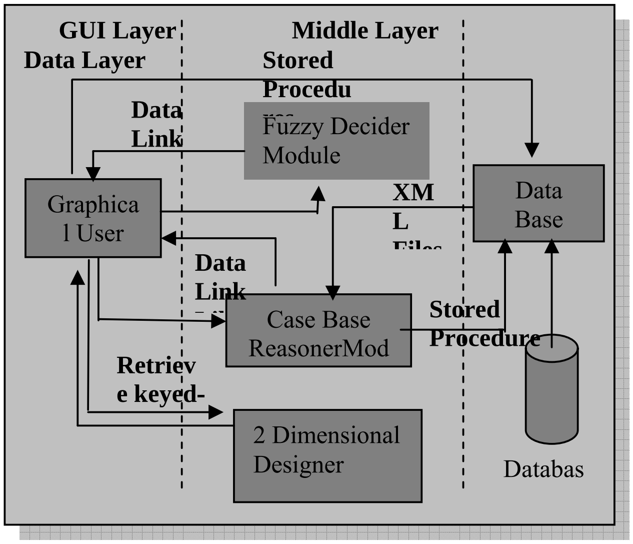


Figure 10.1 – Overall system architecture

The GUI of the prototype has the following capabilities:

[*] For brevity, the GUI designs are not discussed in this paper. However it can be found in reference [16] for interested readers.

- ❑ Downloading or key-in regional climatic information
- ❑ Customer Information and inquires
- ❑ Optimum roof architecture based the designed roof architecture
- ❑ Optimum roof architecture based on user requirements
- ❑ Optimum roof architecture obtained at the renovation stage of the existing building
- ❑ Insulation solution for an Existing dwelling
- ❑ Two Dimensional design portrayed for the selected optimum roof architecture

11. CONCLUSION

It is interesting and not altogether coincidental that this paper opens up a new vista to the concept of "Thermal Comfort dwelling designs through optimum roof architecture", challenging the conventional methodologies by incorporating new technologies to tackle the problem. This has discussed what seems to be impossible, indeed feasible and practical towards the betterment of the architectural community and relevant to most personnel attempting to tackle this problem. Many software tools exist, but most of the tools concentrate on low energy buildings, energy efficiency. Very few tools are designed to target on "Thermal comfort in dwellings" which will it in turn achieve energy efficiency and cost effectiveness.

In order to address the issues on thermal discomforts, energy crises as well as economical issues arises in third world countries, a new dwelling design concept was proposed. The unstructured, heuristic nature of the current methodologies paved the way to the proposed concept "Thermal Comfort achieved in passive houses via roof architecture (RoofOpt)". This framework proposes to develop an intelligent artifact which will determine the optimum roof architecture according to thermal comfort conditions in a dwelling.

The "RoofOpt" is researched in achieving a structured design method and devise a proper solution through a new design mechanism eliminating previous mathematical models. Consequently, the concept would provide optimal roof architecture and the appropriate thermal insulation thicknesses and materials for a given user specification.

REFERENCES

1. University of California "RESFEN 3.1 for Calculating the Heating and Cooling Energy Use of Windows in Residential Buildings", Home page, Jan.10 ,2003 (http://www.resfen.com/)

2. Konstantinos Papamichael. "Building Design Advisor", Home page, June. 21, 2001 (http://www.eere.energy.gov/ buildings/ tools_directory/ software/bda.html)

3. Federspiel, C., R. Martin, and H. Yan "Thermal Comfort Models and Complaint Frequencies, Dec. 20, 2003 (http://www.lema. ulg.ac.be/TOWNSCOPE/townscope.html)

4. "ASHRE Fundamentals Handbook", Dec. 20,2003 (www.ceere.org/beep/docs/ASHRAE/ C29_txt_IP_rev1.doc)

5. Swedish International Development Cooptation Agency. "DEROB-LTH for MS Windows, User Manual (version 99.01+3). [CD-ROM]. Feb. 28, 2004

6. Nugroho Susilo. "Passive Design in Warm-Humid?", Feb. 20, 2004

7. G.K. Garden. "Thermal Considerations in Roof Design", Jan. 12, 2004 (http://irc.nrc-cnrc.gc.ca/cbd/)

8. Marek Obitko, "Introducton to genetic algorithms", Dec. 20,2003, (http://cs.felk.cvut.cz/~xobitko/ga)

9. Morgan Amelia A. Baldwin. "Case base Reasoning", Dec. 20,2003,(http://accounting. rutgers.edu/raw/aies/www.bus.orst.edu/faculty /brownc/aies/news-let/fall95/casebase.htm)

10. Welch Kevin. "Insulation Fact Sheet", Jan. 10, 2004, (http://www.ornl.gov)

11. Jayasinghe M Thishan R , Priyanandana A.K. M "Thermally comfortable passive houses for tropical uplands" ,*Research project* , June 2002

12. Darshan Singh." XML for C# Programmers",Home page, March. 23, 2004 (http://www.PerfectXML.com)

13. Ferguson Jeff, Brian Petterson, Jason Beres , Meeta Gupta. "C# programming Bible", March. 23, 2002

14. Macromedia, Inc, "Welcome to Macromedia Flash MX", Home page, March. 23, 2003 (www.macromedia. com)

15. Nelson Marcos. “Fuzzy Logic Description Detailed”, Home page,Feb. 20, 2004, (http://www.comp.nus.edu.sg)

16. Perera M.S.R., ``Thermal comfort for passive dwellings via optimum roof architecture,'' **Final** *year project thesis*, Informatics Institute of Technology, Wellawatta, Sri Lanka, April 2004

Factors that Impact the Link between Software Development and Maintenance

M. R. S. Perera[1] and K. P. Hewagamage[2]
[1]Creative Solutions (Pvt) Ltd, No 80, Keells Realtors Building, Nawam Mawatha,
Colombo 02

[2]University of Colombo School of Computing,
(UCSC), Colombo, Sri Lanka
e-mail: [1]ruwansurvey@yahoo.com and [2]kph@ucsc.cmb.ac.lk

Abstract

At a time when there is severe price competition for new and existing business in the software industry, there is additional focus on efficiency and cost reduction for survival. Available literature shows software maintenance activities to constitute a major portion of the costs incurred in the software development cycle. This study based on expert interviews and existing literature, identified eight sub-factors as being instrumental in widening the gap between software development and software maintenance activities. Data also showed the importance of proper documentation, skills, communications, vision, and estimation of financial and non-financial costs in reducing this gap. Implications for software companies are discussed.

Keywords: software maintenance, cost estimation, software development, regression analysis, and software life cycle

1.Introduction

Irrespective of the application domain, "maintainability," is one of the key attributes of an effective software product. In fact, the most widely used software development process model, the waterfall process model, introduced by [1] Royce (1970) identified "system maintenance" as one of it's key five stages [2] (Sommerville, 1995). However, most new software development companies tend to handle new software design, implementation, and integration as a separate activity from maintenance. As a result, when new software is developed, the system developers such as the designers and the programmers tend to give less than adequate attention to maintenance phase of the software development cycle. In addition, the main focus of the project most often is system design, implementation, and integration and not the maintenance of the system. This can be costly to the company in the long-term where maintenance costs can be significantly higher than the development costs. For example, [6][11] Lientz and Swanson (1980) found 50% of programming efforts of large organizations to be related to maintenance activities. [4] McKee (1984) found between 65% and 75% of total software efforts to be attributed to maintenance. According to [2] Sommerville (1995), the single most expensive software engineering activity is maintenance.

In legacy systems, software maintenance can be as high as 80% of system life-cycle costs [7] (Banker and Slaughter, 1997). Inadequate emphasis on the maintenance phase of the software developmental cycle can have a negative impact on non-financial aspects of an organization as well. For example, increase in customer complains, negative word-of-mouth, and lack of customer loyalty are all concerns for software companies, especially in Sri Lanka that have lately come under intense competition for new business.

The purpose of this research is to empirically investigate if a gap exists between the development phase and the maintenance phase of the software development cycle so that recommendations can be made to narrow this gap. The study has three specific objectives. First, based on one-on-one interviews with experts in the software development industry in Sri Lanka and literature that is available, an attempt is made to identify the different factors that are likely to result in software maintenance activities either being delayed or altogether fail. Second, the study attempts to determine the relationships on those factors identified in objective 1, taken one at a time. Finally, an attempt will also be made to determine in the presence of all factors, which factors are most likely to impact the gap between software development phase and software maintenance phase. Results from this study will be most beneficial to software companies in Sri Lanka in reducing their maintenance costs so that they can be more competitive in the global software industry.

The organization of the rest of the paper is as follows. In section 2, we discuss the background of the research by introducing the study variables/factors and the conceptual model. We conclude section 2 by providing a brief definition of the study (factor) variables. In section 3 we discuss the methodology that include survey design, sample, data analysis and results. Finally in section 4, discussion and implications of the study results and limitations and direction for future research are provided.

2.Background

2.1 Factors that affect software maintenance activities

[2] Ian Sommerville (1995), maintenance activities in software systems are characterized as a sequence of corrective, adaptive, and perfective actions. While corrective actions relate to new enhancements made to an existing system that may need further corrective action, adaptive and perfective maintenance of the system may arise due to changes in the organization's business processes and user needs [10] (Krishnan, Mukhopadhyay, and Kriebel, 2004). The following discussion attempts to identify the factors that contribute the gap between software development and maintenance that can delay or halt the maintenance activities such as corrective, adaptive or perfective actions.

Using available literature[2] and interviews with individuals who are knowledgeable about the software development industry, three general factors that may contribute to the gap between the software development phase and software maintenance phase were identified. These general factors are:

(1) Access to information and competency of software development and maintenance staff.

(2) Management philosophy.

(3) Cost estimation.

These general factors were multi-dimensional in that several sub-factors were identified within each factor. The general factor access to information and competency of software development and maintenance staff, consists of three sub-factors that were labeled "lack of knowledge," "lack of documentation," and "lack of skills." The general factor management philosophy also consists of three sub-factors. They were labeled "lack of communications," "lack of coordination," and "lack of a long-term vision." Finally, the general factor, estimation of cost contained two sub-factors. They were labeled "financial & non-financial", and "methodology and modelling." In summary, eight sub-factors or dimensions were identified as likely of delaying or halting maintenance activities in a software company. In addition, these eight sub-factors are likely to be instrumental in widening the gap between software development and software maintenance activities, if not managed properly.

2.2 Conceptual Model

Figure -1 visually illustrate the impact the eight sub-factors identified in section 2.1 have on software development and maintenance activities. These sub-factors for the most part are under the control of either the software development or the maintenance company. The figure introduces three additional factors: (1) competitors, (2) technology, and (3) market fluctuations, which may not be controlled by either the software development or the maintenance company. While this study does not investigate the influence of these three factors have on the gap between software development and maintenance activities, they are nevertheless included so that the diagram provides an accurate depiction of the environment, software development and maintenance activities are generally undertaken.

The impact of each sub-factor has on the link between software development and software maintenance activities is dependent on the difference between the required (necessary) level and the available level of each of the eight sub-factors. When the required level is more than what is available for the software to function satisfactorily, the sub-factor will contribute to the gap between software development and maintenance and the gap will increase.

However, when the required level is less than the available level, the sub-factor will reduce the gap between software development and maintenance and the gap will decrease. Obviously, the importance each sub-factor has on the gap between software development and maintenance is dependent on the type of the project that is undertaken. In addition, the larger the difference between required level and available level of a sub-factor, larger the impact the sub-factor will have on the link between software development and software maintenance.

Essentially, eliminating the negative impact each of the sub-factors has on the link between software development phase and maintenance phase is dependent on exceeding the available level of the sub-factor more than it's required level. For example, if the available documentation that is needed to meet the maintenance activities is less than what is required, the factor (documentation) is likely to increase the gap between software development and software maintenance.

Impact of the = Required level of the - Available level of
sub-factor sub-factor sub-factor

An attempt is made below to define and explain each of the sub-factors in terms of a gap that is likely to exist in the maintenance phase of the software development cycle.

As stated earlier, each of these sub-factors are also likely to impact the link between the software development and maintenance activities.

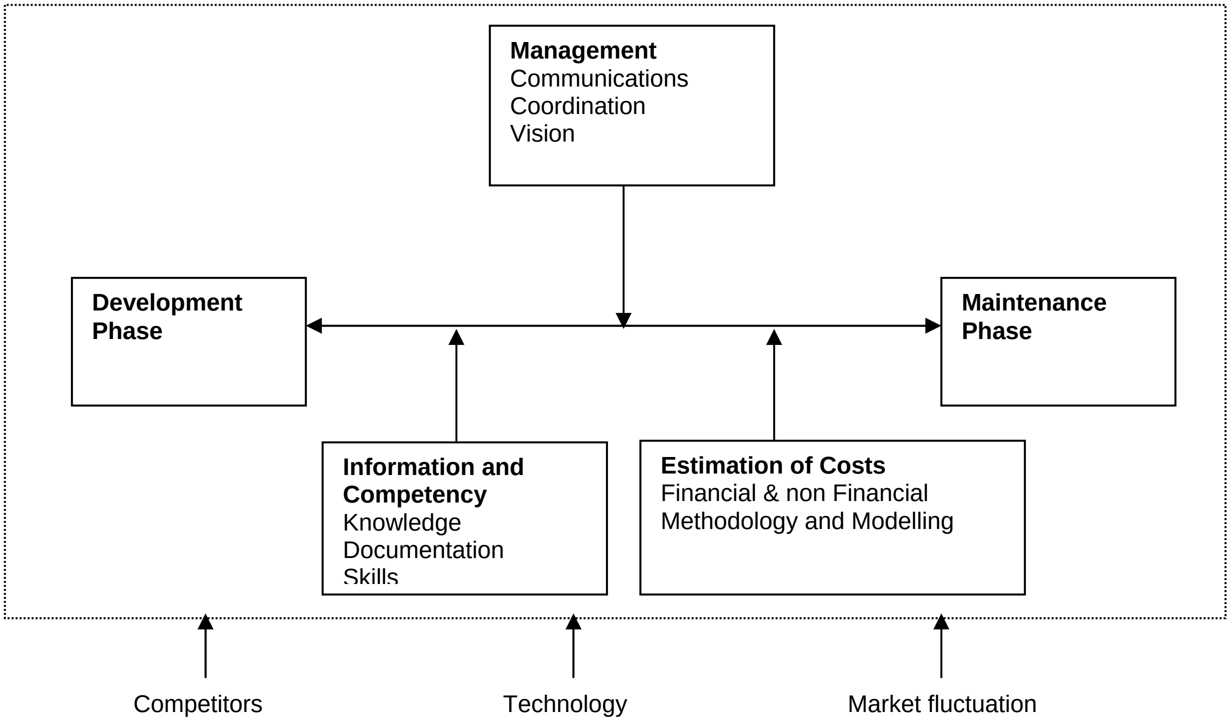


Figure 1 - Conceptual Model

2.3 Brief definition of the sub-factors

Sub-factor - 1: Knowledge Gap. This gap is the difference between the knowledge that is required from the maintenance team to conduct their activities satisfactorily and the knowledge they actually possess. Knowledge here is defined as the business knowledge or experience on the domain and the technology that is used in the software product.

Sub-factor - 2: Documentation Gap. This gap is the difference between instructions (i.e., manuals etc.) that are required for the maintenance team to conduct their activities satisfactorily and the instructions that are actually available. These documents are most often produced by the software development company.

Sub-factor – 3: Skill Gap. This gap is the result of the difference between the skill set that is required of the maintenance team and the skill set that the maintenance team really possess. The skill set is defined as the necessary skills by maintenance team to perform their activities satisfactorily.

Sub-factor – 4: Communications Gap. The communications gap is the difference between the level of communication that is required between the development team and the maintenance team and the actual level of communication that is taking place between the two teams.

Sub-factor – 5: Co-ordination Gap. The coordination gap is the difference between the level of coordination that is necessary among the maintenance, marketing, and management teams in creating the synergy that is required to meet the maintenance goals and the level of coordination that is actually present among the three teams or divisions.

Sub-factor – 6: Vision Gap. The long-term vision that is required by the top management to meet the true potential of a company compared to the present vision of that company. The vision of software maintenance companies can be at many times short-term (i.e., managing a single business transaction), rather than being long-term (developing a business relationship).

Sub-factor – 7: Financial and Non-financial Cost Gap. This gap is the difference between initial financial and non-financial cost estimates and the financial and non-financial costs actually incurred by a company.

Sub-factor – 8: Methodology/Modelling Gap. This is the gap between presently applied methodologies & models in software maintenance phase and the optimal methodologies & models that are required for proper software maintenance activities.

The remainder of the paper attempts to investigate if data from Sri Lankan software industry support the basic proposition of this research that the eight sub-factors impact the development and maintenance phase of the software development cycle.

3. Methodology

3.1 Development and administration of the questionnaire

To determine if the eight sub-factors identified using one-on-one interviews will emerge when tested using a sample, it was first important to develop the items that would measure each sub-factor. The items used to measure each of the eight sub-factors or dimensions were developed using secondary data sources (i.e., software journals, magazines, and books) and reasons mentioned by interviewees for the existence of these dimensions. This resulted in 32 items that measured the eight sub-factors.

The survey also contained the single item, "based on your experience when developing software, how much of a gap do you believe is present between the development phase and the maintenance phase?" This item measured the magnitude of the gap that the respondent identified as present between the development and maintenance activities in the project he or she was involved in at the time the data was collected. Finally, the survey included demographic items for classification purposes. They were: (1) types of projects the respondent has been involved in, (2) the role of the respondent in the current project, (3) years of IT experience, (4) whether the development and maintenance activities took place within the same or different companies, and (5) size of the company.

Surveys were administered electronically using a sample of 35 individuals who are currently employed in the software industry in Sri Lanka. A cover letter that accompanied the survey explained the study objectives as well as instructions that were needed to complete the survey. This data was then analyzed using the SPSS 12.0 (Statistical Software Package for Social Sciences) data analysis software [8] (Norusis, 2004). While 85.7% of the sample was involved in both development and maintenance of software in the past 3 years, a relatively small portion of the sample (8.6% or 3 individuals) had worked exclusively on development activities. Respondents on average had 4.36 years of software work experience and these individuals were employed at 7 different software companies. Other pertinent demographic information is provided in Table 1.

Table -1 Respondents' Three Most Recent Projects: Involvement by Other Companies and Their Location

Where Development & Maintenance Has Taken Place	Project-1	Project-2	Project-3
In Terms of Company			
Same company	60.0%	74.3%	68.6%
Two Different companies	37.1%	22.9%	17.1%
Multiple companies	2.9%	2.9%	5.5%
In Terms of Location			
Same country	34.3%	48.6%	62.5%
Different countries	65.7%	51.4%	51.4%

3.2 Measurement

Prior to investigating the three research objectives using the collected data, it was necessary to measure how reliable each of the 32 items were in measuring the dimension they were supposed to measure. [5],[9] Item-to-total correlation and Cronbach's alpha were used for this purpose. Table 2 provides the items used in the survey, the item-to-total correlation for each of the selected items, and the alpha reliability coefficient for seven of the eight sub-factors or dimensions. Since, only a single item measured the dimension "vision," item-to-total correlation or the reliability coefficient was not calculated for this dimension. There were five other items included in the original survey that is not listed in Table 2. These items were dropped because their item-to-total correlations were low, and therefore, not acceptable.

Table – 2 : Survey Items, Item-to-total Correlations, and Chronbach Alpha

Item	Item-to-total Correlation	Alpha Coefficient
Dimension – 1 Knowledge		0.66
1. The maintenance team had no exposure to the required domain / technology area.	0.48	
2. Sufficient time was not provided to the maintenance team to ramp up in the domain knowledge.	0.46	
3. New maintenance team members did not have sufficient knowledge or did not have access to experts.	0.48	
4. Constraints & changes of the business domain.	0.36	
Dimension – 2 Documentation		0.77
1. The documentation prepared for the development phase did not support the maintenance phase	0.56	
2. Out dated documents and irrelevant information was passed into the documentation that was provided to the maintenance staff.	0.59	
3. The prepared documentation was very complex where the maintenance team was unable to understand the contents	0.57	
4. The maintenance phase documentation was completely ignored by the development team.	0.56	
Dimension – 3 Skills		0.86
1. Limited (inadequate) knowledge of the tech writers in both development & maintenance teams	0.67	
2. Maintenance team skills are not identified at the initial stage.	0.65	
3. The technical details of the software product were not discussed in detail at the maintenance phase	0.77	
4. The complexity of the technology & other software product details were not considered at the maintenance phase.	0.70	
Dimension – 4 Communications		0.76
1. Expectations of the client and what the maintenance team could deliver had not been considered	0.44	
2. Lack of facilities and/or inadequate funds been allocated for communication hardware at the maintenance phase	0.70	
3. Communication protocols and the level of knowledge between the development and maintenance teams were different	0.55	
4. Member(s) of the maintenance project team leaving the company	0.58	
Dimension -5 Coordination		0.84
1. The company Maintenance team, Marketing team and the Company management team motivated (driven) by different objectives.	0.73	
2. Lack of understanding and appreciation for contributions made by different teams for the growth of the entire company	0.73	

Dimension – 6 Vision		____
1. The maintenance team’s company vision being short-term, rather than being long-term	____	
Dimension – 7 Financial & Non-Financial		0.83
1. Applying wrong technical approaches at the maintenance phase that could lead to deviation from the initial cost estimation.	0.58	
2. Granular level tasks (e.g., machine configuring etc.) were ignored or not given ample time at the maintenance phase.	0.61	
3. Estimating the financial and non-financial costs of the maintenance phase based on self-confidence	0.69	
4. Unexpected issues such as the need for additional software/hardware at the maintenance phase	0.65	
5. Maintenance staff exceeding the time estimated	0.65	
Dimension – 8 Methodology & Modelling		0.76
1. Inadequate knowledge in multiple methodologies & modelling techniques by the maintenance staff.	0.62	
2. Fear of applying new methodologies & modelling techniques by the maintenance staff	0.54	
3. Incorrect approaches (ambiguous DFD designs etc.) to methodologies & models used by the maintenance staff.	0.66	

As evidenced in Table 2, reliability coefficients for each of the seven multi-item scales were above or close to 0.70 which is acceptable as reliable as per [3] Nunnally (1978), for exploratory research such as this.

For each of the seven multi-item scales, the items that measured the sub-dimension were summated and divided by the number of items to derive the average score. These average scores were then used in the analyses that followed.

3.3 Analysis and Results

The first objective of the study was to determine if factors identified in the literature review and one-on-one interviews could also be identified using a sample of data. Table 3 provides the mean and standard deviation for each of the eight sub-factors. As can be seen in Table 3, all means were above 2.0 (2=slightly effected) suggesting each of the eight dimensions to play a role in delaying or halting software maintenance activities. The size of the mean can be interpreted as the importance of each dimension in slowing or stopping maintenance activities. For example, the mean for the dimension “knowledge” was the highest suggesting knowledge to be the most important factor in failing or delaying a maintenance project. The dimension “skills” having the lowest mean suggests this factor to have the least effect on failing or delaying a project. One sample t-test (using 2.0 as the test value because a mean more than 2.0 reflects the respondents belief that the factor impact the software development cycle) resulted in each of the means being significantly more than 2.0 at the 0.05 level of significance. These results supported the first research objective in that factors identified in the literature review and interviews were also identified using a sample. In summary, each of the eight sub-factors can play a role in delaying or stopping a maintenance project.

Table - 3: Means, Standard Deviation, and Results from One-sample t-test

Sub-Factor or Dimension	Mean	Standard Deviation	t-value
Knowledge	2.88	0.78	6.68[a]
Vision	2.79	1.22	3.78[a]
Financial & Non-financial Cost	2.75	0.91	4.91[a]
Skills	2.74	0.97	4.56[a]
Documentation	2.74	0.97	4.56[a]
Coordination	2.72	1.16	3.63[a]
Communication	2.51	0.79	3.84[a]
Methodology & Modelling	2.48	0.98	2.86[a]

[a] Significant at the $p < 0.05$ level

The second objective of this study was to determine the relationship each of the sub-factors or dimensions have with the perceived gap identified between the development phase and the maintenance phase of the software cycle. This was evaluated using Pearson's Product Moment Correlation Coefficient and the results are provided in Table 4. This correlation is interpreted as follows: When the coefficient is close to 1, there is a very strong relationship and when the coefficient is close to zero, there is no relationship. A positive coefficient demonstrates a direct relationship while a negative coefficient demonstrates an inverse relationship. As the results in Table 4 show, all correlation coefficients are in the expected direction in that they are positive. When the values of the sub-factors increase, so does the gap between software development and software maintenance. The five sub-factors: documentations gap, available skills gap, communications gap, vision gap, and financial and non-financial gap demonstrated significant correlations at the $p < 0.05$ level with the gap perceived between software development phase and software maintenance phase. The sub-factors knowledge, coordination, and methodology & modelling did not show a significant relationship with gap perceived between software development phase and software maintenance phase.

Table - 4 : Relationships Between the Eight Sub-factors and Perception of the Gap Between the Software Development and Software Maintenance Teams

Dimension	Perceived Gap Between Software Development & Maintenance Teams
Knowledge	0.22
Documentation	0.38[a]
Skills	0.38[a]
Communication	0.41[a]
Coordination	0.24
Vision	0.49[a]
Financial & Non-Financial	0.36[a]
Methodology & Modelling	0.22

[a] Significant at $p < 0.05$

Finally, to determine the presence of all sub-factors, and which sub-factors were most likely to impact the gap between software development phase and software maintenance phase, a Regression Analysis using the stepwise procedure was implemented. Results are provided in Table 5. As can be seen, only the dimensions, company vision and documentation were selected as significant predictors. This result seems to suggest that the size or the magnitude of the gap between the software development phase and software maintenance phase is mostly impacted by the vision gap and the documentation gap, in the presence of all eight gaps. As indicated by the coefficient of determinantion (r^2), these two variables explained close to 32.5% variance in the perceived gap between software development and maintenance phase.

Table – 5 : Results from Multiple Regression Analysis

Dimension	Slope Coefficient	Standard Error	t-value
Vision	0.24	0.10	2.37[a]
Documentation	0.29	0.13	2.16[a]

For the model
F = 7.57[a]
Coefficient of Determinantion (r^2) = 0.33

[a] Significant at $p < 0.05$

4. Discussion, implications, limitations, and direction for future research

Results show all eight sub-factors to impact software maintenance activities to some degree as indicated by the means in Table – 3. In fact, the most important sub-factor was knowledge followed by vision, financial & non-financial costs, skills, documentation, and coordination. Table -3 also seem to indicate the two sub-factors communication and methodology & modeling to have the least impact in delaying or stopping software maintenance activities. The finding taken as a whole show how software maintenance companies need to prioritize their focus in addressing issues that contribute to delaying or halting maintenance activities. For example, narrowing the knowledge gap will be more important than narrowing the methodology & modeling gap in meeting maintenance goals.

A somewhat surprising observation was that these gaps were not wider than what was found (the highest mean was 2.88). One possible reason for this may be that in most projects that were reported (in 60 to 74.3 percent of the projects), both development and maintenance activities took place within the same company. As a result, there was continuity in the software development cycle and the chances of problems getting out of hand may have been avoided.

Study also found five of the eight sub-factors documentation, skills, communication, vision, and financial and non-financial to impact the link between software development and software maintenance activities (Table 4). Clearly, these five factors seem to impact the software development cycle especially at the link between development and maintenance. A software company that has a clear long term vision, that can put together a team that possess the necessary skills, where key personnel have easy access to the necessary documentation, estimate both financial & non-financial costs accurately, and there is good communication within the organization will have the most efficient and effective software development cycle in place. As a result, the maintenance cost for the company will be the lowest.

Finally, results from regression analysis once again show the importance of software companies to have a long term vision and not a short term vision and the necessity for the maintenance staff to have access to proper documentation. In the presence of all eight sub-factors, vision and documentation contributed the most towards the gap between the development phase and the maintenance phase in the software development cycle. For example, it is not unusual for software managers who are often confronted with limited budgets and scheduling pressures to take a short term view by focus on incremental maintenance to systems rather than taking a long term view of reworking the entire software system through major improvements. These are issues that the top management of the software company has to deal with if they are to be competitive.

This study has one limitation, which is the small sample size. Nevertheless, the study makes a significant contribution at a time when there is a need for such investigations, but few if any are available. Finally, the study provides direction for future research. First, research can be undertaken using a larger sample to provide validity to the results of this study. Second, it will be important to determine differential impact the sub-factors identified in this study have on software development and maintenance activities for: (1) small versus large companies, (2) companies that handle both development and maintenance activities versus companies that handle only development or maintenance activities, and (3) companies that are in Sri Lanka versus companies located in other countries.

References

[1] Royce, W. W. (1970). Managing the development of large software systems: concepts and techniques. Proc. IEEE WESTCON, Los Angeles.

[2] Sommerville, I. (1995), Software Engineering FIFTH EDITION, Lancaster University

[3] Nunnally, Jum C. (1978). Psychometric Theory, McGraw-Hill Book Company, New York, NY.

[4] McKee, J.R. (1984). Maintenance as a function of design. In Proc. AFIPS National Computer Conf., Las Vegas, 187-93[660]

[5] Devellis, R.F. (1991). Scale Development: Theory and Application. Sage Publications, Newbury Park, CA.

[6] Swanson, E.B. (1976). “The Dimensions of Maintenance,” Proceedings of the 2nd International Conference of Software Engineers, San Francisco, CA, 492-497.

[7] Banker, R.D., and Slaughter, S. (1997). “A Field Study of Scale Economies in Software Maintenance,” Management Science, Vol. 41(12):1709-1725.

[8] Norusis, M. J. (2004). SPSS 12.0 Guide to Data Analysis, Prentice Hall Publishes, New York.

[9] Carmines, E.G., and Woods, J.A. (2005). “Reliability Assessment,” Encyclopedia of Social Measurement, Vol. 3, Elsevier Inc, Oxford, U.K.

[10] Krishnan, M.S., Mukhopadhyay, T., and Briebel, C.H. (2004). “A Decision Model for Software Maintenance,” Information Systems Research, Vol. 15 (December):396-412.

[11] Enterprise Application Software

http://www.synergy-infotech.com/enterprise_applications.htm
(Last Visited on 05/19/2005)

Data Warehousing and Decision Support-Explore the Sri Lankan Context

R. A. C. P. Rajapakse[1] and W. M. J. I. Wijayanayake[2]
[1]Market Analyst at Madison Maidens Lanka (Pvt.) Ltd, Katunayaka,
Sri Lanka.
[2]University of Kelaniya,
Sri Lanka
Email : [1]chatura@email.com,chathura@jnyi.lk and [2]janaka@kln.ac.lk

Abstract

Data Warehousing is widely used around the world since '80 s, as a means of providing decision-support at complex and ad-hoc analytical situations. It has been considered as a valuable tool that allows organizations to be competitive and productive in their operations as it provides a multidimensional view over business data. However still Data Warehousing is not a popular and widely used decision-support solution in Sri Lanka. This paper is an attempt to understand the factors that affect to the less deployment of Data Warehouses in Sri Lanka and explain the current conditions for Data Warehousing in the Sri Lankan Context through a behavioural model.

Key Words: Data Warehouse, Data Warehousing

1. Introduction

When organizations are operating in a highly dynamic and uncertain business environment with increasing competition among rivals, the accuracy of decisions of the managers is highly concerned. Therefore there is a growing demand for a unified and integrated view over the business data being analyzed. Especially when it comes to the strategic and tactical levels, the information requirement most of the times become ad-hoc or non-routing. There, the accuracy of decisions highly depends on the better availability of the required information in the correct format. In that sense, companies need information systems that support the diverse information and decision-making needs of managers and business professionals [4]. How can the businesses fulfill that information and decision-making needs of their senior managers effectively? One popular solution for that is building a Data Warehouse [7].

The concept of Data Warehousing goes back to early 80's and it has been evolved over the past two decades from online analytical processing to more advanced capabilities like knowledge discovery and predictive modeling [1,15]. Many organizations in most of the developed countries have got the use of Data Warehousing. There also have been done a lot of researches on different aspects of Data Warehousing generating a vast amount of knowledge in this regard. But, as far as the Sri Lankan business organizations are concerned, so far it has not been a popular decision support tool. If we take a look at the industry, there are only few companies who have attempted to get the use of Data Warehousing.

As a developing country, there can be number of factors that lower down the potential of using a Data Warehouse as a decision support tool in Sri Lanka. For example, compared to the global standards, almost all the Sri Lankan firms are falling into the category of small businesses and even compared to the local standards, majority of the firms belong to the small and medium scale enterprises. In case of competitiveness, the number of competitive industries is also limited to industries like banking, apparel, Insurance, etc. But, with their expansion, increasing competition and productivity concerns, organizations may increasingly look for the support of information technology for their decision-making. In that case, Data Warehouse could be a strong candidate in fulfilling those information requirements. On the other hand, even though Data Warehousing has a wide playing area for researchers, there has not been done much researches related to the Sri Lankan context. Therefore, it is important to understand the influencing factors for Data Warehousing in Sri Lanka and their behavior, in order to support the local business and research community.

What is called Data Warehousing

Data Warehouse is the queryable source of data in the enterprise [7]. Even though there are number of definitions for Data Warehouses given by number of authors, the most common and widely accepted one in the academic arena is by W. H. Inman in 1992. According to him, a Data Warehouse is a subject oriented, integrated, time varying, non-volatile collection of data, basically used for management's decision-making process [3]. But, as the technology is getting advanced rapidly, it is not reasonable to stick into a specific definition all the time. For example recently, there has been movement toward virtual data warehouses, which has implications for both information dissemination and improved decision-making. Virtual data warehouses allow users to distill the most important pieces of data from disparate legacy applications, without the time, expense, and risk to data required by traditional data warehousing [16]. Therefore it is more reasonable to say that Data Warehousing is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of part or all of a business [9].

Research approach

The main objective of the research was to reveal the factors that affect the less deployment of Data Warehouses in Sri Lanka and study the effect of their behavior in the local context. In order to achieve the objectives, the research questions were designed basically focusing on the practical aspects of Data Warehousing such as information requirements, costs and benefits, return on investment, structure of the project team, know-how, etc.

Due to the exploratory nature of the research, it was designed as a qualitative interpretive research. Focusing on the above aspects, six Data Warehousing efforts in well-known leading private sector companies in Sri Lanka were studied as cases. They were representing electricity distribution and utilities, banking, tea exports, cellular telecommunication, apparels and software development industries. Other than that, expert views of the professionals in information technology and software development were also taken into consideration.

The key people involved to the Data Warehousing projects at the above six companies and some experienced professionals in Data Warehousing, were interviewed at their companies, during the year 2004. There, semi-structured, open-ended interviews were mainly used, in order to allow participants to give their personal experience and interpretation of the event freely. Apart from that, in some cases, managerial level users of the Data Warehouse were also interviewed. The rationale used here was that the people, who have involved in Data Warehousing, are most capable of providing reliable and accurate information on the practical aspects in the local settings.

The semi-structured open-ended interviews resulted data collected as narrations. In order to identify the influencing factors, those narrations were coded, grouping them into relevant sub-topics based on their similarities. Out of that analysis, the influencing factors were identified and the interrelationship between them was also defined.

Short profiles of key participants

Company L

Company L is a famous company involved in electricity distribution and utilities in Sri Lanka. They developed their Data Warehouse in year 2001/2002, in order to satisfy the non-routing information requirements of the senior managers more effectively. The architecture of their Data Warehouse is very much closer to that of Virtual Data Warehouses. They extract data from their well-integrated database using a cheap Online Analytical Processing (OLAP) tool available in the market and provide OLAP capabilities to the users. It has been a very low-cost solution compared to the other companies and it is successfully in use at the moment.

Company S

Company S is a leading private bank in Sri Lanka. They have developed their Data Warehouse in year 2003, basically aiming at the future needs of Customer Relationship Management (CRM) and Data Mining in a competitive banking environment. It follows the Central Data Warehouse architecture and involves most modern Database Management System (DBMS) software and web technologies (i.e. J2EE, Oracle, etc.). They extract data from number of sources and once cleaned and transformed, presented to the users through web interfaces. It has been a very large investment and according to them it is a continuous development process.

Company D

Company D is a well-known Tea Export company in Sri Lanka. Their intention in developing the Data Warehouse was to support their top management (Directors), in their information needs. It also follows the centralized architecture and involves an investment of about 1.5 millions. They use MS SQL server to develop the Data Warehouse and extract data from their Enterprise Resources Planning (ERP) database. It is currently at the development process and they are very positive about its return.

Company M

Company M is a leading Cellular Telecommunication service provider in Sri Lanka. They developed their Data Warehouse in year 1999/2000 expecting a better decision-support system. They used the Informix Data Warehousing solution with Metacube OLAP tool. It was an investment of about 10 millions but mainly due to operational reasons, it ended up with a failure.

Company V

Company V is one of the leading software development firms in Sri Lanka. They have got the expertise and know-how in Data Warehousing as they have developed Data Warehouses to the US market. Company V participated to the study by providing expert view.

Company U

Company U belongs to a well-known group of companies in apparel industry in Sri Lanka. They found the requirement for a Data Warehouse to obtain information for their Quality Assurance process. It was conducted as a student project but ended up with a failure.

Data analysis

The data collected as narrations during the interviews were analyzed by coding them. Based on their similarities, they were grouped into seven sub categories namely, competitiveness of the industry, availability of reference point, sufficiency of data volume for Data Warehousing, information requirements and the analytical culture, awareness in Data Warehousing, cost of implementation and availability of expertise. From those seven sub-categories, it was then possible to identify seven factors that would potentially affect to the less deployment of Data Warehouses in Sri Lanka.

- ### Competitiveness of the industry

One strong argument stated was the absence of competitive industries in Sri Lanka that requires the kind of intelligence Data Warehouses can bring. The reason is the need of perfect information to take accurate decisions, in a highly competitive industry. Therefore, the main objective of a Data Warehouse is providing that information as the company can gain competitive advantage. As a Data Warehousing professional at Company V says,

"There are very few industries in Sri Lanka, which are suitable for Data Warehousing. Those are, Telecommunication, Banking, Insurance and Manufacturing-apparel. Therefore the fact that 'what industry you are in', plays a vital role in Data Warehousing in Sri Lanka."

Representing the banking industry in Sri Lanka, Company S totally agrees with this and according to their view, even in the competitive industries, only the private sector companies are seeking for this much of intelligence. But, a strong counter argument for this points out that the competitive advantage is not the only thing, which can be expected by a Data Warehouse.
It says, Data Warehousing is suitable even for the organizations in the non-competitive industries as they can achieve high performance levels through accurate decision-making. By that, they can enhance the productivity of the operations of their organization even in non-competitive industries. Company L sets a very good example for this through their Virtual Data Warehouse. According to the IT manager of company L,

"Any organization can develop and get the better of a Data Warehouse. All what they need is an information system running on a well-integrated enterprise database"

- ### Availability of reference point for Data Warehousing

According to Company L's point of view, the only thing a company needs to build a Data Warehouse is an information system with a well-integrated enterprise database. That is going to be the reference point of the Data Warehouse and, on top of which the analytical tools are going to be built-up. The reference point would carry data related to one or more specific functions of the organization.

That reference point is basically coming through the deployment of enterprise solutions like Enterprise Resources Planning (ERP) systems, Customer Relationship Management systems, Supply Chain Management (SCM) systems, etc. As long as the companies studied are concerned, all of them are having that reference point. For example, company L, company U and company D are having ERP systems while company S and company M are having the core banking and telecommunication systems respectively.

But, as far as the environment of the country is concerned, the deployment of enterprise solutions like ERP systems, CRM systems, etc. is very little. The absence of that reference point is also identified as an influencing factor for the low deployment of Data Warehouses in Sri Lanka. According to a professional at company V, *"Another thing a company needs to be having in order to go for Data Warehousing is a reference point to develop the Data Warehouse on top of it. Examples for such reference points are ERP systems, CRM systems and SCM systems. That reference point is still not properly available here"*

But now in Sri Lanka, going for enterprise solutions like ERP systems and CRM systems, has become a general trend among the local business community. That is clearly evidenced by the number of workshops and conferences held on the related topics as well as the number of different representatives operating in the country. But according to company V, still those systems are not stabilized and matured enough for Data Warehousing.

"Even if the enterprise solutions are present, the installations are still getting stabilized and therefore the need for analytics on top of them is quite low at the moment. But when they become matured, there will be a need for Data Warehousing."

Company D is also carrying the same idea and according to them, Sri Lankan firms will soon reach the level in which there are ideal conditions for Data Warehousing. As the IT project manager at company D says,

"Compared to the developed countries of the world, Sri Lankan firms can be one or two years behind. But, the ERP systems are getting more and more popular and soon they will reach that international level"

- **Sufficiency of data volume for Data Warehousing**

Addressing the performance issues related to operational systems arise, as a result of dealing with high data volumes in providing management information, is one of the major benefits expected by a Data Warehouse. But, as pointed out by many interviewees, Sri Lankan firms do not deal with that much of Data Volume as they get the requirement of a Data Warehouse to store that high volume of data out of the operational systems. As a Data Warehousing professional at company V says,

"When assessing a company for Data Warehousing, one major factor to consider is whether they have a sufficient data volume. But the Sri Lankan firms still do not have the data volume, which gears them towards data Warehousing"

Company M provides a strong backup to this argument. According to the project leader of company M, telecommunication is the only industry in Sri Lanka, which generates a sufficient volume of data for Data Warehousing. As he says, all the other industries, even banking and insurance, do not have that much of data.

"If we take a bank with a clientele of 300,000, even 100,000 of them would not do any transaction per a given day. But in telecommunications industry, if the company has a clientele of 300,000, most of them will take or receive at least one call per a day. So it generates a large volume of data everyday and it is impossible to keep data more than one year in operational systems"

But on the other hand, both company L and company D are in industries, which does not generate that much of data volume. They have their ERP systems to do the day-to-day operations smoothly and their Data Warehouses to strengthen the accuracy of the management decisions. Also it is important to note that, both the companies have gone for a Data Warehouse without any issues of poor performance of operational systems resulting from higher data volumes. Therefore, the fact that both the companies have successfully initiated Data Warehouses without a large data volume shows that it is not always a must.

- **Information requirements and the analytical culture**

The nature of the information requirements highly affects the decision to initiate a Data Warehouse. That is, the senior managers' willingness and the need to take accurate decisions based on the results of complex and ad-hoc analytical situations. Data Warehouse becomes ideal when they are used to regular data analysis and expect fast processing of complex queries, interactive decision support systems, flexibility in ad-hoc queries, ability to look at data in different dimensions, etc.

But, a visible factor at many places was the lack of the sort of requirements that can be addressed by developing a Data Warehouse. In many places, the top management is satisfied with some pre-defined standard reports such as annual and monthly reports, and does not demand for advance decision support capabilities. For example at company L, there is a separate MIS system to provide the standard reports to the top management. Even in the case of tactical level, this situation has no difference. There the tactical managers too, are highly satisfied with the reports generated by the ERP system. Therefore, even though they have been given an OLAP system, they don't get requirements to use that very often. As a result, they don't feel the value that OLAP system can add to their operations and are very lazy on using that system to support their decisions. According to a user-manager at company L,

"We can get any report we need from the ERP in what ever the format we want. Also the frequently used reports

are given as standard reports. Therefore, even though the 'Seagate Analysis' [OLAP system] is there, it is very rarely we get a requirement to use that."

Even with a strong approach to a large-scale Data Warehousing project, company S still does not find the real requirements to use that. Though they are in a highly competitive industry, they see the Data Warehouse just as a future requirement. According to them,

"Data Warehouse is not a today's need. But, definitely it is going to be needed in the future banking industry because perfect information will be needed to take accurate decisions in a competitive environment"

Since it is not a today's need, they have the problem of defining the real requirements of the system, as they can't get the participation of the actual users in the implementation. Therefore, there is no point of providing a Data Warehouse when the requirement for a Data Warehouse is not there. As highlighted by company M, their Data Warehouse was failed completely due to the fact that the users did not use the system. According to the project leader of company M,

"In Sri Lanka, even the MIS s are not much popularly in use yet. So, before going to a Data Warehouse, you must be sure that your users demand something more powerful than the MIS. If they are satisfied with the MIS, there is no need to go for such a costly solution."

Therefore, what is implied by these facts is the fact that the Sri Lankan firms still do not have the requirements and the analytical culture to select a Data Warehouse as a mean to provide the information support to the management. But, company L's attitude towards the analytical culture is as something that can be created. According to the IT manager of L,

"Analytical culture is something that can be created. What is delivered has to be delivered to a particular requirement. Without a requirement the users may not use that but at the same time, without an analytical tool they may not feel the requirement"

Company D in contrast to the others shows the impact of having analytical culture and requirements for an initiation of a Data Warehouse. At D, the top management shows their consciousness about the accuracy of decision-making and hence expects information support, which can only be provided through a Data Warehouse. According to the IT project manager of company D,

"Delivering the information needed by the top management in the most cost effective manner, is what we have done through this Data Warehouse initiative"

- ## Awareness in Data Warehousing

Even if the requirements and the analytical culture is present, without a very good awareness in the underlying concepts of Data Warehousing, there will not be a strong motivation to go for such a costly and risky solution. The lack of that awareness is another factor highlighted during the study.

For example, according to an IT related manager of a leading private bank, a Data Warehouse should be something like a 'library' as everyone in the industry can access the data. That is clearly showing the nature of the understanding of some managers in the industry about Data Warehousing and that creation of the required knowledge in the industry has not been done properly. According to a professional at company D,

"The awareness is not created as sufficient. Only the large companies have the knowledge. But if given the knowledge, even the small companies can implement Data Warehouses"

The reason is the number of sources through which that awareness is coming, is very limited in Sri Lanka. One possible source is the software vendors who provide Data Warehousing solutions and represent the leading software vendors of the world. But, as far as the local software firms are concerned, none of them are seeing Sri Lanka as a potential market for Data Warehousing. Therefore, most of them are not providing services to the Sri Lankan market. For example company V, a leading international software development firm operating in Sri Lanka, is catering to the US market but not to the local market.

On the other hand, the representation of the leading software vendors in the world is also lacking in Sri Lanka. Therefore, Data Warehouse is not getting popularised in Sri Lanka as the ERP systems nowadays. According to the IT manager of company L,

"There are couple of companies that drive IT in the world like Informix, Oracle, etc. They have still not decided to send Data Warehousing technology to Sri Lanka through a proper representation. One day, if the local market grew as it is in India, they will send a competent person to Sri Lanka representing them"

So, the awareness is not coming through the vendors like it is coming for ERP systems through the SAP R/3 system. The other sources lacking are the researches in the academic arena and the industry-university collaboration. The very limited number of researches, which are done locally show the less attention of the local research community in Data Warehousing. Also the collaborative work between the industry and the universities like workshops and conferences on Data Warehousing is not done frequently as it is done for ERP systems.

- **Cost of implementation**

While Data Warehouses carry significant benefits for organizations, all those intelligent capabilities are at a premium price. That is, unlike the other information systems, Data Warehouses need a huge amount of resources and implementation effort. Basically, this cost structure includes hardware, software and development (staff) costs. Many interviewees pointed out this huge cost associated and the difficulty of justifying that cost, as another influencing factor in Data Warehousing in Sri Lanka.

For example, both company S and company M have invested a couple of millions for their Data Warehousing projects. So, both the companies see this huge investment as a factor that lowers down the motivation of the Sri Lankan organizations to develop Data Warehouses. According to the system designer of company S,

“It is a very high investment. Therefore, it is very hard to bear that sort of a high investment for most of the companies in a developing country like Sri Lanka”

The other issue associated with cost is the difficulty to justify that huge cost in terms of benefits. Having bad experiences resulted from a failure, company M sees the benefits of Data Warehousing is nowhere closer to the cost of implementation. According to the project leader of company M,

“Compared to the cost of resources and effort put, the benefits are very little. Data Warehousing … is good as a concept but practically you can’t say it is beneficial in Sri Lanka”

All the companies have relied on the qualitative explanations on the benefits to prove the top management that there is a return for the investment. But according to company S, it is very hard to convince the top management with the qualitative facts and that also affects to the less availability of Data Warehouses in Sri Lanka. According to the system designer,

“We also couldn’t prove that there is a return on investment in numbers, like this much of return within this much of a period. We used the qualitative facts and luckily our top management had an idea of what Data Warehousing is and what benefits it could bring. But this is not always present in other places”

So this difficulty of justifying the cost results in reducing the willingness of the top management to select Data Warehouse as a decision support tool for their company. But, another important thing revealed during the study is the possibilities of reducing the cost of implementation. Open source technologies play a major role in this case and that may enable to reduce the licensing cost of software significantly. According to a professional at company V,

“The advantage of using open source is the lower cost of acquisition. Large-scale enterprise applications like Max DB and Ingres are available, making going for open source a compelling case. Clover and Octupus are some other tools in the ETL area, which are also available”

In the case of Extract Transformation and Loading (ETL) tools, none of the companies who are using ETL tools have purchased them from outside. They are using in-house developed ETL tools. This is also a significant cost, as it involves a lot of coding to do the required transformation and the time spent for that counts as development cost as well as opportunity cost. According to company V,

“You can reduce the cost of implementation by going for ETL tools, which are already developed. Developing ETL in-house is very costly because you have to write a lot of codes for the transformation process”

For example, since company S didn’t recruit additional staff for the project, the current staff has to share their time for the project work as well as their other responsibilities.

In the case of the cost of implementation, company L’s approach also carries a significant importance. Their Virtual Data Warehouse is an example for a low cost alternative for the centralized architecture adapted by the others, as long as the requirements can be satisfied with that. That is, if the situation were as such as at company S, that sort of architecture would not do anything for the company. But for small-scale projects, it would be an ideal solution due to the very low cost associated with it.

- **Availability of expertise**

Data Warehousing is a state-of the-art technology that very often becomes a large-scale project, which consumes a lot of resources and time. Therefore, the expert knowledge and experience in Data Warehousing is also vital, to gain success. But as revealed during the study, the expertise in Data Warehousing is highly lacking in Sri Lanka and that also is an influencing factor for Data Warehousing here. Both company S and company D are strongly agreeing to this and according to the system developer at D,

“Getting the know-how is very difficult. The number of books you can find in Sri Lanka on Data Warehousing is

also limited and also you can't get much technical support from the web"

Both the vendors and the academia of Sri Lanka are not doing much contribution in providing the expertise to the industry. In the case of vendors, it is very few operating locally who have the competency in Data Warehousing. But, they are also not serving to the local business community. According to a professional at company V,

"We have the required expertise for Data Warehousing. But we are not targeting at the local market. In general, expertise on Data Warehousing is highly lacking in Sri Lanka"

On the other hand, the contribution form the academic arena is also poor in Data Warehousing. As pointed out in a study on the missing dimensions of Sri Lankan software industry also, training people to use state-of the-art technologies was seen as highly lacking. This could be seen worse in Data Warehousing.

2. Influencing factors for the less deployment of Data Warehouses in Sri Lanka

Out of the above qualitative analysis, it could reveal the following seven factors that possibly have an influence on the less deployment of Data Warehouses in Sri Lanka.

1. Less number of competitive industries

As revealed by the analysis, competitiveness of the industry is one strong driving force to go for a Data Warehouse. When the competitiveness of the industry is high, the need for getting the competitive-advantage demands for better decision-support, in the form of complex and ad-hoc analytical capabilities. But in Sri Lanka, the number of competitive industries is limited and therefore the need of better information support for accurate decision-making is not so common.

2. Lack of the reference point for Data Warehousing

The deployment of enterprise solutions like ERP systems is still not so common in Sri Lanka. Those systems very often provide the most appropriate reference point for Data Warehousing as they concentrate on different specific business functions. But, except in some of the top companies, in most of the places, the reference point for a Data Warehouse is not still available or even if available, still not matured enough as they produce enough data volumes or get advance analytical requirements based on them.

3. Insufficiency of Data Volume in OLTP systems

Most companies in Sri Lanka do not generate a sufficient volume of data to look for a Data Warehousing solution, to address the performance issues arising as a result of that high volume. The possible reasons can be the industry that company is in and the immaturity of their source systems. So, this insufficiency is also affecting to the low deployment of Data Warehouses in Sri Lanka. But, with the maturity of the reference points, some companies may get a sufficient data volume for Data Warehousing.

4. Lack of information requirement and the analytical culture

This is a major influencing factor revealed during the study. In most of the places managers are satisfied with the information support given by their OLTP systems and MIS s. Also they are not used to complex analysis of data in decision-making. Therefore, in most of the places, the demand for complex and ad-hoc analytical capabilities is very low. So as long as the managers are satisfied with the available capabilities, there is no need for an advance solution like a Data Warehouse.

5. Lack of awareness in Data Warehousing

The awareness on the concepts of Data Warehousing is very low in Sri Lanka. In the case of increasing the awareness, unlike for ERP systems, the contribution from the vendors as well as academia is at a very low level. So this lack of awareness among the senior managers on Data Warehousing may stop them going for a Data Warehousing solution even if the requirements are present at their places. But on the other hand, awareness may gradually increase with the evolvement of requirements.

6. Very high cost of implementation

Huge cost of implementation is another major influencing factor for the low deployment of Data Warehouses in Sri Lanka. Most of the companies in Sri Lanka cannot easily bear that sort of a huge investment for a decision support tool. Cost justification on the other hand is another big issue as there are no tangible benefits. Therefore, both these facts associated with cost have an impact on Data Warehousing in Sri Lanka. But the difficulty of justifying the cost is mostly due to the lack of awareness and, may dissolved with the increasing awareness.

