How Dirty is your Data?

Identification of the Effects of Unclean Data and Incorporation of String Matching Techniques to Mitigate these Effects in the Telecommunication Industry

By

P. A. S. N. Jayawardena

13020242

N. M. D. Muthuweera

13020315

This dissertation is submitted to the University of Colombo School of Computing

In partial fulfillment of the requirements for the

Degree of Bachelor of Science Honours in Information Systems

University of Colombo School of Computing

35, Reid Avenue, Colombo 07,

Sri Lanka

December, 2017

Declaration

I, P. A. S. N. Jayawardena, 13020242 hereby certify that this dissertation entitled 'How Dirty is your Data? Identification of the Effects of Unclean Data and Incorporation of String Matching Techniques to Mitigate these Effects in the Telecommunication Industry' is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

…….. ……..
Date P. A. S. N. Jayawardena

I, N. M. D. Muthuweera, 13020315 hereby certify that this dissertation entitled 'How Dirty is your Data? Identification of the Effects of Unclean Data and Incorporation of String Matching Techniques to Mitigate these Effects in the Telecommunication Industry' is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

…….. ……..
Date N. M. D. Muthuweera

I, S. M. K. D. Arunatileka, certify that I supervised this dissertation entitled 'How Dirty is your Data? Identification of the Effects of Unclean Data and Incorporation of String Matching Techniques to Mitigate these Effects in the Telecommunication Industry' conducted by P. A. S. N. Jayawardena and N. M. D. Muthuweera in partial fulfillment of the requirements for the degree of Bachelor of Science Honours in Information Systems.

…….. ……..
Date S. M. K. D. Arunatileka

I, Roshan Rajapakse, certify that I supervised this dissertation entitled ‘How Dirty is your Data? Identification of the Effects of Unclean Data and Incorporation of String Matching Techniques to Mitigate these Effects in the Telecommunication Industry’ conducted by P. A. S. N. Jayawardena and N. M. D. Muthuweera in partial fulfillment of the requirements for the degree of Bachelor of Science Honours in Information Systems.

…….. ……..

Date Roshan Rajapakse

Abstract

Telecommunication organizations in Sri Lanka face the challenge of improving and maintaining data quality in customer data. Customer data are considered as a data category which is exposed to the infusion of dirty data. The problem gets accumulated in customer profiles, creating varied duplicated accounts of the same customer profile. In the current circumstances, data cleansing tools are used to improve data quality, but the suitability of a tool to control the problem is limited. Even though tools provide recommendations on merging duplicate accounts, it is crucial to clerically review on the final linkage status.

This dissertation outlines a detailed analysis of dirty data infused in customer data, which would create an adversative effect on CRM operations and overall decision making process. Further, it incorporates similarity measure techniques to provide a better decision making in the clerical review process, by evaluating the similarity in different attributes of customer data, in order to control and mitigate the effect. Experimental comparisons on a Sri Lankan telecommunication dataset indicates appropriate similarity measure techniques to be applied to full name, address and email address attributes of customer data with a high level of accuracy.

Acknowledgement

We would like to express our sincere gratitude to our research supervisor, Dr. S. M. K. D. Arunatileka, Senior Lecturer of University of Colombo School of Computing and our research co-supervisor, Mr. Roshan Rajapakse, Probationary Lecturer of University of Colombo School of Computing for providing us continuous guidance and supervision throughout the research.

We would also like to extend our sincere gratitude to Dr. H. A. Caldera, Senior Lecturer of University of Colombo School of Computing and Dr. F. H. A. M. Premachandra, Lecturer of University of Colombo School of Computing for providing feedback on research proposal and interim evaluation to improve our study. We also take the opportunity to acknowledge the assistance provided by Dr. T. A. Weerasinghe as the coordinator of Final Year Research Project in Information Systems.

We appreciate the feedback and motivation provided by our friends to achieve the research goals, specially Ms. Maneesha Perera who was with us throughout the research. This thesis is also dedicated to our loving families who have been an immense support to us throughout this journey of life. It is a great pleasure for us to acknowledge the assistance and contribution of all the people who helped us to successfully complete the research.

Table of Contents

List of Figures

List of Tables

List of Acronyms

CRM	Customer Relationship Management
SBU	Small Business Unit
NIC	National Identity Card
GM	General Manager
CLM	Customer Life Management
PVT	Private
PLC	Public Limited Company
BI	Business Intelligence
BA	Business Analysist
GCCO	Group Chief Customer Officer
CE	Customer Experience
IT	Information Technology
J-W	Jaro Winkler
Sgram	Skip Grams
EDist	Edit Distance
MeDis	Modified edit distance
bDist	Bag distance
SeqMa	Sequence Match
ComBZ	BZ2 compression algorithm
ComZL	Comparator measure using the zlib compression library
ComAC	Compression algorithm
LCS2	Longest Common Substring
LCS3	Longest common substring with minimum length of substrings 3
OLCS2	Ontology longest common substring with minimum length of substrings 2
OLCS3	Ontology longest common substring with minimum length of substrings 3
P-Win	Permuted Winkler
S-Win	Sorted Winkler
SWDis	Smith-Waterman distance
SyADi	Syllable Alignment distance

CharHisto Char Histogram

Chapter 1 - Introduction

1.1 Introduction to the Research

Interpretation of data for business decision making in many organizations have become demanding and challenging. At various levels, business stakeholders exploit data to harvest insights to assist in their operations which ultimately contribute in the business's success. However, organizations that utilize data-led insights cannot simply rest, since a minor error or a misunderstanding can adversely escalate to compromise all the work and effort invested. Inaccuracies, misleads, errors, non-integrated and data violating rules, generally known as dirty data would simulate such situations [1]. Therefore, it is crucial for many organizations to realize and understand the gravity of existence of dirty data in the business.

Data entry and acquisition is inherently prone to errors, even though much effort is given to these front-end processes, with respect to reduction in entry error [1], it is common practice in business to tolerate dirty data to a substantial degree rather than to manage or eliminate it [2]. With this situation, dirty data multiplies within the organization undermining the organizational data quality within the domain. Business decisions taken and the analysis made on this erroneous data tend to direct the organization into false route. For existing data, the logical solution to this problem is cleansing the data using tools or some other way. That is to explore the data and rectify the data with any errors, and organizations spend millions of dollars to automate this process using powerful cleansing tools. Even though data quality has been addressed up to a reasonable level, the existence of dirty data is inevitable, mostly due to the addition of dirtiness at various touch points. The tools are good in analyzing the data but with the introduction of new data types, error types and loopholes in the system, tools always fade out. Therefore, a control check is very much needed in the current context to avoid any adverse effects from dirty data. And even though the data is cleansed 100%, control is crucial, since more problems with dirty data tend to arise with time. This cannot be achieved without analyzing the current domain context and understanding how data has become dirty. Therefore, it is important to understand

various implications of dirty data infused to the business processes and the effects it make to the overall performance of the business.

Effects of dirty data are often shown in the telecommunication industry in different contexts. Mainly data entered to customer accounts are infused with dirty data, which will cause difficulties in understanding a single customer, which will then lead into revenue leakages and cost accumulation [3].

Traditionally, dirty data infused customer information is cleaned using data cleansing tools, yet merging and deciding on linking of duplicated accounts always relayed on clerical review. The use of human oversight, to resolve the final linking status, is known as clerical review and in theory, it is assumed that the person undertaking this clerical review has access to additional data or has the ability to seek it out. However, often no additional data is available or it consumes a great deal of time and resources, and the process becomes one of applying human experience, common sense or human intuition to make the decision [4]. Therefore, what is needed are more accurate and automated decision making that will reduce, or even eliminate the amount of clerical review needed, while keeping a high linkage accuracy.

1.2 Statement of the Research Problem

More commonly, the ignorance of dirty data effect, has created inefficiencies in operations and inaccuracies in decision making. Research seek to identify different contexts in which dirty data would mostly generate, in order to anticipate the most affected area with dirty data, with the main aim of analyzing the effect of dirty data, within identified fields of strategic decision making.

The need to address this problem can be understood from available research work which has been contributed in the identification of dirty data and in determining the importance of data cleansing. However, the work conducted by the previous researchers are inconclusive about the effect of dirty data on business decision making in Sri Lankan Telecommunication industry. Consequently, it has been an area which has gained the attention yet not being addressed. Therefore, through the research, the gap in the knowledge is filled by addressing the effect of dirty data on strategic decision making and by discussing approaches to mitigate and control.