7. Lack of expertise to develop Data Warehouses

The expertise available in Data Warehousing is also very low in Sri Lanka. Similar to the case for awareness, here also the contribution from both the vendors and academic institutions is very low. Vendors do not see a market for Data Warehousing in Sri Lanka and hence do not serve for the local industry and the academic institutions do not train people to use this technology. But again it can be seen that the expertise will also be more available with the evolvement of requirements in the industry for Data Warehousing.

3. Behavioral model of the seven influencing factors

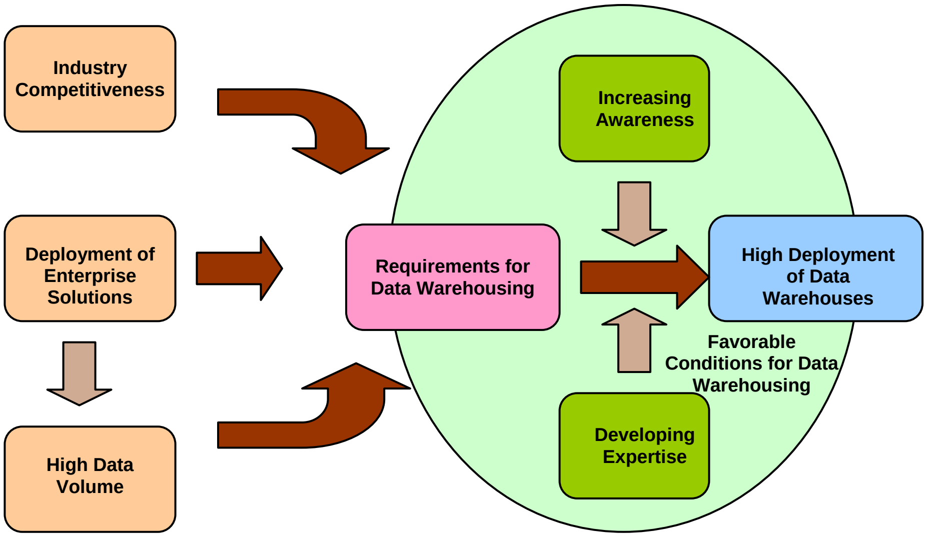
When looking at the above influencing factors, it is possible to see that they make a combined effect on Data Warehousing in Sri Lanka rather than individual effect. Therefore, it is possible to define an interrelationship between those factors, based on the same above qualitative analysis. The interrelationship defined based on the researcher's judgment is given in the table.

	Less Competitive-ness	**Lack of Reference Point**	**Less Data Volume**	**Lack of Requirements**	**Lack of Awareness**	**High Cost of implementation**	**Lack of Expertise**
Less Competitiveness				✓			
Lack of Reference Point			✓	✓			
Less Data Volume				✓			
Lack of Requirements					✓		✓
Lack of Awareness						✓	
High Cost of implementation							✓
Lack of Expertise							

Based on that interrelationship, it is possible to derive a graphical model, which would be called as a behavioral model in this study (See figure below), and it can be used to explain the combined effect of those seven factors on Data Warehousing in the Sri Lankan context. Also it can be used to get an understanding about the Sri Lankan context for Data Warehousing and come to a conclusion about its behavior.

According to the behavioral model, the conditions should become favourable to see high deployment of Data Warehouses. That means there should be high requirements that drive towards Data Warehousing and the required expertise and the awareness should also be highly available.

The primary need to get the favourable conditions is the availability of the requirements for a Data Warehouse. Those requirements for a Data Warehouse are coming from the need to be accurate in decisions, in order to ensure the productivity in operations and/or gain competitive advantage and the limitations of the existing information systems. They may be resulting


Industry Competitiveness
Deployment of Enterprise Solutions
High Data Volume
Requirements for Data Warehousing
Increasing Awareness
High Deployment of Data Warehouses
Favorable Conditions for Data Warehousing
Developing Expertise

The Behavioral Model

from the industry competitiveness, deployment of enterprise solutions and the data volume but may exist even without them.

When there are requirements, which can be satisfied by developing a Data Warehouse, the awareness and the know-how will start flowing into the industry. This would happen, as software vendors start promoting their Data Warehousing solutions in the local market and the academia generates new knowledge through researches. The ultimate result of this is the high deployment due to the favourable conditions. Out of the seven influencing factors, only six are directly present in the behavioral model. In this model, high cost of implementation is a hidden component and the effect of that component depends on whether the conditions are favourable or not. That means, the high cost becomes a problem only when the conditions are unfavorable.

4. The combined effect of the seven factors to the Sri Lankan context

In the Sri Lankan context what is basically identified as the reason for the less deployment is the lack of favourable conditions for Data Warehousing. For that, the less number of competitive industries, less deployment of enterprise solutions and the insufficient data volume of the companies are highly affecting. In the case of analytical requirements and the culture, companies do not get complex analytical requirements, which requires a decision support tool like a Data Warehouse and also they are not used to an analytical culture.

Due to the absence of these requirements, there is no strong motivation to look for advance decision support through Data Warehousing. Also since they are not used to an analytical culture, new analytical requirements are also not arising frequently. Hence the creation of awareness and the development of expertise to use the technologies associated with Data Warehousing is highly neglected. The software vendors does not see a potential market in the country and hence, their contribution is also minimum in creating the awareness and bringing the know-how into the country

Under these unfavorable circumstances, high cost of implementation is a big problem but the situation has become worse due to the lack of awareness and expertise. That is, the required awareness to look for low cost alternatives and the expertise to use the technology effectively is highly missing. So under these circumstances, the cost of a project becomes a problem due to the unnecessary expenses. On the other hand, due to the lack of awareness, it is hard to do the cost justification over the potential benefits. Therefore, the ultimate result has become the less deployment of Data Warehousing as a means for better decision-making.

5. Discussion

As revealed by this study, lack of information requirements among the potential users is a major reason for the unfavorable conditions. Another factor is the high cost and the difficulty of justification of that cost against the potential benefits. These two arguments can be further strengthened, by comparing to a study done by David Sammon and Pat Finnegan in year 2000 [2].

In their study, they have identified ten organizational pre-requisites that need to be satisfied for a successful Data Warehouse implementation. There, the necessity of a business-driven Data Warehouse initiative and the funding commitment was identified as two major pre-requisites. In the Sri Lankan context, these two pre-requisites have become major influencing factors to the less deployment of Data Warehouses in general and the lack of awareness and expertise has made the situation worse.

6. Conclusion

The main objective of the research was to provide an understanding about the local Data Warehousing context by identifying the factors that affect the less deployment of Data Warehouses in Sri Lanka. That has been reached in this research by deriving a behavioral model that can be used to understand and interpret the local Data Warehousing context.

As the first step in the study, seven influencing factors were revealed namely, Lack of competitive industries and competition among the companies, Lack of the reference point for Data Warehousing, Insufficiency of data volume in the OLTP systems, Lack of information requirements and the analytical culture, Lack of awareness in Data Warehousing, Very high cost of implementation, Lack of expertise to develop Data Warehouses.

Those influencing factors were found as interrelated and based on that interrelationship; a behavioral model could be derived. According to that behavioral model, these factors in combination create an unfavorable environment for Data Warehousing in Sri Lanka. That unfavorable environment ultimately results the low deployment of Data Warehousing for better decision-making.

Since the current conditions are unfavorable, it is better to follow somewhat a defensive strategy if an organization is planning to go for a Data Warehouse. The decision should be taken after a careful assessment of the type of analytical requirements, capabilities of available MIS s, current analytical culture, types of OLTP systems available and their maturity level and the ease of access to the required expertise in the local context.

References

1. Christopher Adamson, Michael Venerable, Forwarded by Ralph Kimball, *Data Warehouse Design Solutions,* John Wiley & Sons, Inc. 1998
2. David Sammon and Pat Finnegan, *Ten Commandments of Data Warehousing,* The DATA BASE for Advances in Information Systems - Fall 2000 (Vol. 31, No. 4)
3. Inman W., Building the Data Warehouse, New York; John Wiley & Sons Inc. Quoted from Paul Gray and Hugh J. Watson, *Present and Future Directions of Data Warehousing, 1998*
4. James A. O'Brien, *Introduction to Information Systems – Essentials for the Internetworked E-business Enterprise.* 10th Edition, Prentice Hall 2001
5. Katherine Jones, *An Introduction to Data Warehousing – What are the Implications for the Network,* John Wiley & Sons, Ltd. 1998
6. Ralph Kimball, Margy Ross, *The Data Warehouse Toolkit,* 2nd Edition, John Wiley & Sons, Inc. 2002
7. Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite, *The Data Warehouse Lifecycle Toolkit - Expert Methods for Designing, Developing and Deploying Data Warehouses.* John Wiley & Sons, Inc. 1998
8. Ranjit Bose and Vijayan Sugumaran, *Application of Intelligent Agent Technology for Managerial Data Analysis and Mining*, The DATA BASE for Advances in Information Systems- Winter 1999 (Vol. 30, No. 1)
9. Stephan R. Gardner, *Building the Data Warehouse - The tough questions project managers have to ask their companies, executives and themselves and the guidelines needed to sort out the answers.* Communications of the ACM, September 1998, Vol. 41. No. 9
10. http://www.kenorrinst.com/dwpaper.html/ *Data Warehousing Technology*, Kenn Orr, Kenn Orr Institute 1996, Revised edition 2000.
11. http://www.1keydata.com/datawarehousing/datawarehouse.html/ *Data Warehouse and Business Intelligence*, 2004.
12. http://www.eunis.org/html3/congres/EUNIS97/papers/022802.html/*Data Warehouses and Executive Information Systems - Ignoring the Hype,* Doreen Stevenson 1997
13. http://www.cssl.lk/conference_papers.htm/ *Missing Dimensions of the Sri Lankan Software Industry*, Asoka S. Karunananda, 23rd National IT conference, 2004.
14. http://www.qual.auckland.ac.nz/#Introduction/ *Qualitative Research in Information Systems*, Michael D. Myers, MIS Quarterly *(21:2),* June 1997, Modified 2004.
15. http://www.lac.uic.edu/~grossman/papers/four-gen-dm-v7.htm/ *Supporting the Data Mining Process with Next Generation Data Mining Systems,* Robert Grossman, Enterprise Systems Journal.
16. http://www.darwinmag.com/learn/curve/column.html?ArticleID=50/ *What is a Data Warehouse*, Tom Wailgum, 2001

Review on Current Steganography Technologies

S. G. K. D. N. Samaratunge

Abstract

Steganography is the art and science of hiding information [7]. The goal of steganography is to avoid drawing suspicion to the transmission of a hidden message. The success of steganography depends on the secrecy of the cover carrier. Once the steganographic carrier is disclosed then the security depends on the robustness of the algorithm and the cryptographic methods used. Therefore, to maintain secrecy either we need to make the carrier more robust against steganalysis or discover new and better carriers. This review will discuss the existing steganographic and digital watermarking techniques and their strengths and weaknesses in order to come up with an improved technique to hide data.

1. Introduction

Steganography is a scheme, which includes methods of transmitting secret messages through innocuous covert channels, concealing the existence of the embedded messages, making them untraceable. A covert channel is any communication channel that can be exploited by a process to transfer information in a manner that violates the system's security policy [6]. These methods combine aspects of digital signal processing, cryptography, statistical communication theory and human perception. With the increasing access to digital source media such as World Wide Web, the protection of ownership and the prevention of unauthorized alteration has become a significant concern.

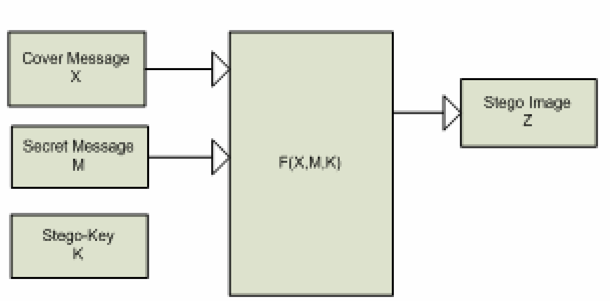


Figure 1 - Steganographic Encoding Process

The above diagram (Figure 1) shows a typical digital steganographic encoding process. The *secrete message* is the data that the sender wishes to remain confidential. This data can be text, images, audio, video, or any other data that can be represented by a stream of bits. The *cover message* is the medium in which the secret message is embedded and serves to conceal the presence of the message. The cover medium is also referred to as the message *wrapper* In the diagram, the image with the secretly embedded message produced by the process is the *stego-image*. The stego-image should resemble the cover image under casual inspection and analysis. In addition, the encoder usually employs a *stego-key* which ensures that only recipients who know the corresponding decoding key will be able to extract the message from a stego-image. Recovering the message from a stego-image requires the stego-image itself and a corresponding decoding key if a stego-key was used during the encoding process. The original cover image may or may not be required; in most applications it is desirable that the cover image not be needed to extract the message. Steganography does not alter the structure of the secret message, but hides it inside a cover. It is a good practice to combine the encrypting and steganographic techniques by encrypting a message using cryptography and then hiding the encrypted message using steganography. The resulting stego-image can be transmitted without revealing that secret information is being exchanged. Furthermore, even if an attacker were to defeat the steganographic technique and detect the message from the stego-image, he would still require the cryptographic decoding key to decipher the encrypted message. To make a steganographic communication more secure, the message can be compressed and encrypted before being hidden in the carrier. Cryptography and steganography can be used together

Cryptography and steganography are different. Cryptographic techniques can be used to scramble a message so that if the message is discovered, it cannot be read. If a cryptographic message is discovered, it is known to be a piece of hidden information and anyone intercepting it will be suspicious. As the message is scrambled it is difficult to understand and decode.

In the last few years the theoretical foundations of information hiding has advanced very rapidly. Modeling the information hiding process as one of communications with side information produced improved information hiding algorithms as well as accurate models of the channel capacity and error rates. At the same time, steganography security, i.e. the ability of information hiding to serve in a scenario where the presence of an *enemy* explicitly aiming at nullifying the hidden information goals, whatever they are, has been recognized as one of the main open issues steganographic techniques face with.

For all the steganographic systems, most vital and elementary requirement is undetectability. [2] The hidden message should not be detected by any other people. More over, the cover message with hidden message i.e. stego-media are indistinguishable from the original ones i.e. cover-media. The cover-media and stego-media should appear identical under all possible statistical attacks and the embedding process should not degrade the media fidelity. The difference between stego-media and the cover-media should be imperceptible for visual attacks.

Steganography uses two types of protocols: secret-key and public-key steganography. In secret-key steganographic model, both sender and receiver share a secret-key before conveying messages. The input message may be in any digital form and be treated as a bit stream. Public-key cryptography requires the use of two keys, one private and one public key. The public-key is used in the embedding process where as the private key is used to extract the hidden message.

Essentially, the information-hiding process in a steganographic system starts by identifying a cover medium's redundant bits (those that can be modified without destroying that medium's integrity). The embedding process creates a stego medium by replacing these redundant bits with data from the hidden message. Steganographic systems, because of their invasive nature leave behind detectable traces in the cover medium. Even if secret content is not revealed, the existence of it may be revealed. Modifying the cover medium changes its statistical properties, so eavesdroppers can detect the distortions in the resulting stego medium's statistical properties. The process of finding these distortions is called *statistical steganalysis*. Some attacks strip away the significant parts of image in a way that facilitates a human searching for anomalies. This is called *visual steganalysis*. *Structural steganalysis* looks for an easily detectable pattern in the structure of data as format of the data file is changed when hidden information is included.

The message embedding technique is strongly dependent on the structure of the cover or the carrier medium and the secrecy of encoding system.

This review will discuss steganographic techniques using different carrier mediums. In most available steganographic techniques, images are used as the covert channel because of the wide availability of images and compression methods. Therefore, this paper will give more weight to image based steganography. Most of the methods used were based on the manipulation of the least significant bits of pixel values or the rearrangement of colors to create least significant bit or parity bit patterns, which correspond to the message being hidden. The paper will also introduce a new method for hiding messages in JPEG images without making any discernible change to the carrier. This will look for the possibility of hiding data in the IPTC information of jpeg images.

2. Steganography History

Steganography has been used in a variety of forms for 2500 years. It has used variously in military, diplomatic, terrorist, personal and intellectual property applications.

According to the writings of a Greek historian, in the 5th century BC, messenger's head was shaved and tattoo a message or image on the messenger's head. After the messenger's hair grew back, he was dispatched with the message. Obviously, this message wasn't especially time constrained [9]. Mary Queen of Scots used a combination of cryptography and steganography to hide letters. Her letters were hidden in the bunghole of a beer barrel, which freely passed in and out of her prison [9]. Another centerpiece of work is a scheme for winding thread through 24 holes bored in an astragal (an astragal is a type of architectural element that separates the capital of a column and the shaft.); each hole represents a letter and a word is represented by passing the thread through the corresponding letters. Another method was to hide text by making very small holes above or below letter or by changing the heights of letter-stokes in a cover text. These dots were masked by the contrast between the black letters and the white paper.

Several stenographic methods were used during World War II. Microdots developed by the Nazis are essentially microfilm chips created at high magnification (usually over 200X). These dots could contain pages of information, drawings, etc. The Nazis also employed invisible inks and null ciphers. Invisible inks were used with much success. An innocent letter may contain a very different message written between the lines. Early in World War II steganographic technology consisted almost exclusively of invisible inks. Common sources for invisible inks are milk, vinegar, fruit juices and urine. All of these darken when heated. Null ciphers were used to conceal the secret message in the third letter of every word in a cover message. The real message is camouflaged in an innocent sounding message.

Cardano grill is another method invented by Girolama Cardano. This is a simple method which uses a piece of paper with holes cut in it. Both sender and recipient should have the same grill. When the grill is laid over printed text, the intended message can be retrieved. In techniques related to the Cardano grill, classical steganography techniques include pin punctures in text (e.g. newspapers), and overwriting printed text with pencil.

Even though considerable no of steganography techniques were in use, study of this subject in the scientific literature goes back to Simmons, who in 1983 formulated it as the "prisoners' problem".[1] Alice and Bob are in jail and want to draw an escape plan. Their entire communications pass through the warden named Wendy. If Wendy finds about their escape plan, they would be thrown into Solitary confinement. So they have to hide the plan in an innocuous cover text. The security of the message depends on the intractability of the hidden message and the secret key that Alice and Bob managed to share.

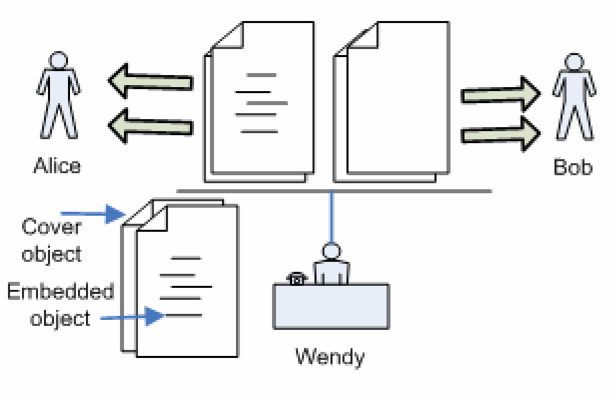


Figure 2 - Prisoners' problem

In the Figure 2, Alice and Bob represents two communication parties, wanting to exchange secret information invisibly. The warden Wendy represents an eavesdropper who is able to read and probably alter the message sent between the communication partners. This model is generally applicable to many situations in which steganography take place.

3. Digital Steganography Methods

The steganography applications range from those that actually hide data, often encrypted, within the file, to those that simply attach hidden information to the end of a file such as Camouflage. The community is concerned with a number of digital technologies, namely, text files, images, movies and audio.

3.1 Textual Steganography

Data may be embedded in files at imperceptible levels as noise. There are a number of techniques that can be employed in textual steganography.

- Open space methods
- Line shift coding
- Word shifting coding
- Syntactic methods
- Semantic methods
- Feature coding

Secret messages can be hidden in documents by manipulating the positions of the lines or the words. When HTML files are written web browsers ignore spaces, tabs, certain characters and extra line breaks. These can be used as locations in which to hide information. Hidden messages can be recovered from text by taking for example the second letter of each word and using them to produce the hidden message. This is called a null cipher or open code. Information can be hidden in the layout of a document for example certain words in a piece of text can be shifted slightly from their positions and these shifted words can then composite the hidden message [3].

3. 2 Image-based steganography

Many steganographic tools in the Internet are available for varied image formats. The fact that images can be usefully subjected to lossy compression methods has suggested that extra information could be concealed in them. Properties of images can be manipulated including luminescence, contrast and colors. A 24-bit color image has three components corresponding to Red, Green and Blue. The three components are normally quantized using 8 bits. An image made of these components is described as a 24-bit color image. Each byte can have a value from 0 to 255 representing the intensity of the color. The darkest color value is 0 and the brightest is 255. Transparency is controlled by the addition of information to each element of the pixel data. A 24-bit pixel value can be stored in 32 bits. The extra 8 bits is for specifying transparency. This is sometimes called the alpha channel. An ideal 8-bit alpha channel can support transparency levels from 0 (completely transparent) to 255 (completely opaque). It can be stored as part of the pixel data.

Structure of digital images is discussed in the *"An evaluation of Image Based Steganography Methods"* by Kevin Curran of Internet Technologies Research Group [3].

Palette-based and JPEG are two most commonly used formats. JPEG uses lossy compression in which the image is not reconstructed in exactly the same way as the original. JPEG data are sensitive to small changes in the image data which results in less capacity for hiding a secret message. A new JPEG version is introduced recently named JP2. The compression method used for JP2 is a lossless compression. Therefore, the original image can be reconstructed after compression.

A palette image format contains a header, a palette and image data. GIF images are popular as the carriers as they are widely used in web pages, recognized by all the browsers and they are easily distributable which ensures that they lend themselves to this type of activity without drawing attention to themselves. GIF uses lossless compression. It reproduces the original image exactly and therefore the original arrangement of bits making up the image is maintained. When a GIF image is displayed, the software paints the specified color from the palette onto the screen at each pixel. If the image has fewer colors than the size of the palette any unused colors in the palette are set to zero. GIF is a bitmap image. Bitmap is a system in which an image is described as a bit pattern or series of numbers that gives the shade or color of each pixel. In palette-based images, directly embedding messages in those indices may lead to major color change, since two neighboring colors in the palette may not look the same. The advantage is that three bits are embedded in each pixel and the color of each pixel does not change radically. However, the palette is modified, and colors in the palette form 32 color groups. This specific pattern can be detected automatically and reveals the existence of the hidden message. Machado proposed a steganographic method in which the palette is not modified. For each pixel, the message is embedded by replacing the index of a color with the one of the luminance-closed color. Since two colors with closed luminance may be totally different, the created stego-image may have perceptual distortion. To lessen this problem, Fridrich proposed a new steganographic scheme using the parity of palette colors. The parity of each color could be assigned randomly or simply by calculating R + G + B mod 2. The closest color with the same parity as the message-bit is used to replace the original color. Since parity bits are randomly distributed, the searched new color never departs from the original one too much. An optimal parity assignment was proposed in order to further improve the scheme and the energy of modification is decreased by 25-35% depending on the image.

For JPEG images, Jsteg embeds the hidden message by modulating the rounding choices either up or down in the quantized discrete cosine transformation (DCT) coefficients. The ability of embedding messages in the JPEG format is an advantage of this tool, since most images are stored in JPEG format and transferred in the Internet. The downgrade image fidelity caused by the embedding process depends not only on the amount of embedding messages but also on the quality factor setting in JPEG compression. If the quality factor is low, the embedding capacity should be limited in order to assure the unrevealed requirement.

A common structure for spread spectrum steganography was proposed by Smith and Comiskey. The main advantage is that it is robust to image modification. However, the shared information before communication is dependent on the length of the secret message, i.e., the payloads of the system is limited. Marvel et al proposed another spread spectrum steganography, called SSIS. For this, the technique used is similar to the spread spectrum watermarking. Because the original cover-image is needed when extracting the hidden message, an approximate estimation that obtained by image restoration process is used. Since the estimate of the embedded message is poor, SSIS should incorporate the use of error-control codes to correct the large number of bit errors. Therefore, most of the payloads in SSIS are wasted on the redundancy for error correction. Another shortcoming is that the stego-images can only be stored in the JPEG format with a high quality setting.

BMP files are uncompressed; therefore the file sizes are much larger than those for other formats. This has advantages from a steganographic point of view. But their corresponding size makes them inappropriate for distribution across a web communication media. The compression and decompression algorithms cause problems in steganography.

Current techniques for embedding of messages into image carriers fall into three categories [10]:

- Least-Significant Bit embedding (or simple embedding),
- Transform techniques,
- Perceptual masking.

3.2.1 Least-Significant Bit Encoding

A digital image consists of a matrix of color and intensity values. In a typical gray scale image, 8 bits/pixel are used. In a typical full-color image, there are 24 bits/pixel, 8 bits assigned to each color components. The simplest steganographic techniques embed the bits of the message directly into the least-significant bit plane of the cover image in a deterministic sequence. Modulating the least-significant bit does not result in a human-perceptible difference because the amplitude of the change is small. Other techniques “process” the message with a pseudorandom noise sequence before or during insertion into the cover image. The advantage of LSB embedding is its simplicity and many techniques use these methods [10]. LSB embedding also allows high perceptual transparency. However, there are many weaknesses when robustness, tamper resistance, and other security issues are considered. LSB encoding is extremely sensitive to any kind of filtering or manipulation of the stego-image [8]. Scaling, rotation, cropping, addition of noise, or lossy compression to the stego-image is very likely to destroy the message. Furthermore an attacker can easily remove the message by removing (zeroing) the entire LSB plane with very little change in the perceptual quality of the modified stego-image.

3.2.2 Transform Embedding Techniques

This technique is used for embedding the message by modulating coefficients in a transform domain, such as the Discrete-Cosine Transform (DCT) (used in JPEG compression), Discrete Fourier Transform (DFT), or Wavelet Transform [5]. Transform techniques can offer superior robustness against lossy compression because they are designed to resist or exploit the methods of popular lossy compression algorithms. Transform-based steganography also typically offer increased robustness to scaling and rotations or cropping, depending on the invariant properties of a particular transform. Spread-spectrum techniques and redundant encoding of the message can be employed in situations where robustness is critical. The message can be thought of as a narrowband signal encoded in a larger frequency band (the cover). By spreading the energy of the embedded message across many frequency bands (such as by frequency hopping) the energy at any particular frequency band is reduced. Therefore the message becomes more difficult to detect or modify without damaging the cover. Error correcting coding can be applied to the message during embedding to allow recovery even when some areas of the stego-image may be damaged or altered.

3.2.3 Perceptual Masking Systems

Recently, a great deal of research has been reported in expanding the hiding capacity and robustness of steganographic techniques by exploiting the properties of the human visual system [8]. The development of accurate human vision models facilitates the design and development of perceptual masking hiding systems [4]. Steganographic techniques designed to be robust to lossy image compression must insert the message into the cover in a manner that is perceptually significant. Techniques that attempt to embed information only in a perceptually insignificant manner, such as LSB embedding techniques, are vulnerable to having the embedded data distorted or quantized by lossy image compression. The masking properties of the human visual system allow perceptually significant embedding to be unnoticed by an observer under normal viewing conditions [4]. “Masking” refers to the phenomenon where a signal can be imperceptible to an observer in the presence of another signal (referred to as the masker.) The masking properties are the reason why it is difficult for one to find a randomly placed needle in a haystack; the needle can be in plain view to an observer (not obscured by any object) yet the observer will have great difficulty locating the needle. Masking (sometimes referred to as image-adaptive [4]) systems perform analysis of the image and use the information to determine appropriate regions to place the message data. Masking systems can also use the analysis to vary the strength (amplitude) of the embedded data based upon local image characteristics to maximize robustness. These systems can embed in either the spatial or a transform domain.

3.3 Sound/Movie File Steganography

The form of steganography that’s getting the most press recently is that of imbedding data in digital audio, video or still images. Sound/ movie require intensive processing to hide data in them as the carrier is stored in tend to be larger. In audio files, small echoes or slight delays can be included or subtle signals can be masked with sounds of higher amplitude. There are number digital audio applications that offer information hiding in various formats such as .wav, .avi, .au, and .mpeg formats. In audio files small echoes or slight delays can be included or subtle signals can be masked with sounds of higher amplitude.

3.4 Other Methods

Information can be hidden in documents by manipulating the positions of the lines or the words. When HTML files are written web browsers ignore spaces, tabs, certain characters and extra line breaks. These could be used as locations in which to hide information. Messages can be retrieved from text by taking for example the second letter of each word and using them to produce the hidden message. This is called a null cipher or open code. Information can be hidden in the layout of a document for example certain words in a piece of text can be shifted very slightly from their positions and these shifted words can then make up the hidden message. The way a language is spoken can be used to encode a message such as pauses, enunciations and throat clearing.

4. The Proposed Method for Steganography

In the last few years, digital images have been one of the most popular carriers. Secret information can be embedded within a digital image in such a way that the image looks unchanged to the human eye. When images are used as carriers, it requires alterations of the carrier media properties which may introduce some form of degradation. If applied to images that degradation, at times, may be visible to the human eye and point to signatures of the steganographic methods and tools used. These signatures may actually broadcast the existence of the embedded message, thus defeating the purpose of steganography, which is hiding the existence of a message. Two aspects of attacks on steganography are detection and destruction of the hidden message. Any image can be manipulated with the intent of destroying some hidden information whether an embedded message exists or not. Detecting the existence of a hidden message will save time in the message elimination phase by processing only those images that contain hidden information. Detecting an embedded message also defeats the primary goal of steganography, that of concealing the very existence of a hidden message. Most approaches have vulnerabilities as they are not as robust as is claimed.

When the available steganographic methods become popular, steganalysts find methods to retrieve hidden or the embedded information. There are numerous tools available for automatic steganalysis. Because of this, requirement for finding new techniques is increasing rapidly. The proposed steganography technique will be using images as the covert channel.

EXIF is a standard for image files created with digital cameras and other input devices. The standard is set by the Japan Electronic Industry Development Association, and formally it is called the Digital Still Camera Image File Format Standard. Uncompressed TIFF images or compressed JPEG images are two types of EXIF files. International Press Telecommunications Council in collaboration with Newspaper Association of America has introduced a standard for comments in digital images. IPTC comments can be saved inside TIFF and JPEG files. From the image formats available, the new method will use JPEG images considering the frequent availability of JPEG images and the ability to store IPTC comments in JPEG images.

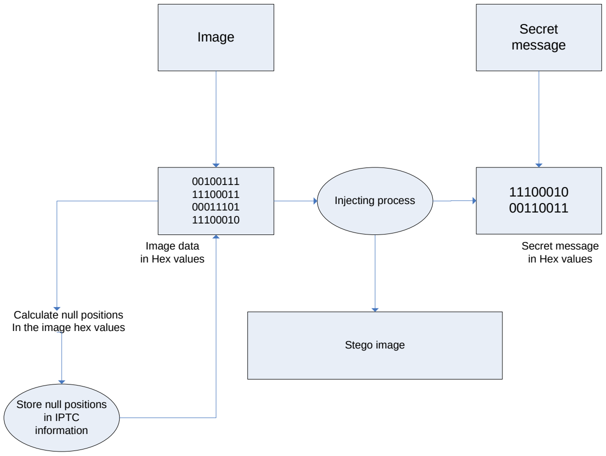


Figure 3 - Encoding process of proposed method

The above diagram shows a model of proposed steganography method. The original image will be converted into hex format and calculate the null values in the hex file. Any no of null values can be inserted into the end of the hex file format without modifying the original image data. An image used once should not be used again as the cover image. Addition of null values and secret message into the image file will change the size of the image. If the same image is used over and over again, due to different file sizes, it may draw suspicious to the eavesdroppers. After calculating the null values in image hex file, the positions of the null values will be stored in the IPTC comments section. This information will be needed when the secret message is retrieved from the image file. For the embedding process, the secret message will be converted into hex format. Then the secret message's hex values will be injected into the positions with null values in image. This process will result the "stego image". Secret message can be encrypted using a good encryption algorithm prior to injecting process.

5. Conclusion

As all the of the methods evaluated required either color reduction of the original images palette or color substitution in the stegoed image, they all had their own weaknesses as the stego-image inevitably suffered some distortion from the steganography process. In the case of color reduction based techniques there were strong tell-tail signs in the palette as well. Overall the color rearrangement technique appeared to be the most resistant to detection as long as suitable images were chosen. Techniques that attempt to maximize the message size that they can store; appear to be the least resistant to detection.

Having carried out a review of steganographic methods it has been found that images are the most widespread carrier medium used. In some methods the purpose is to minimize changes to the image and in some the purpose is to store the message in a random way so as to make it more difficult to detect. In some methods more information can be stored by using more than one bit of the colors representing the pixels. This allows more information to be stored but also results in the creation of more new colors and the need to use an image that is comprised of fewer colors to start with. In some methods the process of storing information can result in the production of new colors in the image by distorting the original image and in some methods the existing colors only are used. Some methods involve the use of a key.

These different characteristics may be used individually or combined to produce similar systems to many of those currently being used. By combining some steganographic techniques and cryptographic techniques, new and strong methods can be implemented. The aim of this study was to discuss the existing methods, their drawbacks and strengths and come up with a new improved solution to hide data. Nevertheless, if humanly imperceptible information is embedded within a cover, then humanly imperceptible alterations can be made to the cover which destroys the embedded information.

References

1. Image Steganography: Concepts and Practice by Mehdi Kharrazi1, Husrev T. Sencar, and Nasir Memon, April 2004

2. Secure Error-Free Steganography for JPEG Images by Yeuan-Kuen Lee, Ling-Hwei Chen, Department of Computer and Information Science, National Chiao Tung University, Taiwan, R.O.C, 2003.

3. An evaluation of Image Based Steganography Methods by Kevin Curran, University of Ulster Karen Bailey, Ireland, 2003.

4. Hide and Seek: An Introduction to Steganography by Niels Provos and Peter Honeyman, University of Michigan, 2003

5. A framework for evaluating the data hiding capacity of image sources by Pierre Moulin, 2002.

6. A Discussion of Covert Channels and Steganography by Mark Owens, March 2002.

7. Steganographic Techniques and their use in an Open-Systems Environment by Bret Dunbar, January 2002.

8. Analysis of LSB based Image Steganography by R. Chandramouli, Nasir Memon, 2001.

9. Steganography: Past, Present, Future by James C. Judge, 2001.

10. A Review of Data Hiding in Digital Images by Eugene T. Lin and Edward J. Delp, Video and Image Processing Laboratory (VIPER), School of Electrical and Computer Engineering, Purdue University, Indiana, 1999

The Sinhala Text to Sign Language Translator

U.L. Perera* and G.P. Seneviratne**
*University of Colombo School of Computing,
35 Reid Avenue, Colombo 7

**Dept of Computation and Intelligent Systems
University of Colombo School of Computing,
35 Reid Avenue, Colombo 7
Email : *umeshaperere@yahoo.com and **gps@ucsc.cmb.ac.lk

Abstract

The Sinhala text to Sign language translator presented in this paper has been designed to introduce a way of interpreting a Sinhala text message in Sinhala Sign language as a video sequence. This system can also be used as a training tool for deaf/mute community and other interested parties. In addition, this could be used to convey messages to deaf/mute people in the absence of a Sinhala Sign language interpreter.

The system consists of a rule-based token extractor, token analyzer and an interpreter. The output is produced as a video sequence while the input may be typed directly through the graphical user interface (GUI) or read from a text file. The technologies used for this were JDK 1.3, Quick Time for Java API and MySQL. It is a low cost system which is affordable to Sri Lankan community. The paper presents the details with practical problems faced during implementation with the solutions to them.

1. Introduction

1.1 The need for a Sinhala sign language translator

The basic need of a *Sinhala text to sign language translator* is to help deaf/mute people to keep at the front as other people, in education, acquiring knowledge and information.

Due to the incapability of most of the normal people who can speak well, to express that same idea in sign language, deaf/mute people who use sign language as their main and, often the only way of communication, get penalized in acquiring that information. The ways of getting special information for them are limited to occasions that use sign language interpreters. Since such facilities are so rare in our society, their scope of knowledge is limited to a small general area. The use of text to sign language translator can benefit such people at audiences, with a little extra work of providing a text file of the lecture.

A word in the Sinhala language has many forms. Those forms differ from its general form basically from the suffixes added. Although the deaf/mute people know basic words in Sinhala language, it is hard to make them understand the differences between those word forms. Apart from that, the teachers who have been teaching deaf and mute students for years are sometimes unaware of these different signs for different forms of the same word. Since they are using the same sign for all the forms of the same word expressing the idea of a sentence or a word in general, causes the deaf/mute students to be weak at their writing. To overcome this, the National Institute of Education (NIE) is in the process creating a new standard set of signs for Sinhala words, starting from the Grade 1 textbook. Since most of the teachers, who teach in sign language, are deaf or mute, it is a challenging task to train them. As Mr. Gunarathna, the vice principal of the Rathmalana Deaf school [2] says, it is hard for children with hearing difficulties to understand another language, such as Sinhala or English. Although they can understand something expressed in sign language, it is hard for them to use another language especially in reading and writing. They face much difficulty in understanding words in reading; also they easily forget the order of letters to use in writing a word. This is a big issue in educating deaf students, because it makes them perform extremely low in written examinations. As further suggested by Mr. Gunarathna, [2], it is important for all the teachers and people in health sector, courts, police, welfare organizations, government offices and training centers to know the sign language since those are the places most disabled people have to move about, and they face many difficulties in interacting with the people in such places. Although it is impossible for all of them to under go a systematic training program in sign language, this text to sign language translator could be useful for this to some extent.

Another purpose of this project is that, to use this translator to standardize the Sinhala sign language. This 'Sinhala text to sign language translator', included with such standard set of signs, can be used to unify the meaning of signs and define a standard way of performing a sign. This could be used by all the deaf schools and special education units to teach the children from the beginning. The advantage of using a system like this for standardizing a sign language is that, it gives a better idea of the signs over the traditional sign language dictionaries, textbooks and manuals that illustrate the signs via drawings. Presenting a sign by a drawing does not give a clear idea of presenting that accurately, especially the correct motions. Since this system is using video clips to present the signs, the correct way of presenting the signs will be easily understood even by the beginners of the sign language.

1.2 Aims & objectives

The main objective of this project is to come up with a system, which can successfully translate a text in Sinhala language in to Sinhala sign language that can achieve following goals:

- Introduce a set of morphological rules to analyze the tokens in the text and to suggest the most appropriate sign or the combination of signs and identify the combination of finger spelling signs needed to illustrate a character.
- Translate word(s) or sentence(s) in a text in to sign language, displaying relevant signs using video clips.
- Achieve a smooth flow of animation in synthesizing.
- A system to help in the process of standardizing the symbols of the Sinhala sign language.
- Introduce a way of teaching disabled students with hearing difficulties and also train their teachers.
- Provide a reference guide for Sinhala sign language.

1.3 Scope of the project

As the basic step, the Grade 1 Sinhala book was used as the reference in developing this system. It is only needed to expand the database for more words to improve the system with its vocabulary.

The font type selected for text input is 'Sandaya Plain' which is a true type font designed for the Wijesekara keyboard, which is the standard keyboard for Sinhala.

2. Literature review

2.1 Types of signs in a sign language

A sign language consists of two parts, direct signs (gestures) and finger spelling. Direct signs are associated with combination of motions. These are the symbols used to interpret a certain word or a set of words of written/spoken language in sign language. In presenting an idea using direct signs, the relative position of hands with respect to the other parts of the body, the order, the direction of motions and the area of motion are very important

A finger spelling is the method of showing a letter in sign language. In sign language, all the words do not have a specific sign related to it. Proper names and some complex words are some examples for this. In the absence of a direct sign for a word, finger spellings are used to express that. In this, finger spellings are shown in the order that word is spelled.

In expressing proper nouns and vague words, facial expressions such as eyebrow movement and mouth shapes are used. According to [3], those facial expressions and mouth shapes that accompany hand gestures are particularly important in expressing the exact meaning of a word, since they can modify the basic meaning of a hand gesture. A gesture in a sign language is equivalent to a word in a written or spoken language and finger spellings are equivalent to letters. Therefore, presentation of a sentence in the written language is, presenting a sequence of gestures or finger spellings in sign language.

According to [4], the deaf and hearing impaired population in Sri Lanka is 73343 and their main communication mode is sign language.

Sinhala sign alphabet has been designed based on single-handed American Sign alphabet. Therefore, Sri Lankan Sinhala sign language is single-handed and, most of the Sinhala finger spellings are just the same as the corresponding sign in ASL (American Sign Language). The Sinhala sign language consists of 16 vowels, 43 consonants and 6 vowel signs ('pili' symbols). These, except for 'pili' symbols, are presented using finger spellings. Since all the vowel signs are associated with an independent vowel, the relevant vowel is used in the place of vowel signs. In finger spellings, some characters are presented as still hand poses, while others have a movement of fingers (dynamic hand gestures).

2.2 Comparison between available products and this work

So far, the Sinhala sign language does not have any system developed for it. Therefore, the comparison has to be done with the sign language software designed for sign languages of other countries.

Most of the available products are sign language dictionaries to look up the signs for given words, numbers and letters or to view the finger spellings of given words. However our system is able to translate sentences, by analyzing a given text and identifying the word forms. Therefore, the functionality of the system presented here is not just a word look up, it can identify and translate number of forms of a word still having only the basic form of that word in the database.

Most of the available sign language software use virtual person as the sign language interpreter, while ‘Sinhala text to sign language translator’ uses video clips for interpretation. Although it is advantageous in the point of view of capacity and the ability to generate gestures just by passing parameters, following an interpretation done by a real person would be more interesting and understandable for users, than staring at a virtual person for a long time. This is one reason to use video clips to present signs in this work.

Another reason to use video clips in this work is to achieve accuracy and the completeness of the sign. That is, a video clip is naturally able to present a sign or a gesture in a more complete manner with all the facial expressions and lip movements, which is impossible to achieve with the use of computer-generated animations.

3. Background

3.1 Character set of the Sinhala language

The character set of the Sinhala language consists of characters (consonants) and ‘pili‘ symbols. According to the Unicode standard version 4.0 [5], Sinhala character set consists of 61 characters (consonants) and 19 dependent vowel signs (‘pili’ symbols) resulting a total of 80 symbols.

Prof. J.B. Dissanayaka states in [6] that, the Sinhala vowels are called ‘svara’ or ‘pranakshara’ in Sinhala and they are positioned at the beginning of words. The consonants are called ‘gathrakshara’ and their original nature is the ‘al’ character of that. Vowel signs are called ‘pili’ symbols and as described in [7], they are associated with the above stated vowels. Some vowels have more than one vowel sign associated with them and some vowel signs related to a vowel, changes with the consonant they are to be associated with. But the vowel ‘අ’ does not have a vowel sign associated with it; instead the ‘al’ symbol is removed from the ‘al’ character when associated with this.

Example:

Vowel	Vowel sign
අ	-
ආ	ා
ඒ	ෙ and ් or ෙ and ්
උ	ු or ු

In presenting different forms (different sounds) of a consonant, the ‘al’ character is associated with one of the vowels. In writing, this is done by placing the relevant vowel signs as suffixes and/or prefixes with the consonants.

Example: Original form of consonant ‘ක’’ is ‘ක්’. In creating different forms, the vowels like අ, ආ, ඇ, ඈ, ඉ, ඊ, උ, ඌ, එ, ඒ, ඔ, ඕ are associated with this original form. Some of those are as follows:

ක් + අ = ක
ක් + ආ = කා
ක් + එ = කෙ
ක් + ඔ = කො
ක් + උ = කු

3.2 Character set of the Sinhala sign language

The character set of the Sinhala sign language consists of 16 vowels, 43 consonants and 6 vowel signs (‘pili’ symbols). These six ‘pili’ symbols are related to special sounds and there are no ‘pili’ symbols related to the above set of 16 vowels. All the signs of the consonants present the ‘al’ character of the relevant consonant. As the same in Sinhala language, these ‘al’ characters are associated with the vowels in presenting the different forms of that consonant. Since the Sinhala sign language does not possess a set of ‘pili’ symbols, in presenting the different forms of a consonant, it is done by presenting the relevant vowel after the ‘al’ character.

Example: ක = ක් + අ
කෙ = ක් + එ

Therefore, any letter except for ‘al’ characters is always represented by two signs (two finger spellings) in sign language.

3.3 Formation of words in Sinhala language

In Sinhala language, every word has a general form and according to [8], there are three types of such general forms in Sinhala language related to nouns, verbs and adjectives.

There are different forms of a word and these indicate the state, time, gender, etc. relative to a word. They are also used to give the forms of the word such as acronym, plural, etc. These different forms of a word are produced by adding sub parts to the general form of the relevant word. As described in [9]**,** these sub parts are called 'upasarga' and they could be added to a word as suffixes or prefixes. Depending on the added 'upasarga' part, a new word is created. In verbs these upasarga parts are used to indicate the time (present tense, past tense, future tense).

Example: Consider the noun පොත (book). Its general form (nama prakurthi) is පොත් and the different forms of that noun is formed as follows:

පොත් + අ	= පොත	(the book)
පොත් + අක්	= පොතක්	(a book)
පොත් + ඉන්	= පොතින්	(from the book)
පොත් + අට	= පොතට	(in the book)

Some of the upasarga parts used to form the acronym are අ, වි, නො, නු, etc. Forming of the acronym of some words with the use of these upasarga parts is as shown below and in the case of acronyms; the upasarga part is used as suffixes.

Example:

අ + ගෞරව	= අගෞරව	(respect → disrespect)
වි + පක්‍ෂ	= විපක්‍ෂ	(support → not support)
නු + පුහුණු	= නුපුහුණු	(trained → untrained)

3.4 Formation of words in Sinhala sign language

The Sinhala sign language has a set of signs to represent words in the Sinhala language. This sign can be called as a general form and in the past, this general sign was used in place of all the forms of that word. Although this use was enough for communication, it caused deaf and hard of hearing people perform weak in writing, reading and understanding a text. As a result, in presenting a word in sign language, it is now used a way of presenting the general sign accompanied with finger spellings in order to give an idea of the form that word. For an example, in representing a Sinhala word in Sinhala sign language, if that word is accompanied with any suffixes, other than being the general word itself, the characters that forming that suffix are shown after the relevant sign of the word.

Example: Consider the same word පොත (book). Now, the general form of this word පොත, which is the most used form of that word. The presentation of the different forms of that word is as follows:

පොතක් → පොත + ක්
පොතක → පොත + ක් + අ (as ක → ක් + අ)

In presenting a verb, the sign of general form of that verb is always followed by a sign presenting the time. There are three signs to represent the past, present and future in sign language. Therefore, one of these signs is shown after showing the general sign.

Example: **Went** = sign for **go** + sign for **past tense**
Going = sign for **go** + sign for **present tense**

3.5 Behavior of words in a sentence

A sentence of the Sinhala language can be in various ways. Therefore, the position of the subject, verb and object can be in different places according to the way the sentence is used in talking/writing. In translating a Sinhala sentence to Sinhala sign language, the signs relevant to the words are presented in the same order as they appear in the Sinhala sentence.

3.6 Relationship between the Sinhala language and Sinhala sign language

Considering the relationship between the Sinhala language and the Sinhala sign language, they are closely related in most instances. But, when it comes to the character set, the character set of the Sinhala sign language is relatively small compared with that of the Sinhala language. Therefore, when it comes to a place of using finger spellings, some of the letters in the Sinhala language cannot be presented in sign language. In such cases alternatives should be used or that word may have to be skipped.

In translating a Sinhala sentence to sign language, the sign to be used for some words is decided according to the context of that word. In translating a speech to sign language, the pronunciation gives a better help to decide this. Therefore, in order to translate a text to sign language it has to be taken in to consideration that, text should be written in an agreed upon standard to indicate the actual state of the word.

In the case of verbs, the Sinhala language has a set of verbs related to a particular action and the words in this set indicate the time and gender of the action. But in sign language, although it is possible to show the time of that action, there are no signs reserved to show the gender of that verb. The gender of a verb is implied by the noun of that sentence. Therefore, the Sinhala sign language has a set of gender-less verbs.

Considering the vocabulary of the Sinhala language and Sinhala sign language, Sinhala language has a massive set of words compared with the Sinhala sign language. Due to these limitations, the Sinhala language cannot be directly mapped to the Sinhala sign language.

The process of translating Sinhala language in to Sinhala language is certainly a process of mapping a larger set to a smaller set.

4. Requirements of the system

As there are no other systems or language tools designed for Sinhala sign language, first and foremost, this system has to fulfill the requirement of being a tool to help Sinhala sign language users. It should be able to translate a given text in Sinhala into Sinhala sign language appropriately such that, it is understood by all sign language users. Translation process involves three steps. They are analysis of text, decision of sequence of signs and displaying the appropriate video clips.

4.1 Analysis of text

In the translation process, it is required to analyze the given text first. In this stage the basic form of a word should be identified correctly. In Sinhala language, words of a certain category may take different styles and it is not always possible to identify the form of a word by sticking to only one form.

Example 1: In Sinhala, the plural of certain words are derived by ending that word with the ‘al’ form of the last letter as given in the table 4.1

Singular	Plural
මල (mala)	මල් (mal)
ගස (gasa)	ගස් (gas)
පොත (potha)	පොත් (poth)

Table 4.1: Plurals ending with ‘al’ letter of the singular

But, there are some other words with ‘al’ endings that do not correspond to a plural of words resulted by its ‘al’ sign removed. ‘සල්’(sal) ‘බත්’(bath), ‘කොස්’(kos) are some examples for such words.

Example 2: In some cases words ending with a letter ‘ව’ (va) or ‘ය’ (ya), the plural is derived by removing the last letter ‘ව’ or ‘ය’ character, as shown in the table 4.2.

Singular	Plural
පුටුව (putuva)	පුටු (putu)
කටුව (katuva)	කටු (katu)
කළය (kalaya)	කළ (kala)
කතන්දරය (kathand`araya)	කතන්දර (kathand`ara)

Table 4.2: Plurals obtained by removing the last ‘ව’(va) or ‘ය’(ya) in the singular

But there are certain words ending with the letter ‘ව’ or ‘ය’, which the removal of the last letter does not results the plural of that. ‘උදය’(ud`aya), ‘කය’(kaya), ‘දුව’(d`uva) are such words. Therefore, in identifying a word, all possible cases have to be considered.

The accuracy of the translation process totally depends on text analysis part. Therefore, it is important to identify the correct form of a word in the text. For this, a certain word should pass through several language rules until its proper basic form is identified. When basic form of a word could not be identified (in cases of proper names and complex words that do not have signs defined for them), it should be identified as a word to be presented with finger spellings.

When it is decided to present a word using finger spellings in the text analysis stage, the correct sequence of finger spelling video clips should be selected. Some sounds of the Sinhala written language is presented by the related character and a combination of vowel symbols. According to [5], there are 18 vowel sounds and they are represented with the use of 14 symbols (්, ෘ, ා, ැ, ෑ, ි, ී, ු, ූ, ෟ, ෲ, ෙ, ෝ, ‿). It is possible for at most of three vowel symbols to be associated with a character to represent a sound. The same vowel symbols may be included in different combinations of sounds depending on the context. Therefore, to identify the vowel sound correctly, it is important to consider the entire set of vowel symbols associated with the character. For this, the range of symbols corresponding to a single sound should be correctly identified. This is a special case compared with a language like English, in which each sound is presented by a single symbol.

Example: The following sounds given in table 4.3 are generated by associating the vowel symbols, ෙ, ා, ‿ with the consonant ක.

Vowel symbol(s)	Combination	Resulted character	Sound	Total no. of symbols in range
ෙ	ෙ + ක	කෙ	Ke	2
ා	ක + ා	කා	Ka	2
ෙ, ා	ෙ + ක + ා	කො	Ko	3
ෙ, ‿	ෙ + ක +‿	ක්‍රෙ	Kre	3
ෙ ‿Y, ා	ෙ + ක +‿ + ා	ක්‍රො	kro	4

Table 4.3: Ranges of symbols correspond to different sounds

4.2 Selecting video clips

After the identification of words, next step is the identification of corresponding video clips. There may be more than one clip related to a word depending on the form it is present. If a word is in its basic form, it is sufficient to present that using one video clip. But, if it is in another form like plural of a noun, a verb in past tense, etc., the relevant clip may be followed by one or more other signs.

Example: The following words are interpreted in sign language as combinations of signs as follows.

නරියාගේ (nariyaage_) = නරියා (nariyaa) + ග් (g)+ ඒ (e_)

මලක් (malak) = මල (mala) + ක් (k)

In the above example, the order of signs is related to the order of characters in the word. But, in presenting verbs, these signs, depend on the time of action.

Example: word(verb) = basic sign + sign to indicate past/present/future

i.e. ran = run + past

In Sinhala sign language, all the characters (sounds) except for 'al' characters in Sinhala written language are presented by two or more finger spellings. Other signs except the 'al' letter sign, are relevant to the vowel sound resulted by the above stated vowel symbols associated with the basic character. The necessary finger spellings and vowel signs are decided in the text analysis phase.

Example: ක් = ක් ('al' sound)

කූ = ක් + ට

ක්‍රා = ක් + ්‍ර + ආ

4.3 Formation of sequence of signs

After selecting the appropriate video clips to present the given text, it is necessary to play these clips in a smooth manner. The presentation of signs through video clips should be as smooth as possible, in order to provide a sense of natural sign language presentation.

4.4 Understandability and usability of the system

The system should be simple enough to be understood by the normal users as well as disabled people. Since the reading ability of some disabled people is considerably weak, it is required to design the interfaces in such a way that, they face no difficulty in using the system. It is important to provide means for them to understand the functionality of the each interface item (buttons, menus, etc.) straight away, rather than understanding them by reading all the words in the interface.