1.3 Significance of the Study

The study will analyze how dirty data affect the operational and strategic decision making in telecommunication industry. It focuses on the dirty data aspect lies within customer data that is to be taken for analysis. Most of the organizations neglect the dirty data component in data sets which are specific to different industry domains. Therefore, the study focuses to analyses the effect of dirty data that might be neglected from customer profiles in the telecommunication industry due to cost factors, time constraints and complexities towards the decision-making process. Further, with identification of these effects, the study aims to mitigate the impact through the association of different controlling approaches.

1.4 Primary Research Questions

What is the Effect of Dirty Data in Strategic Decision Making within the area of Customer Relationship Management in Telecommunication industry?

This problem will mainly focus the context of dirty data lies within the organization. How far inclusion of dirty data within customer data that is used for analysis would create an adversative effect on overall strategic decision making? The problem will mainly address the severity of the impact caused by dirty data over business operational factors, associated in CRM. Therefore, it is fundamental to be aware about the effect that dirty data that would influence in business operations and decision making.

How to Mitigate and Control the Effect of Dirty Data on Strategic Decision Making?

With the identification of dirty data effects on strategic decision making, the next concern arises on how to mitigate and control the inefficiencies and integrity issues arisen by dirty data over operations and decision making. In addressing the problem, in the context of Sri Lankan Telecommunication industry, it is important to evaluate approaches to rectify the circumstances of dirty data, in order to minimize its impact.

1.5 Methodology

The approach followed by the research is the pragmatic approach, where it appears as the best suited approach for the research problem. The use of variety of data sources, use of multiple perspectives to interpret the results and the use of multiple methods to study the research problem are the base for using pragmatic method of research. It is determined to explore using qualitative research and then to use the results of quantitative research to present the findings. Figure 1.1 represents a high level diagram of the proposed research methodology.

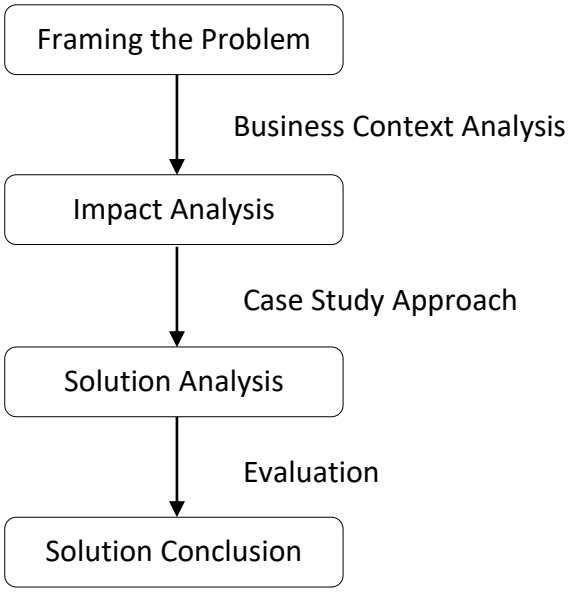

Figure 1. 1: Overview of Research Methodology

As the first step of the research approach, framing the problem, an identification of the problem to be addressed out of many problems given by organizational experts are being addressed. With the selection of Customer Profiling as the most common dirty data infused problem within the telecommunication industry, a wide business context analysis is done. Subsequently, the impact of dirty data would be analyzed in order to showcase the gravity of the effect towards the business decision making. Thereafter, an analysis will be represented through several case studies and each case study will address different business aspects. In order to control and mitigate the effect of dirty data, in the selected problem of Customer Profiling, techniques of similarity measure in a Sri Lankan context, are being considered in preparation of a solution. Finally, each of these solutions are being evaluated to conform the conclusion.

1.6 Aims and Objectives

The main aim of the research is to understand the effect of dirty data lies within customer data over decision making in telecommunication industry. And to analyze different techniques to be used as a solution mechanism to control and mitigate.

The Goals and objectives of the research are as follows:

1. Framing and identifying the most affected problem to be analyzed.
2. Identifying the impact of dirty data in the business context.
3. Designing use cases for identified problems.
4. Analyzing existing techniques applicable for similar problems.
5. Applicability of analyzed techniques on local datasets and use cases.
6. Devise an evaluation mechanism to measure the accuracy of the identified techniques.
7. Suggesting the best optimal technique for the identified problem.

1.7 Assumptions, Limitations and Scope

Assumptions

1. It is assumed that every customer profile is infused with dirty data in different attribute levels.
2. The effect mentioned within the research problem solely determined from the dirty data component while other factors remain constant.
3. Effect of dirty data changes from one industry domain to another.