5. Design of the system

The design of the system can be divided mainly into four categories as below:

i. Design of the text analysis rules
ii. Design of the database
iii. Design of the video clips
iv. Design of the user interface

5.1 Design of the text analysis rules

The text analysis rules fall into three main categories as, numerical analysis, character analysis and word (token) analysis. The Figure 5.1 depicts the design of the text analyzing process.

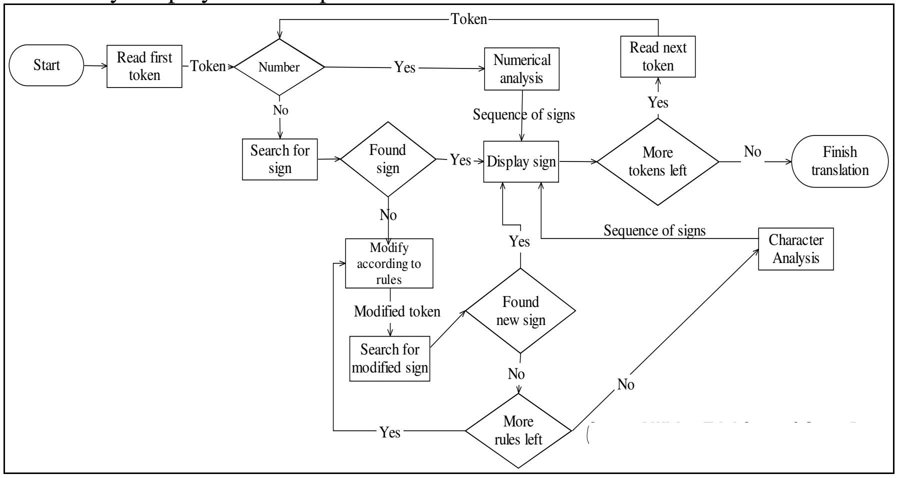


Figure 5.1: Flow chart of the Text analysis

5.1.1 Numerical analysis. In the numerical analysis process, the number of ones, tens, hundreds, etc. the considered number is consisted of is calculated. The sequence of signs to present is decided by this.

5.1.2 Word analysis rules. In this process, when a word (token) is passed through the word analysis process, a search is performed to find out whether there is a sign defined for that word. If the corresponding sign was found, relevant video clip to be displayed is decided and further analysis on that word is terminated. In the absence of a sign defined for that word, the analysis rules are applied on the word. In this, the token is modified according to the rules and at each point, a search is carried out to find the appropriate sign for the modified token. These rules are written considering the end of the word. The end of the word depends on the type of the word (noun, verb, adjective).

Some of the possible endings of words, suggested basic sign and additional signs by the word analysis rules are as shown the table 5.1.

Word	Type	Basic sign	Additional signs
කරන්නේ	Verb	කරන්න	ඒ
කරා	Verb	කරන්න	(past)
කරයි	Verb	කරන්න	ය් + ඉ / (future)
ලස්සනයි	Adjective	ලස්සන	ය් + ඉ
මලක්	Noun	මල	ක්
මල්	Noun	මල	(plural)
අම්මාගේ	Noun	අම්මා	ගේ (belongs to)

Table 5.1: Basic forms and additional signs to represent a word

5.1.3 Character analysis rules. Character analysis of a token is done by examining the characters in the word sequentially. In this, the range of characters related with a sound is identified. According to the rules written for character analysis, there are two possible cases for a word to start; a word starting with an independent vowel or with another character that is not an independent vowel (a consonant or dependent vowel signs)

5.1.4 Words starting with an independent vowel. Although there are 16 independent vowel signs in the alphabet of the Sinhala sign language, considering a text, they can be started from 7 character symbols. The vowel sign represented by them is decided by the subsequent vowel sign. The table 5.2 shows the commonly used such independent vowels and their associated vowel signs.

Symbol of the position		Resulting sound	Examples
1st	2nd		
අ	Not a vowel sign	a	අම්මා
අ	ා	aa	ආවා
අ	ැ	a_	ඇයි
අ	ෑ	a__	ඈත
ඉ	Not a vowel sign	i	ඉන්නවා
ඊ	Not a vowel sign	ee	ඊයේ
උ	Not a vowel sign	u	උඩ
උ	ෟ	uu	ඌරා
එ	Not a vowel sign	e	එන්න
එ	්	ei	ඒවා
ඔ	Not a vowel sign	o	ඔව්
ඔ	්	o_	ඕනෑ

Table 5.2: Independent vowels and associated vowel signs at the beginning of a word

5.1.5 Words starting with a character other than an independent vowel. The character analysis performed for this case is same for all other characters placed within a word. The objective of this analysis is to identify the range of symbols correspond to this sound and to determine the letter and related vowel associated with that sound. For simplicity of analysis rules, this category is divided in to two parts as,

- Symbol range started with a 'kombuva' (ෙ) symbol
- Symbol range started with a consonant

5.1.6 Sounds having its symbol range started with a 'kombuva' (ෙ). In this case, the 'kombuva' (ෙ) is generally followed by a consonant then by zero or more dependent vowel signs, except for two special cases. The rules for identifying the sound are designed by considering at most of three subsequent symbols after the 'kombuva' (ෙ). There are 10 possible cases in this category, considered in designing the text analysis rules and some of them (considering the letter 'ක') are shown in the table 5.3.

Symbols at different positions in the range				Combination	Resulting sound	Examples
1st	2nd	3rd	4th			
ෙ	ෙ	ක		ෙ + ෙ + ක	කෛ (kai)	කෛරාටික
ෙ	ක			ෙ + ක	කෙ (ke)	කෙසෙල්
ෙ	ක	්		ෙ + ක + ්	කේ (ke_)	කේලම
ෙ	ක	්‍ර	ා්	ෙ + ක + ්‍ර + ා්	ක්‍රෝ (kro_)	ක්‍රෝධය

Table 5.3: Consonants and associated vowel signs for symbol ranges starting with 'ෙ'

5.1.7 Sounds having its symbol range started with a consonant. In this category, the symbol range relevant to a sound is started with a consonant and it may be followed by zero or more (up to two) vowel signs. Therefore, the number of symbols considered to analyze a sound may vary between one and three. There are 17 such possible cases in this category) that considered in designing rules, and some of them (considering the letter 'ක' are given in the table 5.4.

Symbols at different positions in the range			Combination	Resulting sound	Examples
1st	2nd	3rd			
ක			ක	ක (ka)	කන්ත
ක	්‍ර		ක + ්‍ර	ක්‍ර (kra)	ක්‍රමය
ක	්		ක + ්	ක් (k)	අක්කා
ක	ා		ක + ා	කා (kaa)	කාලය

Table 5.4: Consonants and associated vowel signs for symbol ranges starting with a consonant

5.2 Design of the database.

In designing the database, information on clips corresponding to numbers, letters and words were kept in separate tables. This is done to improve the efficiency of the searches in each analysis stage. In each table paths of the video clips are stored with the other relevant details.

5.3 Design of the video clips

In designing video clips, the upper body, from hip area to head is used to display signs.

Apart from that basic requirement, steps should be taken in the design of video clips, to achieve the smooth flow in displaying sequences of video clips. To preserve the consistent look in the presentation of video clips, all these video clips should have equal properties. These can be grouped in to following categories.

- Appearance of the interpreter
- Light and background
- Position of the interpreter

5.4 Design of the user interfaces

The interface of the system is designed as simple as possible to avoid it being too complex for disabled users. As help for such people, the objects on the interface are associated with suitable signs.

6. Implementation of the system

The implementation of the system can be mainly divided in to parsing text to the text analysis process, text analysis and selection of video clips and displaying of video clips.

6.1 Implementation of analysis rules

The text to be translated is passed to the text analysis process as tokens. In this, either the whole text or a selected section of it could be passed to the text analysis process. Before passing, a prepossessing is performed on the token to remove unwanted symbols in the token. The text analysis process consists of three parts, numerical analysis, token (word) analysis and character analysis. The numerical analysis is done performing some arithmetic functions on the value of the integer given by the token. In character analysis, symbols are read from the beginning of the token and the range symbols to consider corresponding to a sound is identified. The table 6.1 provides some of the word forms identified by the token analysis within the scope of this work. The required video clip(s) to present a token is identified by these three analyses.

End of token	Possible forms	Example
‘al’ symbol	plural (nouns)	ගස - ගස්
	‘ක්’ ending (passive nouns / adjectives)	ගස - ගසක්
	‘එක්’ ending (active nouns)	ළමයා - ළමයෙක්
‘ක’	Nouns	ගස - ගසක
‘යි’	Adjectives	නරක - නරකයි
	and (nouns)	අම්මයි මමයි
	future tense (verbs)	කරන්න - කරයි
‘ත්’	Nouns	මම - මමත්
‘ඉන්’	Prepositions/adjectives	උඩ - උඩින්
‘ගේ’	belongs to (nouns)	නංගි - නංගිගේ
‘ගෙන්’	From	අම්මා - අම්මාගෙන්
‘මි’	I’m doing (verbs)	කරන්න - කරමි
‘ලා’	past tense (verbs)	කරන්න - කරලා
	plural (nouns)	අක්කා - අක්කලා
‘ඕ’	plural (nouns)	කුරුල්ලා - කුරුල්ලෝ

Table 6.1: Some of the word forms identified by the token analyzer

6.2 The database

The database of the system is included the numbers necessary to present the values up to 10000000 (10^7), Sinhala character set and 518 words in ‘Sandaya’ font. The words database contains all the words that are necessary to translate the lessons in the grade 1 Sinhala book and around 250 extra words that can be used in day-to-day speaking. In creating the word database, generally the basic form of a word was stored in the table, except for the past tense of verbs. Past tense of a verb may vary from its basic form without having a relation to it.

Example:

General form	**Past tense**
එන්න	ආවා
කන්න	කෑවා
යන්න	ගියා
කියන්න	කිව්වා
තියන්න	තිබ්බා / තැබුවා
බොන්න	බිව්වා
අහන්න	ඇහුවා

Since it is impossible to define a simple set of rules to identify the past tense of a verb based on the basic of it, past tense of a verb was inserted to the word database, in addition to its basic form.

6.3 Problems encountered in the implementation

In developing the ‘Sinhala Text to Sign Language Translator’, most of the problems encountered were related with Sinhala alphabet and the keyboard. Since the character set of the Sinhala language has a large number of symbols, special keys of the keyboard are used to type some of those characters. ‘!’, ‘@’, ‘%’, ‘&’, ‘?’ are some of such keys and they are always used to present vowel signs in Sinhala. In the text analysis phase, a token is searched though the database by passing that string to a SQL statement. When it comes to places where that string contains special characters like ‘%’, the result of the SQL statement deviates from the desired. Hence, such characters in a string should be handled before passing to the SQL statement. The approach taken to handle this was to replace the characters in the string with characters that are present in the normal keyboard but generally not used in simple Sinhala text typing. The character ‘ත්‍’ is one of that. These are contracted characters and are called ‘බැඳි අකුරු’ in Sinhala. They are always incorporated with some other characters to present a meaningful letter and never used alone. In general, there is an optional way of presenting these characters with the use of other two normal characters, as shown in the example.

Example: ත්‍ව = ත් + ව
ක්‍ෂ = ක් + ෂ
ත්‍ධ = ත් + ධ

Therefore, words containing such characters could always be written in the optional way in the text. Another fact is that the Sinhala finger spellings do not contain signs relevant to contracted characters. Therefore, the above approach does not affect the contents in the text or the translation process when in it comes to presentation of words using finger spellings.

Another problem faced during this work is the use of homonyms in the language. The desired meaning of such words is determined by the context it is used. In general, interpreters determine this with their prior knowledge, experience in using the language and from physical factors like the pronunciation. Following sentences exemplifies the use of such homonyms that may lead to ambiguity.

Example: The word 'කඩන්න' has several meanings when its neighboring words are considered.

- කඩන්න : break
- මල් / ගෙඩි කඩන්න : pluck
- ලිපිය කඩන්න : open

Therefore, it is necessary to have an agreed upon way of using words. For an instance, 'නෙලන්න' and 'අරින්න' could have been used to mean pluck and open respectively.

Since the system does not incorporate an expert system, to determine the context, it is necessary to use the most appropriate words in an agreed upon way avoiding homonyms. Therefore, the approach used to handle this problem was the introduction of standards to the input text.

The next problem occurred in text analysis is the presence of some words that could be broken in to two words, that are not necessarily related to the combined word. There are no hard and fast rules about such words and in writing some write them as two words while some write them as a one word. The meaning it implies is determined by the context.

Example: The word 'තණකොළ පෙත්තා' which is found in the grade 1 Sinhala book which was used in this project. Here, those two words taken together give the intended meaning of 'grasshopper'. If they were considered as two different words, 'තණකොළ' means grass and 'පෙත්තා' is a name generally used call pet parrots.

Therefore, to handle situations as above, it is necessary to stick to the convention of writing such words as a one whole word.

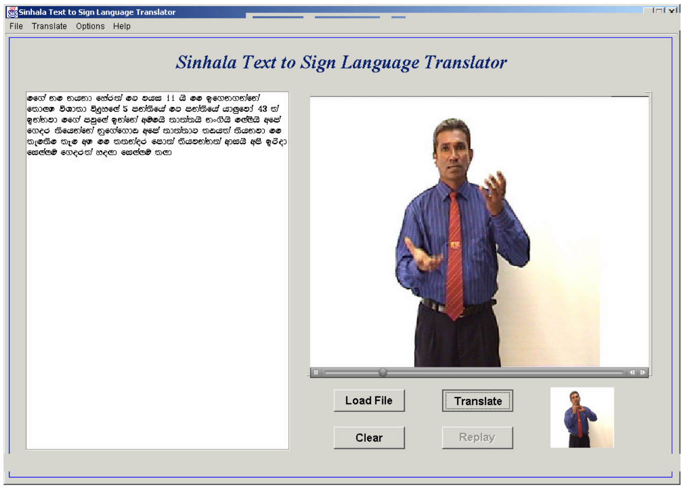
7. Testing and evaluation

Testing of a software system is performed basically to ensure that the system meets its requirements (validation) and to ensure that the system works correctly according to the requirements (verification)

As unit testing, the functionality of each part was tested while developing. The functionality of the numerical analyzer was tested by passing different natural numbers in the range from 0 to 2147483647. This is the range of positive integers allowed to be stored in 32 bit integer variables. The character analyzer was tested by parsing different string tokens through it and obtaining the resulting sequence of consonants and vowels. In the phase of writing text analysis rules, the result of each rule applied on suitable words was tested.

After performing unit testing in each section, the functionality of the system was tested using sentences, where the three numerical, token and character analyzers are combined to perform text analysis. For these sentences taken from the grade 1 Sinhala book was used. This was the integration testing performed on the system.

Completing the verification of text analysis and translation processes, the resulting sequence of video clips of the system was compared with a natural sign language translation process. This was done by comparing the outcome of the system with a translation done by an interpreter for the same sentences. Figure 7.1 is a screen shot taken in this translation process. And the video frame appear in this correspond to an instance of the sign 'කඩය' (boutique) which is a dynamic gesture.


Sinhala Text to Sign Language Translator
File
Translate
Options
Help
Sinhala Text to Sign Language Translator
Load File
Translate
Clear
Replay

මගේ නම නයනා හේරත් මට වයස 11 යි මම ඉගෙනගන්නේ
කොළඹ විශාකා විදුහලේ 5 පන්තියේ මට පන්තියේ යාලුවෝ 43 ක්
ඉන්නවා මගේ පවුලේ ඉන්නේ අම්මයි තාත්තයි නංගියි මල්ලියි අපේ
ගෙදර තියෙන්නේ නුගේගොඩ අපේ තාත්තාට කඩයක් තියනවා මම
කැමතිම කෑම අඹ මම කතන්දර පොත් කියවන්නත් ආසයි අපි ඉරිදා
සෙල්ලම් ගෙදරක් හදලා සෙල්ලම් කළා

Figure 7.1: A screenshot from translation process ('කඩය')

7.1 Observations on test Results

The analysis of text using the numerical analyzer and the character analyzer resulted expected results. But there were some deviations in the results obtained from the token analyzer compared to the expected results. The main reason for this was the complexness in the Sinhala language. The major points where the token analyzer resulted with unexpected translations were homonyms, contracted words and some verb forms deviated from its basic form in an extraordinary manner. They can be further described as follows:

In the word database, each record contains a word and the appropriate sign of that. Since a word can have only one sign associated with it, when it comes to homonyms the sign suggested by the token analysis may get deviated from what was expected. The translation process always presents the same sign for all the homonym words.

Example: The word 'හරි' is used to express the idea 'very' most of the times. But in some sentences it may be used to mean 'correct', 'surprising', 'or', 'ok', etc. considering the following sentences, the system interprets the word 'හරි' as 'very' in all the above cases.

මට හරි බඩගිනියි	- I'm very hungry
ඔබ හරි	- You are correct
හරි වැඩක් වුනා	-Something surprising happened
අම්මා හරි	- Mother or father will come
තාත්තා හරි එයි	
හරි මම එන්නම්	- Ok, I'll come

'පිට', 'කටුව', 'කඩන්න', 'ගහන්න' are some of such words. Wherever, these words are used, only one sign is presented in the translation process.

The words 'දුවයි' and 'දුවලා' are two words that behaved in an unexpected manner in some situations. Considering the word 'දුවයි', it always gave the meaning 'daughter'. That is, 'අක්කයි මල්ලියි දුවයි' is always interpreted as 'sister brother and daughter'. But, it can be used mean 'sister and brother are running'. The reason for this is, in analyzing 'යි' (yi) ended words, the rule to identifying such ended words meaning 'and/also', comes before the rule to identify 'යි' (yi) endings meaning present tense of a verb in the token analysis process. Considering the word 'දුවලා', the system always translates this as 'ran'. That is, a sentence like 'දුවලා එන්න' will be always interpreted as, 'come running', although it could be used to mean 'come daughters'. The reason for this is also as described in the previous case that, the rule to identifying 'ලා' (la) ended verbs is written before the rule for identifying 'ලා' (la) ended plurals.

To handle the above stated issues in translation process, it is necessary to identify the context of a word. Therefore, the solution for this will be associating the system with an intelligent system.

'ඉන්නවා' and 'ගන්නවා' are two special words that could not be identified with the common rules written to identify the present tense of a verb, based on their basic forms 'ඉන්න' and 'ගන්න'. Considering other verbs like 'දුවන්න', 'නටන්න', 'මරන්න', 'ගහන්න', etc. their present tense takes a common type like 'දුවනවා', 'නටනවා', 'මරනවා', 'ගහනවා'. Therefore, the basic form of those present tense verbs can be determined by the following modification.

Verb – 'න් න' + 'වා'

Example: දුවන්න - 'න් න' → දුවන + 'වා' → දුවනවා
නටන්න - 'න් න' → 'නටන' + 'වා' → 'නටනවා'

But, according to the above rules, word 'ඉන්න's present tense would be suggested as 'ඉනවා' while 'ගන්න's present tense is suggested as 'ගනවා'. And the words 'ඉන්නවා' and 'ගන්නවා' would be presented in finger spellings. This issue can be avoided by inserting above two words explicitly to the database.

In the comparison of the system output with a natural sign language translation process, following facts were observed. Generally, sign language has many signs or gestures related with the same meaning. In the normal translation process, the most suitable sign to represent a word is decided by the interpreter at the translation time. This is done with the skill and the experience of the interpreter to identify the context of the sentence he is interpreting. Therefore, some of the signs he used in the natural translation process are different from the sign presented by the system.

Another observation is, in the natural translation process the duration that a sign is displayed vary from time to time. For an example, presenting of finger spellings after a gesture to represent suffixes is done relatively fast than it is done in presenting a proper noun. But, since this translator uses the same set of video clips in both places, the speed of presenting a sign remains the same. Other than this, due to shortcomings in video recording like change in light conditions, the interpreter being moved from the center of the screen and small changes in position of hands in the beginning and the end of the video clip some discontinuities occurred in the sequence of video clips outputted by the system.

8. Conclusion

Considering the aims and objectives of the 'Sinhala Text to Sign Language Translator' and the work done in developing this system, following conclusions can be made.

8.1 Accomplishment of aims and objectives

The main aim of this work was to introduce a way of translating a Sinhala text in to Sinhala sign language. This was achieved by analyzing the text using a set of morphological rules that perform analysis on one token at a time. These text analysis rules are able to suggest the appropriate sign or the combination of signs to interpret that text. The signs, including gestures and finger spellings are presented with video clips and it was aimed to achieve a smooth flow of these to provide an impression of a natural sign language interpretation process. These aims and objectives were generally achieved and beyond the aims and objectives presented for this project, a real-time translation feature is also implemented in this work.

A main difference of this system compared to a natural sign language translation process is that, it does not perform a summarization on the text to be translated. When it comes to normal sign language interpretation of occasions like speech or news that needs the interpretation to be really fast, most sign language interpreters summarize the sentence and present only a small number of signs that is sufficient to give the basic meaning of the sentence. Since this system is developed to emphasize even the minor differences in words and to present all the words in a sentence, the system does not perform such a summarizing process.

The system also achieves the aims and objectives of introducing a teaching tool for deaf students especially in language education, self-practicing tool to for Sinhala sign language that could easily be used by students, teachers and other interested parties, a way of acquiring knowledge for deaf users with reading difficulties. Apart from that, this system has the ability to standardize the Sinhala sign language by using standard signs for the video clips.

8.2 Future enhancements

As a sign language translation system, a voice to sign language translation system could provide better service by automating the task of a human sign language interpreter. Such a system may consist of two major parts, voice to text converter, and text to sign language translator. As the current project is able to perform the work of latter part, one can incorporate a voice to Sinhala text converter to this system to produce a voice to sign language translator.

As this 'Sinhala Text to Sign Language Translator' is not incorporated with an intelligent system that is able to determine the context and decide the meaning of a word depending on that, some translations may produce results deviated from expected idea. Therefore, incorporating the current system with an intelligent system to perform context prediction would be an improvement that can increase the accuracy of the system.

In any sign language, the signs present a general idea of the underlying word. Therefore, there is a possibility for one sign language user of one country to understand what the other sign language user from another country is expressing. But, when it comes to finger spellings, it is almost impossible for sign language users from different countries to understand each other, basically due to the differences in alphabets in different countries. If it is possible to use a standard way to present proper nouns that are presented by finger spellings, a system like this could be used at universal level to express something written in Sinhala to deaf person of any country. One solution for this is the use of ASL signs, which is considered as an international language among sign language users. But, in enhancing this system to such state, it is important to avoid presenting of finger spellings after direct signs to represent the form of the word. For an example, signs for words like 'ගස', 'ගසක්', 'ගසක', 'ගසේ', etc. should be equal. Only the proper nouns like names, cities should be presented with finger spellings. In this process, it is necessary to map Sinhala sounds to English letters in someway like transliteration.

As described in the previous sections, this Sinhala text to sign language translator is designed for the 'Sandaya' font that is designed according to the Wijesekara keyboard layout. Since existing Sinhala fonts are not designed to be consistent with each other, this system will not be compatible with the other font types, except for the ones designed for the same keyboard layout. If it is possible to incorporate this system with a character mapping system that could map each font type to the 'Sandaya' font, this system will be capable of using with any font type.

Another stream of using this system is to use this as a plug in for web pages, so that the disabled people who are more comfortable with the sign language than the text can take advantage of this and get information as fast as the other people do. And also, one can introduce a new way of chatting using this.

The video clips used in this system are, .mov files. Although they results a better quality video presentation, they take a large capacity of the disk in storing. Therefore, a better compression method that does not affect the quality of them could be used in future expansions.

More information on this work including a detailed literature survey, design, implementation, discussion sections is available in [1].

References

[1] Umesha Perera, “The Sinhala text to Sign language translator”, Dissertation submitted for the Special degree in Computer Science, UCSC, 2004

[2] Interview with Mr. U. K. D. Gunarathna, The former vice Principal, Deaf School, Rathmalana.

[3] Synthesis and presentation of the polish sign language gestures. (Piotr Fabian, Jarosław Francik – Gliwice, Silesian University of Technology, Institute of Computer Science)

[4] Depatment of Census and Statistics, Census of Population and Housing – 2001, Information on Disabled Persons

[5] Sinhala Range: 0D80–0DFF
Available at:http://www.unicode.org/charts/PDF/U0D80.pdf

[6] Prof. J. B. Dissanayaka, Basaka mahima 1 (බසක මහිම 1 - සිංහල හෝඩිය)

[7] Prof. J. B. Dissanayaka, Basaka mahima 2 (බසක මහිම 2 - අකුරු හා පිලි)

[8] Prof. J. B. Dissanayaka, Basaka mahima 5 (බසක මහිම 5 - පද සාධනය)

[9] Prof. J. B. Dissanayaka, Basaka mahima 7 (බසක මහිම 7 - උපසර්ග)

Adapting Sinhala Language Facilities Into Database Systems

J.S. Goonetillake, K. P. Hewagamage and S.Perera
University of Colombo School of Computing (UCSC),
Colombo, Sri Lanka
Tel:94-1-2581245-47
e-mail: jsg@ucsc.cmb.ac.lk and kph@ucsc.cmb.ac.lk

Abstract

Though the majority of Sri Lankans are communicating in Sinhala, the lack of software that is based on the native language, Sinahala has lead to a tendency of users accepting any conventional software in English. The English proficiency in Sri Lanka is not very sound. Particularly, the government sector organizations can be considered as the organizations where there are considerably high amount of employees (users) with poor English as well as computer literacy are working. Due to the absence of a natural language interface in accessing database, most non-technical business users also have to spend non-productive time learning a database manipulation language. This paper presents a user friendly Natural Language Interface to a native language database which enables non-technical users to directly access a database using native language phrases (spoken or written). Most significantly, it enables to retrieve data in native language form, without requiring any prior programming language knowledge or English knowledge from the user. The feasibility of this concept is proved through an implemented prototype which addresses the above-mentioned problems considering a database at a Divisional Secretariat Office.

1. Introduction

To get the required information at the correct time, employees must be able to access the database, which would normally require them to have a knowledge and experience in a data manipulation programming language. These languages were designed and developed for use by programmers and selected users from the end-user community, but not for the typical business users. Therefore typical business users end up spending non-productive time trying to learn programming languages rather than concentrating on business affairs. Since these languages are based on English language syntax it would be rather difficult to grasp easily particularly due to the poor English language proficiency in Sri Lanka. It is always observed that a community would prefer their native language for the convenience in communication. Therefore computer operations in native languages bear a paramount importance.

Normally English characters are used to store the text data in text fields in currently available databases. When a manual system is transformed into a computerized information system (Change over) all the operational data will be represented in English language format. Such a representation may cause efficiency problem with respect to a poor English language proficiency and Erroneous Interpretation and Implementation Problems. Storing the data in databases by using native languages can solve these problems. The researches have been done with regard to possibilities in storing data in native language databases. But there is no specific tool for retrieval of data, which were stored in native languages. Currently, the existing standard SQL is used for manipulating the data, which were stored in native languages. Introduction of a new Sinhala Natural Language Interface for databases that is derived from a native language can easily eliminate these problems.

In this paper, authors present an interface tool as well as a mechanism for adapting Sinhala language facilities into database systems. Hence, the logical design of the system which includes SQL query generator, SQL error handling and validation and SQL parser for Sinhala language, are presented.

The paper is organized as follows. Section two describes general mechanisms in accessing database systems such as structured query language (SQL) and natural languages based interfaces. Section three, presents the logical design of the system which includes the design of Sinhala natural language interface (SNLI). Section four covers the implementation and testing of the tool which facilitates SLNI. Finally the section five, discusses the conclusion and possible future work.

2. Database Accessing Mechanisms

2.1 Structured Query Language (SQL)

SQL contains facilities for defining, manipulating, and controlling relational databases. Not only has SQL become the “de facto standard”, but it is also the official standard in relational databases. SQL facilitates the use of queries which can take the form of a quick ad hoc inquiry directly from the keyboard, or it can be included in an application, which uses the database to make routine inquiries such as reports, mailing lists and updates. Furthermore, it provides facility to join relations within a database and thus produce outputs containing attributes from several different tables [Date,1995].

2.2 Natural Language Interface for A Database (NLIDB)

This is a practical application for Natural Language Processing. Natural language is one of many ‘interface styles’ (or ‘interaction modalities’) that can be used in the dialog between a human user and a computer. There is a significant appeal in being able to address a machine and direct its operations by using the same language we use in everyday human-to-human interaction [Androutsopoulso et. al 1994]. LIFER (Hendrie 1978) and INTELLECT (Harris 1977) were some of the early system to be developed. However, the degree of ambiguity in natural language considered too extreme for it ever to be used effectively as an interface style. In order to implement a working natural language system one must usually restrict it to cover only a limited subset of the vocabulary and syntax of a full natural language. This allows ambiguity to be reduced and processing time to be kept within reasonable bounds [Bharati et. al.,1995]. In order to still be considered a natural language interface, most of the positive traits of a general natural language would have to be maintained. To retain the properties of ease of use and ease of remembering, the limitations of the system must somehow be conveyed to the user without requiring them to learn the rules explicitly.

3. Analysis and Design

The system allows users to enter their requirements such as data insertion, deletion and modification as Sinhala sentence (Spoken/Written). The Sinhala sentence could be in any spoken or written language format, familiar to the user. An error message has to be displayed when the Sinhala Sentence contains an error in Table name, Column name or when the requested Information unavailable. The error messages need to be recognized more clearly and the user should be able to rectify the specific statement rather than writing the sentence again. The logical design of the system which is proposed to satisfy these requirements is depicted in figure 1.

3.1 Logical Design of the System

3.1.1 Design of the SQL Query Generator

The SQL query generator’s main responsibility is to generate the relevant SQL query from the native language statement. The query generation is a mandatory requirement, which enable to build up the SQL query for data retrieval from the database. The query generation will be carried out as presented in figure 2.

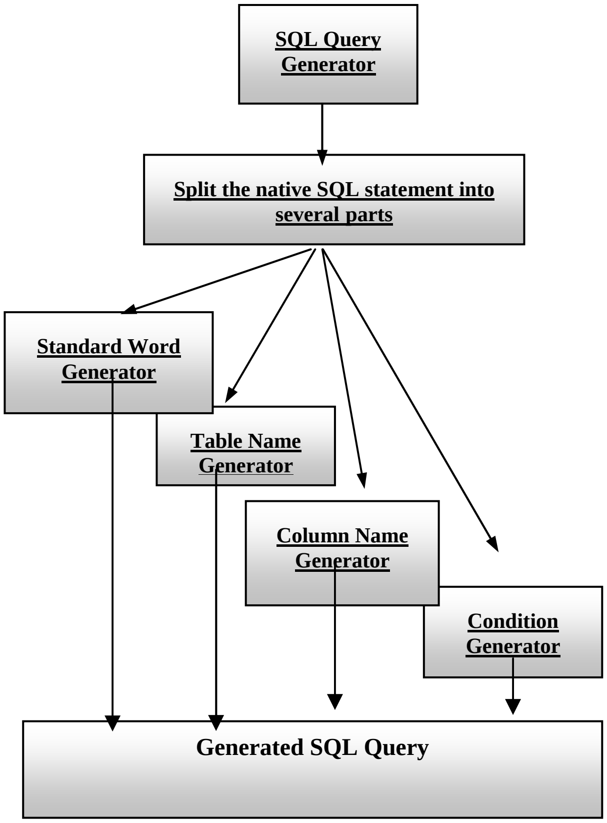


Figure 2 - Steps involved in the SQL query generation

Standard Word Generator is responsible for generating the standard SQL words out of the native query statement. In the process of standard word generation, ‘□□□□□’ table is accessed in order to find out the existence of relevant SQL word. (e.g. SELECT, DELETE, UPDATE – figure 3(a)). Table Name Generator concentrates on the generation of table names from the native query statement. In the process of table name generation, □□□□□□□ table is accessed in order to find out the existence of relevant table name (figure 3(b)).

Column Name Generator generates column names from the native query statement. In the process of column name generation, □□□□□□□ table is accessed in order to find out the existence of relevant column name (figure 3(c)). Condition Generator generates any condition in the native SQL statement that should be generated too. In the process of condition generation, □□□□□□□□□ table is accessed and by means of the foreign key, get the condition from the □□□□□□□□ table.

3.1.2 The SQL Validation and Error handling

The SQL validation is taking place to validate the generated SQL query. The error handling is also taking place for displaying the relevant error massage if the specific validation criteria are getting unsuccessful. The error messages are displayed in Sinhala and the error of the particular statement can be rectified during the execution time.

3.1.3 SQL Parser

A parser is a program, usually part of a compiler, that receives input in the form of sequential source program instructions, interactive online commands, mark-up tags, or some other defined interface and breaks them up into parts (for example, the nouns (objects), verbs (methods), and their attributes or options) that can then be manage by other programming (for example, other components in a compiler) [Peter, 1999]. A parser may also check to see that all input has been provided that is necessary.

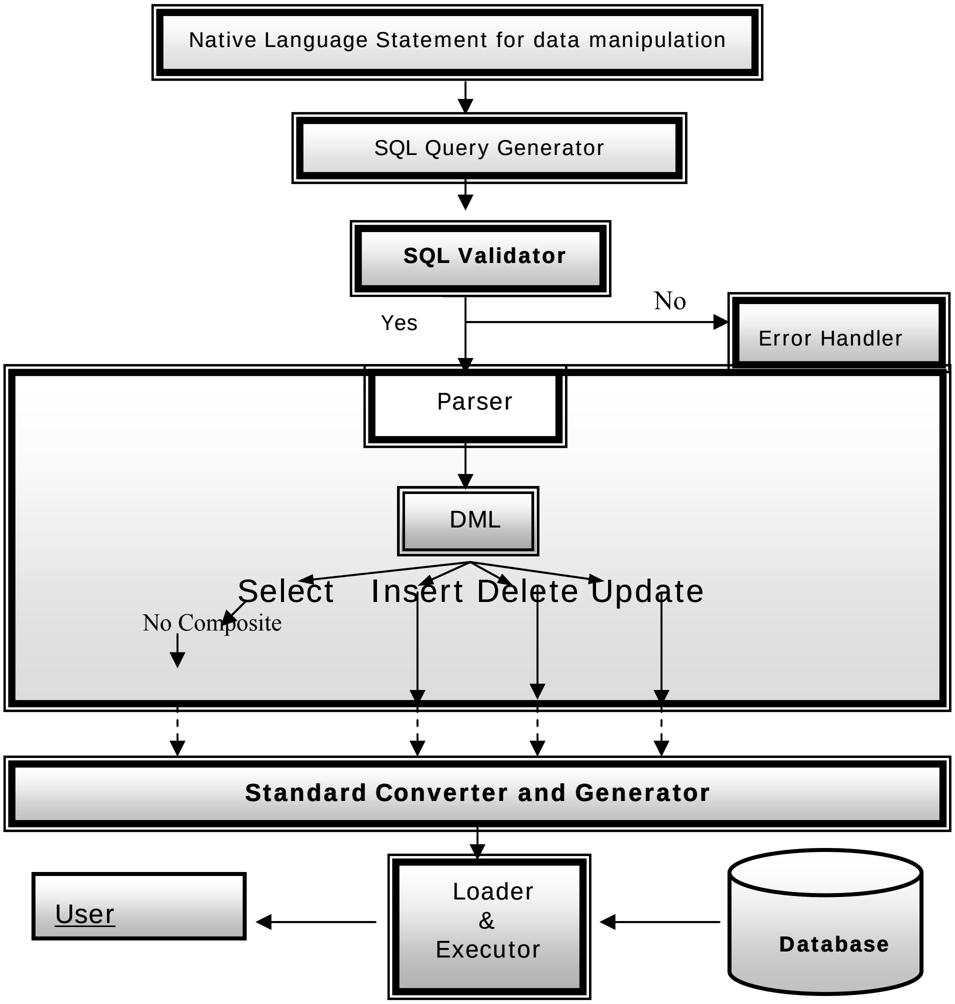
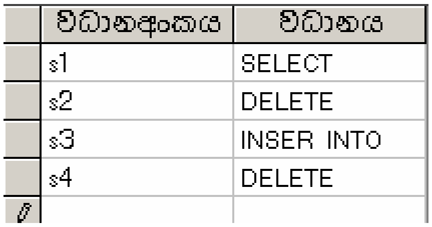
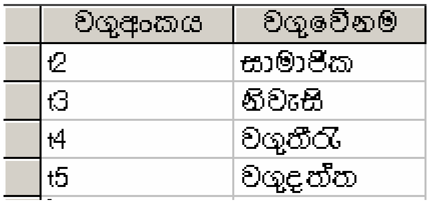


Figure 1 - Design Model of the proposed SNLI



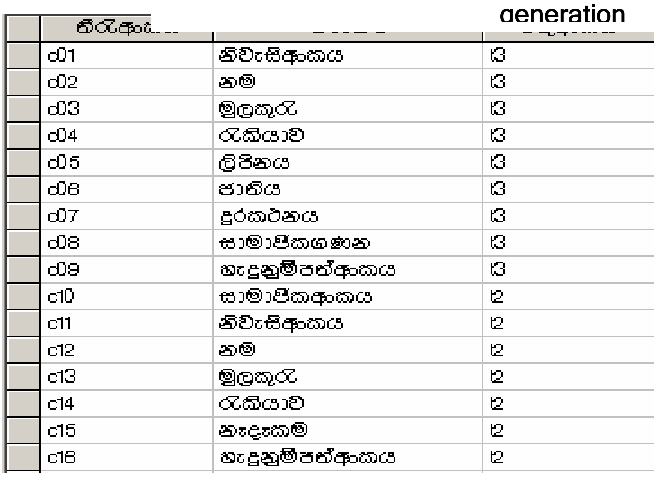
	විධානඅංකය	විධානය
	s1	SELECT
	s2	DELETE
	s3	INSER INTO
	s4	DELETE
✎		

Figure 3(a) - '□□□□□' table for standard



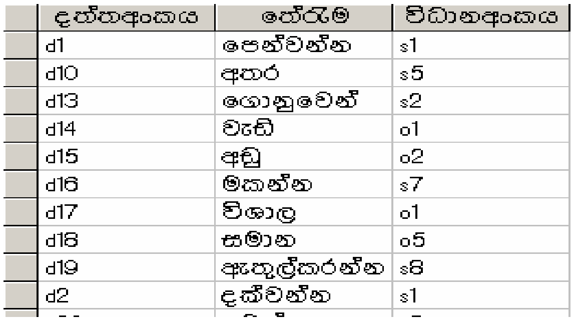
	වගුඅංකය	වගුවේනම
	t2	සාමාජික
	t3	නිවැසි
	t4	වගුතීරු
	t5	වගුදත්ත

Figure 3(b) - '□□□□□□□' table for table name generation

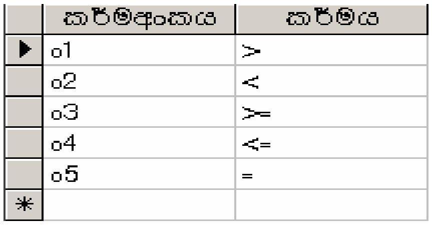


	තීරුඅංක		
	c01	නිවැසිඅංකය	t3
	c02	නම	t3
	c03	මුලකුරු	t3
	c04	රැකියාව	t3
	c05	ලිපිනය	t3
	c06	ජාතිය	t3
	c07	දුරකථනය	t3
	c08	සාමාජිකගණන	t3
	c09	හැදුනුම්පත්අංකය	t3
	c10	සාමාජිකඅංකය	t2
	c11	නිවැසිඅංකය	t2
	c12	නම	t2
	c13	මුලකුරු	t2
	c14	රැකියාව	t2
	c15	නෑදෑකම	t2
	c16	හැදුනුම්පත්අංකය	t2

Figure 3(c) - '□□□□□□□' table for column name generation



	දත්තඅංකය	තේරුම	විධානඅංකය
	d1	පෙන්වන්න	s1
	d10	අතර	s5
	d13	ගොනුවෙන්	s2
	d14	වැඩි	o1
	d15	අඩු	o2
	d16	මකන්න	s7
	d17	විශාල	o1
	d18	සමාන	o5
	d19	ඇතුල්කරන්න	s8
	d2	දක්වන්න	s1



	කාර්මඅංකය	කාර්මය
▶	o1	>
	o2	<
	o3	>=
	o4	<=
	o5	=
✱		

Figure 3(d) - □□□□□□□□□□ and □□□□□□□□ tables for condition generation

4. Implementation & Testing

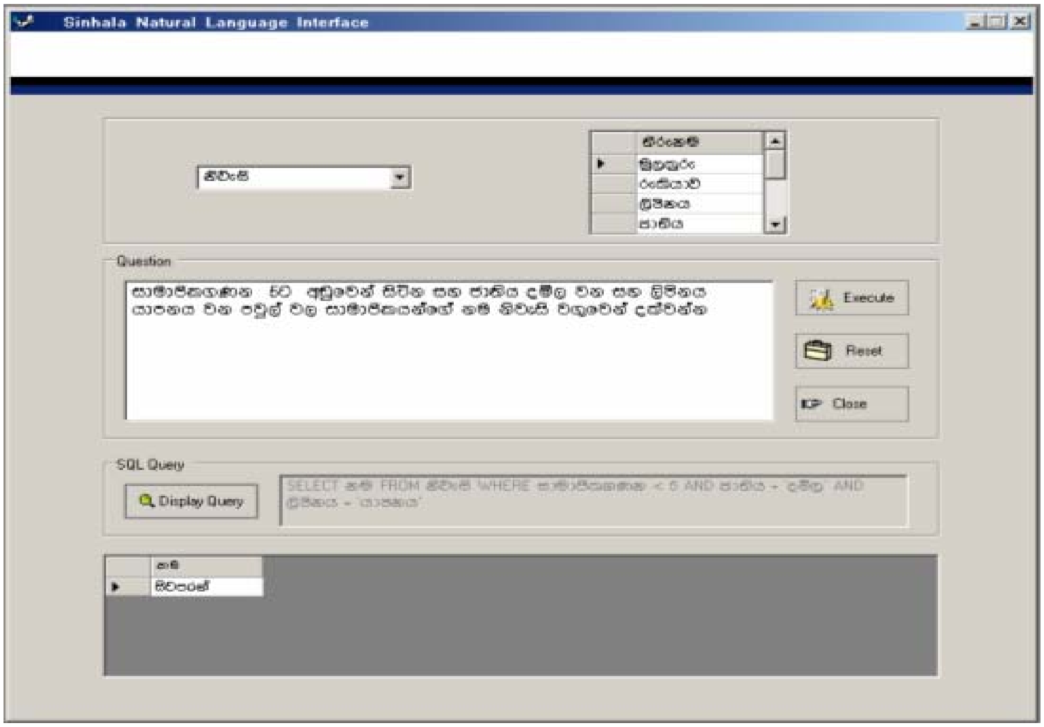
Due to required visual, ease of use (familiarity) and popularity Windows 98 and Windows XP were the front running candidates. However Widows XP was selected due to its enhanced visual application supporting, security and popularity. In selecting the development language the obvious choices are MS VB6, MS Visual C++ and C#. The wide use of these languages is another factor that contributed for the selection. Considering these languages, it was decided to use C# as the development language for the reasons such as its support for most of the backend tools, facility to create a user interface easily and it is easy to learn and fast to write codes in C#. MS Access was selected as the database considering its wide use in the industries.

4.1 Sinhala Natural Language Interface (SNLI)

SNLI is a simple user friendly tool, which runs on MS Windows XP. This provides the user the facility in accessing a database using native Sinhala language. The SNLI is the prototype that is designed to cater the requirements identified in the requirement specification. The prototype was developed according to the design described in section 3 and it consists of three major components namely: Enter Query, Generate the SQL query and Execute the query.

4.2 Testing

The database was tested at the front-end interface was created. Database was populated and entered as necessary to reflect all or most of the possible situations as depicted in figure 4,5.and 6.

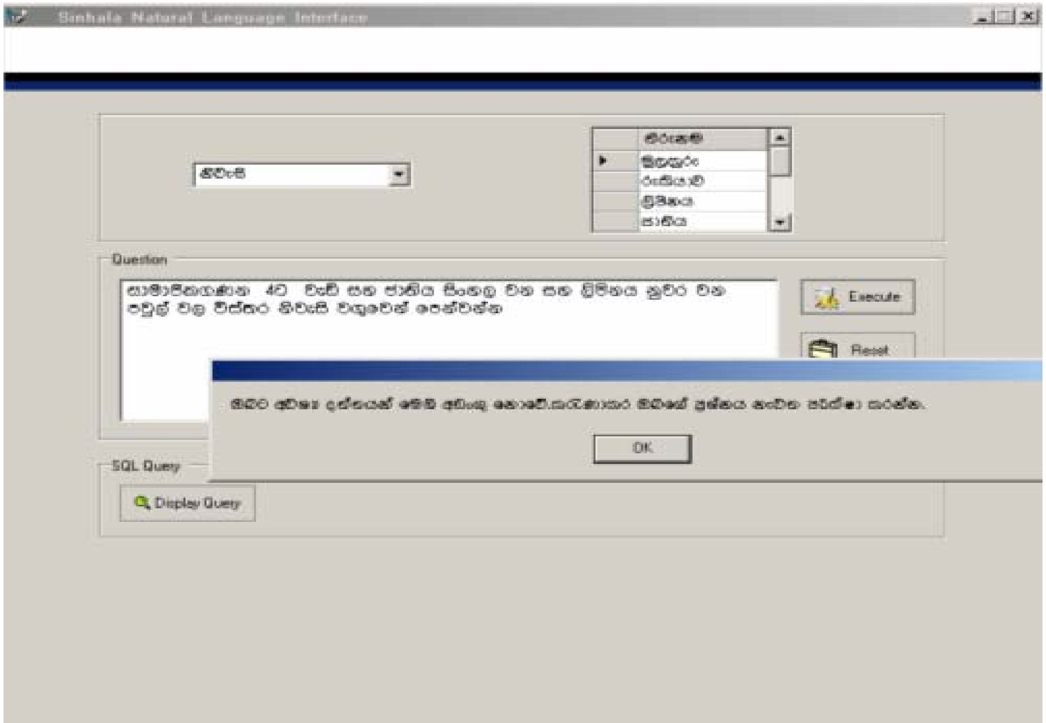


සාමාජිකගණන 5ට අඩුවෙන් සිටින සහ ජාතිය දෙමළ වන සහ ප්‍රදේශය යාපනය වන පවුල් වල සාමාජිකයන්ගේ නම් නිවැසි වශයෙන් දක්වන්න

```
SELECT නම FROM නිවැසි WHERE සාමාජිකගණන < 5 AND ජාතිය = 'දෙමළ' AND
ප්‍රදේශය = 'යාපනය'
```

නම
සිවපාල්

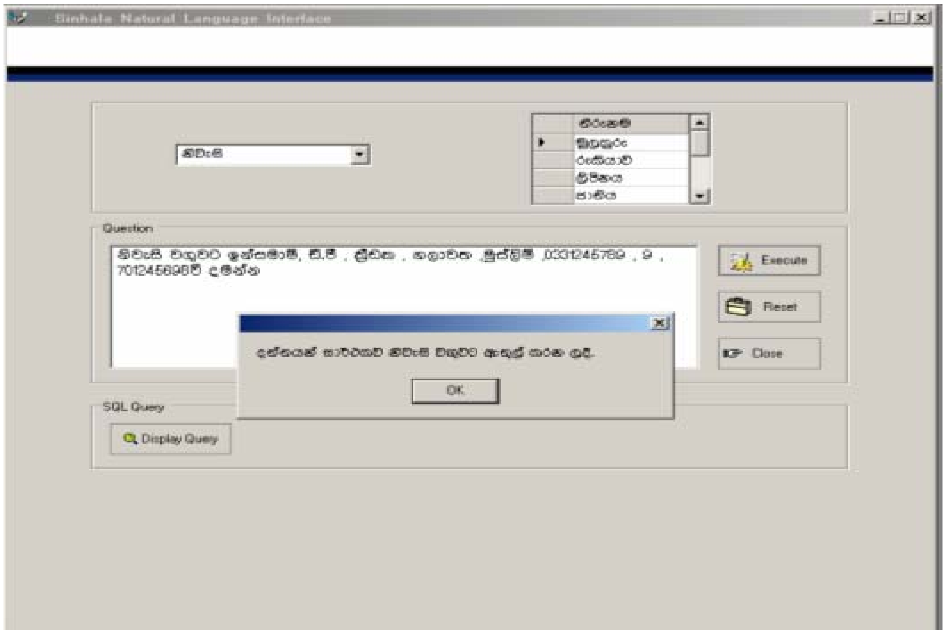
Figure 4: Retrieving data with a condition



සාමාජිකගණන 4ට වැඩි සහ ජාතිය සිංහල වන සහ ප්‍රදේශය නුවර වන පවුල් වල විස්තර නිවැසි වශයෙන් පෙන්වන්න

ඔබට අවශ්‍ය දත්තයන් මෙහි අඩංගු නොවේ.කරුණාකර ඔබගේ ප්‍රශ්නය නැවත පරීක්ෂා කරන්න.

Figure 5: A message for unavailability of data



සිවැසි වගුවට [illegible], පී.පී , [illegible] , [illegible] [illegible] ,0331245789 , 9 , 7012456898වී [illegible]

දත්තයන් සාර්ථකව සිවැසි වගුවට ඇතුළු කරන ලදී.

Figure 6: Insertion of data

Evaluation

The prototype was critically evaluated using a demonstration together with interviews and questioner. A test group was identified for this purpose and 75% considered that the prototype is 'very good' while the other 25% considered that the prototype 'can further be improved'. This evaluation process facilitated to identify the limitations of the current prototype and in turn the future enhancements.

5. Conclusion and Future Work

A successful attempt was made to clear the illusion that the industry had about the native query languages ability in providing a method to store, access and modify the data in native languages. It was possible to meet a complete logical design for the development of a Sinhala Natural Language Interface, facilitating many of the data manipulation operations in catering the requirements, stated in section 1. The outcome of the evaluation made it very clear that the prototype was a success. However, the limitations encountered in the prototype were identified as listed below, which can be overcome in future through further modifications to the prototype.

- The prototype is currently limited to the relational data model MS Access.
- It doesn't support for the mathematical functions such as SUM, AVG and COUNT.
- Complex Query handling is limited in the current prototype. For example select statements with sorting functions such as Group By, Order By, Ascending, descending are not incorporated.

Acknowledgement

The authors would like to convey their sincere gratitude to Mr. Nalaka Jayasena and the staff of UCSC for their constructive comments and suggestions for this project.

References:

- Hellwing Peter (1999*) Natural Language Parsers – A course in cooking,* Heidelbery.
- Harris, Larry R (1997) User Oriented Database Query with ROBORT – Natural Language query system Int. J. of Man-machine studies.
- Chomsky N. (1957) Syntactic Structures, The Haguei Mounton.
- Bharati, A , Chaitanya , V. ,Sangal, R. (1995) Natural Language Processing – A Paninam Perspective, new Delhi. Prentice Hall of India
- Roger S.Pressman (1997) 'Software Engineering – A practitioners Approach, Prentice Hall.
- Date, C.J., (1995) An Introduction for Database system. Addision & Weley Publication Company.
- Tenenbaum, A.S., (1992) Modern Operating System 2nd India reprint, Prentice Hall of India private Limited.
- Androutsopoulso, I., Ritchie, G.D., Thanisch (1994) Natural Language Interface to Databases – An introduction, DAI research Paper No. 709, Endinbury Univesity Scotland, UK.

Development of Standards for Local Language Support for Text Messaging within the GSM Specification

G. K Kulatilleke[†] and S. A. D Dias*
[†] Department of Computer Science & Engineering.
University of Moratuwa,
Sri lanka
* Department of Electronics & Telecommunications Engineering,
University Of Moratuwa,
Sri lanka

Email : [†]tidalbobo@gmail.com

Abstract

The development of a standard for Native language messaging is an important milestone for a country's IT roadmap. We attempt to address this important need in the context of extending the use of ICT to the rural masses in Sri Lank, for which this research was carried out. English is the only language supported at present, making text messaging (which is both economical and easy) somewhat hard to be used by non-English speaking persons. For micro (single proprietor), medium scale as well as non-English literate persons, the limitation of text messaging in English only, has become a barrier. It should be noted that countries such as China, Korea and Thailand have been successful in implementing their local languages in mobile phones.

This research offers an analytical view of the issue and importance of multiple language SMS capabilities as well as outlines the new possibilities and expansions in this area. Considering Asian, and in particular, complex fonts, it examines a range of possible input mechanisms incorporating characters on to a simple virtual keypad, designed as representative of all existing mobile handsets. The work is based on the recent SLS1134 Sinhala Unicode standard. Language encoding, including the generalized standard UTF-8 and many other optimized coding mechanisms for Sinhala are evaluated along with the statistical data, in order to come up with a comfortable and universal configuration. The outcome of the research is aimed at proving an add-on standard to the GSM/SMS and SLS1134, giving operators as well as manufactures the ability to standardize and use the most suitable as well as interoperable implementation.

Thus, we look in to the aspects of implementation, technical as well as social and psychological and propose appropriate methodologies and recommendations for text messaging in local languages within the GSM standard. We also present a protocol and standardization for the native language support features, multiple language support for text messaging focused on "Singlish", recommend a keypad layout and a layered hierarchical model for implementation as well as a migration and transitional framework for manufacturers, users and operators.

The SSMS we propose is capable of sending 184 characters as opposed to 140 on existing SMS specification, a 31% increase.

1 Introduction

1.1 Background

This work addresses an important need in the context of extending the use of ICT (Information Communication technology) to the rural masses in Sri Lanka. The GSM (Global System for Mobile) standard, on which the vast majority of cellular communications systems operate today, has specifications which allow text messaging with data/character encoding in 3 formats[1][5]

- 7 bit character encoding
- Unicode (16 bits)
- Binary

However, English is the only language supported in text messaging in Sri Lanka at present. This makes the facility (which is both economical and easy) somewhat hard to be used by non-English speaking persons. For micro (single proprietor), medium scale as well as non-English literate persons, the limitation of text messaging in English only has become a barrier.

Ad-hoc solutions do not enable Sri Lanka to convince mobile phone manufacturers to include support for local languages as built-in components of their future products.

Thus, we propose to establish appropriate standards for text messaging in local languages within the GSM standard.

The ability for text messaging in local languages will also reinforce the eSrilanka initiative, along with its recent success in standardizing Sinhala character encoding and the accompanying keyboard for computers. The impact of such standardization for mobile phones is much more powerful due to the lower cost of a phone, the higher level of penetration and the expected growth of mobile penetration compared to computers.

Through development of appropriate standards, Sri Lankans both local and aboard will be able to converse in their native tongue, using a common standard irrespective of operators or national borders.

It should be noted that countries such as China, Korea and Thailand have been successful in implementing their local languages in mobile phones.

Native language support has existed for some time in one of the following forms.

Form1:Genuine SMS using Unicode mode of the SMS specification and a language pack provided by the phone manufacturer.

When the ME (Mobile Entity/handset) manufacturer provides a language pack, so that all operations of the ME can be handled using a non-English language, we almost get native language SMS (Short Message Service) for free. How ever this inherently relies on the SMS specification's support for Unicode data transmissions, a flag set in the SMS header. What is sent is a basic 16-bit Unicode message.

Ideally this is the best solution, (free, nothing to be done, interoperable among manufactures/operators/users, standards based - Unicode as well as GSM specification) if not for the inherent disadvantages of the process.

It has to be commercially viable (large market share, big user base, competitive advantage, local legislation, other financial/political reasons). Thus not all languages get the opportunity to be released as a language pack by a manufacturer. Given that Sri Lanka is a small nation, with over 2 million phone users currently based in Colombo and suburbs as well as an increasing number in the rural areas, commercial viability for any single manufacturer to extend effort in this direction is not very high at present. Also, even with the ICTA (ICT Agency) initiative Sri Lanka is not in a strong position to influence politically or legally the availability of such a feature.

Manufactures have not until now released API's or details, for 3rd parties to create such language packs.

Even if manufacturer wants to release such a language pack for Srilanka, there are major gaps in the standards hampering this effort. Though, for the representation of Sinhala the SLS1134 standard can be used, there is no implemented or established standard for a keypad. A manufacture's will not be standards based and has no guarantee of being followed by the others, leading to major adaptation issues from the end users perspective. Thus the gaps in the standards prevent any serious joint effort materializing that will give end user transparency similar to English ASCII SMS available at the moment.

Vender lock in as well loss of control (intellectual property rights, modification rights, maintenance) will also be key issues.

Cost of Unicode is very high given the 140 byte payload size in an SMS. The fact that concatenated SMS is possible does not address the problem, as what matters is that a Unicode SMS character (16) bits[4] 2.28 times lager in bit space as the same ASCII (7) character. This boils down to increased cost of 2.28 times. And considering Sinhala as a linguistic based language[8][9], a full letter representation such as "□k" takes 2 full Unicode letters, or 32 bits, increasing the cost to 4.56.

Form2:As a work-around to actually sending an SMS using non-SMS features, mechanisms and application support on modern ME.

It is unlikely that any popular manufacture will expend recourses with the present infrastructure and environment in Sri Lanka.

With no immediate possibility of built-in support from the ME manufacturers, the burden on providing local language support to end users falls on the operators as well as the national bodies. Being survivors in a competitive market, invariably the operators are normally the first to come out with some form of native language support, as this is the means of gaining a distinctive advantage as well as an untapped market segment. Given the fact that Srilanka has over 5 national mobile operators and also over 70% of the rural mass is English illiterate, the prospect of providing native language SMS is an opportunity not to be missed. How ever in their haste, more broad issues have been neglected.

While operators have adopted several techniques to overcome this problem, they are ad-hoc solutions, leading to problems such as non-interoperability among operators, limitations in the type of mobile phones supporting the services etc. Also, in a technical sense these solutions to not fall in to the category of SMS, but are either MMS or some form of java based proprietary data exchange format. They are not machine-analyzable as SMS text (in case of MMS based images), cost much higher and use too much bandwidth. Messages received are generally not editable. Most solutions require a set of capabilities in the ME that are not required to send and receive plain SMS; these include MMS, java, GPRS and sometimes even smart phones with micro-OS capability. Though arguably such ME are becoming rapidly popular with dropping prices, the this research believes that to be able to provide a interoperable and cheap standards based solution, a pioneering effort has to be taken in many areas, such as input standardization, language mappings as well as addressing some issues particular to Srilanka itself.

Form3:Covering a particular language, by devised by a national body or group, based on interoperable standards.

Some regions and countries have implemented their own language support in to ME. This is generally done through a national body, defining code points on the Unicode Basic Multilingual Plane (BMP)[4], and is accepted as vendor neutral as well as standards compliant.

For example The BMP contains code points used to encode Chinese, Japanese, and Korean (CJK) characters. Although CJK encodings have common character sets, the encodings often used to represent them have been developed separately by different East Asian governments and software companies, and were mutually incompatible. Unicode has attempted, with some controversy, to unify the character sets in a process known as Han unification. A similar task has been done in India.

A standard is consolidation point around which all stakeholders can base solutions. It gives a commonality much needed to realize interoperability as well as conformity. Sri Lanka currently lacks such a national standard, attuned to its language and national needs and characteristics.

The study has found that each nation or area has to address the following points, issues and solve, standardize on these.

- Keypad arrangement for input
- Encoding format for representation of content. (unicode)
- Some method to avoid the Unicode penalty.
- Ensure the above does not compromise international standards and is able to coexist with older/present as well as non-conformant ME
- Address particular issues relevant to the locality. They can be both advantages as well as disadvantages for the acceptance of the standard.

For example the study identified the following salient point in the Sri Lankan context as relevant

1. Low cost handsets popular among most rural as well as low income groups
2. Life time of a ME exceeding 4 years. For any new technology to successfully penetrate all layers of the society, takes 4 years. (In Japan/Singapore it is less than one year as a handset is considered fashion statement culturally, and discarded)
3. Cost of the ME
4. Technology acceptance and reliance factor

The mentioned points have to be addressed by any nation that is interested in developing a native language capability for its ICT needs. It has to be unique, address specific issues of its language and region as well as be impleimentable in a realistic sense. Until now no such initiative has been taken for Sri Lanka, which is this studies main aim.

1.2 Aims

1. Protocol and standardization for the native language support features.

The study presents a protocol/architecture stack to be used for the messaging system that can evolve and extend, is future safe and easy to implement.

This brings up the following issues…

- Identification of language for a particular set of data in the message format
- Encoding languages in a universal format
- Ability for non aware phones to handle these special message information

Major components of the stack are allow for these functionality and are decoupled allowing being replaced by better implementations similar to the TCP (Transmission Control Protocol) stack.

2. National (Singlish) language support for text messaging.

One must be able to type an English word within a Sinhala message, as is commonly done in conversations. This helps the text messaging service to appear natural and easy to use, capable of handling the local requirements.

Issues relevant are the language encoding; the character set from the local languages which must be supported in mobile text messaging (entire character set or a subset determined to be most common in everyday use); maintaining maximum packing density and active compression, so that message size is not significantly reduced from the present length for English of 160 characters.

3. Physical format standards for a keypad layout

The research also analyses a set of design methodologies for a comfortable physical design to incorporate on the ME (the mobile phone) to successfully operate as a local language message input device. This includes the definition of the character locations on the standard numeric phone keypad as well as the mechanism used to do the data input.

4. Migration issues and transitional framework for users and operators

We also look in to migration ease as well as compatibility with non-aware devices. This is aimed at easing the cost and effort of the operators to enable smooth and gradual transition to the new messaging features. Thus it will not require redistribution of SIMs or reprogramming the phones.

2 Architecture

The native language SMS [2] works on top of the GSM 03.40 version 6.0.0 [1], using it as the transport mechanism to deliver content. It also utilizes generally implemented features of the GSM 03.40 version 6.0.0, in an attempt to make the implementation and adherence much mere easier. Rather than completely redefine a new protocol, this layering will isolate the dependencies and differences among phone implementations and provide a much more general and hence more flexible set of guidelines.

2.1 General Description

The native language SMS consists of 2 main elements.

- SC (Service center)
- ME (Mobile entity/handset)

SM MO (Short Message-Mobile Originated), denotes the capability of the GSM system to transfer a message submitted by the ME to another ME via an SC, and to provide information about the delivery of the short message either by a delivery report or a failure report. The message must include the address of that ME to which the SC shall eventually attempt to relay the short message.

SM MT (Short Message-Mobile Terminated), denotes the capability of the GSM system to transfer a message submitted from the SC to one ME, and to provide information about the delivery of the short message either by a delivery report or a failure report with a specific mechanism for later delivery.

The text messages to be transferred by means of the SM MT or SM MO contain up to 140 bytes.

This is the basic transport function of the GSM 03.40 version 6.0.0.

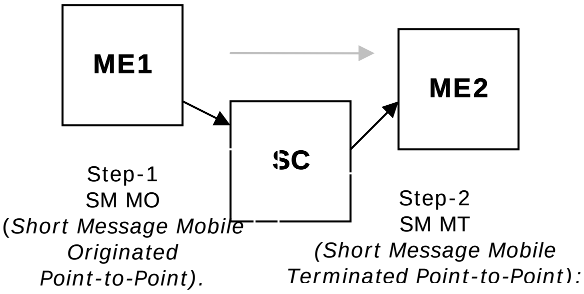

ME1
ME2
SC
Step-1
SM MO
(Short Message Mobile Originated Point-to-Point).
Step-2
SM MT
(Short Message Mobile Terminated Point-to-Point):

Figure 1

Based on this service, implementing additional layers and extensions to facilitate native language SMS can be done at any end points; ME1, SC and/or ME2.

It is possible to make ME1 and SC have the capability to send native language SMS, whist making ME2 totally unaware of such features. This requires that one or more ME's as well as the SC be modified. While almost always at least one ME on the path has to be modified, it is possible to avoid having to make any changes in the existing SC operation.

This can be achieved by making the new features invisible to the SC, by encapsulating the SMS data within the SMS data block of 140 bytes and using GSM standard defined headers and options only.

This study recommends, at this stage, that the latter method be used for the following reasons.

- It is much more safe having to modify individual ME (optional in case a user community does not wish to have such a facility), rather than doing a centralized modification to the SC (which is non-optional and affects almost all users in the SC area)
- Cheaper due to the fact that SC software modifications are proprietary and complex.
- Given the current proposal, having to modify SC internals is redundant.
- A much smoother transition is possible with modifying ME independently and individually.

How ever, due to the fact that most ME will not be able to handle the new feature, some initial support from the SC side will be necessary if non-conformant ME were to be able to display these SMS's. We will present generic guidelines on this issue under section 2.3.

2.2 Components

The study presents a 4 layers implementation stack, to be included in any conformant ME implementation. These components are only in the ME and not in the SC.

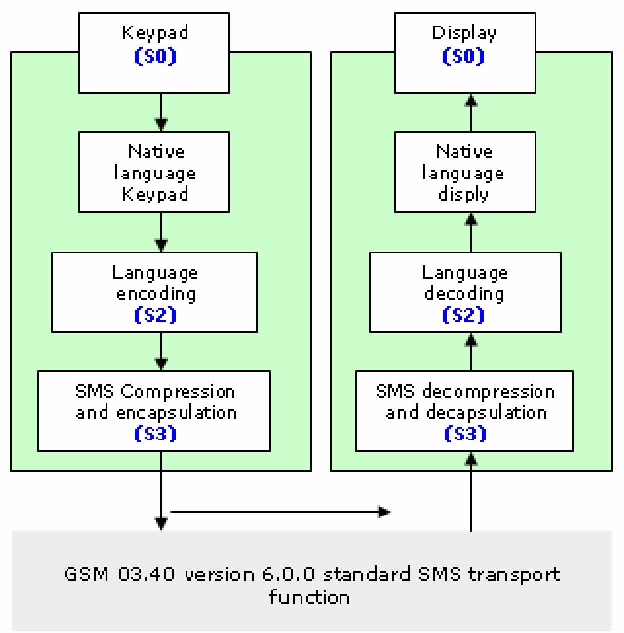

Keypad
(S0)
Native language Keypad
Language encoding
(S2)
SMS Compression and encapsulation
(S3)
Display
(S0)
Native language disply
Language decoding
(S2)
SMS decompression and decapsulation
(S3)
GSM 03.40 version 6.0.0 standard SMS transport function

Figure 2

These are S0, S1, S2, S3 and fall on top of each other in a layered manner. The reason behind this layered design is characteristic of the classic layering models and aimed at clearly categorizing /isolating and simplifying the design as well as providing modifiability.

A layer on top will provide input to the layer beneath, in case of sending a message. (See left side of figure) Each layer will provide a service to the above layer in case of receiving a message. (See right side of figure).

The following is a description of each layer.

2.2.1 S0 : Keypad, The physical input layer

This is a hypothetical minimal set keypad that is expected to be found on any capable ME (phone). The study uses this input device for all its research and simulations. By standardizing on a minimal generally found key set, it is possible to provide an interoperable, identical user- interface and input mechanism across a wide range of ME's.

It will accept the user input in terms of physical key presses and provide a key ID.

Message	Stream of strokes emitted
a	NUM_2
b	NUM_2, NUM_2
aa	NUM_2, NUM_DELAY, NUM_2
bb	NUM_2,NUM_2 NUM_DELAY,NUM_2,NUM_2
ab	NUM_2,NUM_DEALY,NUM_2,NUM_2

Table 1

Table 1 show the key ID's generated.

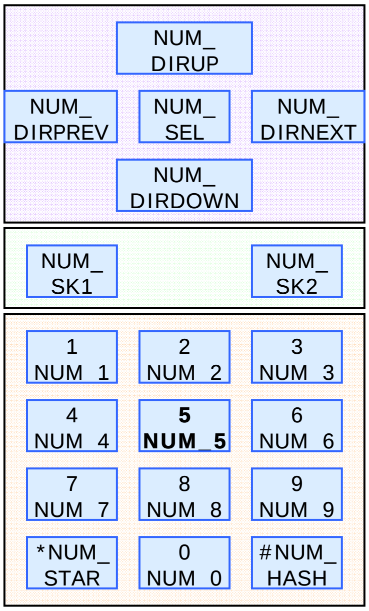


Figure 3

2.2.2 S1 : The Sinhala keyboard, Native language Keypad mapping

The S1 layer accepts the key ID's from S0 and maps them to proper user input, and will emit a stream of standard Unicode characters, based on the mapping used.

The study has identified phonetics and a novel SLS1134 based letter-sound coding scheme. Based on simulated data we recommend the latter, as well as leave space for any vendor conformant implementation, so that future enhanced input mechanisms can be incorporated.

What S1 does is accept user's key presses and generate the appropriate latter/full letter code. For example for the recommended key mapping, it will take NUM_1 and emit the code "ƌ". It will take NUM_1+NUM_1 and emit "ƌ∼∼∼∼`". The output will be SLS1134 compliant Unicode data stream representing characters of the alphabet (Sinhala).

2.2.3 S2 : Language encoding

Here the characters undergo compression + encoding based on the characteristics of Sinhala language. This is used to...

- Increase the compression ratio
- Make SMS based (raw dynamic Huffman) encoding redundant, as this is not a generic SMS feature on most ME

There are 2 modes to choose from.

- Do not use. This will simple let the Unicode from the keyboard (S1) to the SMS compression (S2). Thus no language based compression is done in this case.
- Use some form of encoding to represent in a lesser number of bits: avoid the Unicode penalty.

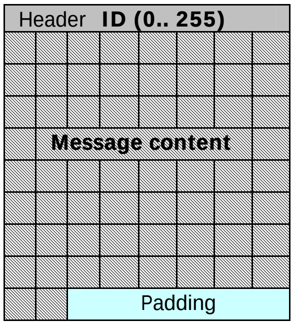


Figure 4 : SSMS (Sinhala SMS block) format

Figure 4 shows the SSMS (Sinhala SMS) block format that is the output from the S2.

Note that in S2, the option selected will dictate the SSMS header format. All SSMS are sent using (GSM 03.40 version 6.0.0: 9.1.2.2) octet representation. The header is invisible to the normal SMS protocol, and is used internally to identify the CODEC (coder/decoder) used for the SSMS. For example we use 0 for the U16 CODEC and 2 for U6+TBL CODEC. The header needs to be set to reflect the CODEC, so that the recipient knows what CODEC to be used for decoding at S2.

Each encoder will have its own unique ID (up to 255 headers, as well as extensible 16 bit headers). Figure 4 shows the output from an encoder, after accepting a Unicode message string. Padding is done to make the size multiple of 8. This entire block is sent as a binary SMS, and its internals are transparent to the SC. On receipt, a decoder that can decode the ID will extract the original Unicode message. Each SSMS will carry a overhead of 8 bits plus up to 7 bits padding. The recipient, on not being able to locate a decoder suitable for the ID, will send a SMS-Failed notification.

The stack makes this component (layer) optional and operator dependant, but recommends an algorithm we refer to as U6+TBL explained ahead. Each operator can have their own coding mechanism. How ever among operators we propose to use direct Unicode (U16) as the SMS delivery mechanism. See Figure 5.

2.2.4 S3 : SMS Compression and encapsulation

This is the SMS compression defined in SMS 3.42. As many MEs do not have this facility built in, it is optional. Any effects of not using compression with Unicode have to be dealt with language compression (S2). Thus an operator will have either S2 and/or S3.

S3 will provide the finalized SMS for delivery to the recipient. As from this point the general SMS takes over providing transport. Validity-Period, Service-Centre-Time-Stamp, Protocol-Identifier, More-Messages-to-Send, Priority, Messages-Waiting, Notifications, Reports, Error messages are handled here providing a reliable transport service.

As the SMS is transparent to the SC, it will be treated as a normal SMS and appropriate action take on its behalf automatically.

2.3 Integration and operation

Operation describes the passage of the SSMS along the delivery path. Though ideally it is the same as of a normal SMS (encapsulating the data block containing the Sinhala encoded message and header) once in the system, some special issues arise when operators want to have both conformant and non-conformant ME operating independently and with each other, as well as using different CODECs at S2.

This means, while the SC transparency can be achieved, in the following cases making the SC aware allows for much more services and features.

As in normal SMS we have the SM-MO and SM-MT. SM-MO is not relevant as sender is not important, assuming the SC has somehow got a valid SSMS for delivery. The recipient ME (ME2) will fall in to one of the following groups.

GROUP 1	In this operators network and has Sinhala SMS protocol
GROUP 2	In this operators network but NO SSMS protocol
GROUP 3	Not in this operators network.

Table 2

For every destination the SC must know what group the recipient is in. If it belongs to the same operator, the distinction of GROUP 1 or GROUP 2 can be made. Like wise, outside networks will be GROUP 3.

A mechanism to detect ME2's SSMS capabilities on registration on the network is proposed. Thus SC will be able to group these correctly. A ME that was installed with SSMS capability simple has to be turned off and on, for the grope change to take effect.

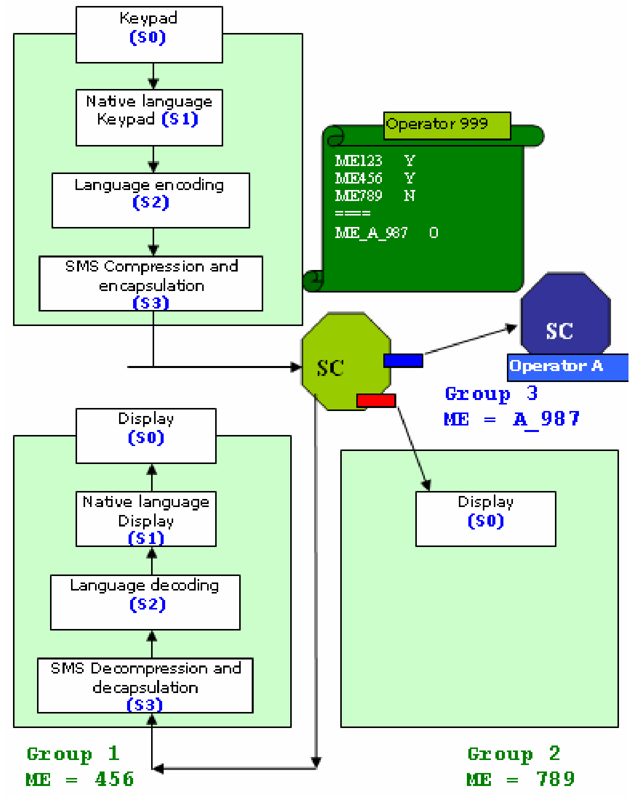


Figure 5

Figure 5 identifies the 3 groups possible.

2.3.1 Group 1 :

No problems. ME2 will decode the SSMS using the same S2 common to the particular operator.

Ideally there has to be only Group1 and Group 2 ME's only. This enables perfect transparency to be achieved, with no modifications needed to be done at the SC.

Also note that a SMS with un-displayable headers will be ignored by the ME (phone etc) so it is entirely possible to imagine a network to be group 1 and use it. In such a case ME2s with the required extensions will display the correct SSMS whilst others will simply ignore it.

How ever in such a case the senders to the non-conformant ME2 will not be able to send SSMS. An operator providing for group 2 and 3 operations will cater for this as well giving a universal service.

2.3.2 Group 2 :

Here the ME does not have Sinhala SMS capabilities.

Following options are available once SC knows that the ME2 has no capability to handle a SMS Sinhala message.

- Convert phonetics and send using English text
- Send and MMS
- Send an indicator to recipient he was just sent a Sinhala SMS and by whom
- Send an indicator to the sender that the recipient has no such support.
- Others…

2.3.3 Group 3 :

Here the ME2 is in another operator's network.

As operators can have their own S2 CODECs, the SC gateway will, optionally, convert the message to standard Unicode (U16) and send to the other SC gateway. If the same S2 (this) is used in the foreign SC, no conversion is needed. The information regarding what CODEC the foreign operator is running at S2 needs to be known, to avoid using U16, which is the failsafe.

Form here on it will be either group 1 or group 2 at the foreign operator's site. If he does not support SSMS he can consider all ME2 at his end to be group 2.

3 Keypad

Due to the complexity in Asian grammars with large, complex, joining character alphabets, the simple 3*4 numeric based keypad layout on the phone cannot be used in a straightforward manner to enter text.

Also, the ABC, DEF grouping which is now the de-facto standard is one of the least efficient and easy input mechanisms around. It does not take usage frequencies of the letters (as in the QWERTY keyboard layout).

Factors that must be considered on this context are…

Modification of the layout

- To be able to put the entire letter set of a non-English alphabet. While this is a nice way of doing it, many practical factors make this a near impossibility.

Ex: In SLS1134 Sinhala there are 41 letters with over 16 sound combinations per letter. Each language will be different, and specifying a unique PHISICAL format for EACH language on the keypad doesn't make commercial sense.

- Such a keypad would be an overkill on a mobile phone, which is a device used to make calls by dialing numeric numbers.
- Cost – special hardware and software to support this
- Elegance, range and selection – not all phones will come in all alphabets. Users will have only a specialized few models to choose from.

Use the existing physical layout (12 key) .

- This can be achieved using software only, with no physical implementation needed. Thus for any regional language, all that is necessary is to modify the software module that interprets the standard 3*4 keypad. This interpretation can come in that language pack itself, so that such customization can be done at regional dealers, operators or user levels.
- It addresses all the problems mentioned previously with minimal hassle.

Based on these arguments, we study some Native language input mechanisms, on top of the traditional 3*4 phone keypad.

.Possible mechanisms include

- Statistical occurrence based positioning on the keypad (one key gives many characters on multiple presses).
- Phonetics
- Menu based (user selects the vowels form a list after typing the character)

Ex: Type k and he/she will get the options of adding @ # $ & ~~ to it, to build the letter.

3.1.1 The physical input layer.

This is a hypothetical minimal set keypad that is expected to be found on any capable phone (or ME). The virtual input device defined here is used as the input for the SSMS data. By standardizing on this it is possible to provide an interoperable and similar user interface and input mechanism across a wide range of ME.

Considering the present as well as older phone layout among manufacturers, the about layout seems to be universal and a very suitable minimal setup. Thus we can base our key positioning and specifications based on such an arrangement. This enables most manufacturers to implement the proposed extensions and mechanisms, giving a standard interface to the users.

3.1.2 Input methodologies.

The research identified the following methodologies as possible implementation for input.

- Phonetics
    1. Normal 1,2ABC,3DEF key pad
    2. Usage frequency based – English
    3. Usage frequency based – Sinhala equivalent
- Sound-Letter
    1. Alphabet based
    2. Usage frequency based

In phonetics we use the English sounds to mimic Sinhala letters. For example “k” is k and “"@k” k+e. In normal allocation we use the standard keypad letter grouping ABC. In the frequency allocated method we use letter groupings based on usage frequencies in English, and in the last, usage frequencies on matching Sinhala sounds.

Ex: “A” phonetically matches the frequency of “a”.

Each key mapping will be completely defined by a key Atlas. An atlas is a collection of maps, each being a representation of a keypad state. Thus By defining any keypad arrangement on a KeyAtlas, the simulator is able to carry out all simulation tests and generate statistics. This open flexible architecture was selected to make future keypad models testable as well as give the opportunity for the simulator to be generic and versatile. Also, it was felt that no optimal solution for a keypad can be found; the best solution to be statistics based scientific as well as trial and error process due to the concepts used.

A sample set of SMS messages are used for the simulation. The sample has to have the following qualities.

- Be representative of SMS sent, not simple language content. SMS content differ from normal text distribution in day to day languages.
- Be sufficiently large so that the statistics are valid
- Cover the entire range of characters, each occurring at least once on the entire collection, so that any omission in the KeyAtlas is highlighted.
- Be in the relevant language: for Singlish the requirement is that all SMS consist of SLS1134 encoding as well as English, punctuation and numbers.

The research used a sample of normal language content for testing all key pad The simulator can test any keypad model for any language and give comparative analysis data.

3.1.3 Testing input methodologies

A successful keypad is on that achieves the following

- Lowest number of key presses
- Lowest time duration

This is not always the same as lowest number of key presses due to the NUM_DELAY being present. (NUM_DELAY is the time gap needed when typing “aa” on a standard hand set.)

- Achieve the highest comfort level

It has to be noted that there can be nor best keypad, but one that is optimal.

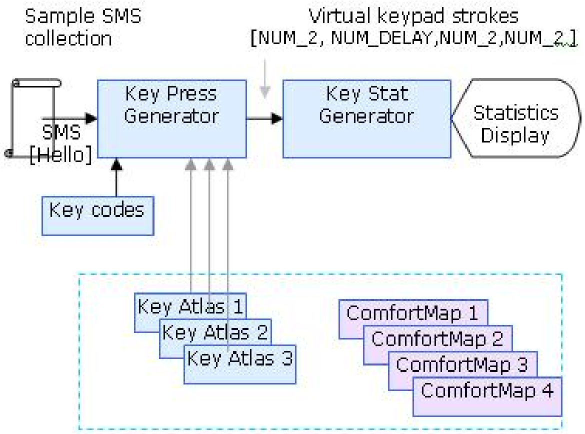


Figure 6: Keypad simulator

The basic operation of the simulator is to mimic a real human user tapping keys on the designed keypad. This is done using the Key Press Generator, which takes in a message and emits a stream of virtual keypad (S0) strokes, identical to what a human would have to press on the defined keypad.

Some additional information is also added automatically beads on the KeyAtlas. This is the NUM_DELEY code used to handle the delay in the keypad when having to enter 2 letters sharing the same key. For example to enter “ab” a user would have to pres NUM_2 for “a” and then wait a moment before successively pressing NUM_2 twice. This information is required for the modeling of time as well as duration for the entire SMS to be input using the keypad. This sequence is dependant on the KeyAtlas defining the keypad.

See table 1.

The Key Stat Generator uses the stroke sequences to generate the statistics. Each key is given a delay based on the location as well as the NUM_DELAY.

3.1.4 Comfort maps

A comfort map assigns comfort levels for each key on the virtual key pad. This models how convenient or easy pressing a certain key is as opposed to some other key. An important consideration in assigning letters to a keypad is the issue of comfort, which is addressed by the comfort maps.

It is not possible to create the ideal keypad due to the problem of comfort zones and mapping being a complex and individual issue. Currently the following comfort models are used to derive a form of comparison to understand the stoke data.

- Equal comfort

All keys are assumed to be of equal comfort (or discomfort).

- LefthandEase

For left handed users, some keys are hard to reach for, based on the assumption that the thumb rests with maximums ease on the NUM_5 key.

- RighthandEase

For right handed users, some keys are hard to reach for, based on the assumption that the thumb rests with maximums ease on the NUM_5 key.

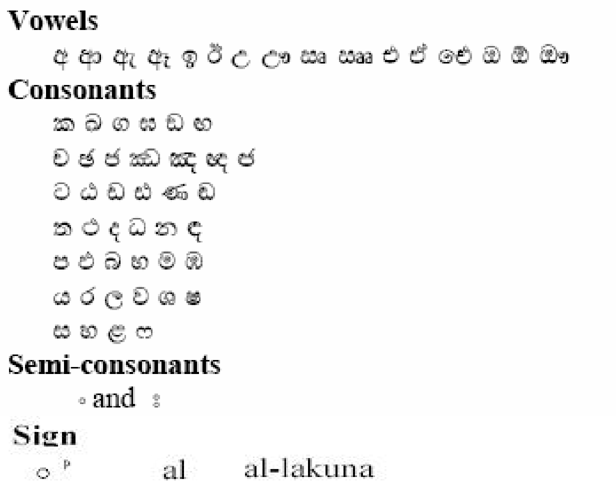
- BothHands
- Weighted

Maps were generated by simple user survey. There is how ever much more comprehensive work on this to be found.

4 Language Encoding

4.1 General comments

Based on the characteristics of the Sinhala language, the following information can be derived. Here we use the SLS1134 , Sinhala Unicode standard. [3]



Vowels

අ ආ ඇ ඈ ඉ ඊ උ ඌ ඍ ඎ එ ඒ ඓ ඔ ඕ ඖ

Consonants

ක ඛ ග ඝ ඞ ඟ
ච ඡ ජ ඣ ඤ ඥ ඦ
ට ඨ ඩ ඪ ණ ඬ
ත ථ ද ධ න ඳ
ප ඵ බ භ ම ඹ
ය ර ල ව ශ ෂ
ස හ ළ ෆ

Semi-consonants

ං and ඃ

Sign

් al al-lakuna

Figure 7

Vowels (Sounds) 16 (+1 + 2)

The standard defines 16 independent vowels. The “al-lakuna” is not a pure vowel, but defined as a sign. The semi-consonants “ang” (ox0D82) and “ah” (ox0D83) gives 2 more.

Consonants (41)

There are 41 identified in SLS1134. The semi-consonants “ang” (ox0D82) and “ah” (ox0D83) are also present, but these are represented along with vowels.

Thus we have 17*41 = 697 distinct letter codes in the language.

A full-letter code is any of “ƙ”, "▯ƙ” etc….

Unlike in English, here we have some unique problems in full-letter representation.

- Too many full-letters to encode
- This will require a minimum of 2^{10} which gives 1024 values as we are interested in Singlish.
- Complex letters such as joined and touching letters are present. If we use the full-letter approach to define a unique value per full-letter, we have to have unique values for all joined full-letter couples as well. This will be an alarmingly larger letter space, of which only a very few will be used in the general case.
- Input mechanism becomes complex.

Whist in English we have 26 characters, having to enter a rich full-letter combination of 697++ characters definitely requires a much more radical and innovative approach

4.2 Available options

The study identified the following methodologies for encoding SLS1134 unicode in to less compact representations, using language specific features. We focused specifically on novel algorithms, as much work is done on Hoffman, LZW etc, but not on ultra small content (140 byte) capitalizing on language features. Each encoder is a CODEC in the S2 of the proposed implementation stack and interchangeable as long as the recipient has the same CODEC, or the SC manages to do a format translation based on the header ID.

4.2.1 OPTION 1 : 16 bit SLS1134 Unicode (U16)

This will use the SLS1134 directly and might take advantage of the SMS compression when available. It uses 16 bits per character and uses linguistic (letter+sound) encoding. Thus effectively one full-letter will use up 32 (2 Unicode “letter + sound” group).

Effectively we will be able to send an average of (140 / 2*2) 35 full letters. In the simulation this was used as the comparison base, in measuring the compressibility of each CODEC.

4.2.2 OPTION 2 : 7 bit truncated modified Unicode (U7)

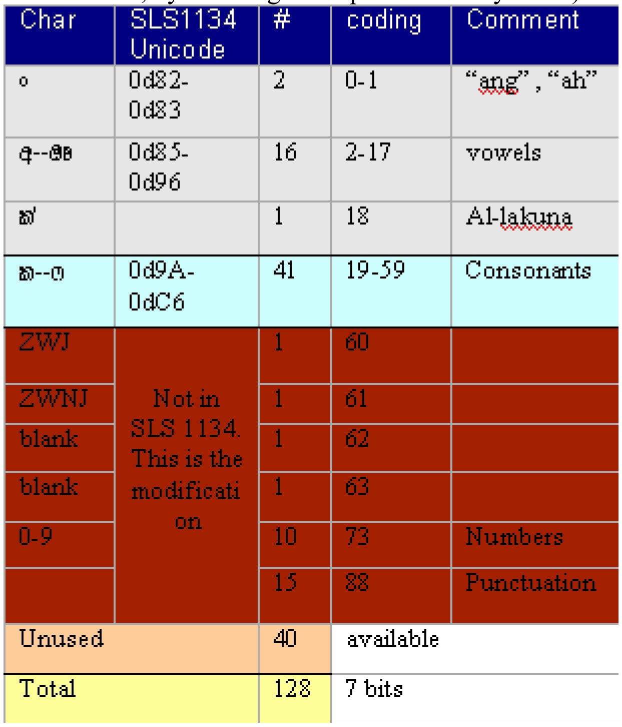
This will also use the SLS1134 directly. How ever it truncates the first 9 bits of the Unicode to make each letter 7 bits. It uses 7 bits per character, and uses linguistic encoding, and has ZWJ, ZWNJ and some punctuations and numerals mapped on to the available 128 value (7 bit space).

Sample modified Unicode is shown in table 3.

Thus effectively one full-letter will use up 14 (2 modified 7 bit “letter + sound” group). This will not need the SMS compression in GSM.

Effectively we will be able to send an average of (140*8 / 14) 80 full letters.

It also has 40 unused bit positions, which can be used to put in an additional alphabet (Ex: an English alphabet with 26*2 characters, by reducing some punctuation symbols).



Char	SLS1134 Unicode	#	coding	Comment
ං	0d82-0d83	2	0-1	“ang” , “ah”
අ--ඖ	0d85-0d96	16	2-17	vowels
ක්		1	18	Al-lakuna
ක--ෆ	0d9A-0dC6	41	19-59	Consonants
ZWJ	Not in SLS 1134. This is the modification	1	60	
ZWNJ		1	61	
blank		1	62	
blank		1	63	
0-9		10	73	Numbers
		15	88	Punctuation
Unused		40	available	
Total		128	7 bits	

Table 3 :

4.2.3 OPTION 3 : Relative 6 bit Unicode with extension tables (U6+TBL)

This is a different version of both the above methods. Here we encode the SLS1134 defined characters in 6 bits, and use extension table to enhance the character space. This will be able to handle all the SLS1134 full-letters as well as punctuation, numerals and English characters as well with the use of the extension mechanism.

Its 3 tables are given in 4-1,4-2 and 4-3.

The escape codes will select which table is used. The current table will be used until and explicit ESCAPE_X code tells to switch to table X (1, 2, 3). Table 1 is the default, and always a message starts in this table.

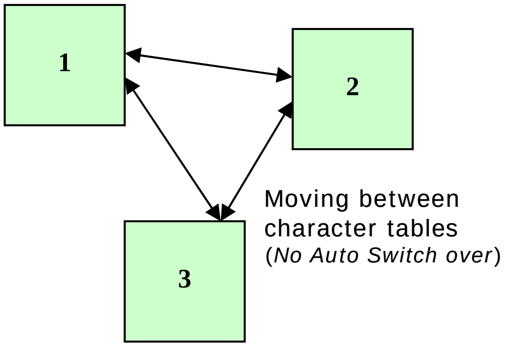


Figure 8

Char	SLS1134 Unicode	#	coding	Comment
ං	0d82-0d83	2	0-1	“ang” , “ah”
අ--ඖ	0d85-0d96	16	2-17	vowels
ක්		1	18	Al-lakuna
ක--ෆ	0d9A-0dC6	41	19-59	Consonants
ZWJ	Not in SLS 1134. This is the modification	1	60	
BLANK		1	61	
ESCAPE_2		1	62	
ESCAPE_3		1	63	
Unused		0	available	
Total		64	7 bits	

Table 4-1: Default

Char	SLS1134 Unicode	#	coding	Comment
0-9	Not in SLS 1134. This is the modification	10	0-9	Numbers
,.;?/“’▯		15	24	Punctuation
ESCAPE_1		1	62	
ESCAPE_3		1	63	
Unused		37	available	
Total		64	7 bits	

Table 4-2 : Numbers and Punctuation

Char	SLS1134 Unicode	#	coding	Comment
A-Za-z	Not in SLS 1134. This is the modification	53	0-52	Alphabet
ESCAPE_1		1	62	
ESCAPE_2		1	63	
Unused		9	available	
Total		64	7 bits	

Table 4-3: English supplementary alphabet

Thus effectively one full-letter will use up 12 (2 modified 6 bit “letter + sound” group). This will not need the SMS compression in GSM.

Effectively we will be able to send an average of (140*8 / 12) 93 full letters including English characters as well, achieving full Singlish capability.

The use of mixed characters is assumed to be a major advantage in the local context. This is due to…

- The actual Singlish versions of Sinhala that some people speak can also be used in SMS.

  mm en@k`t CD ar@gn en~v`.
  aq MATCH ghn~n b$

- Some times we do not know the correct wording (the English word maybe more comprehensible, shorter and meaningful in an SMS)

  TV as opposed to “r$pv`hQnQ"""".

- This SSMS will give the user ability to send English SMS much more efficiently than the “true” English SMS!

((140*8)-8 /6 = 184 characters as opposed to 140)

- As the SSMS can send both Sinhala *and* English, the user can use the same application/service without having to switch between both. Much simple and centralized.
- Flexibility and ability to add Tamil as well, so that a single add-on will be able to send the whole range of languages applicable for the nation

4.2.4 OPTION 4 : Letter-Sound statistical grouping with extension tables (LS)

This takes in to account the relative frequencies of the letter usage. Letters (consonants/vowels) are sorted by usage frequency and put in to maps. Map 0 has the most occurring set. Each map also has escape sequences to access all other maps. All other maps auto returns to map 0, which has the highest probability. Map size was tested from 5 to 10.

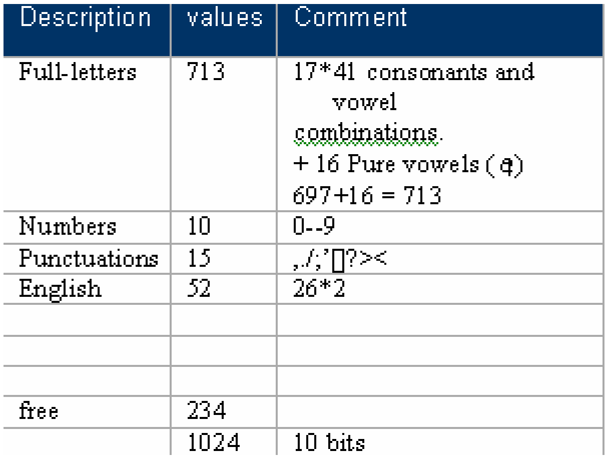
4.2.5 OPTION 5 : Flat allocation mapping (FL)

This is the opposite of LS concept. Instead of using letter+sound to create full-letters such as “□k”, we consider “□k” as a stand alone single complex character.

Given the fact that we have 17*41 = 697 distinct letter codes in the language, it is possible to simple provide each full-letter with a unique value if we have a 1024 (10 bit) allocation space. For simulation, we use maps as in LS, and test from 6 bit to 11 bits.

Table 5 shows how the allocation map is made up.

Interestingly now each full-letter will take only 10 bits giving an effective average of (140*8/10) 118 characters. We are effectively encoding 32 bits using 10. This is by far the highest character count we have in theory. In practice, how ever the values differ. Please see the results chart.



Description	values	Comment
Full-letters	713	17*41 consonants and vowel combinations. + 16 Pure vowels (ɋ) 697+16 = 713
Numbers	10	0--9
Punctuations	15	,./;’□?><
English	52	26*2
free	234	
	1024	10 bits

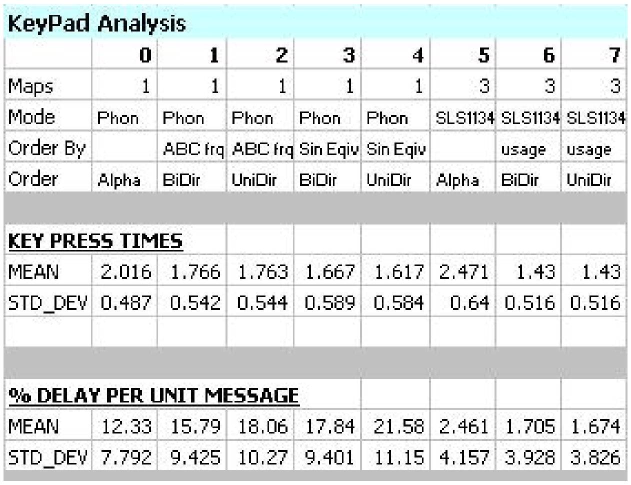
Table 5

Phon indicates a phonetic (using abc) text input, whilst SLS1134 is the developed letter+sound input method. Keypad 0 is the ITU E 1.161 International Standard Key Pad, found on most handsets. Keypad 1 has English letters allocated based on usage frequencies allocated bi-directionaly: 2 is the same but allocated unidirectionaly.

Bi-directional allocation is used to minimize the joint probability of letters appearing on the same key. Figure 9 shows the allocation method to minimize key presses. First top horizontal row is the 1-key press letters, next row the 2-key press letters etc. If we allocate on descending frequency from 2..9, to minimize the joint probability the next row has to be allocated from 9..2, aligning the least possible letter of the 2nd row with the most possible on the 1st. This minimizes the delay such as encountered when typing M<delay>N.

5 Results

5.1 Keypad simulator data



KeyPad Analysis

	0	1	2	3	4	5	6	7
Maps	1	1	1	1	1	3	3	3
Mode	Phon	Phon	Phon	Phon	Phon	SLS1134	SLS1134	SLS1134
Order By		ABC frq	ABC frq	Sin Eqiv	Sin Eqiv		usage	usage
Order	Alpha	BiDir	UniDir	BiDir	UniDir	Alpha	BiDir	UniDir
KEY PRESS TIMES								
MEAN	2.016	1.766	1.763	1.667	1.617	2.471	1.43	1.43
STD_DEV	0.487	0.542	0.544	0.589	0.584	0.64	0.516	0.516
% DELAY PER UNIT MESSAGE								
MEAN	12.33	15.79	18.06	17.84	21.58	2.461	1.705	1.674
STD_DEV	7.792	9.425	10.27	9.401	11.15	4.157	3.928	3.826

Table 6

KEY PRESS TIMES (KPT)

NumKeysPresses/letters in message

% DELAY PER UNIT MESSAGE (DEL)

*(DelayWaits0/Letters in message)*100*

For the tested keypad designs, KPT is significantly lower in frequency ordered layouts, taking usage in to consideration; this is common to phonetics as well as SLS1134. Bi-directional allocation shows better KPT in phonetics mode, but not in the SLS1134 mode. This may be due to the letter+sound method alternating between maps using DIR_NEXT and DIR_PREV keys. In fact, it shows the opposite in DEL from what we normally expect, suggesting that the bi-directional hypothesis does not apply in case of SLS1134 type arrangement.

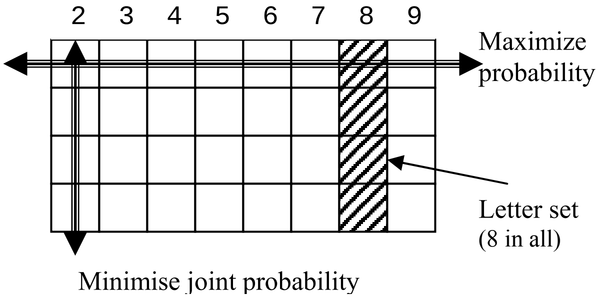


Figure 9

5.2 Keypad allocation (comfort mappings)

Key	LEFT	BOTH	RIGHT	AVG
0	16	4	16	17
1	15	7	9	4
2	8	1	1	2
3	9	11	15	5
4	17	8	10	11
5	10	2	11	12
6	11	12	17	15
7	18	9	12	19
8	12	3	14	16
9	13	13	18	18
*	19	10	13	13
#	14	14	19	14
S1	1	5	7	3
S2	2	6	8	7
<	3	15	5	1
>	4	16	6	9
Up	7	17	2	8
Sel	5	19	3	10
Down	6	18	4	6

Table 7 [1= easiest, 19=hardest]

Using the keypad simulation data we have obtained the 8 letter sets (Figure 9). These sets each will give a combined occurrence probability as the sum of its component probabilities. Based on the comfort maps, allocation is done to position the sets on the relevant keys, most occurring sets on the most comfortable keys.

We do not recommend a definitive comfort map in this study, but outline the methodology for the implementation.

5.3 Language Encoder simulator data

COMPRESSION TIMES (CT)

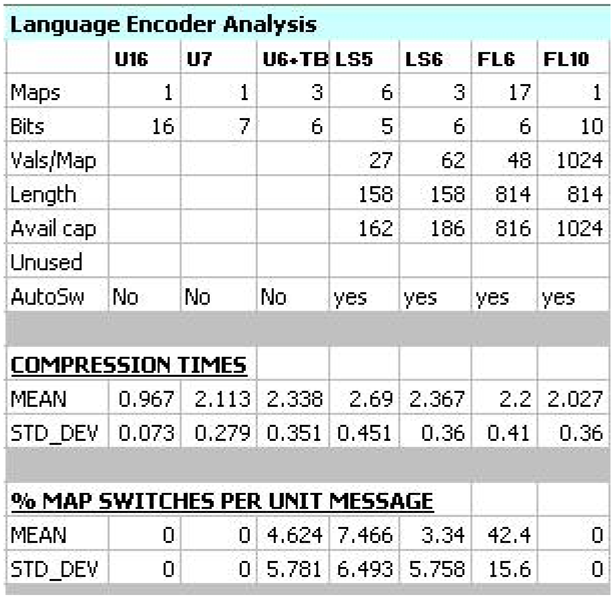
inputLineLength(bits) / compressedlength(bits)

% MAP SWITCHES PER UNIT MESSAGE (MSW)

*(TableSwitches /lineLength)*100%*

Thus a higher CT would indicate better compressibility. LS5, LS6 and U6+TBL give the highest compression times.

The concept of table switches was used to understand the values of compressibility. A high percentage of table switches would mean that the overhead in the compression is high. LS5 has a 7% switch rate, each costing a 5 bit chunk, but this is offset by the shorter per letter bit size. FL6 shows poor compressibility due to the 42% overhead incurred in table switching, indicating that table size for such large content (816) is not optimal with maps of 6 bits.



Language Encoder Analysis

	U16	U7	U6+TB	LS5	LS6	FL6	FL10
Maps	1	1	3	6	3	17	1
Bits	16	7	6	5	6	6	10
Vals/Map				27	62	48	1024
Length				158	158	814	814
Avail cap				162	186	816	1024
Unused							
AutoSw	No	No	No	yes	yes	yes	yes
COMPRESSION TIMES							
MEAN	0.967	2.113	2.338	2.69	2.367	2.2	2.027
STD_DEV	0.073	0.279	0.351	0.451	0.36	0.41	0.36
% MAP SWITCHES PER UNIT MESSAGE							
MEAN	0	0	4.624	7.466	3.34	42.4	0
STD_DEV	0	0	5.781	6.493	5.758	15.6	0

Table 8

6 Conclusion and Future work

The research has been successful in identifying the important categories to look at in implementing a localized input mechanism and SMS capability for communication. It also presents a stack based architecture for a generic ME as well as outlines the SC operations for smooth transition to the new features.

6.1 Key Pad

We recommend the SLS1134 unidirectional usage based allocation as the keypad, based on the simulated data, which differs from normal expectations, as one might expect bi-directional allocation to improve the performance. The delay count also shows this effect.

The key pad has 3 modes.

1. Letter mode (L)
2. Sound mode
   a. Altering sound mode (AS)
   b. Pure sound mode (PS)

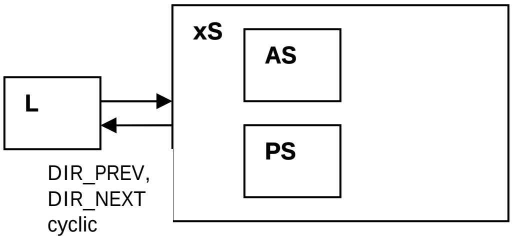


Figure 10

Figure 10 shows the mode switching arrangement. The key pad is initially in the L mode. This can enter any of the 41 consonants as well as ZWJ, number and punctuation. Switching to sound mode is done by the direction keys. If the last entered character is a consonant, the key pad enters to AS mode, allowing the consonant to be added with a sound, else it goes to PS, so that a user can enter a vowel. This decision is made automatically.

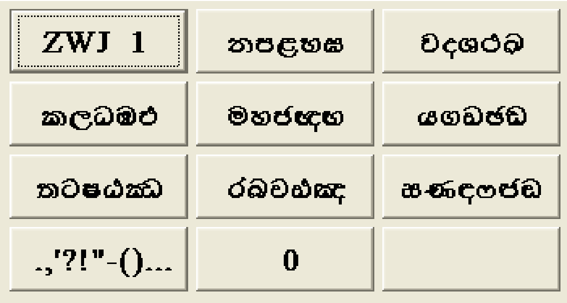


Figure 11-1 Key pad in L mode

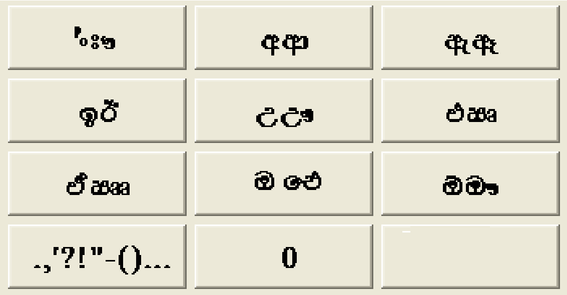


Figure 11-2 Key pad in S mode

6.2 Language encoder

Based on the simulation and other features we recommend the use of U6+TBL mode CODEC for S2. This gives the following advantages.

- This SSMS will give the user ability to send English SMS much more efficiently than the "true" English SMS!

(184 characters as opposed to 140)

- As the SSMS can send both Sinhala *and* English, the user can use the same application/service without having to switch between both. Much simple and centralized.
- Flexibility and ability to add Tamil as well, so that a single add-on will be able to send the whole range of languages applicable for the nation
- More SLS1134 Conformant
- Smaller CODEC (encoder/decoder) resources required at ME
- Complex letters such as joined and touching letters are present. If we use the full-letter approach to define a unique value per full-letter, we have to have unique values for all joined full-letter couples as well. This will be an alarmingly larger letter space, of which only a very few will be used in the general case.

6.3 Future work

The work discussed can be further extended in some areas. Particularly, the alphabetical ordering of the SLS1134 keypad could be looked upon and an optimal (constrained) alphabetical layout devised. This will help novice users a fairly easy adaptation. How ever, we must bear in mind, unlike the English alphabet, not many are familiar with the SLS1134 alphabet, seriously undermining the advantages or the ordering. This area has to be further looked in to.

Dictionary based predictive methods, such as T9 from Tegic Inc., iTap from Motorola, and eZiText from Zi Inc., only require one key press for each character inputted. For Singlish such method might be best.

In our studies, we used language based concepts to compress 140 byte size data amounts, standardized algorithms such as Huffman, LZW were not considered, as this work is already covered elsewhere.

References

[1] GSM 03.40 version 6.0.0

[2] SMS version 3.42

[3] SLS1134 Documentation (www.fonts.lk)

[4] Unicode standards documentation homepage (www.unicode.org)

[5] [ETSI 1996] European Telecommunications Standards Institute, "Digital cellular telecommunications system (Phase 2+), Technical realization of the Short Message Service (SMS), Point-to-Point (PP)", GSM 03.40 version 5.4.0, November 1996

[6] Gihan V. Dias "Representation of Sinhala in Unicode" ICT Agency (www.fonts.lk)

[7] The Unicode Web Site; www.unicode.org

[8] The ICTA Language website, (www.fonts.lk)

[9] Sri Lanka Standards Institute; “Sri Lanka Standard 1134 – Sinhala Character Code for Information Interchange”, Revision 2, 2004.

[10] Gong J., & Tarasewich P , “Testing Predictive Text Entry Methods with Constrained Keypad Designs”

[11] Gong J., & Tarasewich P. (2004). “Guidelines for handheld mobile device interface design.”

[12] Jeffery Smith, “Thumbscript™: Designing a general solution to the problem of text input in small devices.”