Limitations

1. Inability to mitigate the effects of dirty data to an absolute level.
2. Limitation of access to data sets due to privacy issues and regulatory obligations.

Scope

1. Different use cases that impact the telecommunication industry in Sri Lanka will be taken into consideration.
2. The focus of the research will be on the Customer data category and the area of CRM.
3. Suitable techniques to mitigate Dirty Data components will be suggested by analyzing the existing techniques.

1.8 Outline of the Dissertation

The dissertation is structured as follows. Chapter two explores the existing research related to the domain of dirty data management and controlling. Chapter three describes the proposed research design and methodology. Potential ways of addressing the research problem is discussed in this chapter. Chapter four demonstrates the implementation details of the proposed methodology. Chapter five presents the evaluation model and the evaluation results of the proposed approach. The last chapter, chapter six demonstrates the conclusion of the thesis and outlines the future work.

1.9 Summary

This chapter laid the foundations for the dissertation. It introduced the research, research background, research problem, research questions and aim and objectives of the research. Then the research was justified, the methodology was briefly described, the dissertation was outlined, and the limitations were provided. Second chapter describes about the related literature with regard to the problem domain.

Chapter 2 - Literature Review

2.1 Introduction

This chapter, mainly focuses on the related work which has been carried out by fellow researchers over the years. Also, it has given right level of significance to elaborate the concepts that are discussed within the research. Many techniques, approaches and methodologies that are used to rectify the similar domain problems are being depicted within this chapter.

2.2 Dirty Data and Causes of Dirty Data

Simply, dirty data are elaborated as erroneous data, it is also explained as data that have within the memory but not have loaded into the database. There are different types of dirty data can it can have classified under the taxonomy of data [5]. For a basic understanding few of the most common dirty data types can be understood accordingly. Duplicate data a very common dirty data type, incomplete data can occur when the fields values are created outside the valid range of values. Inaccurate values are created through technically correct but inaccurate when considered the business context etc. [5].

In the literature, dirty data is being classified vibrantly and Table 2.1 shows the categorization [5].

<table>
<tr><th colspan="4">Dirty Data</th></tr>
<tr><th>Missing Data</th><th colspan="3">Not Missing but</th></tr>
<tr><td rowspan="2">Missing data where there is no Null not allowed constraint.
Missing data where Null-not-allowed constraint should be enforced.</td><td colspan="2">Wrong data due to</td><td>Not wrong but unusable data</td></tr>
<tr><td>Non-enforcement of automatically enforceable integrity constraints.</td><td>Non-enforceability Of integrity constraints.</td><td>Use of abbreviations Different representations of compound data.</td></tr>
</table>

Table 2. 1 : Classification of Dirty data

Also, some of the researchers define dirty data using three different anomalies, syntactical anomalies, semantic anomalies and coverage anomalies [6]. Within the syntactical anomalies it includes lexical errors, domain errors and also irregularities concerning the non-uniform use of values. Semantic anomalies include integrity constraints violations, contradictions, duplications, invalid tuples, while coverage anomalies will include missing values and missing tuples.

Also through related work many dirty data types such as missing value, missing tuple, unique value violation, exact duplicates, inconsistent duplicates, implausible range, unexpected low high values, outdated temporal data, wrong data format, misfield values, embedded values, domain violations, incorrect derived values are being listed down [7].

Dirty data has been caused due to many reasons and those reasons are being stated in Table 2.2 [8],

Type of the Dirty Data Cause	Examples
Data Sources	Selection of candidate data references dismally cause data quality problems (sources which do not comply with business rules).
	When the time from the reference increase and nearness from the source increase, the possibility for getting correct data decrease.
	Insufficient knowledge of inter dependencies among data sources encompasses data quality problems.
	Incapacity to cope with ageing data cause data quality problems.
	Dissimilar timeliness of data sources.
	Lack of Accepted routines at sources causes data quality problems.
Issues at data profiling	Data quality issues are formed by improper data profiling of data references.
	Manually procured information about the data Contents in operational systems reproduce poor data quality.
	Due to data quality issue, inappropriate selection of Automated profiling tool arises.
	Inadequate data content analysis against external reference data causes data quality problems.
	Inadequate structural analysis of the data sources in the profiling stage.