[13] R. L. Deininger , “Human Factors Engineering Studies of the Design and Use of Pushbutton Telephone Sets”

Accessibility for the Blind in Ubiquitous Environment Using Binocular Vision System

Janaka Chaminda Balasuriya,* Chandrajith Ashuboda Marasinghe,**
Keigo Watanabe,* and Kiyotaka Izumi*
*Department of Advanced Systems Control Engineering
Graduate School of Science and Engineering
Saga University, 1-Honjomachi, Saga 840-8502, Japan
**Software Engineering Lab, Department of Computer Software
University of Aizu, Aizu-Wakamatsu, Fukushima, Japan
*jcbala@lycos.com, {watanabe,izumi}@me.saga-u.ac.jp
**ashu@u-aizu.ac.jp

Abstract

This paper describes an attempt to construct a mechanism for guiding blind people in ubiquitous environment. Features of this research include, real time environment monitoring (with static and dynamic objects) by single 'binocular vision system' and intelligent decision making by 'adaptive neuro fuzzy inference system (ANFIS)'. Hence, necessary information throughout the surrounding area (ubiquitous environment) is distributed to the blind people. The system is comprised of gathering data using the binocular vision camera to feed to a prediction system to analyze the future trends of the dynamic objects in the concerned area. This new set of data is used to assign a shortest and safest path for a blind person to travel to a specific place or meet another human (who is in motion). Although the main aim of this research project is to construct an intelligent system that guides blind people, once completed, it may also be possible to use this system in other areas like air traffic navigation systems and guiding ubiquitous robots.

Keywords

Fuzzy decision making, Forecasting, Prediction, Ubiquitous environment, Binocular vision system, Obstacle avoidance, Path planning

1 Introduction

Target tracking, path planning and guiding are used in many fields such as robotics, intelligent systems, autonomous vehicles, human guidance systems, etc. There are many approaches in the areas of research to fulfill the dream of a fully functional automated guiding system in a real-time environment. The simplest form of a tracking and guiding system can be considered as guiding an agent such as a blind person or a robot to a stationary position in a static environment. But, there will be many critical issues to consider when the target is moving and the environment is dynamic. According to Luo et al. [1], in addition to the sensory detection module in a general position based tracking system, there are many limitations to address, such as estimation/prediction of the object position from noisy sensory measurements, motion control of the tracker to track the moving object, etc.

When the target is a dynamic object with a complex nature of movement (or rather having a non specific motion pattern) or target tracking is performed in a dynamic environment, accurate detection and fast responses are a must. These need much more advanced features such as fast response sensors, high speed processing equipment, etc. On the contrary, if the sensors have a slower response time, the estimation is no longer sufficient for a real-time target tracking [2].

Hence many real-time tracking and guiding systems use intelligent prediction strategies based on fuzzy logic decision making to predict the potential object position at the next time period. Information gathered in such a way is used to speed up the detection procedure and to control or assist the motion parameters of an agent.

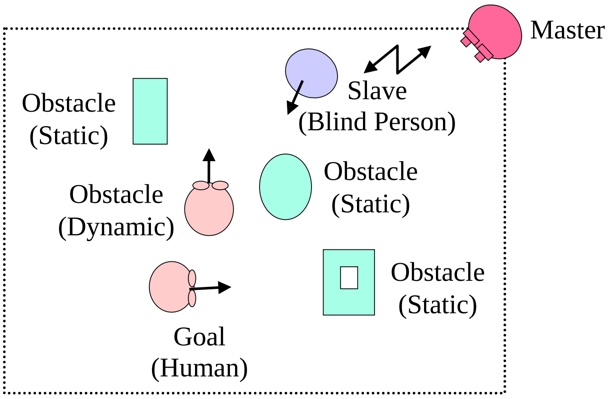


Figure 1: Navigation in dynamic environment, where master, central intelligent system consistes of binocular vision system and fuzzy logic decision making

2 Target System

Navigate a tracker (e.g. a blind person or a robot) to meet a goal (e.g. human) in motion while avoiding static and dynamic objects in the surrounding environment according to commands issued by a master (intelligent system consisting of fixed binocular vision system and fuzzy logic decision-making).

The environment is continuously monitored by the binocular vision system to identify static and dynamic objects. Here, objects like desks, chairs, standstill humans, robots, etc. are identified as static objects and on the other-hand moving humans, vehicles, animal, etc. are identified as dynamic objects. Next a goal (e.g. a human, vehicle, etc.) and a tracker (e.g. a robot, a blind person, autonomous vehicle, etc.) are identified within this environment. Once a goal and a tracker are identified, all the other objects (static or dynamic) in the environment will be considered as obstacles. Continuously analyzing the changes in the environment, master guides the tracker (slave) until the tracker reaches the goal by fuzzy logic decision-making as shown in Figure 1.

As the initial stage of this research project, when an obstacle is nearby or moving towards the blind person, it will be considered just as an ‘obstacle’, i.e. *‘there is an obstacle to your left’* or *‘an obstacle is moving across your path’* etc. But with the progress of the research, it will be possible to identify the objects as they are and give more information to the user, such as *‘a human is coming towards you’* or *‘a table is in straight ahead’* etc.

3 Localization and Target Tracking

Target tracking and path planning are widly used in robotics and intelligent machines. Robots, finding its own path by means of visuals, sounds or any other method, have many problems due to the limitations of the sensory elements. Robots with binocular vision system have many advantages in this regard; but there are some limitations such as reduction in visible area [5] once the cameras are focused to obtain a closer view of an object as shown in Figure 4. This gets worse when there are moving objects in the surrounding area. There may be possible collisions due to blank angles and hence necessity of a continuous environmental awareness system arises.

There are platforms that estimate relevant quantities in the vicinity formed by combining information from multiple distributed sensors. For an example, robots in a team estimate their relative configuration by combining the angular measurements obtained from all of the ominidirectional images and performing triangulation operation as described by Spletzer et al. [6]. There are other variants such as the system proposed by Guirnaldo et al. [7] for controlling the perceptual process of two cooperative non holonomic mobile robots by formalism called perceptual anchoring. Their system enhances the awareness of the system by employing an anchor based active gaze control strategy to control the perceptual effort according to what is important at a given time. But such a system is of little use or not adequate for those (blind people) whose main intention is to interact with humans or any other objetcs in the real environment. The situation gets critical, when such a person requires to interact with another human while in motion. To place many sensory elements in the

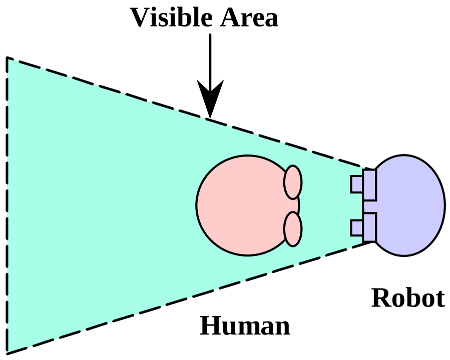


Figure 2: Limitation of vision

vicinity just to obtain an idea of the surrounding will be redundant and expensive. According to Spletzer et al. [6], questions of the quality and of how informative the gathered data are also arise, because they are obtained from individual sensor units. In addition, there may be other issues like how the sensor units should be deployed in order to maximize the quality of the estimates returned by the set, because the data required are needed for motion analysis.

On the other hand, a different set of questions arises when one considers the problem of integrating information from a number of fixed distributed sensors such as cameras. Cost associated with transmitting and processing data, sensors that should be used to form an estimation for a given time, coordination among the sensors, automatically relating events among each other, sensor geometry, effects due to characteristic differences, etc. are some of the problems to be solved. In multiple video streams generated by multiple distributed cameras, finding correspondence is the key issue as observed in Lee et al. [8]. Hence a newer, simpler yet versatile localization and tracking system is required.

4 Existing Techniques and New System

Navigation and path planning using modern technology are not so new to the robotics research, though, there can be very few examples that can be shown in the case of using it in guiding blind people. But it can be seen that such technologies are even exist in exploration of new planets such as 'Rover mission' in Mars exploration [9],[10], guiding autonomous vehicle in unknown terrains [11], tour guide robots in many indoor conditions [12], etc.

While in motion, sensing of current position to the objective position while avoiding collision with the surrounding objects can be considered as the two most considerable facts. The basic techniques that are used to obtain collision free navigation can be categorized into using stereo-vision, land marks identification, collaborative techniques of various sensory elements, etc. Although there are several incidents that had reported successful results, using the similar mechanisms to guide blind people is rarely stated. It may be due to complexity of the natural environment, ambiguity in motion patterns of humans, guiding a human (blind person) is not so easy as driving an object such as a robot, etc.

But, as Luo et al. [1] performed with an autonomous mobile robot 'Chung Cheng-1', the present research project tries to guide humans as trackers with necessary alterations to the system with additional key features such as binocular vision camera fixed in the environment, decision making by adaptive neural fuzzy inference system (ANFIS), voice guidance to the tracker unit (blind person), etc. Further, by generalizing, such a system will be in helpful to any environment by identifying the target/tracker pairs autonomously, facilitating many at the same instant.

5 Intelligent Space

In an era of 'intelligent objects', it is also possible to think about 'intelligent space', as proposed by Hashimoto et al. [3]. Intelligent space is an environmental system, which is able to support humans in informative and physical ways. Most of intelligent systems interact with humans in a passive space, but in intelligent space, an environment containing humans and artificial systems is capable of self-sufficient. Artificial systems like sensors, cameras, etc. become agents of intelligent space and simultaneously humans and artificial systems like robots become clients of intelligent space. Since the whole environment is an intelligent system, it is capable of monitoring and providing services to clients easily. Specific tasks that cannot be achieved by using only intelligent space or only humans are accomplished by utilizing its clients. For an example, intelligent space provides computer monitors to give information to the humans while robots are utilized to give physical services to humans. Humans as well as robots are supported by intelligent space as it is required. Consider a situation where a human is lacking in vision sensors to navigate around an intelligent space. At such instance he can get the necessary guidance from the intelligent space very easily. The ultimate goal of intelligent space is to accomplish an environment that comprehends intentions of humans and to cater for them [4]. In addition to guide blind people, there are numorous other applications of intelligent space.

Possible applications of intelligent space are as follows:

- Warehouses that can track inventory, the use of machinery, and the movement of people and objects in and around a busy workplace,

- Homes that can monitor their own environment, including the movement of people, their safety and well being, and the presence of any potentially dangerous situations,

- Shipping companies that can monitor the movement of freight, the current location of all trucks and rail cars, and how all assets are being used and treated,

- Stores that know who their customers are as they walk through the door, and can respond with the product offers and type of customer assistance tailored to that person's preferences and buying habits assignment to different cluster.

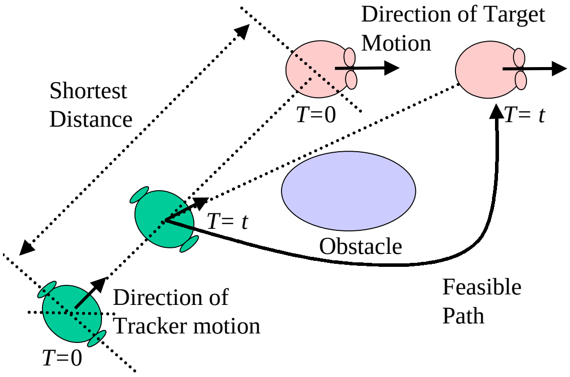


Figure 3: Feasible path planning

6 Feasible Path Planning

As Chung et al. [13] and Lin et al. [14] described, tracking an object is required to guide the tracker in shortest and safest possible path. But when the target is in motion, it may not be possible to achieve the shortest possible path in every time due to many reasons such as static obstacles in the path, dynamic obstacles moving towards the path, etc. At such times, safest and shortest possible paths should be compromised. As shown in Figure 3, at $T = t$, regardless of the shortest possible path, feasible (and shorter) path should be taken.

7 Decision Making

The ability to perform current situation analysis and appropriate decision making is an essential feature of many real-world applications. At this right moment, this ability is used in emergency medical treatment in intensive care units, air traffic navigation, and many other autonomous systems [15]. Existing formalisms and methods of inference have not been effective in real-time applications where tradeoffs between decision quality and computational tractability are essential. In practice, an effective approach to time-critical dynamic decision making should provide explicit support for the modeling of temporal processes and for dealing with time-critical situations.

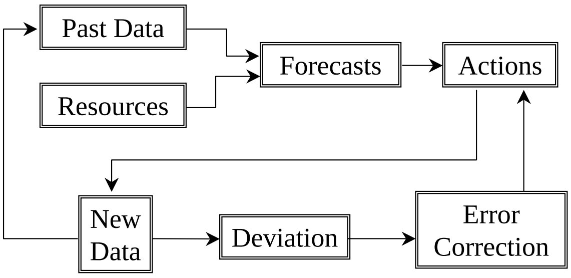


Figure 4: Forecasting

Almost all decisions are based on forecasts. Every decision becomes operational at some point in the future, so it should be based on forecasts of future conditions.

Forecasts are needed continually and throughout an operation, and they should certainly not be produced by an isolated group of data nor forecasting ever 'finished' (i.e. forcasting will be continued until the end of the operation). Forecasts are needed continually, and as time moves on, the impact of the forecasts on actual performance is measured, original forecasts are updated, and decisions are modified, and so on. This process is shown in Figure 4.

The decision-maker uses forecasting models to assist in decision-making process. The decision-making often uses the modeling process to investigate the impact of different courses of action retrospectively; that is, 'as if' the decision has already been made under a course of action. That is why the sequence of steps in the modeling process, in the above figure must be considered in reverse order. For example, the output (which is the result of the action) must be considered first.

7.1 Forecasting and prediction

Forecasting is a prediction of what will occur in the future, and it is an uncertain process [16]. Because

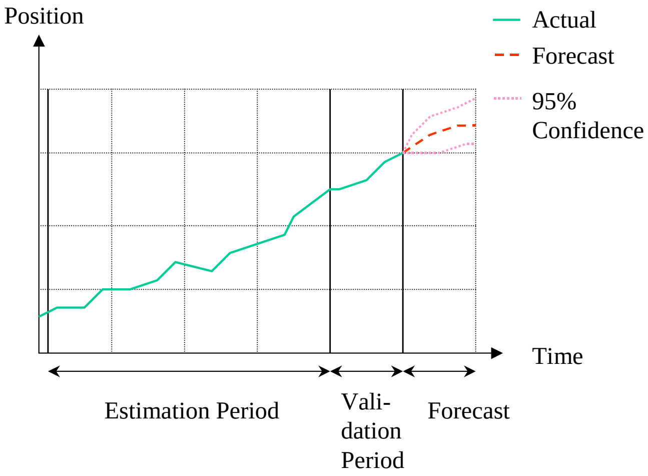


Figure 5: Graph of estimation, validation, and forecast

of the uncertainty, the accuracy of a forecast is as important as the outcome predicted by forecasting the independent variables $X_1, X_2, \cdots, X_n$. A forecast control must be used to determine if the accuracy of the forecast is within acceptable limits. Two widely used methods of forecast control are a tracking signal, and statistical control limits.

7.2 Modeling the causal time series

With multiple regressions, we can use more than one predictor. It is always best, however, to be parsimonious, that is to use as few variables as predictors as necessary to get a reasonably accurate forecast. The forecast takes the form:

$$Y = C_0 + C_1X_1 + C_2X_2 + ... + C_nX_n \qquad (1)$$

where C_0 is the intercept and $C_1, C_2, \cdots, C_n$ are coefficients representing the contribution of the independent variables $X_1, X_2, \cdots, X_n$. Since validation is used for the purpose of establishing a model's credibility, it is important that the method used for the validation is, itself, credible. Features of time series might be revealed by examining its graph with the forecasted values, and the residuals behavior.

An effective approach to modeling forecasting validation is to have a specific number of data points for estimation validation (i.e., estimation period), and a specific number of data points for forecasting accuracy (i.e., validation period). The data, which are not held out, are used to estimate the parameters of the model, the model is then tested on data in the validation period, if the results are satisfactory, and forecasts are then generated beyond the end of the estimation and validation periods. As an illustrative example, the graph of Figure 5, depicts the above process.

In general, the data in the estimation period are used to help to select the model and to estimate its parameters. Forecasts into the future are 'real' forecasts that are made for time periods beyond the end of the available data. The data in the validation period are hold out during parameter estimation. One might also withhold these values during the forecasting analysis after model selection, and then one-step-ahead forecasts are made. A good model should have small error measures in both the estimation and validation periods, compared to other models, and its validation period statistics should be similar to its own estimation period statistics. Holding data out for validation purposes is probably the single most important diagnostic test of a model. It gives the best indication of the accuracy that can be expected when forecasting the future. It is a rule-of-thumb that should have at least 20% of data for validation purposes.

8 Fuzzy-Logic in Real Environment

Guiding and tracking system in real environment should include artificial intelligence due to the facts that the knowledge needed to have a solution is incomplete, the sources of the information may be unknown at the current time that the solution is achieved, the environment might be changing and cannot be anticipated through analytical design, etc. [17, 19]. Artificial intelligence encompasses a number of technologies and fuzzy logic is a very successful approach, as shown in Mamdani [18].

Consider the motion as given in Figure 3, where there is a tracker following a target. Although it seems to be very simple as it looks (i.e. to travel the 'shortest distance' to reach the target), it may also be complex when considered in real environment. As it shows at a specific time (at $T = 0$), next instance (at $T = t$) may need decision making. Reasons such as the target moves in a peculiar pattern or there are static objects in the path or other dynamic objects in the vicinity coming closer or due to many other reasons, etc. may forced the tracker to withdraw taking the shortest point to point distance to the target.

In order to find the feasible path, not only the current position of an obstacle, but as well as the

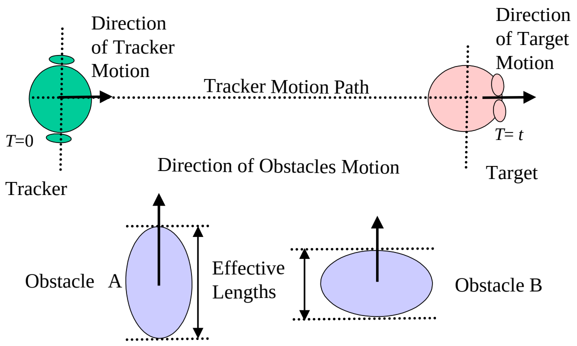


Figure 6: Effective length of an obstacle

direction of its motion (in case of a dynamic obstacle) should be analyzed. This can be simply explained as follows. Consider the obstacle in Figure 3 again. The feasible path, as shown in the diagram, will be valid only if the obstacle is static. But what happens if it is in motion, say, in a direction of up or down? If it moves as such, feasible path may still be the shortest distance, since at the time of the tracker coming to the position of the obstacle (where the obstacle was), obstacle may have already moved to a new position which may be far away from the shortest distance path permitting the tracker to go straight ahead.

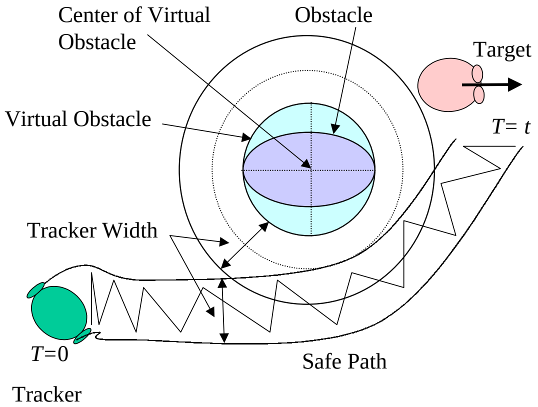


Figure 7: Generalization to nearest shape

9 Re-shaping of Obstacles

When considering obstacles around vicinity, it is necessary to get an idea of their sizes, distance, and rate of change of distance. In order to avoid any possible collision, tracker motion should be adjusted accordingly. In order to estimate the size of an obstacle, there should be a proper mechanism to convert any shape into a generalized shape for simplicity. In most of the situations, this will be a must since only that matters to a collision free path is obstacle's *'effective length'* with reference to the target tracker path. This is shown in Figure 6. Although it is always better to identify an object as its true shape (in order to inform the user about the nature of the obstacle such as, *'there is a moving dog in front of you'* etc.), it may be also acceptable to warn *'there is a big obstacle moving towards, turn little left and slow down'* as it alone will handle the situation. (But in a latter stage, it may be advisable and desirable to identify the obstacle itself as well).

For such generalization, convert to the nearest circular shape (in a case of 2D analyzing) or to the nearest cylindrical shape (in case of 3D analyzing) will be useful as shown in Figure 7.

10 Simulated Work

Once obtained the system data (current position, velocity, and rate of change of velocity, that is, acceleration) for a specific instance from the binocular vision system, decision-making process starts. This includes feeding the data thus obtained into a fuzzy logic system for a next instance position prediction of the dynamic objects and planning shortest and safest path for the tracker (blind person) to move. One such decision-making system to forecast tracker motion is analyzed, having two input variables 'Target motion' and 'Distance to target' with one output variable 'Tracker motion'. Some of the graphs, membership function plot for input variable 'Target motion' (Figure 8), firing of fuzzy rules for some input values (Figure 9), and surface graph for the system (Figure 10) are shown.

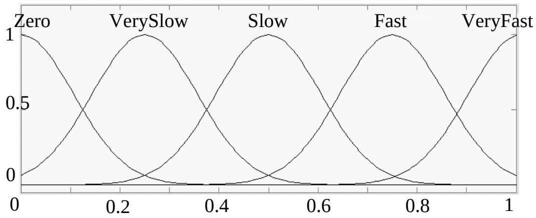


Figure 8: Membership functions of target motion

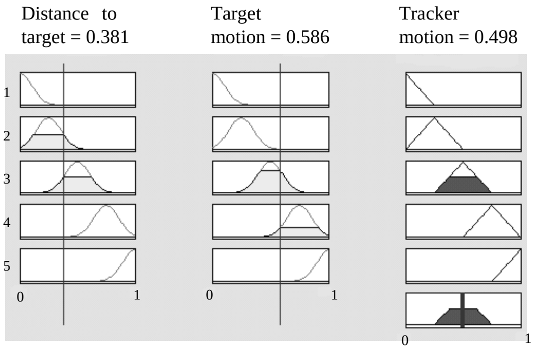


Figure 9: Firing of fuzzy rules

Initially, output variable 'Tracker motion' is categorized into three membership functions such as 'Fast', 'Slow', and 'Zero' with respect to the motion of the target and obstacles in the vicinity. But once applied the system to the real world, it may be necessary to fine-tune the speed so as to compensate many environmental conditions such as turning in corners, ground conditions, etc. These with many other similar conditions are currently under consideration.

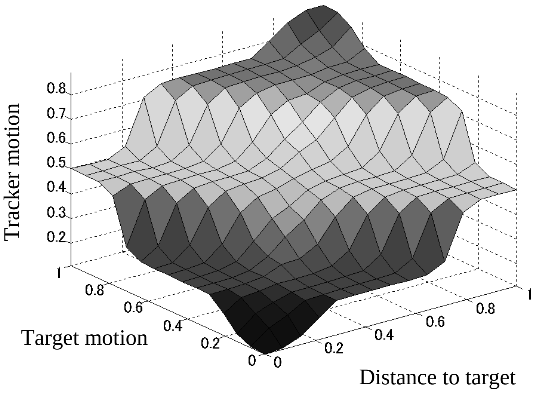


Figure 10: Surface view of the tracker motion fuzzy system

11 Summary

Attempt of implementing a fuzzy decision-making system for guiding a blind person in ubiquitous environment using binocular vision system has been presented. Though much work has to be done in the future in this regard, current simulations gave encouraging results for a better navigation system. Safe and quick path planning was an essential part in guiding blind people. There exist many approaches and currently there are many researchers working in similar fields to find a feasible solution. Instead of making each guiding unit be intelligent, it may be better to have a central intelligent unit that will support many people at the same time. This will help to reduce the cost of planning and constructing each guiding unit and help to maintain a standard among such systems. Application of a wireless media like 'Bluetooth Wireless Technology' may make the implementation be more and more easier.

References

[1] R. C. Luo, T. M. Chen, and K. L. Su, "Target tracking using hierarchical grey-fuzzy motion decision-making method," *IEEE Transactions on Systems, man, and Cybernetics - Part A: Systems and Humans*, vol. 31, no. 3, May 2001, pp. 179–186.

[2] R. C. Luo and T. M. Chen, "Target tracking by grey prediction theory and look-ahead fuzzy logic control," in *Proc. IEEE Int. Conf. Robotics and Automation*, vol. 2, Detroit, MI, USA, May 1999, pp. 1176–1181.

[3] H. Hashimoto, P. T. Szemes, and K. Morioka, "Intelligent space — Robots and spaces —," in *Proc. International Symposium on Electronics for Future Generations*, Tokyo, March 2004, pp. 383–388.

[4] H. Hashimoto, J. H. Lee, and K. Morioka, "Human robot interaction via intelligent space," in *Proc. International Conference on Control, Automation and Systems (ICCAS2002)*, Jeonbuk, Korea, October 2002, pp. 512–517.

[5] D. J. Coomb, "Real time gaze holding in binocular robot vision," *Technical Report, University of Rochester*, Dept. of Computer Science, 1992.

[6] J. Spletzer and C. J. Taylor, "Sensor planning and control in a dynamic environment," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA '02)*, vol. 1, May 2002, pp. 676–681.

[7] S. A. Guirnaldo, K. Watanabe, and K. Izumi, "Enhancing awareness in cooperative robots through perceptual anchoring," in *Proc. of the 9th International Symposium on Artificial Life and Robotics (AROB)*, vol. 2, January 2004, pp. 523–526.

[8] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: Establishing a

common coordinate frame," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, Aug. 2000, pp. 758–767.

[9] I. Nourbakhsh, E. Hamner, D. Bernstein, K. Crowley, E. Porter, T. Hsiu, B. Dunlavey, E.M. Ayoob, M. Lotter, S. Shelly, A. Parikh, and D. Clancy, "The Personal Exploration Rover: The Ground-up Design, Deployment and Educational Evaluation of an Educational Robot for Unmediated Informal Learning Sites," *Tech. Report CMU-RI-TR-04-38*, Robotics Institute, Carnegie Mellon University, December 2004.

[10] E. Hamner, R. Gockley, E. Porter, and I. Nourbakhsh, "The Personal Rover Project: The comprehensive design of a domestic personal robot," *Robotics and Autonomous Systems: Special Issue on Socially Interactive Robots*, vol. 42, no. 3–4, March 2003, pp. 245–258.

[11] A. C. Schultz and J. J. Grefenstette, "Using a genetic algorithm to learn behaviors for autonomous vehicles," in *Proc. of the American Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation and Control Conference*, Hilton Head, SC, August 1992. http://citeseer.ist.psu.edu/alan92using.html

[12] W. Burgard, A. B. Cremers, D. Fox, D. Hahnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "The interactive museum tourguide robot," in *Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI 1998)*, Madison, USA, 1998, pp. 11–18.

[13] J. Chung and H. S. Yang, "Fast and effective multiple moving targets tracking method for mobile robots," in *Proc. IEEE Int. Conf. Robotics and Automation*, Nagoya, Japan, vol. 3, May 1995, pp. 2645–2650.

[14] Z. Lin, V. Zeman, and R. V. Patel, "On-line robot trajectory planning for catching a moving object," in *Proc. IEEE Int. Conf. Robotics and Automation*, Scottsdale, AZ, vol. 3, 1989, pp. 1726–1732.

[15] B. R. Chang and S. F. Tsai, "An intelligent prediction method for short-term time series forecast on engineering education," in *Proc. International Conference on Engineering Education*, Manchester, U.K. August 2002.

[16] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, Prentice-Hall, New Jersey, USA, 1994.

[17] M. Lopez, F. J. Rodriguez, and J. C. Corredra, "Fuzzy reasoning for multi-sensor management," in *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics*, Vancouver, Canada, vol. 2, October 1995, pp. 1398–1403.

[18] E. H. Mamdani, "Application of fuzzy algorithm for control of simple dynamic plant," *Proc. IEEE*, vol. 121, no. 12, 1974, pp. 1585–1588.

[19] M. A. Teixeira, G. Zaverucha, V. N. A. L. da Silva, and G. F. Ribeiro, "Fuzzy bayes predictor in electric load forecasting," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, vol. 4, 2001, pp. 2339–2342.

[20] Y. P. Huang and C. C. Huang, "The integration and application of fuzzy and grey modeling methods," *Fuzzy Sets and Systems*, vol. 78, 1996, pp. 107–119.

[21] T. Brown, *Stereoscopic Phenomena of Light and Sight: A Guide to the Practice of Stereoscopic Photography and its Relations to Binocular Vision*, Reel Three-D Enterprises Publications, 1994.

[22] E. Paulos and J. Canny, "Delivering real reality to the world wide web via telerobotics," in *Proc. of IEEE International Conference on Robotics and Automation*, vol. 2, April 1996, pp. 1694–1699.

[23] D. Schulz, W. Burgard, A. B. Cremers, D. Fox, and S. Thrun, "Web interfaces for mobile robots in public places," *IEEE Robotics and Automation Magazine*, vol. 7, no. 1, March 2000, pp. 48–56.

[24] B. R. Miller and C. Bisdikian, *Bluetooth Revealed: The Insider's Guide to an Open Specification for Global Wireless Communication*, 2nd Edition, Prentice Hall, 2001.

The EPSO Algorithm : An Efficient Global Search Technique for Neuro Evolution

S. Jayatilaka[1] and R.Weerasinghe [2]

Department of Computation & Intelligent Systems
University of Colombo School of Computing (UCSC)
37, Reid Avenue, Colombo 7,
Sri Lanka

[1]sudeepaj@yahoo.com and [2]arw@ucsc.cmb.ac.lk

Abstract

The backpropagation algorithm has been widely used to train feed forward neural networks. However being a gradient decent local search technique it often yields sub optimal solutions having being trapped in local minima in the fitness landscape. Research effort has recently been expended on training neural networks through global search techniques such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO).

GA and PSO are mutually distinctive population based paradigms that are based on radically different approaches to explore the problem space. Thus it is interesting to see the hybrid performance of both of these techniques in neural learning.

This paper proposes a novel hybrid algorithm of GA and PSO called EPSO (Evolutionary Particle Swarm Optimization) as a powerful global search technique to evolve the weight set of a feedforward neural network The performance of the technique has been demonstrated through a frequently used benchmark problem for learning algorithms and compared to that of GA and PSO. The results show that the EPSO algorithm outperforms the sole performances of GA and PSO on our benchmark problem.

1. Some background

Neural networks are statistical models of cognition that are capable of learning decision surfaces in pattern space. They have been used to solve a range of problem classes such as classification, clustering, regression and optimization. Some basic material on neural networks can be found in [1].

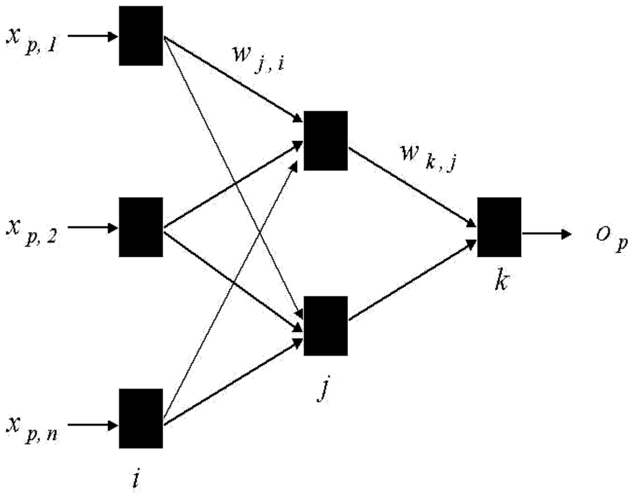


Figure 1.1: A feedforward neural network

Figure 1.1. depicts a famous neural network model called the feedforward architecture in which the processing elements are arranged as vertices of an acyclic directed graph.

1. 1. Training neural networks through error backpropagation (BP)

The BP algorithm [2] has been widely used as the learning algorithm for feedforward neural networks. Feedforward neural networks trained by BP are frequently referred to as multilayer perceptrons. It has been shown that multilayer perceptrons are universal function approximators [3].

The learning process of a multilayer perceptron through BP proceeds by modifying network weights to minimize the mean squared error between the desired and actual outputs of a network in a direction that corresponds to the negative gradient of an error measure.

The Backpropagation Algorithm

```
start with randomly chosen weights
while (mse is unsatisfactory & computational bounds not exceeded) do
for each input pattern $x_p$, $1\le P \le P$,
compute hidden node inputs ($net^{(1)}_{p,j}$)
compute hidden node outputs ($x^{(1)}_{p,j}$)
compute inputs to output nodes ($net^{(2)}_{p,k}$)
compute network outputs ($o_{p,k}$)
compute error between ($o_{p,k}$) and desired output ($d_{p,k}$)
update weights between hidden and output nodes according to:
$\Delta w^{(2,1)}_{k,j} = \eta(d_{p,k} - o_{p,k})S'(net^{(2)}_{p,k})x^{(1)}_{p,j}$
modify weights between input and hidden nodes according to:
$\Delta w^{(1,0)}_{j,i} = \eta \sum_{k} \left((d_{p,k} - o_{p,k})S'(net^{(2)}_{p,k})w^{(2,1)}_{k,j}\right) S'(net^{(1)}_{p,j})x_{p,i}$
end ( for)
end ( while)
```

1. 2. The evolutionary computing paradigm as a substitute for BP learning

The problem of deciding on an optimal set of weights in an artificial neural network has been shown to be NP complete. Consequently there has been a recent interest in utilizing global search paradigms such as evolutionary algorithms (EA) to evolve one or more aspects of neural networks in order to obtain better convergence. These aspects include the weights, structure & the learning algorithm [4], [5] and [6].

1. 2. 1. Advantages of EA over BP

BP applies the steepest decent method to update the weight vector. Subsequently it converges slowly and often produces sub optimal solutions [7]. On the other hand EC techniques perform global search and prove capable of finding global minima on the function landscape.

The BP algorithm presumes the presence of a differentiable transfer function. However EA are not bound by such a limitation and even can be used with discontinuous transfer functions, like the step function. Furthermore, EA can be used to train networks having different transfer functions, some thing which is not possible in the case of BP [4].

In contrast to BP, EA can be used across many network topologies and in circumstances where gradient or error information is not available.

Further the fitness function of an evolutionary neural network need not be differentiable and can be defined in a way appropriate to the problem at hand.

The generalization capability of evolutionary neural networks has been shown to be superior to that of human designed ones trained with error BP [8].

1. 2. 2. Disadvantages of EA against BP

Evolutionary paradigms like GA perform global search (exploration) quite satisfactorily but perform much poorly in local search (exploitation) [5].

Further it is important to choose an appropriate chromosomal representation scheme in order to encode the weights. The ordering of the weights in a chromosome is particularly important if the EC algorithm incorporates crossover or recombination operations.

Operators and their parameter values of an EC algorithm need to be selected in a suitable way for the problem.

EA such as GA in practice may perform much slower than gradient descent BP because of their higher computational complexity. However contradictory results have been reported [9], and hence it is our conclusion that the comparative performance rather depends on many factors such as parameter values of the algorithm and the nature of the problem space.

The crossover operator in an evolutionary algorithm might produce child chromosomes, which are not closer to any of their parents with respect to fitness. Thus EA like GA, which uses crossover, may bring problems of convergence on multimodal function surfaces [9].

1. 2. 3. GA as an EC paradigm

Formally theorized and introduced by John Holland in 1962, GA forms a major sub field of computational intelligence [10].

Genetic algorithms simulate the *survival of the fittest* principle of Darwinian evolution. They maintain a population of solutions and evolve the population by applying genetic operators such as crossover, mutation, inversion, and selection. The evolution process results in a better population of solutions at each generation. For a thorough introduction to GA see [11].

Steps of a Genetic Algorithm

```
begin GA
  g:=0  % generation counter
  Initialize population P(g)
  Evaluate population P(g)
  while not done do
    g:=g+1
    Select P(g) from P(g-1)
    Crossover P(g)
    Mutate P(g)
    Evaluate P(g)
  end while
end GA
```

GA for neural learning

GA for neural learning has a rich collection of literature. Several examples of training neural networks using genetic algorithms could be found in [12], [13], and [14].

[15] contains an interesting comparative study on the performances of GA, BP and their hybrid [see section 1. 4. 5.] on several benchmark problems.

1. 3. Particle Swarm Optimization (PSO) & Neural Networks

1. 3. 1. An overview of PSO

PSO is a population based collaborative optimization technique introduced recently by James Kennedy & Russell Eberhart [16]. The PSO algorithm has been inspired by the group dynamics of cultural animals such as bird flocks in search for food.

The philosophy behind PSO has its roots in social psychological theory, which suggests that members are influenced in socio cognitive space by past experience and successful behavior of others [18], [19]. Gradually the members of the population shift towards optimal regions in socio cognitive space.

The potential solutions in PSO are called particles, which fly through the problem space as influenced by their previous best positions and the position of the best member in the population.

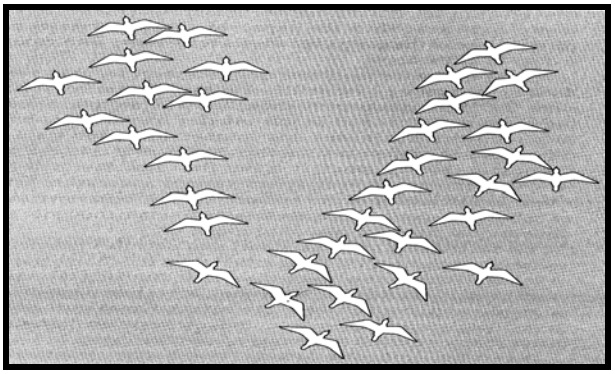


Figure 1. 2. : Choreography of a bird flock in search for food [17].

1. 3. 2. The PSO algorithm

```
for each particle
    initialize particle
end
do
  for each particle,
      calculate fitness value
      if the fitness value is better than the best
      fitness value pBest in history,
      set current value as the new pBest
  end
  choose the particle with the best fitness value
  of all the particles as the gBest
  for each particle,
      calculate particle velocity using
        $v_t[\,] = v_{t-1}[\,] + \varphi_1(\text{pbest}[\,] - \text{position}_{t-1}[\,]) + \varphi_2(\text{gbest}[\,] - \text{position}_{t-1}[\,])$
        ascertain $v_t \in (-v_{max}, +v_{max})^{\dagger}$
        update particle position using
        $\text{position}_t[\,] = \text{position}_{t-1}[\,] + v_t[\,]$
  end
while total number of iterations or error criteria is not
attained
```

Comments

v [ ] is the velocity that directs the flying of the particles.

position [ ] is the particle position.

φ_1 and φ_2 are control parameters.

† Particle velocities on each dimension are clamped to a maximum velocity. This helps the trajectory of the swarm to lie within finite bounds of a useful region in the solution space thus avoiding explosion.

1. 3. 3. PSO for neural learning

PSO has been emerged as an alternative to GA for neural learning in the recent years. Computer numerically controlled milling optimization [20], Human tremor analysis [9], and State of charge estimation of a battery pack of an electric vehicle [21] are some examples. Iris flower classification is an example for a classification benchmark problem, which has been solved by PSO [1].

1. 4. A comparison of PSO & GA

Comparisons between GA and PSO can be found in [22] and [23].

1. 4. 1. Similarities between PSO & GA

Both GA and PSO are naturally inspired stochastic global search techniques, which do not guarantee success. They both proceed from a population of random solutions and evolve the population members through generations, searching for optima on the fitness landscape.

1. 4. 2. Differences between PSO & GA

The population members of the PSO system possess memory compared to the members of a genetic based system in order to store the previous best (pBest) values.

PSO and GA use different information sharing mechanisms among population members. The population members of a GA share information between each other, encouraging the population to move as a group towards optimal regions. The PSO algorithm however operates in a one way information sharing fashion, from the gBest to the rest of the membership.

GA and PSO are mutually distinctive population based paradigms that use radically different approaches to explore the problem space. PSO deploys collaborative learning among the population as opposed to competitive learning in GA, which are based on the *survival of the fittest* principle of evolution.

1. 4. 3. Advantages of PSO

Unlike GA, PSO can be easily implemented and only few parameters need to be adjusted. PSO is also less complex and computationally efficient than GA. Furthermore, it has been observed that PSO operates an order of magnitude faster than the traditional genetic based approaches on benchmark problems. Compared to GA, the weights of a neural network can be represented in a more straightforward way in PSO.

GA could theoretically reach any point in the problem space through mutations. However because the probability of survival gradually decreases by time as the population fitness increases, mutated chromosomes rarely survive selection. Subsequently, GA has a significant chance of not finding an optimal solution in practice. PSO on the other hand is very much tolerant to get caught in local minima because, the particles move in a stochastic oscillatory trajectory exploring the problem space more thoroughly, given enough iterations and the right parameter values [9].

It has also been shown that PSO performs quite satisfactorily on multimodal problem surfaces. See figure below.

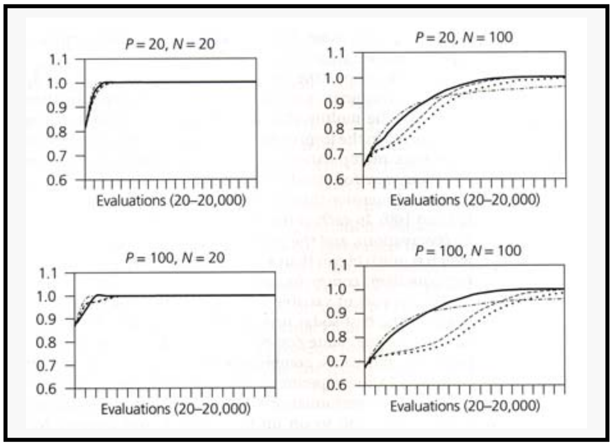


Figure 1. 2. : Performance of PSO (---) against evolutionary algorithms on multimodal function surfaces [9]

Thus PSO seems to overcome the mentioned drawbacks of evolutionary computing approaches [see section 1.2.2.] in the context of neural learning.

1. 4. 4. Drawbacks of PSO

PSO is a recent search paradigm, which is still in its juvenile years of maturity. More research needs to be done in order to further our understanding of the particle swarm behavior.

1. 4. 5. Hybrids of PSO & GA

Several authors have predicted that hybrids of PSO and GA could lead to better performance than PSO or GA alone [22], [23].

Selection has been incorporated to the PSO algorithm by [24] in such an early attempt. This has been done by replacing the worst half of the membership by the replicas of the best half of the membership at each round. In social psychological terms this would simulate some thing like the replacement of deviant members of a group by conformant ones. Encouraging results have been obtained on the functions of the De Jong test suite by this new technique. The same philosophy has been used for distribution state estimation in [21], producing more accurate results than when the PSO algorithm is used alone.

Another hybrid particle swarm algorithm has been introduced in [25], based on the ideas of breeding and subpopulations.

Breeding has been incorporated in to the main loop of the PSO algorithm as shown below.

```
while (not terminate condition) do
    begin
        evaluate
        calculate new velocity vectors
        move
        breed
    end
```

Breeding has been done by selecting two parent particles randomly from a pool and then performing arithmetic crossover between them to produce the offspring particle positions for the next iteration. The velocity vector of the offspring population is calculated similarly by using arithmetic crossover between the parent velocity vectors.

The above model has been further extended by incorporating the notion of subpopulations to it. Particles now breed with different probabilities in the native subpopulation and between foreign subpopulations. This helps diversity to grow in the population by restricting the gene mix, which in turn would help the system to escape from sub optimal solutions.

Literature reports hybrids of GA and PSO in which they have been used one after the other for optimization tasks. This has been done by using the population of the first algorithm as the initial population for the second algorithm, when performance level offs [26]. In this approach the PSO followed by GA hybrid has yielded the best performance while the GA followed by PSO hybrid has only been superior to GA alone.

2. The Evolutionary PSO (EPSO) Algorithm

GA and PSO are based on mutually distinct cognitive philosophies and employ radically different approaches to explore the problem space. Thus it is interesting to see their hybrid performance on neural learning.

This section proposes a novel hybrid algorithm of GA and PSO called EPSO for neuro evolution. The hybrid algorithm has been implemented by optimizing the population members of the PSO algorithm by a genetic search at each iteration to produce the swarm of the following iteration. The hybrid has been implemented and tested in steps.

In the first step the PSO algorithm was hybridized using a selection operator. Here the best half of the population was cloned to replace the worst half of the population at each iteration. This simulates the metaphor of least fit individuals being replaced by fitter ones in a social network.

In step 2 we incorporated both selection and crossover. The concept was inspired by the Lamarckian philosophy of evolution. Lamarckism was the prevailing belief of evolution before Darwin. Lamarck believed that evolution came about through the inheritance of acquired characteristics. For instance a body builder would transmit his muscular features to his children according to Lamarck. However, this is not how it happens in nature.

The PSO algorithm searches for optima in the phenotype space (i.e. good behaviors are learnt by others). On the other hand a GA operates on the genetic composition of an individual (genotype). Our algorithm transforms the phenotypes of the population members of PSO (swarm) to genotypes on which a genetic search is performed. The genotypes are in turn expressed to form the phenotypes of the swarm of the following iteration. Thus this forms an unbroken cycle of consecutive phenotype ↔ genotype transformations and brings together the power of two different search techniques into a single algorithm.

The performance of EPSO has been demonstrated through a frequently used benchmark problem for learning algorithms.

3. The benchmark problem used

The Iris dataset [27], [28] has been used as the benchmark to demonstrate the performance of the algorithms. It is a frequently used benchmark for learning algorithms. The dataset consists of 150 patterns of Iris flowers belonging to 3 species and the idea is to classify the flowers using 4 attributes : sepal length, sepal width, petal length, and petal width.

4. Experiments

EPSO was tested along with its constituent GA and PSO algorithms on the Iris benchmark. The objective was to study the comparative performances of the techniques being investigated. Our hypothesis was that EPSO would outperform GA and PSO on this problem.

EPSO algorithm

```
for each particle
    initialize particle
end

do

  for each particle,
      calculate fitness value
      if the fitness value is better than the best
fitness value pBest in history,
set current value as the new pBest
  end

choose the particle with the best fitness value
of all the particles as the gBest

for each particle,
    calculate particle velocity using
$v_t [\ ] = v_{t-1} [\ ] + \varphi_1 ( pbest [\ ] - position_{t-1} [\ ] ) + \varphi_2 ( gbest [\ ] - position_{t-1} [\ ] )$

ascertain $v_t \in (- v_{max} ,+ v_{max} )^{\dagger}$

update particle position using
$position_t [\ ] = position_{t-1} [\ ] + v_t [\ ]$
  end

% Phenotype to genotype mapping

Select particles for breeding
Produce offspring through crossover

% Genotype to phenotype mapping

Assign offspring to the swarm

while total number of iterations or error
criteria not attained
```

4. 1. Experiment setup

The dataset for the experiment comprised of 150 patterns partitioned into a training set of 100 patterns and a test set of 50 patterns while the neural network consisted of 4 inputs, 3 outputs and 5 hidden nodes.

4. 2. Experiment strategy

Two approaches were taken to report performance. In the first approach the number of iterations and the time taken to converge was reported whereas in the second approach the performance accuracy was reported after a certain number of iterations. This way, both the speed of learning and the performance accuracy of the algorithms could be apprehended.

Under each approach we ran each algorithm 10 times using mean square error as the performance measure and reported the (1) best (2) mean and the (3) worst of the performance. A 'winner takes all' approach was adopted to interpret the real valued results obtained as outputs, in order to determine the species to which a particular flower belongs.

4. 3. Parameter values

Table 1 contains the parameter values used in the algorithms. See Appendix 1 for a justification of these. Parameter values common to both GA and PSO are remained the same in order to facilitate comparison.

Parameter	**GA**	**PSO**	**EPSO**
Error goal	0.02	0.02	0.02
Maximum epochs	1000	1000	1000
Population size	50	50	50
Representation	Real	Real	Real
Selection technique	Roulette wheel	--	Roulette wheel
Generation gap	0.5	--	0.5
Crossover type	Single point	--	Single point

Crossover rate	0.7	--	0.7
V_{max}	--	2	2
Cognitive acceleration	--	2	2
Social acceleration	--	2	2
Constriction coefficient	--	1	1

Table 1 : Parameter values

4. 4. Experiment environment

All the experiments were performed using Neural Network Toolbox 4 of MATLAB 6. 5. 1. on a 512 MB 2378 MHz dual processor machine running Windows XP professional. The Genetic Algorithm Toolbox version 1. 2. [URL1] has been used for the experiments with genetic algorithms.

5. Results

Performances of different population based algorithms on the IRIS dataset are shown in the following tables.

Measure	GA	PSO
Average mean square error	0.1315797	0.01124578
Mean population fitness	0.0990086	0.05196525
Average learning accuracy (%)	79.45 %	98.3 %
Average test accuracy (%)	78.3 %	94.8 %
Epochs (best case)	67	134
Epochs (average case)	Did not converge	378.35
Epochs (worst case)	Did not converge	1000
Mean CPU time (sec.)	∞	31.41

Measure	EPSO 1	EPSO 2
Average mean square error	0.010067187	0.009932547
Mean population fitness	0.02098565	0.0243499
Average learning accuracy (%)	98.2 %	98.15 %
Average test accuracy (%)	96 %	96 %
Epochs (best case)	39	31
Epochs (average case)	129.7	137.95
Epochs (worst case)	297	237
Mean CPU time (sec.)	16.03	17.22

6. Discussion

Firstly it is important to note that PSO outperforms GA in all the aspects. GA's average learning accuracy of 78.3 % has been improved to 94.8 % with PSO. Further, PSO seems to converge well when compared to GA, which did not converge in several runs. Speaking on the hybrids, EPSO 1 outperforms PSO in all the criteria measured. With EPSO 1 the average test accuracy has been improved to 96 % and the number of epochs and convergence time have been brought down to half the original values. EPSO 2 did not show an improvement over EPSO 1 however.

7. Conclusion

Based on the above discussion it is evident that EPSO outperforms standard versions of PSO and GA on the Iris problem.

8. Future work

As future work, it is interesting to see other ways of forming hybrids of GA and PSO. Some of the other directions would be to try out subpopulations and migration as variants of GA and different neighborhood topologies as variants of PSO. It is also interesting to see hybrids of PSO and other search techniques such as simulated annealing and taboo search. Techniques discussed in this dissertation evolve weights of a neural network. It is also interesting to evolve other aspects such as topology, transfer function and inputs.

9. Acknowledgements

The first author of this paper thanks all who contributed in order to produce this publication including the rest of the authors.

References

[1] Eberhart, R. C., Simpson, P. K., and Dobbins, R. W. (1996). *Computational Intelligence PC Tools.* Boston: Academic Press.

[2] Gori, M., & Tessi, A. (1992). On the problem of local minima in backpropagation. *IEEE Trans. Pattern Analysis and Machine Intelligence, 14,* 76-85.

[3] Rumelhart, D. E., McClelland, J. L., and the PDP Group (1986). *Parallel Distributed Processing: Exploration in the Microstructure of Cognition. Vol. 1: Foundations.* Cambridge, MA: The MIT Press.

[4] Schaffer, J.D., Whitley, L.D., and Eshelman, L.J. (1992). Combinations of genetic algorithms and neural networks: A survey of the state of the art. In L. D. Whitley and J.D. Schaffer (Eds.), *COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks*, 1-37. Los Alamitos, CA: IEEE Computer Society Press.

[5] Yao, X. (1995). Evolutionary artificial neural networks. In A. Kent and J. G. Williams (Eds.), *Encyclopedia of Computer Science and Technology.* New York: Marcel Dekker.

[6] Fogel, D. (1998). Evolutionary computation: *The Fossil Record.* Piscataway, NJ: IEEE Press.

[7] Gori, M., & Tessi, A. (1992). On the problem of local minima in backpropagation. *IEEE Trans. Pattern Analysis and Machine Intelligence, 14,* 76-85.

[8] Schaffer, J. D., Caruana, R. A., and Eshelman, L. J. (1990). Using genetic search to exploit the emergent behavior of neural networks. In S. Forrest (Ed.), *Emergent Computation,* 244-248. Amsterdam: North Holland.

[9] Kennedy, J., and Eberhart, R. C. (2001). *Swarm Intelligence.* San Francisco, CA: Morgan Kaufmann.

[10] Holland, J. H. (1962). Outline for a logical theory of adaptive systems. *Journal of the Association for Computing Machinery, 3,* 297-314.

[11] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Reading, MA: Addison-Wesley.

[12] Whitley, D. (1989). Applying genetic algorithms to neural network learning. *Proceedings of the Seventh Conference of the Society of Artificial Intelligence and Simulation of Behavior,* 137-144. Sussex, England: Pitman Publishing.

[13] Montana, D. J., and Davis, L. (1989). Training feedforward neural networks using genetic algorithms. *Proceedings of the 11th Annual Joint Conference on Artificial Intelligence, 762-767.* San Francisco: Morgan Kaufmann.

[14] Seiffert, U. (2001). Multi layer perceptron training using genetic algorithms. *Proceedings European Symposium on Artificial Neural Networks 2001(ESANN2001),* 159-164. Bruges(Belgium).

[15] Alba, E., & Chicano, J, F. (2004). Training neural networks with GA hybrid algorithms. *Genetic and Evolutionary Computation Conference 2004 (GECCO – 2004).*

[16] Kennedy, J., and Eberhart, R. C. (1995). Particle swarm optimization. *Proceedings of the IEEE International Conference on Neural Networks, IV,* 1942-1948. Piscataway, NJ: IEEE Service Center.

[17] Kennedy, J., and Eberhart, R. C. (2001). *Swarm Intelligence.* San Francisco, CA: Morgan Kaufmann.