Issues at data staging and ETL stage	Data quality is pretentious by the data warehouse architecture undertaken.
	Relational or non-relational type of staging areas affect the data quality.
	Variety data sources and its different business rules creates problem of data quality.
	Business rules lack currency bestows to data quality problems.
	Lack of periodical refreshing of the integrated data storage (Data Staging area) cause data quality degradation.
Issues at data warehousing and schema design	Incomplete or wrong requirement analysis of the project lead to poor schema design which further cause data quality problems.
	Lack of currency in business rules cause poor requirement analysis which leads to poor schema design and contribute to data quality problem.

Table 2. 2 : Types of Dirty Data Causes

2.3 Effects of Dirty Data

Due to the existence of dirty data, many problems are faced by domain experts. One of the main issue faced are adverse effects on data quality aspect. Furtherly when the quality has been reduced it has indirectly affected the decision-making process of the organization. It is understood that quality data is a significant factor for an organization to success. When the dirty data is existed among the data sources it will create poor quality of data, indirectly it will cause inefficient impacts over strategic decision-making process. When the business is considered there are mainly three types of data [8]. Namely those data types are master data, transactional data and historical data. These categories can be further explained using examples,

- Master data: customers, employees, suppliers, these can be identified as the basic characteristics of the organization.
- Transactional data: orders, invoices records, deliveries and storage data will be classified under this section.
- Historical data: Categories that will be within historical aspects of the data included within other categories.

The errors that are affecting the master data category includes vast data issues. Main cost effect that has caused through dirty data within master data category would be the cost factor. From pricing the products to bundling any information that is poor in quality will arise adverse financial cost issues. Another colossal issue faced by the many multinational companies would be the inconsistencies in understanding data definitions, data formats, and data values which lead into the unfortunate situation of understanding the use of key data. Dirty data existing within the company can create substantial social and economic matters. This will lead into customer dissatisfaction, increase in running costs, lower performance and lowered job satisfaction among the employees will be created [8].

Another aspect that affects by dirty data would be increments within the operational costs. Many resources and time will be allocated flawed manner due to inaccurate managerial decisions. Due to the dirty data availability as mentioned above the quality of data will be gradually decreased, due to this the trust with regard to company data will also get reduced. This situation will lead into lack of user acceptance will arise. Many of the industry experts have done many surveys to identify more of errors that is occurring due to the dirty data existence. Industry experts include Gartner Group, Price Waterhouse Coopers and The Data Warehousing Institute, which claim to identify a crisis in data quality management and a reluctance among senior decision-makers to do enough about it. The findings from such surveys are summarized into the following bullet-points [1];

- "88 per cent of all data integration projects either fail completely or significantly over-run their budgets".
- "75 per cent of organizations have identified costs stemming from dirty data".
- "33 per cent of organizations have delayed or cancelled new IT systems because of poor data".

- "$611bn per year is lost in the US in poorly targeted mailings and staff overheads alone".
- "Bad data is the number one cause of CRM system failure".
- "Less than 50 per cent of companies claim to be very confident in the quality of their data".
- "Business intelligence (BI) projects often fail due to dirty data, so it is imperative that BI-based business decisions are based on clean data".
- "Only 15 per cent of companies are very confident in the quality of external data supplied to them.
- "Customer data typically degenerates at 2 per cent per month or 25 per cent annually".
- "Organizations typically overestimate the quality of their data and underestimate the cost of errors".
- "Business processes, customer expectations, source systems and compliance rules are constantly changing. Data quality management systems must reflect this".
- "Vast amounts of time and money are spent on custom coding and traditional methods - usually fire-fighting to dampen an immediate crisis rather than dealing with the long-term problem".

2.4 Data Quality in Telecommunication Industry

Maintenance of data quality has been challenging for many industries due to the complexity of the systems and functions involved and the volume of the data to be managed. The problem gets heightened due to the stiff competition, advancement of technology and frequent initiation of offers and services by the competitors. Since there are many more problems in the industry to worried about, no organization could afford to lose its share due to poor data quality. Research has summarized the business challenges produced by data quality and integrity issues for telecommunication industry into areas such as minimizing revenue leakage, maintaining and delivering on Service Level Agreements and customer retention [3]. Even though Telecommunication spend millions of dollars on data quality issues, most Telecommunication companies still struggle to address this. The main reasons behind this failure lie due to the utilization of legacy systems or discrepancies in merging databases, difficulty to focus or correct data quality issues whilst main business operations are online, time consumption and the

prioritization for more urgent issues. Moreover, the data quality mindset of different users affects as a front user may not have same interest and incentive to maintain data quality as a business decision maker. Corrupted data can be entered with each entry level as more interfaces are built with complex systems and as data flows to other systems unnoticed. Poor data can be suppressed at these touch points of entry if every user is kept informed. As it is stated, data quality issues begin to surface when the data is integrated, summarized, standardized and used to arrive for a business decision [3].

2.6 Concepts of Dirty Data Management

Data analytics and quality of data have been attracting a significant amount of research, industry lately. Merging large databases acquired from different sources with heterogeneous representations of data has become an increasingly important and difficult problem for many organizations [9]. The integrity of the data used to operate and make decisions about a business affects the relative efficiency of operations and quality of the decisions made. Scholars, have pointed out the importance of data quality for organizations, since it puts companies into competitive disadvantage [10]. Many managers are unaware of the quality of data taken for decision making and perhaps assume that IT ensures that the data are perfect. Although poor data quality appears to be the usual due to dirty data, rather than the exception, organizations have largely ignored the issue of poor quality of data [10].

Research on Dirty Data conducted by scholars has set out new methodologies for data quality management that encouraged data to be seen and managed as a corporate resource. The practical importance of data cleaning is well reflected in the commercial marketplace in the form data cleaning tools and services, as the industry has been forced to deal with this on a regular basis. Data patterns, algorithms and methods have been found through research in addition to the current solutions for data cleaning, involving many iterations of data “auditing” to find errors, and long-running transformations to fix them [11].

Instances of this problem appearing in the literature have been called data cleansing problem and mostly it has taken the attention of the database researchers. There the data quality problems that are addressed by data cleaning have been classified and an overview of the main

solution problems are provided [12]. Furthermore, researchers have outlined the major steps for data transformation and data cleaning, providing data cleaning approaches [12]. Studies are done in problems of merging multiple databases to a single data collection and accomplishing data accuracy by eliminating duplicate information [9], which is known as the merge/purge problem in business organizations. Further, the data cleansing problem has been addressed as a crucial step in Knowledge Discovery in Databases process [13]. And research has been on detecting informative patterns and cleaning data [13].

After the observation of work that were carried out within the field of dirty data, it was understood that different techniques, tools and models were used by organizations to address the issue of dirty data. ETL tools and dirty data management methodologies were used to manage dirty data up to an extent. Many of the ETL tools are encompassed with algorithms such as data profiling which cleanse data. Nevertheless, these tools are highly affected with dirty data where results that are generated are not 100% accurate [5]. Often these techniques such data profiling ignore typographical dirty data which has an influence over decision making. Hence within suggested research above concern will be adhered. Also, research in the area discovered many tools that supports data cleaning. Potter's Wheel: An Interactive Data Cleaning System which supports transformation and discrepancy detection through simple specification interfaces with minimal delays is such a tool [11]. It supports interactive transformation, which allows users to construct transformations gradually, adjusting them based on continual feedback [14].

2.7 Dirty Data Management

Data quality has been a subject of longstanding discussions and there are even software products that can help cleanse dirty data on the market. However, only now it is beginning to be recognized that an inordinate proportion of data in most data sources is dirty and before any data analysis applications are applied against any data, the data must be cleansed to remove or repair dirty data. But, it is difficult to know with a high degree of confidence the quality of business intelligence derived from data warehouses and the quality of decisions made on the basis of such business intelligence [5].

A large variety of tools is available on the market to support data transformation and cleaning tasks. Some tools concentrated on a specific domain, such as cleaning name and address data, or a specific cleaning phase, such as data analysis or duplicate elimination. Due to the restricted domain in specialized tools, even though they perform well, to address a broad spectrum of transformation and cleaning problems specialized tools should be complemented by other tools [2]. Other tools, e.g. ETL tools, provide comprehensive transformation and workflow capabilities to cover a large part of the data transformation and cleaning process.