[18] Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice Hall.

[19] Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

[20] Tandon, V. (2000). Closing the gap between CAD/CAM and optimized CNC end milling. Master's thesis, Purdue School of Engineering and Technology, Indiana University Purdue University, Indianapolis, IN.

[21] Shigenori & others. (2001). Practical distribution state estimation using hybrid particle swarm optimization. *Proceedings of IEEE Power Engineering Society Winter Meeting.* Columbus, Ohio, USA.

[22] Eberhart, R. C., and Shi, Y. (1998). Comparison between genetic algorithms and particle swarm optimization. *Evolutionary Programming VII: Proceedings of the 7th Annual Conference on Evolutionary Computing.* Berlin, San Diego, CA: Springer-Verlag.

[23] Angeline, P. J. (1998b). Evolutionary optimization versus particle swarm optimization: Differences in philosophy and performance differences. In V.W. Porto, N. Saravanan, D. Waagen, and A.E. Eiben (Eds.), *Evolutionary Programming VII: Proceedings of the 7th Annual Conference on Evolutionary Programming.* Berlin: Springer-Verlag.

[24] Angeline, P. J. (1998a). Using selection to improve particle swarm optimization. *Proceedings of the 1998 International Conference on Evolutionary Computation,* 84-89. Piscataway, NJ: IEEE Press.

[25] Lθvbjerg, M., Rasmussen, T, K., and Krink, T. (2001). Hybrid particle swarm optimizer with breeding and subpopulations. *Proceedings Genetic and Evolutionary Computation Conference 2001 (GECCO – 2001).* San Francisco, CA: Morgan Kaufmann.

[26] Robinson, J., Sinton, S., and others. Particle swarm, genetic algorithm, and their hybrids: Optimization of a profiled corrugated horn antenna. *Dept. of Electrical Engineering, University of California, Los Angeles, California.*

[27] Anderson, E. (1935). The IRISes of the Gaspe Peninsula. *Bulletin of the American IRIS Society, 59,* 2-5.

[28] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7,* 179-188.

Web references

[URL1] http://www.shef.ac.uk/cgi-bin/cgiwrap/ ~gaipp /gatbx-download

[URL2] http://www.engr.iupui.edu /~eberhart/ web/ PSObook.html

An Intelligent System for Smart Thinking in Clinical Diagnosis

D.S.K Mendis[1], A. S. Karunananda[2] and U. Samarathunga[3]

[1]Department of Information Technology,
Advanced Technical Institute,
Labuduwa.
Sri Lanka
Tel: 0112625291
kalanaatil@mail.com

[2]Department of Mathematics & Computer Science,
Open University of Sri Lanka,
Nawala.
Sri Lanka
asoka@maths.ou.ac.lk

[3]Gampaha Wichramarchi Ayurveda Institute,
University of Kelaniya,
Yakkala,
Sri Lanka

Abstract

Computerized clinical guidelines can provide significant benefits to health outcomes such as classification and reasoning; however, their effective implementation presents significant problem in present mechanism of clinical diagnosis in Ayurveda medicine. Vagueness and ambiguity inherent in clinical guidelines is not readily available for formulating consistent methodology of diagnosis of patients in Ayurvedic medicine. Fuzzy logic in Artificial intelligence facilitates to formalize the treatment of vagueness in clinical diagnosis of Ayurvedic medicine. This paper discusses a novel methodology for clinical diagnosis in Ayurvedic medicine. We consider how fuzzy logic can be applied and give a set of classifications in clinical knowledge for addressing uncertainty in practice guidelines. We describe the specific applicability of expert system technology in Artificial Intelligence for reasoning in classifications derived from fuzzy classifications. The intelligent hybrid system has been tested and gained 87% accuracy and presented as a clinical diagnosis system for General Practitioners in Ayurvedic medicine

Key words: Clinical diagnosis, Fuzzy logic, Expert systems, Principle component analysis

1. Introduction

It has been investigated that Fuzzy logic has a history of application for clinical problems including use in automated diagnosis [1], control systems [2], image processing [3] and pattern recognition [4]. Liu and Shiffman [5] have demonstrated the application of fuzzy logic to model the imprecision of a published clinical practice guideline, which is cited by Zielstorff [6] as a promising direction for future development of computer-based decision support in medicine; however, the work is still quite far from revealing a general approach to applying fuzzy logic to clinical guidelines. Further Information overload in medicine has long been acknowledged and remedies sought [7]. One option is to devise medical expert system programs that reason for the doctor. A more modern approach is not to supplant but to support human reasoning (i.e., to build decision support systems, DSS). There fore such systems may still be expert systems in the sense of having highly sophisticated reasoning capabilities [7].

Ayurvedic medicine has proposed a different approach to western medicine for diagnosing of diseases and prescribes the treatments. In particular system of Ayurvedic medicine takes infidelity of patients into consideration. Stated in another way, treatments are based on individual's mental physical constitents. In this sense, Ayurveda postulates tenfold examination for classification of individuals [10]. Under tenfold examination, the *prakurthi pariksha* (examination of the constitution of the individual) is very important before prescribe the drugs. The *prakurthi pariksha* is subjective. By using a computation model evaluation has been computed objectively. The model avoids repetitions and inconsistent information. This paper provides the development of intelligent hybrid system, which emulates Ayurvedic knowledge of classification of individuals. The intelligent hybrid system developed can be used as both a diagnosis tool and a learning system for Ayurvedic medical students.

2. Clinical practice guidelines in Ayurvedic Medicine

Ayurvedic medicine has a very strong bearing on the concept of *Prakruti,* which means nature (natural form) of the build and constituents of the human body. According to Ayurveda the path to optimal health is different for people depending on their *Prakruti.* For individuals the *Prakruti* is defined as a combination of *(Vata, Pitta* and *Kaph*a). A balanced state of the *Prakruti* makes a healthy and balanced person (Physically and mentally). Since we all have different combinations of the *Prakruti,* The diagnosis of *Prakruti* offers unique insights into understanding and assessing one's health. It is not merely a diagnostic device but also a guide to action for good health. It assesses the, dominance of *Prakurti* and gives advice for preventive and primitive health care. The ancient science of Ayurveda is the oldest known form of health care in the world.

Recognition of human constituent in Ayurveda, is currently based on a standard questionnaire on subjective criteria based on ancient theories of Ayurvedic scholar *Charaka,* 1000 BC and *Susruta,* 600 BC. Questions in concerned are very much user-friendly and based on medical theories of Ayurveda, used for finding constituent type, has probes such as repeating questions and classification of constituent type. This has been used for classification of individuals for many centuries. In this research presented the subjective data in to objective measurements. By using Artificial Intelligence methods, the constituents of individuals illustrate as percentages which postulates a novel methodology for existing system of clinical guideline in Ayurveda.

3. Methods

We propose an Intelligent Hybrid system for developing an approach for modeling clinical knowledge as shown its procedure in figure 1. The intelligent hybrid system is involved with artificial intelligent techniques, namely fuzzy logic and expert system technology. We primarily used fuzzy logic together with statistical technique of principle component analysis for modeling tacit domains.

3.1 Fuzzy Logic

Fuzzy logic handles situations, where conclusions do not fall into one extreme. As compared with classical logic, fuzzy logic can handle real world problems, which deal with more than two truth-values. In fuzzy logic, everything is a matter of degree. Therefore fuzzy logic can be used to make decisions in domains with tacit knowledge [8]. Individual classification in Ayurveda is a classic example, where the decision has more than one possible truth-value.

In our research we have used fuzzy logic for addressing the vagueness involved in tacit knowledge relevant to clinical diagnosis. For example, vagueness (dependencies & inconsistencies) involved in Ayurvedic classification of individuals has been manipulated using fuzzy logic.

3.1.1 What is a Fuzzy Set?

A fuzzy set can be simply defined as a set with fuzzy boundaries. Let X be the universe of discourse and its elements be denoted as *x.* Fuzzy set *A* of universe *X* is defined by function $\mu_A(x)$ called membership function of set *A.*

$$\mu_A(x):\ X \longrightarrow [0.1]$$

This degree, a value between 0 and 1, represents the degree of membership, also called membership value, of element *x* in set *A.*

3.1.2 Fuzzy Rules

A fuzzy rule can be defined as a conditional statement in the form:
IF *X*= A THEN *Y*= B
Where *X* and *Y* are linguistic variables; and *A* and *B* are linguistic values determined by fuzzy sets on the universe of discourses *X* and *Y*, respectively.

3.2 What is Principle Components Analysis (PCA)?

The concept of PCA is based on the derivation of linear combinations of the *p* measured variables X_1, X_2..X_p to produce 'derived variables', that are uncorrelated and are such that explains a different 'dimension' within the data[3.] Such derived variables are referred to as principal components (PCs) [11]. As there are *p* response variables within the data set, *p* principal components can be derived. The first PC, denoted PC1, is expressed in the form

$$PC_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + ... + \alpha_{1p}X_p$$

Where the α terms refer to the weights of each variable within this principal component PC_1. The weights of each PC_i represent the eigenvector solution, which maximize the variance of each PC_i, where i is the number of components.

3.2.1 Extracting Principal Components (PC)

The importance of each PC, in terms of level of data variation explained, is specified by its eigen value, the λ term, with Σ λ representing the total of the *p* eigen values. A measure of the proportion of data variation accounted for by each PC, based on the equivalence of eigen value and PC variance, is provided by the expression λ/(Σ λ).

Generally, it is required to select those PCs, which account cumulatively for at least 80% to 90% of the data variation. In addition that each PC must exceed eigen value more than 1.However, if nearly all the correlations are less than 0.25, then there is probably not much point in carrying out a PCA. But to reduce even that much of interdependency PCs can be computed.

3.3 Expert Systems

An expert system is a computer program, which emulates the decision-making ability of a human expert [9]. As compared with other techniques, the expert systems are capable of giving alternative answers, explaining reasons for conclusions, handling incomplete information and uncertainty in answers. An expert system contains knowledge derived from an expert in some narrow domain. Over the last many decades, his technology has been successful for modeling the domains with explicit knowledge. However, there is a need for developing expert systems for domains with tacit knowledge too, since expert's knowledge is always not explicitly defined. Since the knowledge-modeling framework using fuzzy logic is integrated with an expert system, next we briefly describe major components of an expert system.

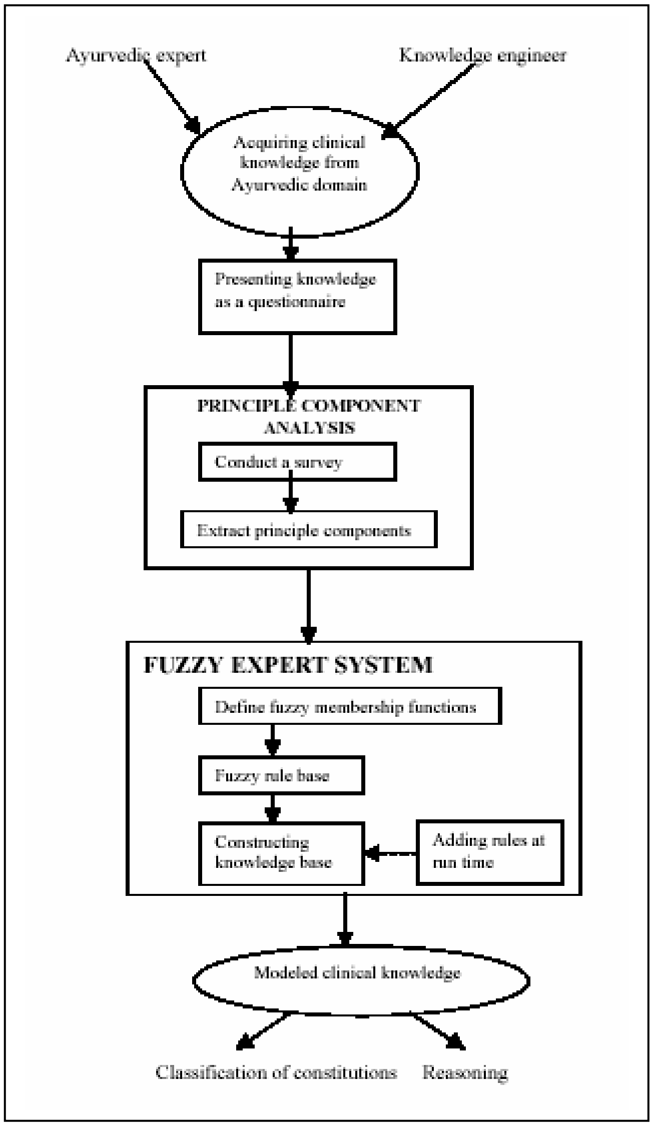

Ayurvedic expert
Knowledge engineer
Acquiring clinical knowledge from Ayurvedic domain
Presenting knowledge as a questionnaire
PRINCIPLE COMPONENT ANALYSIS
Conduct a survey
Extract principle components
FUZZY EXPERT SYSTEM
Define fuzzy membership functions
Fuzzy rule base
Constructing knowledge base
Adding rules at run time
Modeled clinical knowledge
Classification of constitutions
Reasoning

Figure. 1 - Clinical knowledge modeling procedure

3.3.1 Components of an Expert System

An expert system consists of these three components, namely, user interface, knowledge base and inference engine**.**

User interface facilitates the interaction between user and the expert system. This is most of the time a natural language interface.

Knowledge base consists of domain knowledge. This knowledge can be represented in the form of rules, diagrams, etc. Knowledge base is the knowledge reservoir of an expert system. Once we have modeled the tacit knowledge, such knowledge will also be stored in the knowledge base. Our expert system consists of set of fuzzy rules on Ayurvedic domain.

Inference engine explores the knowledge for finding answers. An inference engine consists of various problem-solving strategies including search techniques, and forward/backward chaining mechanisms. All special features of expert systems such as explanation ability, handling uncertainty are embedded in the inference engine of an expert system.

4. Results

In the exciting system, the method of analyzing constituents is not consistent. Although Ayurvedic practitioners use a questionnaire but leads several problems like dependencies among the questions in the questionnaire and analysis of the constituent type. We addressed these problems to solve using following stages.

4.1 PCA for removing dependencies

It is consisted of 72 questions to analyze *vata*, *pita* and *kapha.* We have done a pilot survey for 100 no. of students for statistical modeling using the questionnaire. Principle component analysis has been used to remove dependencies in the questionnaire. It has been identified 25 principal components using SPSS [25] as shown in matrix given below. Here V1, V2..V24, K1, k2..K24, P1, P2..P24 denotes question-numbering system in the questionnaire.

$$M = \left(\begin{array}{llrrrrr}
 & & 1 & 2\ .. & & 24 & 25 \\
V1 & & -0.228622 & 0.249362 & . & -0.073945 & 0.058179 \\
V2 & & 0.08431 & 0.20654 & . & -0.097192 & -0.112795 \\
. & V= & & & & & \\
. & & & & & & \\
V23 & & -0.645803 & 0.232312 & . & 0.0067 & -0.083959 \\
V24 & & -0.222147 & -0.06453 & . & -0.073514 & 0.084404 \\
K1 & & 0.012511 & -0.096332 & . & 0.141314 & 0.25113 \\
K2 & & -0.005642 & 0.268145 & . & -0.179992 & 0.111715 \\
. & K= & & & & & \\
. & & & & & & \\
K23 & & 0.409442 & 0.073812 & . & -0.115118 & -0.056431 \\
K24 & & 0.696973 & 0.126679 & . & 0.098213 & 0.045471 \\
P1 & & 0.430044 & 0.14608 & . & 0.023669 & 0.09045 \\
P2 & & 0.243781 & 0.373485 & . & -0.040468 & 0.149644 \\
. & & & & & & \\
. & P= & & & & & \\
P23 & & 0.009727 & 0.012529 & . & -0.072224 & 0.177827 \\
P24 & & -0.378091 & 0.096985 & . & 0.158006 & 0.069821
\end{array}\right)$$

Co-matrix computed in principle component analysis is given below.

$$V= \begin{pmatrix}
-0.228622 & 0.249362 & . & -0.073945 & 0.058179 \\
0.08431 & 0.20654 & . & -0.097192 & -0.112795 \\
 & & & & \\
-0.645803 & 0.232312 & . & 0.0067 & -0.083959 \\
-0.222147 & -0.06453 & . & -0.073514 & 0.084404
\end{pmatrix}_{24*25}$$

4.2 Fuzzy logic for clinical guidelines

Human constituents can be computed in to *vata, pita* and *kapha* in percentages as shown below. Membership functions for *vata, pita* and *kapha* have been constructed using fuzzy logic based on out puts of principle component analysis.

4.2.1 Membership function for classifying *Vata* constitution

Boundary values of membership function have been constructed using the output of the principle component analysis.

$$\because X_L = 1 \sum_{i=1}^{25} \sum_{j=1}^{24} a_{ji} = 8.510004 \qquad (1)$$

$$\because X_U = 6 \sum_{i=1}^{25} \sum_{j=1}^{24} a_{ji} = 51.06002 \qquad (2)$$

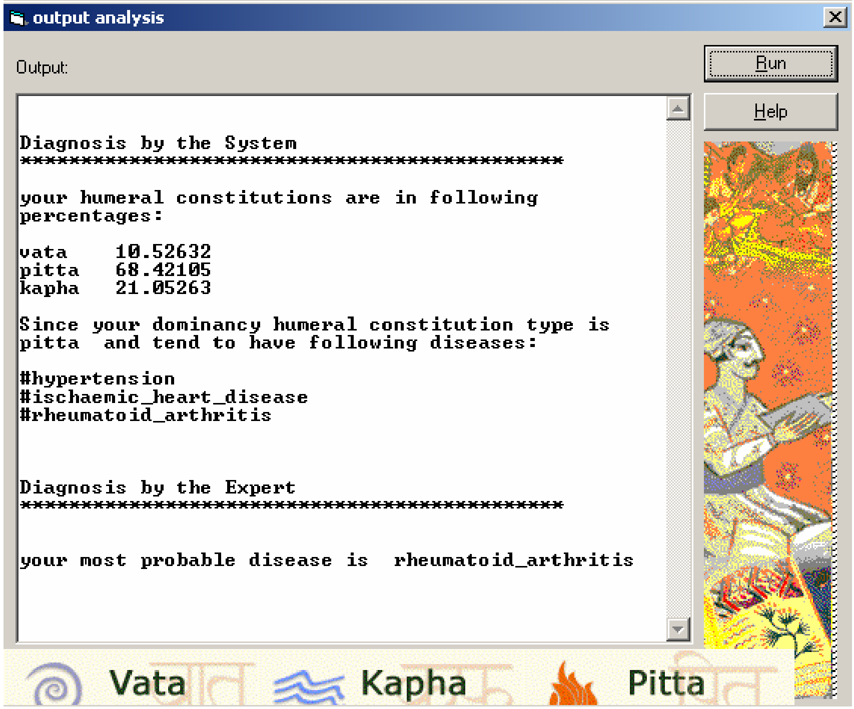
Here X_L denotes lower bound value at the minimum level of evaluation scale (Does not apply) in the questionnaire. X_U denotes upper bound value at the maximum level of evaluation scale (Applies most) in the questionnaire.

$$\mathbf{V\ (X)} = \begin{cases} \mathbf{0} & \mathbf{X=<X_L} \\ \mathbf{(X-X_L)/(X_U-X_L)} & \mathbf{X_L<X<X_U} \\ \mathbf{1} & \mathbf{X=>X_U} \end{cases}$$

V(x) denotes membership function for classifying *vata* constitution. This has been constructed using Visual Basic.

4.2.2 Expert system technology for reasoning

Explanations for output generated by the fuzzy logic have been processed using fuzzy rules in the knowledge base in the expert system. The knowledge base has been implemented using FLEX expert system shell [13], which embedded in WinProlog. In relation to Ayurvedic domain, possible diseases can be occurred due to dominated constituent type. It is illustrated as shown in Figure 2.



```
Diagnosis by the System
*********************************************

your humeral constitutions are in following
percentages:

vata    10.52632
pitta   68.42105
kapha   21.05263

Since your dominancy humeral constitution type is
pitta  and tend to have following diseases:

#hypertension
#ischaemic_heart_disease
#rheumatoid_arthritis


Diagnosis by the Expert
*********************************************


your most probable disease is  rheumatoid_arthritis
```

Figure. 2 - Explanation windows

In the first place we have tried statistical technique of Principle Component analysis for recognition of any dependencies among classification of individuals. We have used 100 participants for the gathering data. Although the experiment identified 25 principal components, it was revealed that those components are not statistically significant to consider. However, this decision does not match with the real world experience, as there were obvious miss match between conclusion through the questionnaire and the actual observations by Ayurvedic physicians. So, it appears that principle component analysis cannot handle the issue we have.

As such we decided to resolve the problem with the help of Artificial Intelligent techniques (AI). It is well known fact that AI techniques are better at solving real world problems, which cannot be solved otherwise. In particular AI techniques can be used to models domains with less formal knowledge. Among other AI techniques, we have used Fuzzy logic to fine tune the results obtained from PC analysis. Finally the system has been developed as an Expert System, which models Ayurvedic classification of individuals. With this technology the system has added features such as incorporating new knowledge, explaining reasons for answers given.

5. Dealing with uncertainty in clinical guidelines

The expert system developed using this approach was tested with a group of 35 persons of Ayurvedic experts and Ayurvedic general practitioners. The evaluation was conducted to see far the answers generated by the system matches with the identification by Ayurvedic experts and the general practitioners. Further, the system's ability to fine-tune the answers was also tested. It is investigated that 87% of conclusions matches with the system and expert using descriptive statistics. The system facilitated to derive constituents types in percentages while Ayurvedic experts obtain only the constituent type. As recommendation given by the Ayurvedic experts, determining constituent's types in percentages is an important criterion for prescribing drugs for a disease. Further, our system can be developed to provide as an option to find out possible diseases. In generally, the system can be used as a self-assessment for finding constituents. The human constituents can be computed by the system as a percentage of *vata, pitta* and *kapha*. So it would help to find the effectiveness of minimum constituent type.

6. Conclusion

The Intelligent Hybrid System developed can be used as a tool for supporting Ayurvedic medical practitioners for recognition of human constituents. The system is user friendly; therefore the ordinary persons to prevent the diseases can use it. Further, Ayurvedic medical students can use the system as a learning system. The users of the system are not expected to hold knowledge in statistical or artificial intelligence techniques. This system can also maintain history of patients for research related human constitutes. It should be noted that with the help of Artificial Intelligence technologies we have improved the correctness of the decision making process in relation to the use of traditional questionnaire. This eliminates the inconsistencies and repetitiveness of answers and also provides a means for explanation of reasons for answers. According to Ayurvedic medicine, the food and regiments can be done easily by knowing the constituent type. The system can be further developed as a comprehensive learning system with access to the Internet. It can also be able to expand to predict about more susceptible disease for an individual depending on the constituents.

References

[1] Adlassnig, K-P. (1986), Fuzzy set theory in medical diagnosis, *IEEE Tr. On Syst., Man, and Cybernetics* **16**(2), March/April, 260-265.

[2] Mason, D., Linkens, D. & Edwards, N. (1997), Self-learning fuzzy logic control in medicine, *Proc. AIME'97*, (E. Keravnou et al., eds.), *Lecture Notes in Artificial Intelligence* **1211**, Springer-Verlag: Berlin, pp. 300-303.

[3] Lalande, A. et al., (1997). Automatic detection of cardiac contours on MR images using fuzzy logic and dynamic programming, *Proc. AMIA Ann. Fall Symp.*, pp. 474-478.

[4] Zahlmann, G., Scherf, M. & Wegner, A. (1997), A neurofuzzy classifier for a knowledge-based glaucoma monitor, *Proc. AIME'97*, (E. Keravnou et al., eds.), *Lecture Notes in Artificial Intelligence* **1211**, Springer-Verlag: Berlin, pp. 273-284.

[5] Liu, J. and Shiffman, R., (1997). Operationalization of clinical practice guidelines using fuzzy logic, *Proceedings of AMIA Ann. Fall Symp.*, 283-287.

[6] Zielstorff, R.D. (1998). Online practice guidelines, *JAMIA* **5**, 227-236.

[7] Jim W, Gleb B, Berend v. Z (2000) Fuzzy logic in clinical practice decision support systems, *Proceedings of the 33rd Hawaii IEEE International Conference on System Sciences* –pp. 1-10

[8] George J.K, Yuan B.(1995), Fuzzy sets and Fuzzy logic, prentice hall of India, pp. 280–300.

[9] Jonson L (1988), Expert system Architectures, Kopan Page Limited

[10] Dubey G.P , (1978)The Physiological concepts in Indian medicine, Science and Philosophy of Indian medicine, Shree Beldyanath Ayurved Bhawan Ltd.

[11] Chatfied C(1996)," Introduction to Multivariate Analysis", Chapman and Hall.

[12] Coppin G, Skrzyniarz A (2003), "*Human–centered processes : Individual and distributed decision support*", IEEE Intelligent systems, pages 27–33.

[13] Dave Westwood, Flex reference guide, LPA, U.K

[14] Mendis D. S. K., Karunananda A. S. and Samarathunga U. (2004), *Multi-Techniques Integrated tacit knowledge modelling system*, International Journal of Information Technology, Vol 9, pp 265-271

[15] Mendis D. S. K., Karunananda A. S. and Samarathunga U. (2004), *An Expert system for analysing Aurvedic human constituents,* Shamisha – Journal of Ayurveda

[16] Richards D. and Bush P., "Measuring, Formalizing and Modeling Tacit Knowledge" IEEE/Web Intelligence Conference (WI-2003) Bejing.

[17] Ross Dawson, (2001), "Developing Knowledge-Based Client Relationship", Butterworth Heinemann.

[18] Tripathi S.N (1978), "Clinical Diagnosis", Science and Philosophy of Indian medicine.

[19] XpertRule Knowledge Builder, www.attar.com.

[20] Noak V (2000), "Discovering the world with Fuzzy logic", A Springer – Verlag Company, PP. 3 – 50.

[21] Rajeev K (1995), "Integrating knowledge bases in expert system shells: an open system approach", International Journal of computer application in technology, pages 78–89.

[22] Hayashi Y, Tazaki E, Yoshida K, Dey P (1988), “Medical diagnosis using simplified multi-dimensional fuzzy reasoning”, In: Proceedings of IEEE 1988 International Conference on Systems, Man and Cybernetics, Beijing, China, 1988. p. 58–62.

[23] http://kmi.open.ac.uk/knowledge-modelling/people.html

[24] Kolousek G (1997), ″The system architecture of an integrated medical consultation system and its implementation based on fuzzy technology”, Doctoral thesis, Technical University of Vienna, Austria.

[25] Matei Ciobanu Morogan, (1997), SPSS for windows, release 8.0.0, DSV KTH⁄SU, Sweden

On the performance of large dimensional parallel matrix multiplication algorithms on the computational grid

D N Ranasinghe*, K P M K Silva*, K Munasinghe** and M Jayewardena*

*University of Colombo School of Computing
**Department of Computer Science, University of Ruhuna

Abstract

In this paper the parallel matrix multiplication problem is revisited from a multi architectural perspective. Performance is obtained for well known algorithms of O(n^3) complexity on two distinct computational paradigms: the shared data space and the grid. Results show that algorithms with linear speedup are slow and those with out linear speedup are fast. Fast algorithms use efficient message broadcast systems unlike others. From the results obtained it can be concluded that the shared data spaces paradigm with an efficient message broadcast system can be a strong contender for message passing grid systems with the added advantage of load balancing resulting in dynamic selection of optimal number of processors.

1. Introduction

Numerical high performance computing and in particular, matrix multiplication algorithms forms a core part of many numerical problems and has a long history of investigation [1, 3]. There has been a considerable amount of work done in deriving efficient matrix multiplication algorithms, both sequential and parallel for various architectural platforms, and the latter mainly for proprietary supercomputers [1,2,3,5,6,7,8,9]. In any of the previous studies, a direct comparison of parallel implementations of shared data spaces and message passing architectures has not been made. With the emergence of the computational grid concept [10,11], the possibility of having the shared data spaces paradigm [2] as a grid service as an alternative to message passing has occurred. Further, the grid middleware is expected to provide fault tolerance and load balancing to the tasks it execute. We examine how shared data spaces could satisfy such a need. In our work, we compare the run time performance of four efficient parallel matrix multiplication algorithms all of O(n^3) complexity, on Nordugrid [11] running MPI and Linda as an application. We obtain a comprehensive set of run times for matrix dimensions ranging from 100 to 4000 with compiler and cache optimization. We found that two of the algorithms that utilize efficient broadcast mechanisms achieve a fast execution, though unable to obtain a linear speed up. This efficient broadcast may be explicit in the algorithm or may be implicitly provided in the middleware. When provided in the middleware the added advantage of it of offering load balancing enables shared data spaces to be a strong contender for a grid service. The next sections briefly discuss the algorithms studied, the shared data spaces model, and the parallel implementation. Finally, the results are evaluated and commented upon.

2. Class of O(n^3) algorithms

It is well known that matrix multiplication yields easily towards parallel implementation due to its inherent divisibility into subtasks that can be computed simultaneously on multiple processors [1, 3]. However, whether there is a linear speedup that is proportional to the number of processors will depend on the algorithm and the architecture for which it is derived. The vast majority of MIMD architectures on which these parallel algorithms have been tested belong to message passing, either loosely coupled or tightly coupled, with the former emerging as a cheaper alternative. The naïve sequential algorithm has a complexity of O(n^3) and the Strassen's algorithm and its close variants have a complexity of around O(n^2.8) [8]. Parallel versions of these two classes of algorithms have existed in the literature for a considerable time, some of which are optimized for well known architectures like the mesh, torus or the hypercube. Cannon's algorithm is optimized for a torus architecture [1]. In our exercise we narrow down our investigation into O(n^3) parallel algorithms which are run on linear clusters. One of our main objectives is to implement a virtual shared memory on top of distributed memory architecture in the form of a shared data space and to observe the special properties of parallel algorithms in such a paradigm. Literature also mentions of mixed mode algorithms that work efficiently at both large and small dimensions using two distinct algorithms that switch over at a threshold dimension. Our work can be extended later to cover both mixed mode algorithms and the O(n^2.8) algorithms on shared data spaces.

3. Shared data spaces model

The characteristic feature of Linda [2] as a widely used shared data space is its efficient tuple replication mechanism across distributed memory architectures supported by an efficient message broadcast mechanism. By virtue of it's publish-subscribe property the programmer eliminates explicit message passing and is forced to think from a different algorithmic view point. The programming model is a master-slave one with slaves acquiring data from the shared space to do the computation and return results to the space. This offers dynamic load balancing where the computation is carried out by an optimal number of processors at any given time. There are instances in parallelization where the apriori knowledge of the optimal number of processors is required to obtain the best performance. In such a case Linda paradigm would be of much use.

4. Parallel implementation

We have selected four parallel algorithms of O(n^3) complexity: the naïve algorithm, that multiplies rows and matrices, Linda version of the naïve algorithm[2], Cannon's algorithm [1] and an efficient broadcast algorithm[9]. All algorithms are compiled with relevant optimization flags, and where applicable cache optimizations have also been applied. The latter ensures that sub block distributions do match the cache size on the machines. Algorithms have been run under the following scenarios. Linda version has been run on a 4 node cluster (dual Intel Xeon 2.4 GHz CPU, RedHat Linux 9 (2.4 Kernel), 512KB cache, main memory 512MB on a 100Mbps switched Ethernet) and the rest on a grid architecture closely emulating a high speed backbone connected cluster (Fedora Core3 as the OS (2.6 kernel), dual Intel Pentium 4, 3 GHz, 1MB cache 2GB main memory, with arbitrary number of nodes). Matrix dimensions ranging from 100 to 4000, all symmetric matrices, have been attempted. Due to the limitations of algorithmic data structures and memory it was not possible to go beyond a dimension of 4000 in certain instances. The free Linda version available to us only permitted a maximum of 4 nodes to be utilized. The parameter NP refers to the total number of processors utilized between the master and the slaves. In certain algorithms such as the one based on Linda, the slaves are the real workers, and in others, all processors join in the computation. Where each node consists of dual or quad processors, MPI will utlise at least one processor of each node depending on the requested number of processors.

5. Performance evaluation

In the following plots, the speedup measure is the ratio between the execution times for a particular algorithm on a single worker to that of multiple workers. It can be clearly seen that Cannon's algorithm (Fig 3) and the raw broadcast algorithm (Fig 2) have linear speedup curves, especially for large dimensions. The apparently super linear speedup seen in Fig 3 is due to the fact that master too is performing some computations. The raw broadcast algorithm gives a significant speedup even at low dimensions compared to Cannon's algorithm due to latter's high communication overhead caused by 'broadcast-multiply-roll' property targeting the torus architecture. However the execution times of Cannon's algorithm are magnitude order higher than those of other algorithms. This clearly shows the architecture dependency of Cannon's algorithm. The other two algorithms, Linda based one (Fig 4) and the efficient broadcast one (Fig 1) have in common a non-linear speedup as well as lowest timings of the four. Assuming that these times are the most efficient ones realizable, the tapering of the speedup curve is realistic and shows that for any given matrix size, the existence of an optimal number of processors. For example, for a size of 4000, the number of processors is around 8 (Fig. 1). This property was not exhibited by either Cannon (on bus architecture) or raw broadcast as the timings are high enough to show a linear speed up. In Linda's case, the tapering is yet to occur, as the number of processors is insufficient limited by the license.

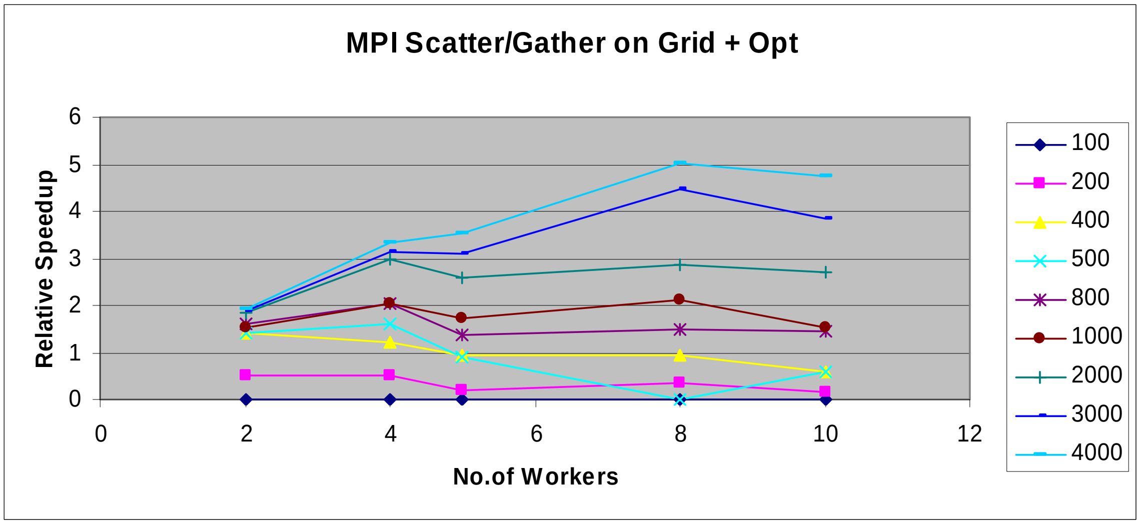


Fig 1. Scatter/gather MPI algorithm on grid

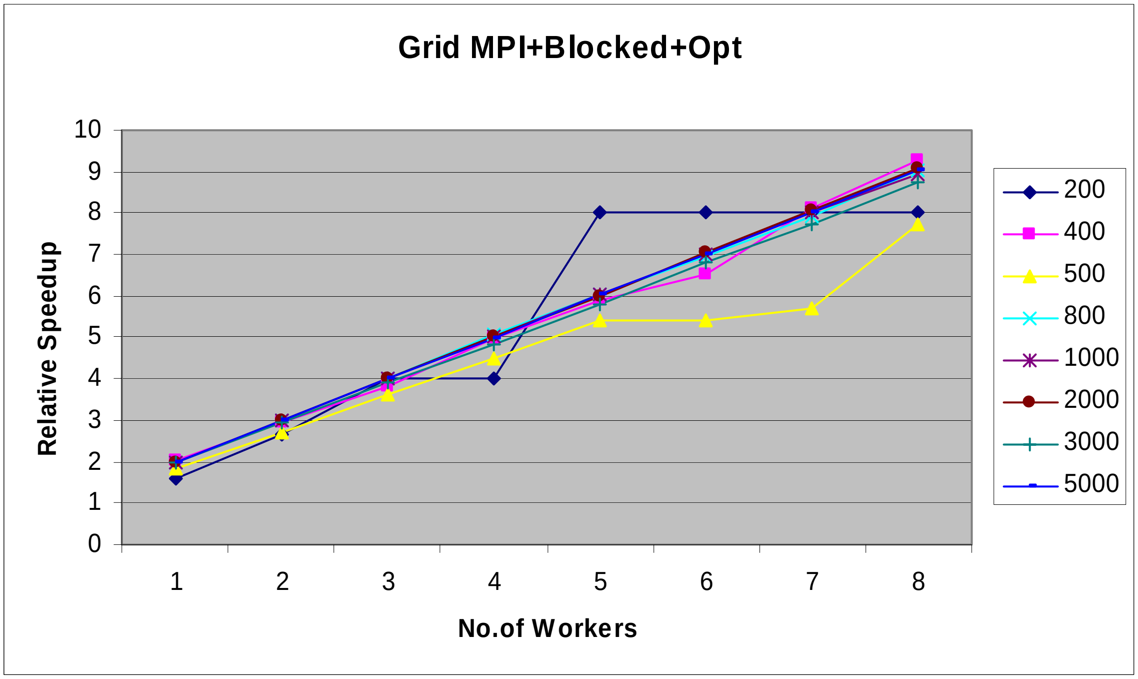


Fig.2. naïve MPI algorithm on grid

The Linda algorithm and the scatter/gather MPI algorithm are characterized by implicit and explicit efficient message broadcast mechanisms respectively. In Linda's case it is implemented in the middleware for replica management. The inherent low times are due to this fact.

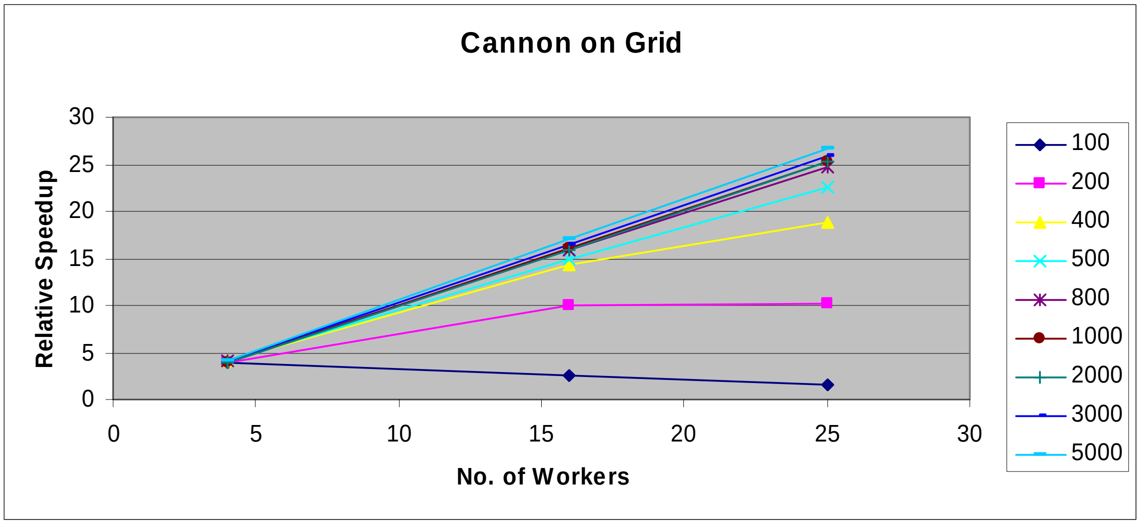


Fig.3. Cannon's algorithm on Grid

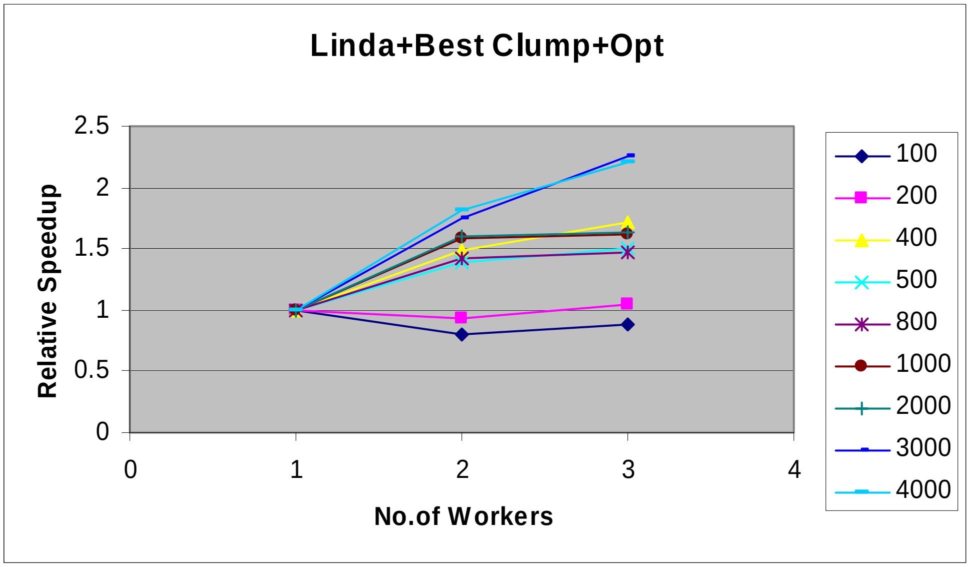


Fig.4. Linda on Cluster

	Time in seconds		
Matrix Size	**NP=2**	**NP=3**	**NP=4**
100	0.031	0.039	0.035
200	0.085	0.092	0.081
400	0.481	0.325	0.281
500	0.721	0.521	0.48
800	2.544	1.793	1.734
1000	5.019	3.162	3.096
2000	30.993	19.377	18.96
3000	102.19	58.721	45.463
4000	258.07	142.057	117.27

Fig 5. Linda timing

	Time in seconds		NP=1	NP=4		
Matrix Size	**NP=1**	**NP=4**	**(UU)**	**(UU)**	**NP=16(UU)**	**NP=25(UU)**
100	0.77	0.39	0.73	0.19	0.29	0.46
200	11.85	3.08	5.88	1.47	0.59	0.58
400	50.74	24.66	47.78	11.88	3.32	2.53
500	186.22	48.15	94.12	23.22	6.31	4.18
800	767.97	197.08	385.9	95.87	24.3	15.64
1000	801.74	384.83	756.01	189.42	46.92	30
2000	11944.36	3079.12	6056.29	1513.83	379.66	239
3000	40353.61	18603.28	21120.84	5123.91	1279.72	818.89
5000	148085.2	48080.65	101168.77	24155.26	5912	3784.89

Fig. 6. Cannon timings

	Time in seconds					
Matrix Size	**NP=1**	**NP=2**	**NP=4**	**NP=5**	**NP=8**	**NP=10**
100	0	0.01	0.01	0.01	0.01	0.01
200	0.01	0.02	0.02	0.05	0.03	0.06
400	0.17	0.12	0.14	0.18	0.18	0.28
500	0.34	0.24	0.21	0.37	error	0.56
800	1.35	0.83	0.66	0.98	0.9	0.92
1000	2.65	1.75	1.29	1.53	1.24	1.72
2000	20.66	11.26	6.91	7.98	7.2	7.68
3000	69.8	37.08	22.38	22.48	15.59	18.12
4000	164.11	85.98	49.23	46.59	32.65	34.61

Fig 7. Scatter/gather timings

Matrix Size	**NP=2**	**NP=3**	**NP=4**	**NP=5**	**NP=6**	**NP=7**	**NP=8**	**NP=9**	**NP=10**
100	0.01	0.01	0	0	0	0	0	0	0.01
200	0.08	0.05	0.03	0.02	0.02	0.01	0.01	0.01	0.01
400	0.65	0.32	0.22	0.17	0.13	0.11	0.1	0.08	0.07
500	1.08	0.59	0.4	0.3	0.24	0.2	0.2	0.19	0.14
800	5.07	2.54	1.7	1.26	1	0.84	0.73	0.64	0.56
1000	9.3	4.66	3.11	2.33	1.85	1.54	1.33	1.16	1.04
2000	76.9	38.42	25.61	19.25	15.36	12.83	10.94	9.56	8.48
3000	252	127.66	85.88	64.78	51.91	43.29	37.02	32.52	28.87
5000	1172.77	588.07	391.31	293.13	235.53	194.92	167.13	146.14	129.65

Fig 8. MPI on Grid

The present study restricts it self to algorithms of O(n^3) complexity. The lower complexity algorithm class comprising the parallel variants of Strassen's and others are most likely to emulate a similar behaviour on grids demonstrating a clear division between the efficient broadcast based implementations and others. As suggested in literature, mixed mode algorithms which combine a parallel execution at large dimensions and a sequential execution at small dimensions with a static switch over yields a further reduction in execution timing. We believe that Linda like subsystems can realise two important requirements in parallel implementations of similar systems: the detection of the optimal number of processors and the dynamic detection of the threshold of switchover between two modes in mixed mode algorithms. It would be worthwhile to see Linda installed as a grid service than a mere application in grid, for the use of parallel applications.

5. Conclusions

We have revisited the large parallel matrix multiplication problem from a grid perspective restricting our investigation to the O(n^3) class of algorithms. Two algorithms that use an efficient broadcast mechanism and two that use raw message passing have been investigated. The latter two demonstrate linear speed ups where as the former two demonstrate fastest execution times. From a grid perspective, the desirable features of an algorithm are scalability, load balancing and fault tolerance. We have shown that an algorithm targeted for a shared data space middleware offers all these as well as algorithmic specific properties such as the detection of the optimal number of processors and dynamic thresholds if any for mixed mode algorithms. As such Linda like subsystems is suitable candidates for grid middleware. As further work we have to corroborate our observations for low complexity algorithm class as well as there is a need to implement a Linda like subsystem on grid middleware as well

Acknowledgements

The authors wish to thank Dr D Atapattu of Scientific Computing Associates, USA, for his help in untangling Linda. The work was performed under a grant from the Swedish SPIDER project.

References

[1]Parallel Programming; Barry Wilkinson; Prentice-Hall

[2] Linda - Scientific Computing Associates, USA

[3] Solving Problems on Concurrent Processors, Vol.1; G Fox et al; Prentice Hall

[4] Go for both types of data locality; M Silva, R Wait; Proceedings HPC Asia 2002, India

[5] Multilevel Hierarchical Matrix Multiplication on Clusters; Proceedings of the 18th annual international conference on Supercomputing; pp136-145; 2004

[6] Ultra-fast matrix multiplication: An empirical analysis of highly optimized vector algorithms; B Kakaradov; surj.stanford.edu/2004/pdfs/kakaradov.pdf

[7] Fast Parallel Matrix Multiplication - Strategies for practical hybrid algorithms - Erik Ehrling http://www.f.kth.se/~f95-eeh/exjobb/background.html

[8] A scalable parallel Strassen's matrix multiplication algorithm for distributed-memory computers, SAC '95: Proceedings of the 1995 ACM symposium on Applied computing, 1995, pp 221-226, Qingshan Luo and John B. Drake

[9] SRUMMA: A Matrix Multiplication Algorithm Suitable for Clusters and Scalable Shared Memory Systems, Manojkumar Krishnan and Jarek Nieplocha, 18th International Parallel and Distributed Processing Symposium, IPDPS 2004

[10] Globus: www.globus.org/

[11] Nordugrid: www.nordugrid.org/

Optimization by Parallel Hopfield Neural Network on GPU Graphics Card

Zheng He [1] and Koichi Harada [2]
[1]Graduate School of Engineering, Hiroshima University, Japan
ethen@hiroshima-u.ac.jp
[2]Department of Integrated Arts & Sciences, Hiroshima University, Japan
harada@mis.hiroshima-u.ac.jp

Abstract

Parallel Hopfield neural networks are often implemented on parallel machines or distributed systems. This paper describes how Parallel Hopfield neural network can be mapped to programmable graphics hardware found in commodity PC. Our approach stores weight matrix, neuron input and output vectors in texture memory on graphics card. All operations in parallel Hopfield neural network are implemented entirely with fragment programs executed on graphics processing unit in parallel. We demonstrate the effectiveness of our approach by solving TSP. Our proposal draws the advantages of parallel Hopfield neural network on low-cost platform.

1. Introduction

Since Hopfield and Tank used Hopfield network to solve Traveling salesman problem in 1982[1], Hopfield network has become a new tool for combinational optimization. From that time on, people has being applying the Hopfield network to solve a wide class of combinatorial optimization.

In a discrete-time version, the Hopfield network implemented local search. In a continuous-time version, it implemented gradient decent. Both algorithms suffer the local minimum problem and many optimization problems in practice have lots of local minima. Furthermore, the Hopfield-Tank formulation of the energy function of the network causes infeasible solutions to occur most of the time [2], [3]. People also found that those valid solutions were only slightly better than randomly chosen ones. To guarantee the feasibility of the solutions, the most important breakthrough came from the valid subspace approaches of Aiyer et al [4] and Gee [5]. To escape from local minima, many variations of the Hopfield network have been proposed based on the principles of simulated annealing [6]. Three major approaches are Boltzmann [7], Cauchy [8], and Gaussian Machines [9]. In theory, simulated annealing can approach the global optimal solution in exponential time. However, it is not guaranteed and is very slow to make it effective in practice. As a general purpose optimizer, the Cauchy Machine has the advantage of exhibiting full parallelism. It can be implemented on parallel computer or other hardware with parallel computing capability. The goal of this paper is to implement Cauchy machine by utilizing graphics hardware in pc.

The graphics processors (GPUs) on today's commodity video cards have evolved into an extremely powerful and flexible processor. Modern GPUs perform floating-point calculations much faster than today's CPUs [10]. Furthermore, instead of offering a fixed set of functions, current GPUs allow a large amount of programmability [11]. These desirable properties have attracted lots of research efforts to utilize GPUs for various non-graphics applications in recent years [12][13][14][15][16][17][18]. Previous research works have already shown that GPUs are especially adept at SIMD computation applied to grid-based data [14]. Therefore, we can envision that Cauchy machines should be a good fit for commodity programmable GPUs.

In this paper, we present a novel implementation of parallel Hopfield network on the GPU. Weight matrix is represented as 2D texture maps; bias, neuron input and neuron output are represented as 1D texture maps. We perform each step of a parallel Hopfield network by executing one or several fragment programs on every pixel at each step in a SIMD-like fashion. Hopfield iterations can be performed entirely on GPU. We will demonstrate the effectiveness of GPU implementation by applying it to solve TSP problem. Relative to software implementation, a distinct speedup is observed when the number of cities is larger than 24.

The rest of the paper is organized as follow: the subsequent section gives some details about parallel Hopfield network; Section 3 introduces some background about graphics hardware to facilitate understanding of our implementation; and we describe the proposed GPU based implementation of Hopfield network in Section 4; Section 5 presents our implementation performance by solving TSP, and the paper concludes with suggestions for future work in Section 6.

2. Parallel Hopfield neural network

2.1. The Hopfield neural network

In order to solve the NP-completed (Non-determined Polynomial-time) constraint satisfaction problems such as traveling salesman problems and linear programming problems, a neural network was introduced for computation using a pre-determined energy function for the neurotic interconnectivity. The usual approach for solving a discrete optimization problem using Hopfield neural network techniques is to formulate the cost function and the constraints of the problem in terms of the minimization of a quadratic energy function which is designed over the binary space $\{0,1\}^n$ and has the following form:

$$E(\vec{v}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j w_{ij} - \sum_{1=1}^{n} v_i\theta_i \qquad (1)$$

In which v_i devotes the output for i-th neuron in network, w_{ij} is the weight between i-th neuron and j-th neuron and θ_i is the bias for i-th neuron .Here, $w_{ij} = w_{ji}$ for every $i=1, ..., n, j=1,..., n$ and $w_{ii}=0$ for every i.

However, because such deterministic neural networks can inherently descend down to a local minimum unless all of the local minima are removed, the converged state may not be one of global minima. Therefore the Simulated Annealing (SA) methodology has been used to obtain the hill-climbing property on an energy surface. The Hopfield neural network attains the minimization of the above energy function based on the Boltzmann Machine which extends the discrete Hopfield model and the analog Hopfield network. The operation of the Boltzmann Machine integrates the dynamics of the discrete Hopfield networks with the SA algorithm [18]. At each step, a unit i is randomly selected and the energy difference ΔE_i that would be caused by a change in the state of this unit is computed as follows:

$$\Delta E_1 = (2v_i - 1)(\sum_{i=1}^{n} w_{ij}v_j + \theta_i) \qquad (2)$$

If $\Delta E_i < 0$, the change is accepted; otherwise it is accepted with the probability that depends on the quantity *exp* $(\Delta E_i/T)$, where the temperature parameter T decreases according to a specialized annealing schedule.

Because the Boltzmann Machine optimizer introduced above contains inherently sequential nature as the general case with SA, a synchronous discrete Hopfield neural network was proposed by applying a step threshold function to the analog Hopfield network in place of the typically used sigmoid function, in which each unit has analog input and discrete output. The Integration of the synchronous discrete Hopfield network with a fast SA method results in the Distributed Cauchy Machine which provides stochastic hill-climbing capabilities and avoids convergence to false local minima.

2.2. The distributed Cauchy machine

A Cauchy machine uses a hybrid neuron (binary neuron input and analog neuron output), and the Cauchy noise version of the Metropolis acceptance function, as well as the Cauchy noise generating randomly states [8]. If each network unit i is characterized by analog input u_i (taking any real value) and binary output v_i (taking values in {0, 1}), the input-output behavior of each unit at time t is based on:

$$v_i(t) = \begin{cases} 1 & \text{if} \quad u_i(t) > 0, \\ 0 & \text{if} \quad u_i(t) \le 0. \end{cases} \qquad (3)$$

During operation, the value of u_i is updated at each time step t by using the following motion equation:

$$\frac{\Delta u_i}{\Delta t} = -\frac{\Delta E_i}{\Delta v_i}. \qquad (4)$$

Here ΔEi is the difference in the energy of the network that would be caused by a change in the output of unit i. Simulation of the dynamics is based on the first-order discrete approximation of

$$u_i(t + \Delta t) = u_i(t) + \Delta u_i. \qquad (5)$$

Equation (4) means that each unit follows gradient descent dynamics which in general leads the synchronous discrete Hopfield network to an equilibrium state. In the case where the energy function is given by Eq(1), the motion equation of each unit i is given by:

$$\frac{\Delta u_i}{\Delta t} = \sum_{i=1}^{n} w_{ij}v_j + \theta_i. \qquad (6)$$

A binary vector $(v_1, ..., v_n)$ constitutes an equilibrium state of the synchronous discrete Hopfield network if for each $i=1, ..., n$:

$$\begin{aligned} &v_i = 1 \quad and \quad \Delta u_i \geq 0, \\ &\qquad\quad or \qquad\qquad\qquad\qquad (7) \\ &v_i = 0 \quad and \quad \Delta u_i \leq 0. \end{aligned}$$

That means once the network has reached an equilibrium state it will remain there forever, since no change is possible in the output of any unit.

In order to provide the synchronous discrete Hopfield network with stochastic hill-climbing capabilities, the output of each unit *i* is stochastically updated at each time step *t* based on the Cauchy distribution:

$$s_i(t) = \Pr\{v_i(t) = 1\} = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{u_i(t)}{T_c(t)}\right). \qquad (8)$$

In Eq(8), $T_c(t)$ is an artificial temperature parameter that is adjusted according to a fast annealing schedule as a function of time *t*:

$$T_c(t) = \frac{T_0}{1+\beta \cdot t}. \qquad (9)$$

Here T_0 is the initial temperature value and β is a value in the range (0, 1) that controls the speed of the schedule. If the value of T_c is equal to zero, we have the case of a deterministic synchronous discrete Hopfield network. In general, the distributed Cauchy machine can be expressed as in Fig.1:

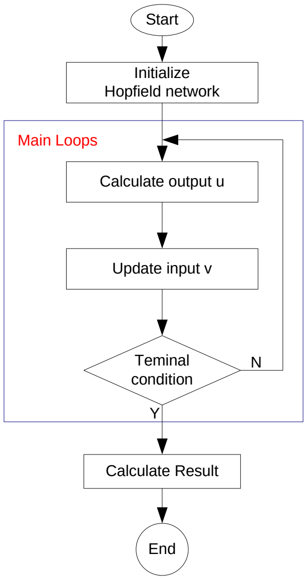


Fig.1: main flow of distributed Cauchy machine

3. Graphic Processing Unit (GPU)

Graphics hardware is originally designed for accelerating images rendering, which has already been used for many algorithms in various areas including computational geometry, scientific computation as well as image processing and computer graphics. Its extended capabilities in supporting complex operations become useful in non-graphics applications; practically the advent of a programmable vertex processor and fragment processor enables flexible functions for general computation. Moreover, GPU can produce a better performance than a general purpose CPU in areas where repeated operations are common since it is designed for high-performance rendering where repeated operations are common and thus more effective in utilizing parallelism and pipelined.

The mechanism of general computation using a GPU is shown in Fig.2 shows. Firstly, commands, textures and vertex data are received from the host CPU through shared buffers in system memory or local frame-buffer memory. The vertex processor allows for a program to be applied to each vertex in the object, performing transformations and any other per-vertex operation the user specified. Vertices are then grouped into primitives, which may be point, lines, or triangles. Next, rasterization is the process of determining the set of pixels covered by a geometric primitive. After this, the results of rasterization stage, a set of fragments, are processed by a program which runs in the programmable fragment processor. Meanwhile, the programmable fragment processor also supports texturing operations which enable the processor to access a texture image using a set of texture coordinates. Finally, the raster operations stage performs a sequence of per-fragment operations immediately before updating the frame buffer.

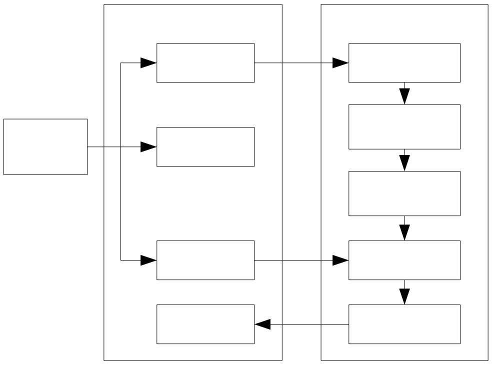


Fig.2: the programmable graphics pipeline

4. Implement Parallel Hopfield network on GPU

To utilize GPU to run algorithm we need two main steps: the first step is storing of the data which is necessary during calculation into a video memory; and the second step is processing those data. In this paper we adopt fragment processor to carry out the general calculation, for which the main process codes were given in Cg code [10].

4.1 Representations

According to the introduction in section 2.2, a parallel Hopfield network contains the following conventional terminologies: weight matrix *ω*, bias vector *θ*, neuron input vector *v* and neuron output vector *u*. In addition, in order to simplify literal expression a buffer is shortened as *buf* and a texture to store the random seed as *random*.

4.1.1 Matrix representations

Given a Hopfield network containing *N* neurons, then the size of weight matrix is $N \times N$, which is symmetry and can be defined as a dense matrix. Many documents suggest that such a dense matrix should be stored as a group of 2D texture [13][19][20] due to hardware limitation, and in video memory, the maximum size of 2D texture is limit to 4096×4096 for floating value [21]. Therefore they first treated the matrix as a combination of such vectors as column, row or diagonal, and then stored each vector into one 2D texture. Consequently, it is necessary to create *N* textures to store a dense matrix containing $N \times N$ elements and upload those contained data from system memory to video memory.

However, there are some serious problems during implementation when using this method. Firstly, creating pbuffer in video memory is a relatively time-consuming work compared to creating buffer in system memory; when the matrix dimension *N* exceeds 1000, pbuffer creation might cost several seconds. Secondly, changing textures during matrix-matrix or matrix-vector operation is also time-consuming since we have to operate one of textures contained one column (or row, diagonal) at one time, and then repeat the same operation to another texture.

In this research, the weight matrix w is expressed by a single 2D texture, which can avoid the above disadvantages; and to attain greater efficiency, the data is stored into all four channels of RGBA. Assume that the number of element that a row in the matrix w can store is Nc, so if N is the multiple of 4, $N_c = N/4$; otherwise $N_c = INTEGER(N/4) + 1$. Figure 3a shows how to store N elements into $1 \times N_c$ 1D RGBA texture. Moreover, in the case of $4N_c > N$, there is no proper weight value to fill in and those elements are filled by zero. So the matrix containing $N \times N$ elements can be successfully stored in the 2D texture of $N_c \times N$ as figure 3b shows.

4.1.2 Vector representation

Corresponding to the representation of weight matrix introduced above, all the vectors in this system are represented by 1D texture and are defined in detail as the following.

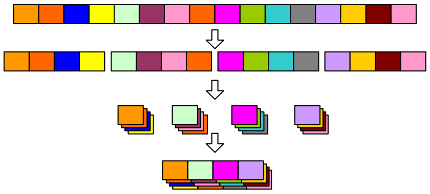


Fig.3a *N* element stored into $1 \times N_c$ 1D RGBA texture

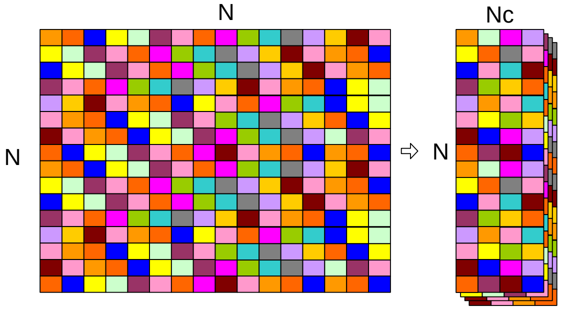


Fig.3b $N \times N$ elements stored in a $N_c \times N$ 2D RGBA texture

u: represented by a 1D RGBA texture with size of N_c. In GPU process, one texture can only be read or written in a program, but u should be read and update at the same time in this parallel Hopfield network. So we create textures with two surfaces according to the Ping-Pong algorithm introduced in [22].

v: represented by a 1D RGBA texture with size of N_c.

θ: represented by a 1D RGBA texture with size of N_c.

Random: represented by a 1D RGBA texture with size of N_c.

buf: Represented by a 1D single (Red) channel texture. Whether or not each neuron reaches equilibrium state should be known in this system to check the terminal condition, which is transformed to check the number of neurons which have already reached equilibrium state in parallel. So a texture is created as a buffer memory to implement the reduction operation, and the size of 1D texture must be the power of 2 for this purpose. Take the symbol of $2^{N_{buf}}$ to indicate the size of *buf* texture, so if log_2N_c is integer, then $N_{buf} = log_2N_c$; otherwise, $N_{buf} = INTEGER\ (log_2N_c) + 1$.

4.2 operations

In parallel Hopfield model, there are three main parts in the main loop; each calculates output, updating input and checking terminal condition which will be described separately in detail in the following subsections. Besides, the random number generator is also explained since the random number plays an important role in updating the process.

4.2.1 Calculation of output for each neuron

To calculate the output u_i of each neuron, the output increment Δu_i should be obtained first according to Eq(6), and then add this increment to the original value to create a new output according to Eq(5). In particular, this calculation on GPU should make sufficient use of parallelism property of GPU, thus Eq(5) and Eq(6) are transformed into other format:

$$\begin{aligned}
& u_i(t+\Delta t) \\
& = u_i(t) + \Delta u_i \\
& = u_i(t) + (\sum_{j=1}^{n} w_{ij} v_j + \theta_i)\Delta t \qquad (10) \\
& = u_i(t) + \theta_i \Delta t + \sum_{j=1}^{n} w_{ij} v_j \Delta t
\end{aligned}$$

If define that $\Delta u_i^j = w_{ij} v_j \Delta t$, then Eq(10) becomes:

$$u_i(t+\Delta t) = u_i(t) + \sum_{j=1}^{n} \Delta u_i^j + \theta_i \Delta t \qquad (11)$$

The above two steps in this process can be expressed by pseudo codes as the following:

Step1:

```
For each column in weight matrix w_j and corresponding element in neuron input vector v_j
    For each element in each column of weight matrix w_ij
        $\Delta u_i^j = w_{ij} v_j \Delta t$
        $u_i = u_i + \Delta u_i^j$
    End
End
```

Step2:

```
For each element in neuron output vector u
    $u_i = u_i + \theta_i \Delta t$
End
```

The step1 should be noted here. One inner loop can be performed through a rendering pass. Because the weight matrix is stored in RGBA mode, four columns of this matrix can be processed by one single rendering pass. So in outer loop, we only need N_c rendering passes to accomplish the calculation.

The Cg program for step1 and 2 of neutron output calculation are also given below.

Step1:

```
float4 SumDeltaU( float2 texCoord : WPOS,
                  uniform samplerRECT  matrix,
                  uniform samplerRECT u,
                  uniform samplerRECT v,
                  uniform float deltat
) : COLOR
{
float4 mt,vt,ut,OUT;
mt = texRECT(matrix,texCoord);
vt = texRECT(v,texCoord);
ut = texRECT(u,texCoord);
OUT =  mt.xyzw*vt.xyzw;
OUT = ut + OUT*deltat;
return OUT;
}
```

Step2:

```
float4 UpdateU(float2 texCoord : WPOS,
               uniform samplerRECT u,
               uniform samplerRECT theta,
               uniform float deltat
) : COLOR

{float4 thetat,ut,OUT;
thetat = texRECT(theta,texCoord);
ut = texRECT(u,texCoord);
OUT = ut + thetat*deltat;
return OUT;
}
```

In both steps, texture *u* has to be read and written. But because one texture must be read-only or write-only in rule, a double-surface texture *u* is necessary. A flag is imported to indicate read-only/write-only surface, and a flag will be flipped accordingly based on a so-called ping-pong scheme.

4.2.2 Update of input for each neuron

Updating neuron input based on Eq(8)is a simple task, and can easily be implemented in parallel by GPU. The Cg cod of updating input is shown as below:

```
float4 CauchyDistributionUpdate( float2 texCoord : WPOS,uniform samplerRECT u,uniform samplerRECT random,uniform float t,
      ) : COLOR
{
float4 value = texRECT(u,texCoord);
float4 rand = texRECT(random,texCoord);
float T0= 10;
float beta = 0.2;
float Tc = T0/(1 + beta * t);
float4 s = 0.5 + atan(value/Tc)/3.14159260564;
return step(rand,s);
}
```

4.2.3 Check the terminal condition

In GPU, the branch command is more expensive than mathematical operations. Generally speaking, eight mathematical operations can be finished in one clock, but an “if” branch will take 4 clocks [23]. According to Eq(7) which checks equilibrium state of each neuron, we transform the terminal condition into the following equation in order to avoid using “if” statement:

$$\delta_{v_i} = \begin{cases} 0 & v_i < 0.5 \\ 1 & v_i \geq 0.5 \end{cases} \quad and \quad \delta_{\Delta u_i} = \begin{cases} 0 & \Delta u_i < 0 \\ 1 & \Delta u_i \geq 0 \end{cases} \qquad (12)$$

$$then \quad \delta = \left| \delta_{v_i} - \delta_{\Delta u_i} \right|$$

δ is introduced to check whether or not one neuron reaches the equilibrium state and $\delta=0$ indicates that the neuron have reached the equilibrium state.

In addition, the Cg compiler provides a mathematical function named “step” which can carry out numeric comparison at low cost. This function is defined as [24]:

$$step(a,x)=\begin{cases}0 & x<a,\\ 1 & x\geq a.\end{cases}$$

We record the equilibrium state onto the buf texture, Cg code is shown below:

```
float4 Terminal( float2 texCoord : WPOS,
                        uniform samplerRECT wu,
                        uniform samplerRECT ru,
                        uniform samplerRECT v

) : COLOR
{
float4 valuew = texRECT(wu,texCoord);
float4 valuer= texRECT(ru,texCoord);
float4 diff = valuew - valuer;
float4 valuev = texRECT(v,texCoord);
float4 result = abs(step(0.5,valuev)-step(0.0f,diff));
return float4(result.x+result.y+result.z+result.w,0,0,0);
}
```

Notice that each fragment program can process 4 neurons simultaneously, so we store the sum of these 4 neuron's state in *buf* texture to simplify process.

Terminal condition depends on equilibrium states of all neurons in stead of any single one. Therefore after checking the equilibrium state of each neuron, we should find out whether the whole network containing all the neurons is in equilibrium state or not. In order to combine all entries of the equilibrium state into a single judgment index, the reduction operation stated in [20][25] is adopted to combine the vector entries in *buf* texture in multiple rendering passes by recursively combining the result of the previous pass. At the end of the pass, we can get the combined judgment index which is stored in the first element of *buf* texture. Then the index is downloaded from video memory to the system memory and checked by CPU whether the main loop should terminated or not. In other words, if this index equals to 0, the whole calculation will stop.

As we know, data transfer between video memory and system memory is expensive, we only check the terminate condition every 10 cycles or more to save time.

4.2.4 Random number generator

As described above in Section 4.2.2, the random numbers are involved in updating the neuron input. However, because the current graphics hardware does not provide the function for generating random numbers, the Linear Congruential Generator (LCG) is used to generate pseudo-random numbers [26]:

$$R_{i+1}=a\cdot R_i+c\bmod m \qquad (14)$$

Where m is called the modulus, and a, c is multiplier and the increment, respectively. LCG can be implemented in a simple fragment program. A vector of random numbers is stored in a random texture which is updated once by the fragment program after finish each main loop.

5. Simulation & Result

In order to evaluate the effectiveness of this parallel Hopfield network, a Test on Traveling Salesman Problem (TSP) by GPU was carried out. Performance difference between GPU and CPU is also given.

5.1 Traveling salesman problem

The TSP is regarded as one of the standard benchmark problems for evaluation of optimization algorithms, because of its two main features: the existence of much local minima in solution space which makes it really hard, and its practical meaning. It is usually stated as follows: a group of cities to be visited and distance between any of two cities are given, the purpose is to find the shortest tour course that visits each city only once and then return to the starting point.

Assume that the number of cities is N and the distance between the cities x and y is d_{xy} with $x, y \in \{1, \ldots, N\}$, The Hopfield net approach to the TSP uses an array of $N\times N$ neurons, each row and each column is associated respectively to a particular city and visit order in the tour. The output of the neuron in the *i*-th row and *j*-th column is denoted by v_{ij}. If the *i*-th city is visited in *j*-th positon,v_{ij}=1.otherwise v_{ij}=0.

The generic forms of the energy function for the TSP as the one presented in [1] is used. Namely, the following energy function is used:

$$\begin{aligned}E&=\frac{A}{2}\sum_{x=0}^{N-1}\sum_{i=0}^{N-1}\sum_{j=0,j\neq i}^{N-1}v_{xi}v_{xj}\\&+\frac{B}{2}\sum_{i=0}^{N-1}\sum_{x=0}^{N-1}\sum_{y=0,y\neq x}^{N-1}v_{xi}v_{yi}\\&+\frac{C}{2}\left(\sum_{x=0}^{N-1}\sum_{i=0}^{N-1}v_{xi}-N\right)^2\\&+\frac{D}{2}\sum_{x=0}^{N-1}\sum_{y=0,y\neq x}^{N-1}\sum_{i=0}^{N-1}d_{xy}v_{xi}(v_{y,i+1}+v_{y,i-1}).\end{aligned} \qquad (15)$$

To map the TSP onto the Hopfield network, the expressions of weight and bias are rewritten as:

$$\begin{aligned}w_{xi,yj}&=-(A\delta_{xy}(1-\delta_{ij})+B\delta_{ij}(1-\delta_{xy})\\&\quad+C+D(\delta_{i,j-1}+\delta_{i,j+1})d_{xy}),\\ \theta_{xi}&=CN,\end{aligned} \qquad (16)$$

$$\text{where},\delta_{mn}=\begin{cases}0 & m=n,\\ 1 & m\neq n.\end{cases}$$

5.2 Setting for test

Tests were performed for TSP problems with different size from 10 to 32 cities in the increment of 2 cities. The distances between the cities were initialized with uniformly random numbers within the range of [0.0, 1.0]. It is noted that a random choice of a path should result in an average distance of 0.5 between any two neighboring cities. For each size 10 instances were created, and for each instance 5 runs were carried out both on GPU and CPU.

To initialize our TSP, we set $A=B=50, C=20, D=1$ in Eq(16). In the annealing schedule of Eq(9) of parallel Hopfield network, parameter sets were: $\beta=0.8$ and $T_0=2.0$, and the time step Δt was selected to a very small value (0.001) to ensure slow annealing.

The computation time of GPU includes the data upload time to video memory, data download time from video memory and main cycles.

The simulations were performed on a Pentium 4 2.8GHz CPU with 512 MB RAM and an NVIDIA GeForce 6800GT GPU.

5.3 Result & analysis

The average distance between cities optimized by CPU and GPU were almost the same which is around 0.27 (a fairly good result comparing with the result obtained by other algorithms [27]). Fig. 4 shows the comparison of the calculation time by CPU and GPU.

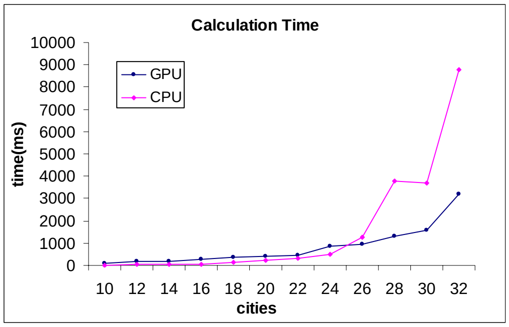


Fig.4. Calculation time of TSP

From Figure4, it can be concluded that when optimizing the TSP with cities less than 26, parallel Hopfield network run faster by CPU than by GPU; and as the number of cities increases, GPU demonstrates its superiority on parallel algorithm compared with the CPU case. The test results show: The process capability of fragment processor is much lower than that of CPU, and its high performance is achieved only when plenty of data runs on fragment processor in parallel. Consequently, GPU is not useful over CPU in a small data flow.

6.Conclusion

This paper presents a novel implementation of parallel Hopfield network on commodity graphics hardware. In our approach, a representation of weight matrix of Hopfield network suitable for GPU processing is demonstrated. All the Hopfield operators have been implemented on GPU. Tests on TSP show that, the larger the problem size is, the better speedup over the software implementation can be achieved. Our work has provided a powerful but inexpensive hardware platform for implementation of parallel Hopfield network. The solution is promising because programmable GPUs are on a much faster performance growth than CPUs.

There is one constraint in our approach. We can only use GPU to solve optimization problems with small size, because the weight matrix is stored in one single 2D texture and thus the size of Hopfield network was limited by graphic hardware. Future work is to find better dense matrix representation methods in addition to the single 2D texture.

Reference

[1] J.J. Hopfield, and D. W. Tank, "Neural Computation of Decisions in Optimization Problems," Biological Cybernetics, 52,141-152, 1985.

[2] K. A. Smith, “Neural networks for combinatorial optimization: A view of more than a decode of research,” INFORMS Journal on Computing, vol. 11, no. 1, pp. 15–34, Winter 1999.

[3] G. V. Wilson and G. S. Pawley, “On the stability of the tsp algorithm of Hopfield and tank,” Biological Cybernetics, vol. 58, pp. 163–70, 1988.

[4] S. V. B. Aiyer, M. Niranjan, and F. Fallside, “A theoretical investigation into the performance of the hopfield model,” IEEE Transactions on Neural Networks, vol. 1, pp. 204–215, 1990.

[5] A. H. Gee and R. W. Prager, “Limitations of neural networks for solving traveling salesman problems,” IEEE Transactions on Neural Networks, vol. 6, pp. 280–282, 1995.

[6] Kirkpatrick, C. Gelatt, and M. Vecchi, “Optimization by simulated annealing,” Science, vol. 220, pp. 671–680, 1983.

[7] G. Hinton, T. Sejnowski, and D. Ackley, “Boltzmann machines: Constraint satisfaction networks that learn,” Carnegie Mellon University, Tech. Rep. CMU-CS-84-119, May 1984.

[8] H. Jeong and J. H. Park, “Lower bounds for annealing schedule for boltzmann and cauchy machines,” in Proceedings IEEE International Joint Conference on Neural Networks, 1989, pp. 581–586.

[9] Y. Akiyama, A. Yamashita, M. Kajiura, and H. Aiso, “Combinatorial optimization with gaussian machines,” in Proceedings IEEE International Joint Conference on Neural Networks, 1989, pp. 533–540.

[10] Fermando, R., Kilgard, M.J.: The Cg Tutorial. Addision-Wesley (2003)

[11] Thompson, C.J., Hahn, S., Oskin, M.: Using modern graphics architectures for general-purpose computing: a framework and analysiy. In: Internaltional Symposium on Microarchitecture, Istanbul, Turkey, IEEE Computer Society Press (2002).

[12] Krger, J., Westermann, R.: Linear algebra operators for gpu implementation of numerical algorithms. ACM Transactions on Graphics 22 (2003) 908–916 .

[13] Harris, M.J.: Real-Time Cloud Simulation and Rendering. Dissertaion, University of North Carolina at Chapel Hill (2003).

[14] Bolz, J., Farmer, I., Grinspun, E., Schroder, P., Schrder, P.: Sparse matrix solvers on the gpu: Conjugate gradients and multigrid. ACM Transactions on Graphics 22 (2003) 917–924.

[15] Hillesland, K.E., Molinov, S., Grzeszczuk, R.: Nonlinear optimization framework for image-based modeling on programmable graphics hardware. ACM Transactions on Graphics 22 (2003) 925–934.

[16] Govindaraju, N.K., Lloyd, B., Wang, W., Lin, M., Manocha, D.: Fast computation of database operations using graphics processors. In: International Conference on Management of Data. (2004) 215–226.

[17] M. Muller, L. McMillan, J. Dorsey, R. Jagnow, Real-time simulation of deformation and fracture of stiff materials, in: Proc. of the Eurographic workshop on Computer animation and simulation, 2001, pp. 113-124.

[18] Aarts. E. and Korst. J, "Simulated Annealing and Boltzmann Machines A Stochastic Approach to Combinatorial Optimization and Neural Computing", J.Wiley & Sons.

[19] K.Fatahalian, J.Sugerman, and P. Hanrahan. Understanding the efficiency of gpu algorithms for matrix-matrix multiplication. In ACM SIGGRAPH Graphics hardware, 2004.

[20] MORAVANSZKY A.: Dense matrix algebra on the GPU, 2003. http://www.shaderx2.com/shaderx.PDF.

[21] Aaron.L, Joe,.K and John.O. Implementing Efficient Parallel Data Structures on GPUs. In Matt Pharr, editor, GPU Gems 2, pages 521–546. Addison Wesley, 2005.

[22] Dominik.G," Playing Ping Pong with Render-To-Texture", 2004.

[23] Emmett.K and Randima.F. The Geforce 6 Series GPU Architecture. In Matt Pharr, editor, GPU Gems 2, pages 471–491. Addison Wesley, 2005.

[24] NVidia, "The Cg toolkits-user's manual",2004.

[25] Daniel.H. Stream Reduction Operations for GPGPU Applications. In Matt Pharr, editor, GPU Gems 2, pages 573–590. Addison Wesley, 2005.

[26] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C++: The Art of Scientific Computing. Cambridge University Press (2002).

[27] G. Serpen, A. Patwardhan and J. Geib, "The Simultaneous Recurrent Neural Network Addressing the Scaling Problem in Static Optimization," International Journal of Neural Systems, Vol. 11, No. 5, pp. 477-487, 2001.

Learner centred ontology driven approach to e-learning

A.S. Karunananda
Department of Mathematics & Computer Science
Open University of Sri Lanka
Nawala, Nugegoda
asokakaru@yahoo.com

Abstract

This paper argues that e-learning could be made more humanized through an approach to learner's perspective driven ontology for learning. We have researched into Buddhist philosophical study on learning Ontologies and exploited the concept of 'Iddhipada' for developing a novel e-learning ontology. The Iddhipada ontology proposes to realize the learning process through four mental factors, namely, desire, effort, thought and investigation. The paper has presented a design and implementation of an ontology driven e-learning Agent that exploits the concept of Iddhipada. The iddhipada ontology kinks up learning components such as learning materials, assessment material and further assistance in an e-learning environment, and maintains a learning ontology in terms of factors in the Iddhipada. The Iddhipada ontology is customisable for learners and evolvable during the learning session.

1. Introduction

E-learning has introduced a new dimension for distance and open learning. This trend has particularly been influenced by the rapid growth of learning resources available on the Internet. However, from educational viewpoint, it would be interesting to discuss whether the mere availability of massive collection of learning materials is adequate for a successful learning process. Undoubtedly, huge collection of learning materials is a bonus for learning, yet they must be supplied in a manner which is agreeable with pedagogical aspects in education. In fact, this question reminds of us the real world experience that persons with huge collection of knowledge do not guarantee to be necessarily good teachers. On the other hand, it is already evident that learning by navigation/browsing through a large collection of web-based materials is rather uncomfortable and one may even loose the track in the middle of the learning process. As such, unless a learner planes very carefully on how to navigate through material, learning from web-based materials will be counter productive.

Obviously, by default, web-based materials are fine for browsing and finding answers quickly through surface learning, but not for learning consciously and analytically. Furthermore, educational researches have shown that learning by reading is less effective than learning from a teacher. There is also a world-wide trend that students are moving away from habit of reading on the average. Stated another way, may students wish to acquire knowledge though a process, which encourages of minimal reading. From educational viewpoint, reading comprehension is a must for students regardless of modern technologies used for education. Therefore, we argue that e-learning should be concerned about pedagogical aspects of education, rather than mere provision of huge collection of learning materials with a little guidance to navigation through techniques such as keeping bookmarks and even the use of the modern Agent technology at a very superficial level.

We fundamentally believe that the best way to device a successful e-learning model is to gain the insight from the practices that are used by good teachers for their teaching tasks. It is a well-known fact that good teachers always use a learner centred approach to teaching. In such a teaching process, the teacher associates a learner's knowledge with the appropriate materials, conducts assessments, train through additional support and also provide psychological boots to the learner. A successful teacher always brings a student into an psychological environment within which the student can learn properly. In this sense, we postulate that learning is a process, in which the learner evolutionary develops his own ontology of the subject of learning.

In this paper, we present a design of an e-learning Agent that mimics the role of a good teacher. Our e-learning Agent uses a learner's perspective driven ontology approach to drive the teaching process. The theoretical basis for constructing the e-learning ontology has been supported by the studies about learning Ontologies, which are presented in Buddhist Philosophy. Our research has exploited the particular learning ontology known as Iddhipada.

The rest of the paper is organized as follows. Section 2 discuss e-learning from an educational viewpoint. Section 3 explains why it is important to realize learning as an ontology driven process. Section 4 is about Buddhist philosophical studies on ontology. This section describes various learning ontologies in Buddhism and point out importance of 'Iddhipada' ontology for driving the process of learning. Section 5 describes design and implementation of the "Iddhipada' ontology with the other components of an e-learning environment. Section 6 concludes the paper with a not on further work of our research.

2. E-learning – paradigm for education

Education intends to produce citizens with high quality in their lives. In achieving this education generally expects learners to gain knowledge, develop attitudes, acquire skills and also practice what is learned. As such education is necessarily more than gaining of knowledge. In the past, access to education was so restricted and learners are supposed have a close association with the teachers during the learning process. Despite this approach could cater for a selected group of persons in the society, it has ensured the production of quality persons, who have improved through almost all dimensions in education. However, as time goes on gaining knowledge has become more important than achieving other aspects of education. As such, people become more interested in gaining knowledge without going to a teacher necessarily. In other words, reducing the physical gap between the teacher and the learner has become a major challenge in education. In this sense, various educational technologies including distance and open learning were developed to disseminate knowledge by breaking the barrier of distance between teacher and the learner. Of course, distance learning concept has also appreciated the value of teachers' involvement in the learning process and encouraged at least limited number of sessions, where learners directly interact with their teachers. This is why all Open Universities world-wide practice concept such as Day-schools and Tutor Clinics, where students meet their teachers time to time. It's our experience that two hours of face-to-face session has a great impact on improving students, who could not clarify various points through mere self-study process for several months.

It is evident that, computer-based e-learning has shifted the traditional textbook-based distance learning into a new direction. E-learning is also a cost effective means of education for various reasons. In fact, e-learning can support the education through various means such as animations, simulations, etc., which cannot be provided by a teacher in an ordinary class room scenario.

With the approach to detaching learner from the teacher, mankind produces more and more knowledgeable persons fast, yet the quality of life of such persons appears below the expectation. This has resulted in several issues such as graduate unemployment, lack of communication skills, human qualities/values, etc. Modern trends in educations including e-learning, web-based learning, etc. have also fundamentally caters for producing only knowledgeable persons through education. Unless it is carefully address, e-learning cannot effectively support other aspects of educations such as development of attitudes, skills and practicing of what is learned. In some disciplines, e-learning will have a minimum use. For example, in general, a medical student cannot be given skills as a medical surgeon though e-learning. This is why many modern educationists reemphasise on the importance of humanization of education process.

However, it should be pointed out that modern approaches to education can be benefited by developments in areas such as Artificial Intelligence (AI), for humanizing the process of education to a large extent. In other words, while restricting the teacher's involvement in the learning process, AI techniques can emulate some of the role of good teachers into e-learning.

We argue that Agent technology, which is also known as a modern approach to AI [12], can be used for supporting an e-learning process, by emulating the role of a good teacher. We also argue that Agent must be powered by a learning-ontology driven approach, as ontology can be developed as a comprehensive and customizable structure of generic elements that describes the process of learning. Theoretical foundation of learning ontology will be provided by Buddhist philosophical studies of learning Ontologies.

3. Learning from ontological viewpoint

Ontology means the theory of what exists in a particular domain. Ontology of a particular domain provides the basis for communicating about the domain. Ontology is a relatively new term in Computer Science and there are so many viewpoints about what ontology is.

For example Guarino [5] has discussed at least seven different viewpoints about ontology. A working definition for ontology can be treated as a description of what exists in a particular domain. In this sense, ontology is a comprehensive structure comprising of concepts, relationships, associated theories, etc. of a domain. It provides a more holistic view than what the concept of objects states about a particular domain. Perhaps, Gruber's definition [4] about ontology, i.e. "explicit specification of a conceptualisation" is the one, which has been commonly used by the computing community. However, as Sowa [15] has correctly pointed out ontological studies in Computer Science still lack a theoretical basis and need to gain insight from philosophical studies about ontology.

The concept of ontology is now evolving as a new paradigm for computing too. As such, it has been shown that some of the real world problems can be modelled from an ontological viewpoint, which is more comprehensive than object-oriented thinking of real world problem solving. For example, online bookstore may have ontology that describes, details of a book, related books, other books by the same author, customer assessments of the book, shipping details of the book, etc.

At present, ontology servers such as Ontolingua server [3] that holds Ontologies for various domains, are available on the Internet and one can download Ontologies exactly the same way we download objects and classes for our applications.

We argue that e-learning should also be thought in terms of Ontologies. There are several reasons as to why e-learning should be looked from an ontological concern. Firstly, learning is a process, which is associated with so many mental attributes such as desire, effort, curiosity, fear, anger, etc. Each of these factors is associated with various values and some may also negatively contribute to the learning process. Secondly, values of attributes change during the learning process and there cannot be an evolving learning if these attributes and values do not change. As such learning should be recognized as an evolving process. Thirdly, when facilitating the e-learning of different learners, we should be able to customize the support through the appropriate learning attributes. Fourthly, a successful e-learning session must maintain learners' profile and guide through the course material, assessment of performance, etc. accordingly. It is clear that a successful e-learning session cannot treat the above aspects as disjoint elements, but highly interrelated and dependent. This is why we argue that it is appropriate to look at learning in terms of ontologies, which provide a holistic view.

Undoubtedly, a successful teacher deals with all the above and even more aspects, but he may not be able to guide individuals depending on the status of learning ontology of each student at the same time. Instead, teacher may use a general learning ontology, which may apply to majority of the students in a classroom scenario. It is our experience that generic approach to teach all the students may not be appropriate in some situations. Therefore, consideration of individual learner's perspective is of great importance for a successful teaching session.

However, ontology driven Agent can perform even better than an average good teacher, because Agent can specifically facilitate e-learning process depending on the user. In addition, Agent can also consider learning ontology of many individuals and evolve a more general ontology. Ideally, learning ontology of an individual can be used as an identity of a learner. Further, learning Ontologies of different persons will be able to use by others to learn how learning should be done. Characteristics of a good learner are of great importance for students and researches in education too. This can be learned from learning ontology of individuals.

Thus far discussion exemplifies the importance of defining a comprehensive ontology for e-learning. Undoubtedly, such ontology should not be proposed in an ad-hoc manner, yet requires a theoretical basis from an educational viewpoint. Next section describes how we exploit Buddhist philosophy for defining a simple ontology for facilitating e-learning.

4. Ontological concern in Buddhism

Despite Buddhism is known as a religion it has a strong philosophical basis, which fundamentally discuss how people understand or learn about reality. Therefore, Buddhism is a philosophy about learning. Buddhism has already been used for various scientific studies on logic, reasoning, ontology, psychology and construction of models of mind. For examples, many researches in psychology have investigated Buddhist philosophical studies of mind [14]. Furthermore, Karunannda [6] has developed a computer model of Buddhist theory of mind. A theoretical foundation for some heuristics used in Artificial Neural Networks has also been provided with use of Buddhist theory of mind [7]. Buddhist philosophical viewpoints of Ontology have also been a research theme for realizing ontological modelling in Computer Science and Information Systems [8].

More importantly, in the context of philosophy, Buddhism emphasises on mechanisms for understanding of the reality rather than believing in some supernatural and eternal power in the universe. This distinguishes Buddhism from most other religions. In particular, Buddhist philosophy has provided an in-depth study about the mental phenomenon of understanding, which is an essential element of learning. Buddhism necessarily identifies learning as an ontology driven process.

From Buddhist philosophical viewpoints, Ontologies are highly personalized and liable to be evolved during the interaction in the learning process [9]. It's a unique contribution from Buddhist philosophy that Ontologies can be treated as evolvable structures, which can capture individual perception of a domain. This is a radically different viewpoint, as the most current approaches ontological modelling considers ontology as eternal and comprehensive structures, that are virtually fully defined and no rooms for evolution or capturing individual perception. Buddhism has not only discussed importance of individuality for ontology, but also described the use of mental factors for capturing individually into Ontologies. Research has also been conducted to develop individual perception-based evolvable Ontologies [9]

Buddhist philosophy encourages an individual to begin with own-partially complete ontology and let the ontology to evolve through learning or interaction. Buddhism very strongly emphasises on the fact that different persons understand differently. As such learning process is different from person to person, so is learning ontology. In fact, Buddhism very clearly says that unless the appropriate Ontologies is used, an individual can understand nothing properly. There are almost 20,000 discourses presented on individual basis, in Sutta pitaka in Buddhism [11, 13]

4.1. Learning Ontology in Buddhism

While emphasising on individuality of understanding, Buddhism has presented several major top level Ontologies. One can associate his/her current understanding with those ontology(ies) and develop on the ontology. Buddhism also encourages interaction among different Ontologies as a powerful way of learning.

It should be noted that Buddhism presents so many top level learning Ontologies and some of they are too specific and scope of Buddhism. For example, *Panchanevarana*, *Satipattana*, *Saptavisuddhi*, *Eight-fold path*, *paticcasamuppada, pattana, Bojjanga* and some parts of *Bodhipakkshiya dhamma* are very specific to key scope of Buddhism [3, 11, 16] and too general for developing a computer model.

Our research shows that Buddhism provides several other simple Ontologies that can be used to develop computable learning Ontologies. These learning Ontologies are presented as a set of mental factors that contributes to learning and understanding. Next we discuss these major learning Ontologies with a particular emphasis on what is known as ontology of Iddhipada or accomplishment.

4.1.1 Iddhipada learning ontology

Iddhipada ontology comes as major section within *Bodhipakkshiya dhamma* in Buddhist philosophy [1]. The term Iddhipada means the factors that make things possible with a particular emphasis on the phenomena of understanding, which is essential to successful learning. Perhaps the term 'accomplishment' would be the closest English translation for the term Iddhipada[11]. The iddhipada ontology includes four mental factors, namely, of *desire*, *effort, thought*, and *investigation*. These factors positively contribute in a learning process. Note that, beyond Iddhipada, Buddhist philosophy provides an in-depth study of mental factors, which are counted as 52 factors. These factors are of great research interest in Western Psychology too [14]. It is beyond the scope of this paper to discuss all such mental factors, yet interesting reader may refer to sources such as [1] for additional information about mental factors. A brief description of elements of Iddhipada is as follows.

Desire: In the iddhipada ontology, the factor of desire has been identified as a key mental factor that makes things possible. This suggests that a successful learning session should maintain a continuous desire regarding the subject matter and the learning process. The importance of desire as a learning factor is unarguable and also matches with our day-to-day experience too. Buddhism has also emphasised on the importance of balance of desire for successful learning. For example, too much desire can give negative effect in learning.

Effort: Effort is another mental factor contributing to the success of learning. This finding is also compatible with our real world experience that mere desire is not adequate yet the effort in the learning process is important. As such maintaining a continuous effort in a learning process is very crucial for success in learning. Balance in the effort has also been recognised as a key factor for learning. According to Buddhism too much effort leads to restlessness in the mind and may distract the potential to learn.

Thought: The factor thought (or thoughtfulness) refers to one's ability keeping the mind in the track until he/she achieves the goal. This is an important factor since one can have both desire and effort maintained properly yet may be compelled to involve in other activities by giving the second priority for the learning task at hand. Of course, we are faced with this issue quite often. This is a rather difficult factor to improve as we are affected by so many things such as social commitments, fear, anger, etc. Obviously, a good learner should to be able keeping the track in the learning process without unreal breaks.

Investigation: The investigation is yet another factor that contributes to improve a learning process. In particular, investigation is a crucial factor when one wants to perform learning in an enthusiastic and analytic manner. It should also be noted that from day-to-day life experience investigation is associated with resolving doubts, logical analysis, restlessness, etc. It's our experience that investigation is an immensely crucial factor in the context of self learning. In a broader sense, e-learning is also an version of self-learning. As such modelling of e-learning through the factor of investigation is of trivial importance.

4.1.2 Other learning Ontologies

Buddhism has identified various other Ontologies, which are related to a learning process. Some ontologies identifies the factors that gives a negative effect in learning, while others cause to give positive results. For example, as we have already explained, the Iddhipada identifies mental factors that work positively in a learning process. In contrast, *Pachaneevarana* ontology defines five factors, namely, *attachment, anger, laziness, restlessness* and *doubt*, which give a negative effect in learning. On the basis of such an ontology, it can be argued that a learning ontology can be developed to realise a learning process through the opposite of those mental factors. Stated another way, such ontology should comprise of attributes, namely, non-attachment, non-anger, activeness, etc. Further more, Buddhism provides *attachment, aversion* and *ignorance* also as the factors of yet another ontology, which describes the negative features of a learning process [3].

We argue that it would be more appropriate to develop a computable learning ontology through the positive factors of learning, rather than on the basis of negative factors. Obviously, it would be computationally counter productive to compute negative learning effects and then compute the progress of learning. In Buddhism, there are more learning Ontologies with positive factors than those with negative factors. Below are some more examples on Buddhist ontologies with positive learning factors.

Ontology known as *five forces* identifies the factors of *confidence, effort, mindfulness, concentration* and *wisdom* [11] as another set of factors, which makes a positive effect in learning. The same factors are also known as *five faculties*. Although this ontology is as simple as the Iddhipada ontology, some of factors are difficult to implement in a computer model. For example, the factor mindfulness, which refers to being aware, cannot be readily implemented on a computer. On the other hand factors such as concentration and wisdom cannot also be readily implemented on a computer. Therefore, we do not select five forces ontology as a computable learning ontology at this stage. Note that the factor such as effort has been a positive learning factor as identified in iddhipada, five-forces and five faculties. As we explain shortly, the effort has been such an important learning factor, which appears in several other Ontologies too.

Sapthabodhyanga also proposes yet another set of positive mental factors that leads to proper understanding or learning. These factors include *mindfulness, investigation of the truth, effort, rapture, quietude, concentration* and *equanimity*. This particular ontology also provides several factors such as mindfulness, quietude, concentration and equanimity that cannot be readily realized through a computer model of learning ontology. Therefore, this ontology cannot also be readily computable without some further research.

However, we believe that Ontologies such as five forces, five faculties and sapthabodhyanaga and even *eight-fold-path* can provide more comprehensive learning Ontologies provided that we will be able to develop mechanisms for computer implementation of elements of these Ontologies. Nevertheless, it is obvious that much research is required before proceeding into computer implementation of such Ontologies.

It should also be noted that ontology with negative learning factors (e.g. *Pachaneevarana*) might not be suitable for developing a model of learning ontology due to two main reasons. Firstly, as we have already mentioned, such ontology describes a learning process through a set of negative aspects of learning. Secondly, some factors such as restlessness and doubt are not readily computable. Therefore, learning ontologies with positive mental factors, which are computable would be the fundamental requirement for exploiting a learning ontology from Buddhism.

5. Agent with *Iddhipada* Ontology

As per with our thus far discussion on research into Buddhist philosophical studies on learning ontology, we have selected Iddhipada learning ontology for developing an e-learning Agent in a novel way. The major reasons for choice of iddhipada ontology for development of a computable e-learning ontology are summarised as follows.

- Learning factors in Iddhipada ontology match with our real world experience
- Iddhispada ontology provides a set of positive learning factors
- Iddhipada ontology is comparatively small, yet comprehensive enough for describing a learning process
- Learning factors in iddhipada ontology are easily computable

5.1 Design of novel e-learning ontology

Having exploited Iddhipada ontology from Buddhism, now we proceed to postulate an e-learning ontology. It should be noted that Iddhipada is only the backbone of the proposed learning ontology, yet it must be associated with other elements of a learning system to form a workable and comprehensive learning ontology. For this purpose, we postulate that Iddhipada shall be associated with the following essential components of a general e-learning environment.

- Learner profile
- Learning materials
- Assessment materials
- Additional learning support

The attributes of the Iddhipada ontology links up with the above components and forms a comprehensive e-learning ontology. Figure 1 shows top level design of the proposed ontological environment for e-learning

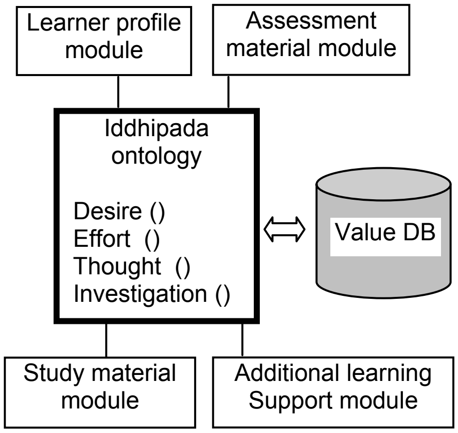


Figure 1 - Ontology-based e-learning system

The proposed e-learning environment must have modules for learner profile, learning materials, additional learning materials and assessment materials. Each attribute (e.g. desire, effort) of iddhipada has a reference to elements in the above four components/modules in a particular e-learning environment. These components are not necessarily located in a client, yet can be distributed on the Internet. However, Iddhipada ontology together with the Value DB preferably run on the client side.

In that sense, we begin with the assumption that there exists an e-learning environment and our research is to bring modules of a typical e-learning system, through Iddhipada into a special kind of e-learning environment. In association with attributes of Iddhipada we maintain a database called Value DB to store e-learning ontology of each learner. Below is a description of each module, with a more complete elaboration to our recent work of this research that was presented in [10].

5.1.1 Iddhipada Ontology

This is the backbone of the proposed e-learning ontology. It has no meaning in isolation. The Iddhipada ontology links up components of an e-learning environment (here with four modules) in terms of the four factors; desire, effort, thought and investigation. Note that depending on various supports available in a particular e-learning environment, the Iddhipada ontology links those through the four factors.

5.1.2 Learner profile module

The learner profile module stores information about e-learning clients. This information may include static information such as user ID, name, educational qualifications, subject of interest, and the previously created e-learning ontology of the learners who use the e-learning system. This is a typical database and can be implemented at the client side. Each user profile has a reference to attributes of iddhipada ontology and written to value database.

5.1.3 Learning material module

The learning material module consists of various e-learning materials, which are preferably available as hypertext documents. The study material module(s) can be seen as a web server located at an e-learning site. Note that this module is not necessarily located in a client machine of e-learning environment. For each learner, there are set of references to sections in study materials to the attributes of iddhipada and value, and also written to value database.

5.1.4 Assessment material module

The assessment material module is a typical question bank with various test kits. This is also preferably a collection of web pages. The learning material and assessment material modules can be seen as components of a typical e-learning environment. The assessment materials also refer to attributes of iddhipada and their values, and are written to value DB.

5.1.5 Additional learning support module

The additional learning module is also same as the learning material modules, yet facilitates the access to additional learning materials. They can be further reading materials and special materials suitable for different learners. In general a good e-learning environment should include a module of this nature. Our proposal is to refer those materials with the iddhipada. These are also linked up with value DB, exactly the same manner to handling learning materials module. Note that some of these materials can be outside a particular e-learning sever. Those resources can be some popular web sites, other e-learning systems, e-books, etc. The contents in such resources may also be linked up with the iddhipada as appropriate. However, it would be more practical not to go into detailed association of additional learning materials with iddipada.

5.1.6 Values DB

The value DB stores dynamic information attributes and vales referring to materials in four modules with respect to each learner. A value of an attribute of iddhipada describes how a particular learning material, assessment, etc. being associated by the learner. For example, we may say that a particular learner has studied section 4 of the study material with low effort. In this scenario, the value of the attribute of the effort is *low.* Of course, these values can be defined as either qualitative or quantitative measures. It is beyond the scope of this paper to discuss how these values can be calculated. However, it would be essentials to point out the importance of these values from Buddhist philosophical viewpoint.

One of the most important finding from Buddhism is that factors in iddhipada ontology may apply to two persons with two different levels of importance. Stated in another way, for example, effort may be the dominating learning factor for a particular person. Another person may have learning process, which is mostly driven by, say investigation. Sutta pitaka in Buddhism has provided various examples to show how different factor in Iddhipada contribute learning of different individuals. For example, Chandawimala [1], has quoted the story of Rattapala bikkhu whose realization of the truth was dominated by the learning factor of desire. Currently we are in the process of researching into quantification of values for attributes of Iddhipada ontology.

Note that the design model presented here describes only how the iddhipada ontology can be applied on an e-learning system. It is not complete yet good enough to demonstrate the philosophy behind our approach to support a process of e-learning. In a similar manner with this basic design, iddhipada can be linked up with other modules in a complex e-learning system. The need for relating the learning elements in an e-learning system, with the factors in iddhipada is the key emphasis that is shown in the basic design of the proposed approach.

5.2 Implementation

We propose to implement the novel e-learning approach as a customizable software that can run on client side of an e-learning session. More importantly, e-learning ontology will be built into a software Agent. With this approach we will be benefited by the features of the Agent technology for the use of the proposed e-learning ontology. Within the Agent, it is proposed to develop iddhipada ontology as a class, whose attributes are the learning factors of *desire, effort, thought* and *investigation.*

Each attribute has a reference to learning/assessment materials and also the appropriate value. For this purpose, the attribute value database can be constructed through a simple data structure as follows.

AttributeName(*ModuleName*(Ref), Val)

For example, *desire*(*learning_material*(#2), good) can be an entry in a Value DB, with regard to a particular learner, who has used section #2 of learning material module with the value good. Since the learner profile is maintained separately, we can minimize the data stored in the Value DB. Note that Value DB consists of only the dynamic information, while the learner profile includes all historical details of a learner.

For a given learner, a particular learning session initiates an e-learning ontology with reference to the learner profile. The learning session begin with last or the recent ontology created in the learner profile module, and evolve its values during the current learning session.

It's obvious that components of Iddhipada ontological system can be implements using object oriented concepts. For example, a simple implementation of user profile class and Iddhipada class can be written as shown in Listings 1 and 2

```
public class UserProfile
  {
      String name;
      String userID;

  public void createSession()
    {
  Iddhipada I = new Iddhipada(name,userID);
    }
  }
```

Listing 1 - User profile class implementation

```
public class Iddhipada
  {
     int desire;
     int effort;
     int thought;
     int investigation;
     String sesID;
     String UserID;

     public Iddhipada();
     public Iddhipada(String newSesID,String   UID);
     public void changeDesire(int Value);
     public int returnDesirevalue();
  }
```

Listing 2 - Iddipada class implementation

Note that Listing 2 shows only the functions of attributes of desire, and similar functions can be written for other attributes too.

It should be noted that, nowadays there are tools available for construction of ontologies without going into programming level. We also intend to use the famous Protégé 2000 [17] for construction of our ontological system.

In a learning session, e-learning ontology can be used to drive the entire learning process. For example, if the learner is weak in particular section, he would be suggested to do additional materials. Further, a fast learner may be directed with additional materials, which might improve the learner's curiosity or investigation. In this manner, e-learning ontology looks after improvements of desire, effort, thought and investigation of a learner. Emerging ontology of a particular learner can also be used as a means for monitoring the progress of a learner.

Since the e-learning ontology is developed as an Agent, it can also work proactively and find additional support for a particular learner in his absence. More importantly, the Agent can be improved to learn from e-learning Ontologies from different learners and recognize how the ontologies of good learners have evolved and use them for guiding learners with low performance. This is something that a good teacher hardly applies in his teaching process.

In general e-learning systems do not provide a principled approach to guide a learner through the e-learning materials. However, we claim that our e-learning Agent guides a learner with reference to development of the learner's desire, effort, thought and curiosity. The Agent can always guide the learner such a way that low profile iddhipada will be improved through the learning session. The Agent can also use the inherently strong iddhipada of a learner as a means for effective learning. For example, if investigation has been the most effective learning factor of a learner, such a person may be given more and more additional and challenging materials. Further, if a learner is much thoughtful, the system will provide materials for relaxation and taking the learner out of the main theme for a while. This would enable a leaner to refresh and develop desire through the learning session.

6. Discussion

This paper has argued for the importance of a learner centred approach to e-learning. The research was inspired by the lessons from the practices by the good teachers, who teach students by taking individuality into consideration. We argue that this approach makes the learning process more humanized as emphasised by research in education. In implementing the idea of learner centred learning, we have devised an ontology driven e-learning approach that captures learner's perspective as the basis for driving the learning session. Theoretical basis for construction of an e-learning ontology is provided by the Iddhiapda ontology, which is identified from Buddhist philosophy through our research into studies on learning ontologies. According to Iddhipada, learning process is driven in such a manner that a learner's desire, effort, thought and investigation will be improved during the learning session. We have presented design and implementation of computable model of Iddhipada ontology as a means of developing a customizable e-learning ontology that can be linked up with any standard e-learning system.

Further work of this research includes quantification of attributes of Iddhiapada ontology and development of a workable model of designed e-learning ontology. Its is intended to quantify the attributes through a survey research on how these attributes are used by some selected set of good teachers. It is our observation that teachers are quite capable of understanding students' desire, effort, thought and investigation in a very simple and qualitative manner. For example, a teacher may use success rate in assessments as a measure for development of effort. Therefore, we believe that insight from how good teachers realize students' performance in terms of attributes in Iddhipada would be of great relevance to our further work.

We have also planned to use Protégé 2000 [17] for construction of the iddhipada ontology. This strategy speeds up the development works without going into programming from the scratch. On the other hand, as Protégé 2000 has been used for development of many ontological systems, our system will be able to access a large set of ontologies on the Internet. It amounts to improve interoperability of our ontological system. Further, since Protégé is an open source it can be used for development and experimental purposes free of charge. Therefore, further work of this research work will be relatively less expensive.

Reference

[1] Chandawimala, R (1997), Description of Bodhipakshika Dhamma (in Sinhala), Sri Chandawimala Dharmapusthaka Sanrakshaka Mandalaya

[2] Farquhar, A., Fikes, R. and Rice, J. 1997. The Ontolingua Server: a tool for collaborative Ontology construction, *International Journal of Human-Computer Studies,* **46,** 707-727.

[3] Gorkom, N.V. 2004. *Abhidhamma and Practice*, http://www.abhidhamma.org/

[4] Gruber, T. A. (1993), *A Translation Approach to portable Ontology Specification*, Knowledge Acquisition, 5(2) 199-220

[5] Guarino, N. (1997), Understanding, building and using ontologies, *International Journal of Human-Computer Studies*, **46**, 293-310

[6] Karunananda, A.S. (1993), *Computer modelling of the thought process*, International Journal of Computer Applications in Technology 6(2/3), pp135-140.

[7] Karunananda A. S. (2002), *Using an Eastern Philosophy for providing a theoretical basis for some heuristics used in Artificial Neural Networks*, Malaysian Journal of Computer Science, Vol 15 No. 2, pp 28-33.

[8] Karunananda A.S. & Rzevski G. (2004), *Relevance of Buddhist Philosophy to ontological modelling in Information Systems and Computer Science,* In proceedings of the 2nd Annual Conference of Sri Lanka Association for Artificial Intelligence

[9] Karunananda A.S. & Rzevski G. (2005), *Ontological Modelling: State of the Art Unresolved Issues and New Research Directions*, Article submitted to the International Journal of Intelligent Systems

[10] Karunananda A. S. (2005), Learner's Perspective driven ontology for e-learning, Proceedings of the Eight International Conference on Humans and Computers, Japan, pp 143-149

[11] Narada Maha Thera (1956*), Manual of Abhidhamma*, BBD Power Press, India, 1956

[12] Russell S. & Norvig P. (2003), *Artificial Intelligence: A Modern Approach*, Prentice Hall

[13] Rahula Walpola (1974), *What the Buddha Taught*, Grove Press, Revised edition 1974.

[14] Robert A.F. Thurman (1993), *A Tebetian Perspective,* MindScience: An East-West Dialogue, Edited by Daniel Goleman & Robert A. F. Thurman, Wisdom Publications, Boston

[15] Sowa John F. (2000), *Knowledge Representation: Logical, Philosophical, and Computational Foundation*, Brooks Cole Publishing Co., Pacific Grove, CA

[16] www.buddhanet.net

[17] http://protege.stanford.edu/

Multimedia Based MCQ Composer for Evaluation of Students in the Medical Stream; in Sri Lankan Medical Schools

N.K.V.M.R Kumara,
Department of Biochemistry, P.O. Box 70, Karapitiya, Galle,
Sri Lanka.
E-mail: ruvinkumara@yahoo.co.in

Abstract

Information graphics effectively convey comprehensive information quickly and improves the metacognitive process of a person. Metacognition promotes higher-level thinking, allowing the mind to compare and contrast stored known data in the mind and relate them to the information being processed. This is very important in evaluating student's skill and the knowledge. A text based examination paper is the interface between student and the evaluator in evaluating student's knowledge which evaluate real learning experiences superficially. This is critical when the teaching and evaluating medium is not based on their mother tongue.

In this investigation we developed a program to replace conventional text based examination paper with the multimedia based computerized MCQ system with flexible computation to calculate total marks at the end of the MCQ paper which evaluates real learning experience effectively within limited time. The program includes image editing facilities like drawing tools, drawing property changing tools, effects tools, text tools, different types of filters and other tools required for editing images and text. Video & audio playing, editing with single video frame extraction, animation and spellchecker facilities are also incorporated with this application. Evaluation and comparison of PPT-MCQ and the MMB-MCQ prepared from this software revealed that the total number of pass students was significantly increased in MMB-MCQ test (24.3%) whereas the total number of failed students was significantly decreased (25.9%) compared to the PPT-MCQ papers.

Key words : MCQ, multimedia database

1. Introduction.

The text based examination paper is a mainstay in evaluating student's knowledge in Sri Lankan educational system despite of consideration in gathering building, recalling analyzing and processing their knowledge based on stored information in the memory in combination of vision, hearing, touching and sensing experiences from various sensory organs.

The text based examination paper commonly known as paper-and-pen testing (PPT) [1] is the interface between student and the evaluator in evaluating student's knowledge. This is critical when the teaching and evaluating medium is not based on their mother tongue.

English is the only medium in teaching and evaluating students in medical stream in Sri Lankan Medical schools. Although English is a very good medium in gathering information, developing their knowledge and becoming a life long learner, it is a barrier in evaluating their knowledge developed within 5 years, efficiently using text based examination paper within limited time.

Conventional existing evaluating systems of student knowledge are based on text based examination paper which is the main interface between the student's knowledge and the evaluators. It is obvious that one image or audio clip can replace thousands of words and one video could replace thousands of images.

Information graphics effectively convey comprehensive information quickly and improve metacognitive process of a person [2]. Student's knowledge and the skill could be evaluated by computer based testing (CBT) [3,4,5] and it could be achieved efficiently within limited time by introducing MCQ based on images, audio and video clips, interactive video and animations which are more close to real learning experiences and reduce the amount of text used in the evaluation process. Text based paper-and-pen testing MCQ (PPT-MCQ) paper may be compared with earlier computers with DOS command system and multimedia based MCQ (MMB-MCQ) with the operating system based on graphic user interface (GUI) which is very useful even for the beginner to understand and operate the computer easily.

Commercially available OCR-MCQ markers used in marking MCQ paper are very expensive and sometimes accuracy is not hundred percent reliable. We have developed a user friendly software to compose multidisciplinary multiple choice MCQ paper based on multimedia and flexible computing method that suit for different departmental requirements. Teachers in the medical school are very busy with their clinical work and it is bit difficult to make them to follow a image editing program like Adobe® Photoshop, Corel® Draw and video editing applications like Adobe® Premier. Our program comprise of image editing facilities like drawing tools, drawing property changing tools, effects tools, text tools, different types of filters and other tools required for editing images. Audio & video editing, playing and single video frame extraction, video creation and animation facilities are also incorporated within this application.

2. Method

The Visual Basic 6 language with structured multimedia database was used to develop this program according to architecture shown in fig.1. The program consists of two separate applications connected to a single database. MedEdCom (Fig.2) was created for teachers to compose multimedia based MCQ using Gdi32.lib and GDI plus converted from GDI plus of Microsoft platform SDK[6]. This consist of a MDI (fig 2), an image & text editor with bilingual (Sinhala &English) menus and drawing tools like curve, polygon, filter brush, different types of brushes, text, fill, rectangle, square, rounded rectangle, rounded square, ellipse, circle, pencil, eraser and pick tool. Selection tools consist of move, cut, copy, paste, delete, crop, apply effects and applying filters. Drawing property changing tools consist of foreground color, fill color, fill style, draw, width, border style and font. Effects tools consist of resize, flip horizontal, vertical and rotate. Different filters include black and white, blur, brightness, crease, darkness, diffuse, emboss, gray and white, invert colors, replace colors, sharpen, snow and wave. A flexible menu system has been introduced in this application that can be used either Sinhala or English. Sinhala menu works with the DrRuvinThin font and it should be installed in to the system. MCQ creator (Fig.3) with spell checker (Fig. 3) for the teachers was connected to a structured multimedia database. MCQ paper (Fig. 5) for the student consist of real MCQ paper connected to the same database and FlexGrid based answer recorder which was used to re-correct their answers with the flexible computation facility. Both MCQ creator and the MCQ paper were design in bilingual format which facilitate conduction of examinations in English as well as in Sinhala medium. The video frame extractor (Fig. 4), video creator & animator were programmed with the help of libraries kernel32, gdi32, avifil32, winmm and MSVFW32 in the modules and class modules. Video player was created using multimedia MCI control.

The system was evaluated using randomly selected medical students (n=66) and compared with the conventional text based paper-and-pen testing MCQ paper [6] at the end of the term. Normally all medical students are ready to face term test which cover all sections that they have studied during the term. Biochemistry test paper was prepared in both methods to cover all sections that they have learned. Twenty five MCQ were prepared in conventional text based paper-and-pen testing method and convert same questions to multimedia based MCQ by MedEdCom as far as possible using relevant Images, videos or animations to evaluate student's knowledge about sections that they have learned during the term. Sixty six medical students (1st year) were randomly selected and divided in to 2 groups of 33 each. The group A was given conventional text based PPT-MCQ paper and the group B was given MMB-MCQ with 75 min. time allocation for each paper. At the end of the paper only group B received their results. The two groups were crossover and the above procedure was repeated with another paper prepared from same sections that they have learned through ought the term. The standard of the questions was similar to the previous paper and Group B received conventional text based PPT-MCQ while the group A was given MMB-MCQ. Results of all four tests were compared using Student's-t-test. At the end of the examination questionnaire was given to all students participated in this investigation and collect data about their responds about both examination methods. The system also was evaluated with the students those who have failed (n=20) the 2nd MBBS examination and ready to face the repeat 2nd MBBS examination in Biochemistry. The paper was consist of 30 questions prepared as PPT-MCQ and converts them in to MMB-MCQ as far as possible with 90 min time allocation. The software was demonstrated to all the staff in the faculty of medicine University of Ruhuna Karapitiya and get their responses.

3. Results and Discussion

3.1 MedEdCom

Image and Text Editor.

Images like X-ray film, pathological slides, blood picture, pictures of patients, specimens were edited according to the requirement and inserted them in MCQ which evaluate the student's practical knowledge effectively and efficiently in wide range of study area(Fig2). Images also were created for animation purposes and send them to the video creator &animator.

Video Player.

This (Fig 5) is useful in inserting video clips to MCQ which can evaluate anatomical dissections, physiological testing, surgery, ultra sound, CT and MRI scan, animations etc. It also played MCQ bound video clips in student paper.

Video frame extractor.

This module (Fig 4) extracts the required frame from a video or animation and edited by an image editor as required for the MCQ. This application was used to extract a frame from ultra sound, CT & MRI scans, animations and create the MCQ.

Video Creator & Animator.

This is very useful in creating real time video from still pictures of bmp, gif and jpg format (Fig. 4). It also create animations from still pictures created from image editor MedEdCom itself. This is very useful in creating videos to explain all metabolic pathways, like cholesterol, carbohydrate, protein and drug metabolism, biochemical reactions, molecular mechanisms of all organs, all biological activities like nerve conductions, muscle action, hormonal actions heart mechanisms cell divisions…etc.

Multi Disciplinary MCQ Creator.

This is a real MCQ creator (Fig 3) which was used to create MCQ question, multiple choices and their answers very easily even without any previous experiences. Images were bound to the data base by pressing add button and right clicking on the image viewer in front of the MCQ. Video and audio also were bound to the database or in the separate folder relevant to the MCQ. The spellchecker (Fig. 3) is bound with the MCQ creator for spelling and grammar corrections.

Student's MCQ paper.

This shows (Fig 5) MCQ bound with images, video & audio clips, and interactive animations connected to the database created by MCQ creator. Students have many options for their answers like true/false, no mark which is achieved by left click, double click and right click on the check boxes. Normally the true option gets a plus mark, a false option gets a minus mark and the no mark get zero marks. This flexibility is very important because some departments carry over minus marks throughout the paper whereas other departments won't carry over and restrict minus marks within the question.

The FlexGrid answer recorder reflects all activities performed on the MCQ paper and provide maximum flexibility for answering and re-correcting the MCQ paper.

The students changed their answer at any time followed by concomitant computational adjustments automatically to give total marks at the end of the examination that can be printed through a network printer.

3.2 Evaluation of the system

Computer aided learning laboratory improves computer literacy of medical students in the faculty of medicine, University of Ruhuna. All medical students participated in this investigation were familiar with this software after the demonstration since their computer literacy was acceptable.

The total score of the students participated in both PPT-MCQ and MBB-MCQ were grouped in to four groups and calculated mean score.

	Score%	>60	60-50	50-40	<40
group A PPT-MCQ paper 1	Student (n) %	24.2%	18.1%	45.4%	13.1%
	Mean Score%	68.5±4	54.6±3	46.1±3	27.6±15
group A MMB-MCQ paper 2	Student (n) %	33.3%	42.4%	18.1%	6.0%
	Mean Score%	72.3±7	56.5±2	45±3	39.2±6
group B PPT-MCQ paper 2	Student (n) %	24.2%	30.3%	36.3%	9.09%
	Mean Score%	67.2±4	53.2±2	43.8±3	32.3±8
group B MMB-MCQ paper 1	Student (n) %	33.3%	36.3%	18.8%	9.09%
	Mean Score%	71.4±6	56.8±4	47.6±3	35.3±9

Table1. Percentage mean of the total score obtained by the students of group A and B for the MCQ paper 1 & 2 prepared in both PPT-MCQ and MMB-MCQ methods.

The number of students who scored marks between 60-50 was significantly increased in the group A (24.3%) sat for the MMB-MCQ compared to the PPT-MCQ of paper 1. Whereas significantly decreased number of students who scored marks between 50-40 were observed in the group A (27.3%) and group B (17.5%) sat for the MMB-MCQ compared to the PPT-MCQ of both paper 1 and paper 2.

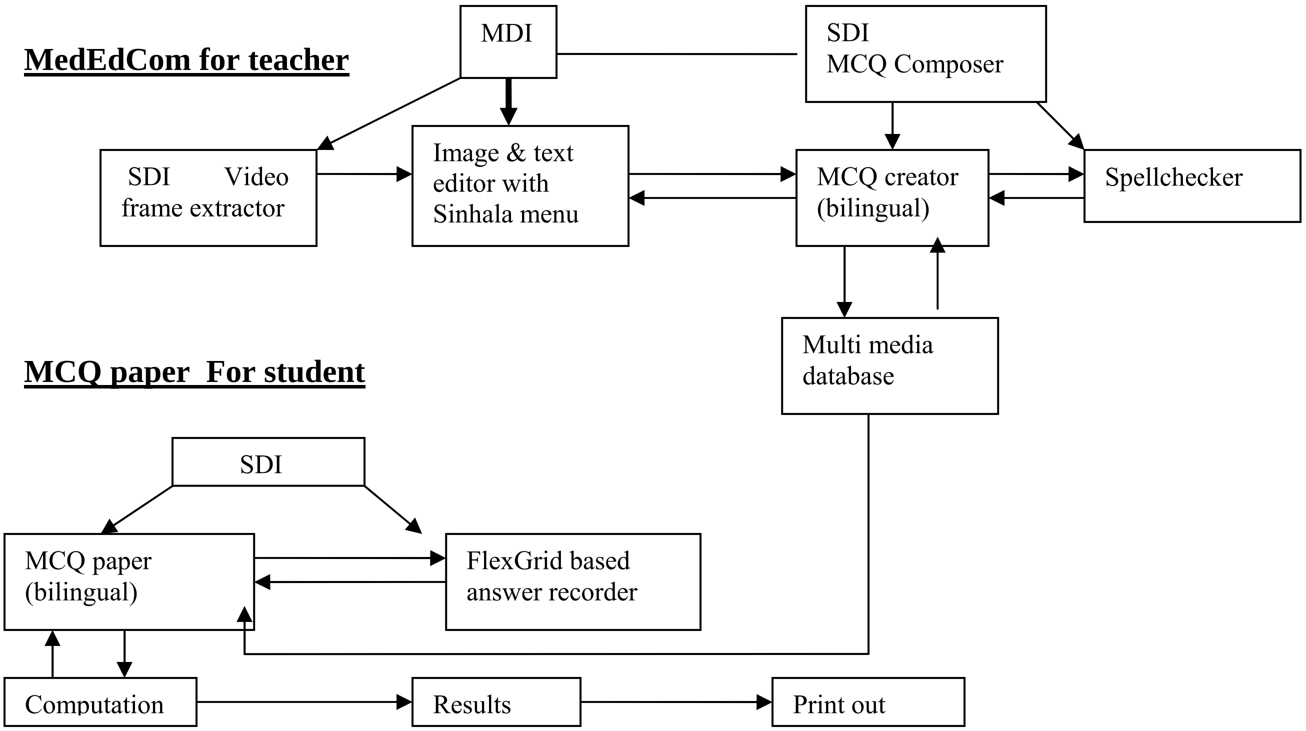


Fig. 1 Architecture of the program

The total number of pass students was significantly increased in MMB-MCQ test (24.3%) compared to the PPT-MCQ of both papers. Whereas the total number of failed students was significantly decreased in MMB-MCQ test (25.9%) compared to the PPT-MCQ of both papers.

The students who scored more than 60% range and 40-50% range obtained maximum benefit from MMB-MCQ (Table.1).

Significant changes were not observed in average score in all the groups whereas the number of students moved from lower score range to higher score range were significantly improved in MMB-MCQ system created from the newly developed software MedEdCom. All the students those who have failed their 2nd MBBS xamination and sat for this evaluation system obtained more than 50 marks and pass the test. Analysis of the questioner revealed that 70% of the students those who have used this MMB-MCQ system categorized as a good system for the student evaluation.

4. Conclusion

MMB-MCQ system evaluates a student's skill and the knowledge in wide range in the study area effectively and efficiently rather than the text based PPT-MCQ paper. Seventy percent of the sample student population agreed that the multimedia based computerized MCQ paper evaluates the student's knowledge effectively and more close to the real learning experiences compared to the conventional text based MCQ paper.

Acknowledgement.

I wish to acknowledge Prof. Susirith Mendis the Dean of the Faculty of Medicine for encouraging me to develop this program and giving me the opportunity to demonstrate and evaluating the program in the Faculty of Medicine, University of Ruhuna, Sri Lanka.

References

1. Luecht R.M, Hadadi A, Swanson.D.B, Case.S.M.(1998) A comparative study of a comprehensive basic science test using paper-and-pencil method and computerized formats. Acad Med;73 (octomber suppl); S51-3.

2. Boyle T., (1997). Design for Multimedia Learning, 45-63. London; Prentice Hall Europe.

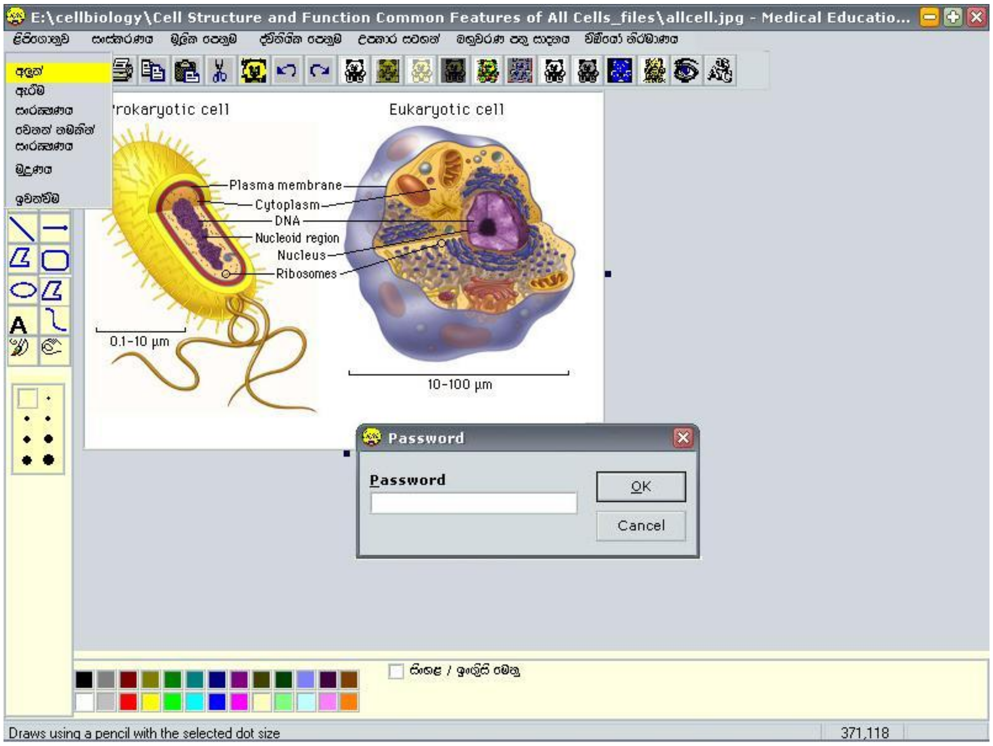
3. Rook C., (1994).Computers and the Collaborative Experience of Learning, 98; 105-137. London.

4. Whittlestone K., Williams J, (1995). CAL Scribe for school book, University of Bristol, Educational Technology Service.

5 Forker J.E, McDonald M.E (1996) methodological trends in the healthcare profession: computer adaptive and computer simulation testing. Nurse Educ; 21;13-14.

6. Genghis Khan(2003) Website: http://www.itkhan.com

7. Henri Wei; (1999), Computer Based Testing (CBT) and the USMLE. Medical Computing Today; September suppl; 19-21.



E:\cellbiology\Cell Structure and Function Common Features of All Cells_files\allcell.jpg - Medical Educatio...

Password

Password

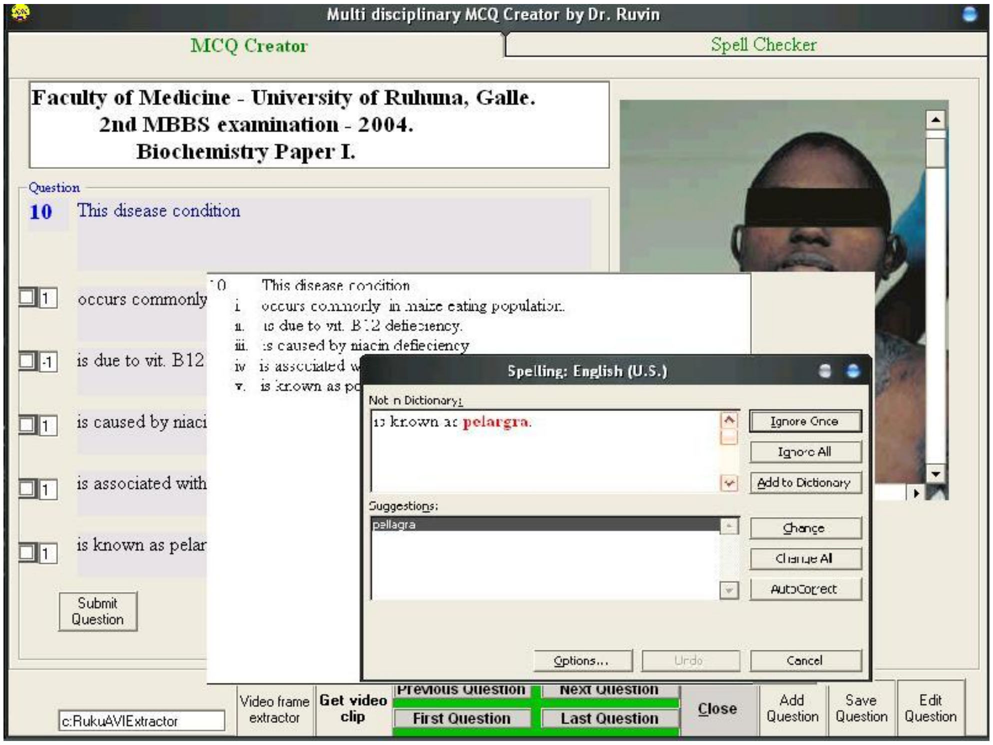
OK

Cancel

Draws using a pencil with the selected dot size

371,118

Fig. 2. MDI Image & text editor with password protection



Faculty of Medicine - University of Ruhuna, Galle.
2nd MBBS examination - 2004.
Biochemistry Paper I.

Question
10 This disease condition

□ 1 occurs commonly
□ -1 is due to vit. B12
□ 1 is caused by niaci
□ 1 is associated with
□ 1 is known as pelar

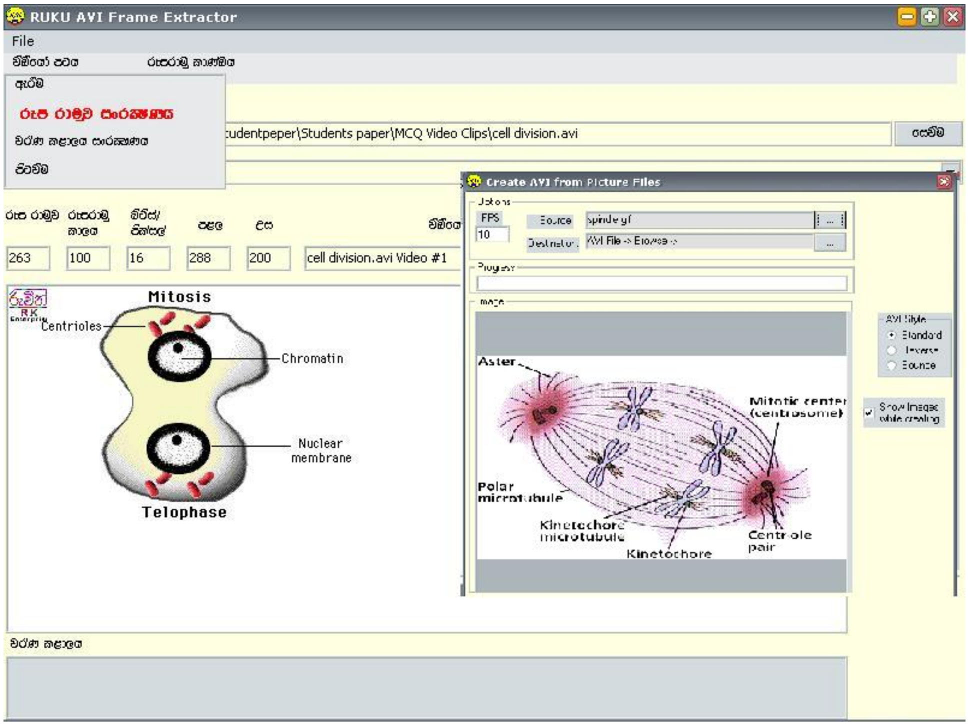
10 This disease condition
i occurs commonly in maize eating population.
ii. is due to vit. B12 defieciency.
iii. is caused by niacin defieciency
iv is associated w
v. is known as po

Spelling: English (U.S.)

Not in Dictionary:
is known as pelargra.

Suggestions:
pellagra

Fig. 3. MCQ Creator with spellchecker

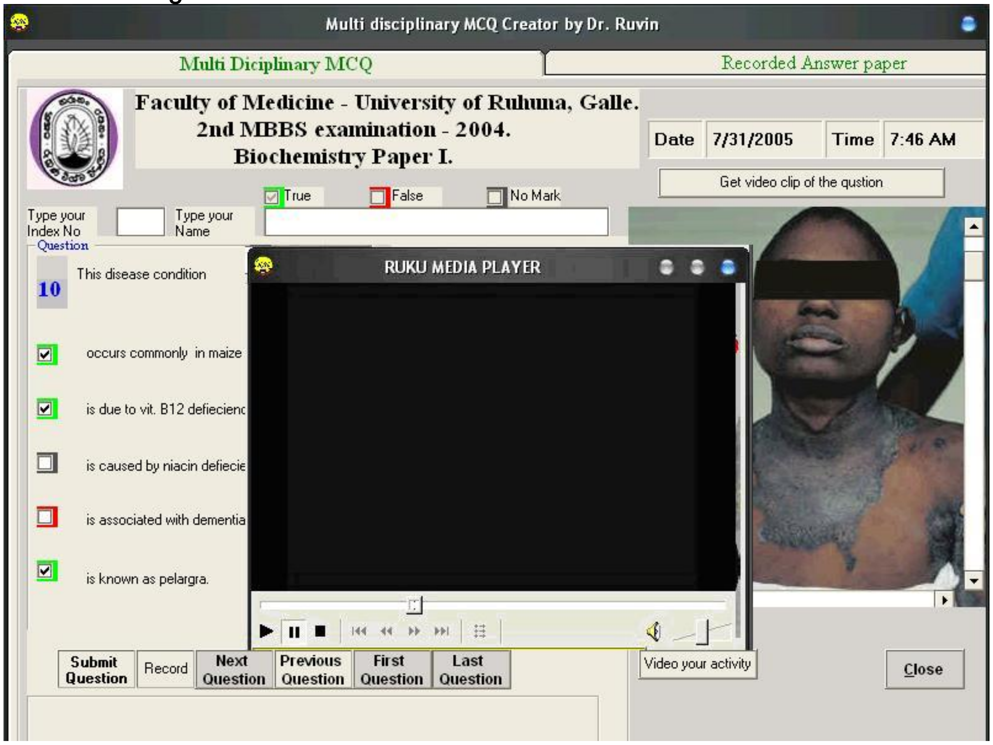


:udentpeper\Students paper\MCQ Video Clips\cell division.avi

රූප රාමුව	රූපරාමු කාලය	බිට්ස්/ පික්සල්	පළල	උස	විඩියෝ
263	100	16	288	200	cell division.avi Video #1

FPS: 10
Source: spindle.gif
Destination: AVI File -> Browse ->
AVI Style: ● Standard ○ Inverse ○ Bounce
☑ Show Images while creating

Fig. 4. Video Frame Extractor and Video creator & animator



Multi disciplinary MCQ Creator by Dr. Ruvin

Multi Diciplinary MCQ | Recorded Answer paper

Faculty of Medicine - University of Ruhuna, Galle.
2nd MBBS examination - 2004.
Biochemistry Paper I.

Date 7/31/2005 Time 7:46 AM

☑ True ☐ False ☐ No Mark

Type your Index No
Type your Name

Question

10 This disease condition

☑ occurs commonly in maize

☑ is due to vit. B12 defieciend

☐ is caused by niacin defiecie

☐ is associated with dementia

☑ is known as pelargra.

Fig.5. Video player bound to student MCQ paper with the answer recorder.

Improving Usability of E-Learning Systems by Using Ontologies

R. Heiyanthuduwage[1] and D. Karunaratne[2]

[1]IDM Computer Studies (Pvt.) Ltd
25, Visaka Road, Colombo-4,
Sri Lanka.
[2]University of Colombo School of Computing
35, Reid Avenue, Colombo 7,
Sri Lanka
Email:[1] rohitha@pgd.idm.edu and [2]ddk@ucsc.cmb.ac.lk

Abstract

Our research is aimed at improving the usability of Learning Management Systems (LMS) by using ontologies. In this paper we propose an ontology-based architecture for e-learning systems to enhance the typical features provided by them. Thus we believe that by using our architecture usability of e-learning systems can be improved, and hence increases the acceptance of the Learning Management Systems by the learners and trainers.

Key words: Ontology, e-learning, Semantic Web, LMS, OIL, Content Management

1. Introduction

"E-Learning is just-in time education integrated with high velocity value chains. It is the delivery of individualized, comprehensive, dynamic learning content in real time, aiding the development of communities of knowledge, linking learners and practitioners with experts" [14]. Successful management of an e-learning system requires an efficient LMS. LMS supports e-learning by providing a software platform, both to manage learners and learning resources. The contents of an e-learning system need to be made easily accessible to the users according to their learning or teaching requirements. For that the contents need to be well-organized based on user level concepts and should be associated to concepts in user requirements. The organization of the e-learning system contents and mapping the user competencies to learning contents is possible with ontologies.

The ontology proposed for the e-learning management environment can be viewed at the conceptual level by a conceptual model, and then at logical level, it can be represented in Ontology Inference Language (OIL), then, at physical level, it can be mapped into the semantic web. Semantic web technologies are typically being used to annotate resources, (web content) in the e-learning systems with uniformity. As some standards for semantic web Resource Description Framework (RDF, 2002), eXtensible Markup Language (XML, 2003) [15], Simple HTML Ontology Extensions (SHOE), and MPEG-7 can be considered. The ontologies mapped in to semantic web help satisfy learning requirements of the learners by providing the required learning objects efficiently and effectively through the semantics of the ontology given in semantic web.

In this research our main aim is to improve the usability of e-learning systems by using ontology based architecture.

This paper has been organized in several sections. The section one gives an introduction to the paper and section two gives an overview of ontology related to e-learning and usability. Section three gives a brief overview of LMSs and several related topics to our research such as SCORM, Sharable Content Objects (SCOs) ontologies, why ontologies are used in LMSs, section four provides an overview of semantic web, XML, RDF, SHOE and MPEG-7. Then, section five gives the proposed architecture of the e-learning system what uses ontology to increase the usability of e-learning and the section six describes the implementation of the architecture and the conclusion in section seven.

2. Usability of Ontologies in e-Learning

Ontology is a specification of a conceptualization [1]. Ontologies are applied in different domains varying from AI (artificial intelligence) [11] to desktop applications. They are applied in World Wide Web, agent technologies [17] e-commerce applications to search engines [12].

One of the main reasons for the use of ontologies in e-Learning systems is to share a common understanding of the structure of information among people or software agents, to enable reuse of domain knowledge, to make domain assumptions explicit, to separate domain knowledge from the operational knowledge, and to analyze domain knowledge [1].

Ontology consists of concepts, slots, facets and relationships between them. Domain ontology can be considered as detailed descriptions about the domain. This collection of descriptions definitions of concepts, entities, attributes and processes related to a given application domain [3]. In this research, our domain is e-Learning systems.

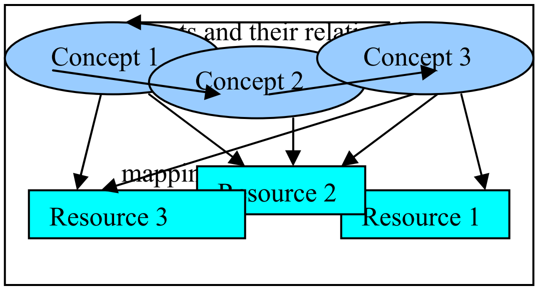


Figure 1: Mapping between the ontology and the resources

Ontologies provides the ability of finding the related concepts at a higher level through the relationships between the concepts and it is possible to find low level resources from the higher level concepts through mapping between them *(figure 1*).

We expect to improve several aspects related to the usability of this ontology based e-learning system: (i) effective selection of learning resources based on learning requirements of the learners, (ii) proposing learning resources to the learner based on learner's competency, (iii) allowing content developers to effectively search of existing SCO to satisfy their learning requirements, (iv) allowing content developers to easily create new SCOs by amalgamating existing source files (text, audio, video and etc), (v) generating effective learning sequences in e-learning for students, (vi) allowing content developers to create new SCOs with the help of existing SCOs (reusing and extending them), (vii) increasing the reusability of source files (text, audio, video and etc) and SCO , (ix) providing a low bandwidth solution.

3. Learning Management Systems (LMS) and Metadata Standards

Advanced Distributed Learning (ADL) is an initiative of Department of Defense (DoD) and the White House Office of Science and Technology Policy (OSTP). This aims at proving high quality education and training, tailored to individual needs, delivered cost effectively anytime and anywhere. SCORM is ADL initiative's goal. Its aim is to foster creation of reusable learning content within a common technical framework for computer and web based learning [4]. SCORM use XML for representing course structures using metadata. It enables the reuse of Web-based learning content across multiple environments and products. It focuses on distribution of packaged material [5].

ADL is going to be made a reality by using LMS and SCOs (*Figure-2)*. There are many different contemporary web-based Learning Content Management Systems (LCMS), such as ATutor, WebCT and etc. Only SCORM compliant LMSs can access the SCOs.

In SCORM the technical framework of e-learning systems is described by a set of guidelines, specifications and standards [4]. The standards and specifications in SCORM are described by set of technical books. There are three main categories of SCORM books. They are, (i) Content Aggregation Model (CAM) which covers, assembling, labeling and packaging of content, (ii) Run-Time Environment (RTE) which describes, LMS's Management of the Run-Time Environment which includes launch, content to LMS communication, tracking, data transfer and error handling, (iii) Sequencing content and navigation (SN) which covers activity tree, learning activities, sequencing information, navigation data model [4].

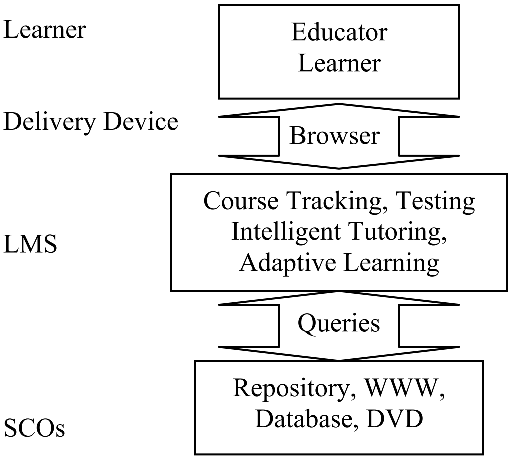


Figure 2: An overview of a LMS

A SCO is a collection of one or more assets that represent a single learning resource *(figure 3)* that utilizes the SCORM Run-Time Environment (RTE) to communicate with Learning Management Systems (LMS) [4]. SCO should be independent of its learning content so that it can be reused across multiple learning contexts.

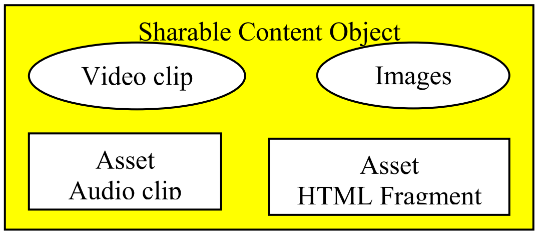


Figure 3: Sharable Content Object

4. Semantic Web and Metadata Standards

The vision of Semantic Web is to enable machines to interpret and process information in the World Wide Web (WWW) in order to better support humans in caring out their various tasks with the web.

XML is used as a mark up language and a Meta markup language. XML is used to define the structure of data. An XML document consists of a properly nested set of open and close tags. Each tag can have a number of attribute value pairs [12]. XML documents are validated against Document Type Definitions (DTD) of XML schema to standardize the structure of data and to increase the reusability.

RDF is a Metadata standard to describe the resources in the WWW, which is recommended by the W3C. Basic constructs of RDF are object-attribute-value triple. It is written as A(O,V), which represents object O has an attribute A with the value V [12].

SHOE language allows users to define extensible vocabularies what humans can understand and to associate machine understandable meaning of those vocabularies [16].

MPEG-7 is the standard for describing multimedia content that provides the richest multimedia content description tools for applications ranging from content management, organization, navigation, and automated processing. It consists of a set of core description tools. It addresses many different applications in different environments [7]. In MPEG-7 the root element is <Mpeg7> and the description metadata header is <DescriptionMetadata> and either a description unit <DescriptionUnit> or a complete description <Description>.

There are three groups of top-level elements, Content Management top-level elements, Content Entry top-level elements, and Content Abstraction top-level elements. Content Management includes user description, media description, creation description, usage description, classification scheme description, Content Entry includes media content such as image, video, audio, audio-visual, multimedia content, multimedia collection, signal, ink content, analytical edited video and Content Abstraction includes semantic description, model description, summary description, view description and variation description. This allows sending only the required part for the request from an application.

OIL tries to define a joint standard to support exchange of information based on ontologies. OIL ontology is a structure made up of several components, some of them are themselves structured. Components can be mandatory, optional or repeated. In OIL we can define classes, slots, slot constraints and etc related to the ontology [19].

5. Proposed Systems Architecture

The main components in this proposed architecture *(Figure 4)* are user profile & training manager what is responsible for capturing user profiles and storing them in the database. The assets manager and SCO manager help to maintain and add new assets and SCOs to the system. Ontology manager can be used to create and to maintain the ontology. Query processor gets all requests to store or retrieve user profiles, assets, SCOs and ontologies. It uses indices to increase the efficiency of this process. The database stores the user profiles, assets, SCOs, ontology and XML files with annotations made according to SCORM and/or MPEG-7.

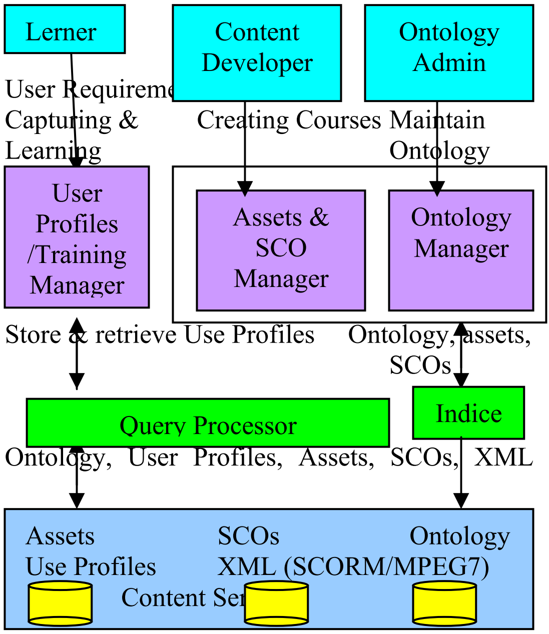


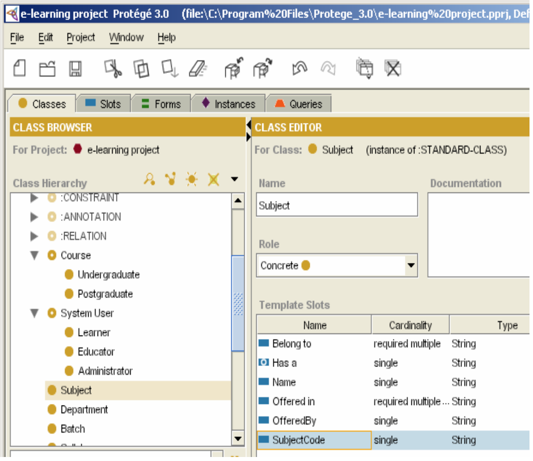
Figure 4: Proposed Architecture for the e-Learning System

When the users log on to the system again these user profiles can be improved. User profiles help the system to identify the learning requirements of the Lerner and the competency of the learner. Lerner's competency assessed based on the competency given as a part of domain ontology [13].

When learners access the system user profiles are created and they are stored within the system. These user profiles can be improved each time they access the system by acquiring their learning requirements. Content Developers can create SCOs by combining assets and metadata is provided for SCOs and assets according to SCORM to make them available for SCORM compliant LMSs. Learners and educators access them using the ontology which is not transparent to them. This ontology-based access of resources is assisted by SCORM or MPEG7 annotations specified with the ontology. MPEG-7 helps to make annotations to the assets. Here MPEG7 becomes more useful and efficient to access any multimedia data used as assets in the system [6]. The ontology helps to do some mapping between the user requirements and the course material. The indices can support the ontology in searching for the SCOs or learning objects specially when there are several contents for the same ontology concept. For example, if a student wants to learn about multimedia, the system can have several multimedia courses, some can be basic courses and some can be advanced courses. Based on the ontology we can find the required concept (class), and then correct material or learning object can be presented to the user with the help of the index.

6. Implementation

A part of the e-learning ontology of the university undergraduate degree system we propose has been given in a conceptual model (figure 5).


e-learning project Protégé 3.0 (file:\C:\Program%20Files\Protege_3.0\e-learning%20project.pprj, Def
File Edit Project Window Help
Classes
Slots
Forms
Instances
Queries
CLASS BROWSER
CLASS EDITOR
Class Hierarchy

For Project: e-learning project

:CONSTRAINT
:ANNOTATION
:RELATION
Course
Undergraduate
Postgraduate
System User
Learner
Educator
Administrator
Subject
Department
Batch

For Class: Subject (instance of :STANDARD-CLASS)

Name: Subject

Documentation:

Role: Concrete

Template Slots

Name	Cardinality	Type
Belong to	required multiple	String
Has a	single	String
Name	single	String
Offered in	required multiple ...	String
OfferedBy	single	String
SubjectCode	single	String

Figure 6: A part of e-learning ontology represented in protégé

In this project Protégé is used as a tool for ontology development. It is a CASE tool developed at Stanford University, USA. It allows the users to design ontologies and it is an open source providing an extensible architecture for the creation of customized knowledge based applications. The e-learning ontology given in the conceptual model (*Figure 4) then* can be represented in protégé (*Figure 6)*

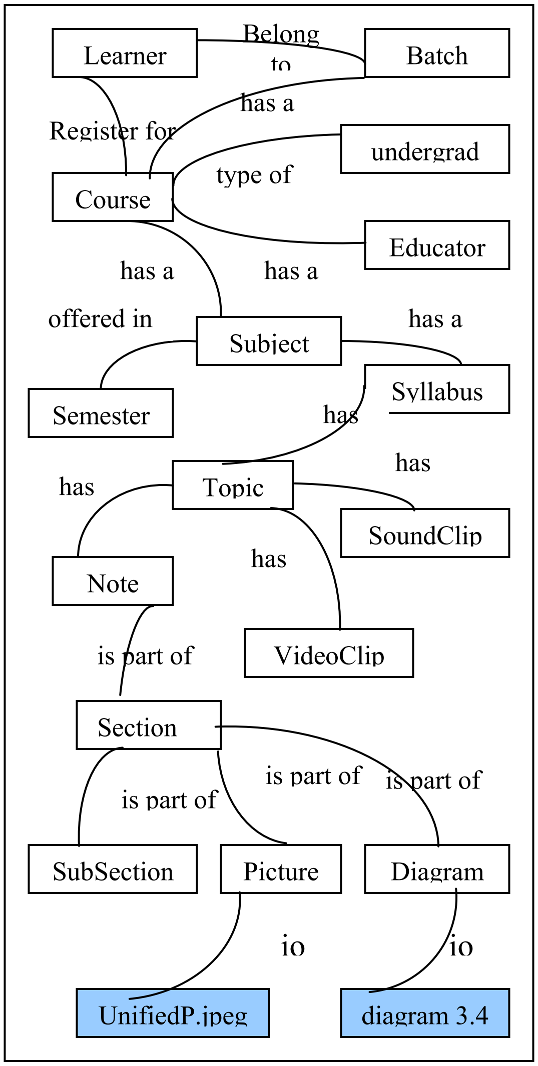


Figure 5: A part of the proposed ontology

Now, if we would see how we could apply semantic web technologies for our e-learning system, the ontology defined in protégé for the e-learning system can be represented in Ontology Inference Language (OIL) as follows, (*figure 7)* [12].

XML works as a markup language as well as a meta markup language for arbitrary document structures. The e-learning system ontology defined in OIL then easily serialized in XML [12] (*figure-8*).

```
class-def e-learningUser
class-def Lerner
            subclass-of e-learningUser
            slot-constraint studies-for
            value-type Course
            slot-constraint registers-for
            value-type Batch
class-def Educator
            subclass-of e-learningUser
            slot-constraint teaches-for
            value-type Course
class-def Course
            slot-constraint followed-by
            has-value Learner
            slot-constraint taught-by
            has-value Educator
            slot-constraint has
            value-type Batch
class-def Batch
            slot-constraint registered-for
            value-type Learner
class-def Subject
            slot-constraint is-part-of
            has-value Course
            slot-constraint has a
            value-type Syllabus
class-def Syllabus
class-def Topic
            slot-constraint is-part-of
            has-value Syllabus
class-def Note
            slot-constraint prepared-by
            has-value Educator
            slot-constraint prepared-for
            has-value Syllabus
```

Figure 7: A part of e-learning ontology defined in OIL

In relation to our e-learning system the following examples can be given,

hasName('http://www.bit.lk/OOSD/Note1', "Introduction to Object Oriented Concepts"),
isPartOf('http://www.bit.lk/diagram-1',
"http://www.bit.lk/OOSD/Note1").

The type of the object can be given in RDF format as follows,

```
<rdf:Description about="www.bit.lk./OOSD/Note-1">
            <rdf: type
resource="http://www.bit/schemea/Note">
</rdf:Description>
```

```
<rdf:Description about="www.bit.lk./OOSD/diagram-1">
            <rdf: type
resource="http://www.bit/schemea/diagram">
</rdf:Description>
```

```
<?xml version="1.0"?>
<e-learning ontology>
   <class-def>
         <name>e-learningUser</name>
   </class-def>
   <class-def>
      <name>Lerner</name>
      <sub-class-of>
         <class name="e-learningUser"/>
      </sub-class-of>
         <slot-constraint>
            <slot name="studies-for"/>
            <value-type>
                  <class name="Course"/>
            </value-type>
         </slot-constraint>
   </class-def>
         .
<e-learning ontology>
</xml>
```

Figure 8: A part of the e-learning ontology serialized in XML

When data/resources are represented in XML or/and RDF associated XML schema or DTD and RDF schema have to be provided to consider them as valid XML or valid RDF documents by the parser.

7. Conclusion

In this paper we described how an ontology can be integrated in to a typical LMS environment to improve its usability. Part of the ontology design has been represented in OIL and XML, which is a part of the proposed architecture. This architecture achieves a higher level of usability such as efficient and effective searching, easy to use due to the ontology and layering the components in this architecture and additional benefits such as reusability of components and the ontology, easy maintenance, flexibility, evolving user profiles, evolving domain ontology and it is a low band width solution.

It is possible to implement this architecture in heterogeneous platforms to achieve the organizational objectives of different universities and institutes that are involved in E-learning. Also this model can be applied for educational and training organizations but also for industrial businesses that need thorough trainings. We have evaluated some aspects of the proposed architecture using diffident usability criteria and it produced good results.

To apply the proposed ontology in semantic web we used XML and RDF. Further improvements to them can be done using MPEG-7 and SHOE as they provide better annotations to the assets. MPEG-7 addresses many different applications in different environments [7].

References

[1] N. F. Noy and D. L. McGuiness *Ontology Development 101: A Guide to Create Your First Ontology,* Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001

[2*]* C. Guangzuo, *OntoEdu: Ontology-based Education Grid System for e-Learning*, Modern Education Technology Centre at Peking University 100871, The Official Journal of Global Chinese Society FOR Computers in Education,2004,pp59-72

[3] A. De Nicola, M. Missikoff, *Towards an Ontological Support for e-Learning Courses*, Rome, Italy.

[4] *The Sharable Content Object Reference Model (SCORM), Acquisition Guidelines for U.S. Military*, August 20, 2004.

[5*]* A. Rockley, S. Manning, *E-Learning, single sourcing and SCOR*M, http://www.stc.org/confproceed/2002/PDFs/STC49-00018.pdf.

[6] S.C. Premaratne, D.D. Karunaratne, G.N. Wickramanayake, K.P. Hewagamage, and G.K.A. Dias, *Profile based video segmentation system to support E-Learning*, University of Colombo School of Computing, in proceedings of 6th International IT Conference 2004, (IITC 2004), November 2004, Colombo, Sri Lanka.

[7] J. M. Martinez*, MPEG-7: Overview of MPEG-7 Description Tools*, 2002 IEEE, reprinted for IEEE Computer Society, July-September 2002

[8] M. Baldom, C. Baroglio, V. Patti, and L. Torasso, *Reasoning about learning object metadata for adapting SCORM courseware,* Torino, Italy

[9] G. Angelova, O. Kalayaydjiev, and A. Strupchanska, *Domain Ontology as a Resource Providing Adaptively in e-Learning*

[10] H. Passier and J. Jeuring, *Ontology based Feedback Generation in Design-Oriented e-Learning Systems,* Faculty of Computer Science, Netherlands, IADIS e-Society 2004 Conference, 16-19 July 2004, Avila, Spain. http://www.ou.nl/eCache/DEF/11/857.html

[11] H. Knublauch, *An AI tool for the real world, Knowledge modeling with Protégé*, June 20, 2003

[12] S. Decker, F. V. Harmelen, J. Broekstra, M. Erdmann, D. Fensel, I. Horrocks, M. Klein, S. Melnik, *The Semantic Web – on the respective Roles of XML and RDF*, IDIMS Report, February 2003, http://pan.nuigalway.ie/code/docs/Report_on_SW.pdf.

[13] D. Woelk, *e-Learning, Semantic Web Services and Contemporary Ontologies*, Elastic Knowledge Solutions, ED-MEDIA World Conference on Educational Multimedia, Hypermedia and Telecommunications, Denver, CO, June 2002.

[14] L. Stojanovie, S. Staab, R. Studer, e-*Learning based on the Semantic Web,* University of Karlsruhe, Germany

[15] N. Henze, P. Dolog, and W. Nejdl, *Reasoning and ontologies for personalized e-learning in the Semantic,* http://www.kbs.uni-hannover.de/Arbeiten/Publikationen/2004/ifets_final.pdf

[16] Heflin, J. and Hendler, J. Searching the Web with SHOE. In *Artificial Intelligence for Web Search. Papers from the AAAI Workshop.* WS-00-01. AAAI Press, Menlo Park, CA, 2000. pp. 35-40.

[17] ***www.agentland.com***

[18] ***www.atutor.ca/***

[19] D.Fensel, I Horrocks, E. Van Harmelen, S. Decker, M. Erdman, and M. Klein, *OIL in a nutshell,* http://citeseer.ist.psu.edu/horrocks00oil.html

Author Index