Specialized cleaning tools typically deal with a particular domain, mostly name and address data, or concentrate on duplicate elimination. In special domain cleaning, names and addresses are cleansed using a number of commercial tools, such as ID CENTRIC (FirstLogic), PURE INTEGRATE (Oracle), QUICKADDRESS (QAS Systems), REUNION (Pitney Bowes), and TRILLIUM (Trillium Software), since finding the customer is very important for customer relationship management. They provide techniques such as extracting and transforming name and address information into individual standard elements, validating street names, cities, and zip codes, in combination with a matching facility based on the cleaned data. Sample tools for duplicate identification and elimination include DATA CLEANSER (EDD), MERGE/PURGE LIBRARY (Sagent/QM Software), MATCHIT (HelpIT Systems), and MASTER MERGE (Pitney Bowes) [15].

A large number of commercial tools support the ETL process for data warehouses in a comprehensive way, e.g., COPYMANAGER (Information Builders), DATASTAGE (Informix/Ardent), EXTRACT (ETI), POWERMART (Informatica), DECISIONBASE (CA/Platinum), DATATRANS FOR MATIONSERVICE (Microsoft), METASUITE (Minerva/Carleton), SAGENT SOLUTION PLATFORM (Sagent), and WAREHOUSE ADMINISTRATOR (SAS) [16].

There is no data analysis support to automatically detect data errors and inconsistencies. However, users can implement such logic with the metadata maintained and by determining content characteristics with the help of aggregation functions (sum, count, min, max, median, variance, deviation etc.). The provided transformation library covers many data transformation and cleaning needs, such as data type conversions (e.g., date formatting), string functions (e.g., split, merge, replace, sub-string search), arithmetic, scientific and statistical functions, etc.

Extraction of values from free-form attributes is not completely automatic but the user has to specify the delimiters separating sub-values [16].

2.8 Data Quality Management Models

Specifically, when models are concentrated, a four stage methodology for total data quality management is one particular model that can be used under the domain of dirty management. It is known as four stage methodology of dirty data management because the stages involved. Audit the data health check, clean the data detox, error prevention and compliance fit for the future are the four stages that were mentioned [1].

2.9 Data Linkage Techniques

Computer assisted data linkage and deduplication techniques have been used in different industry sectors and the process of data cleansing, standardization and data linkage have various names in different user communities. Statisticians and epidemiologists declare of record or data linkage and by the computer scientists and database community refers the same process as data or field matching, data scrubbing, data cleansing, preprocessing or as the object identity problem and it's sometimes called, merge/purge processing, data integration or ETL (extraction, transformation and loading) in commercial applications [4].

Data may be recorded or captured with various errors, obsolete formats, missing data items, therefore, data cleansing and standardization are important preprocessing steps for successful data linkage where data can be used for further analysis. The cleaning and standardization of names and addresses is especially important to have no misleading or redundant information produced or used in organizations. Names are often recorded differently with different written forms of proper names, missing middle names, swapped names which might lead to the availability of redundant data and duplicated data. Names are important pieces of information when the databases are duplicated and when there is no unique identifier to link two data sets [18]. Since there are many spelling variations for personal names, applying exact string matching leads to poor results. In order to improve the matching quality of names many techniques have been developed and are still being invented [19,20]. The record pairs can be compared by

applying variety of comparison techniques to one or more or a combination of attributes of the records. And each comparison returns a numerical value which clarifies the record pairs into matches, non-matches and potential matches depending on the decision model used [18].

Name matching can be defined as the process of determining whether two name strings are instances of the same name. Mostly name matching is used in instances where a unique identifier is not available in the datasets to be linked. Then the linkage process must be based on the existing common person identifiers such as name and date of birth and demographic information like address. But still the linkage is quite cumbersome since these attributes can contain typographical errors, swapped names and missing information. The traditional probabilistic linkage pairs records as matches if their common attributes predominantly agree and as non-matches if they predominantly disagree [4].

The data linkage process has been improved with the exploration of new techniques originated from machine learning, data mining, information retrieval and database research [18]. Many of these approaches are based on supervised learning techniques and assume that training data is available. It has been used SQL like language that allows approximate joins and cluster building of similar records and decision functions that determine if two records represent the same entity [18]. Many of the researches in the domain have used the approach of learning distant measures for approximate string comparison. The authors of [4] present a framework for improving duplicate detection using trainable measures of textual similarity. They argue that both at the character and word level there are differences in importance of certain character or word modifications (like inserts, deletes, substitutions, and transpositions), and accurate similarity computations require adapting string similarity metrics with respect to the particular data domain. They present two learnable string similarity measures, the first based on edit distance (and better suitable for shorter strings) and the second based on a support vector machine (more appropriate for attributes that contain longer strings). Their results on various data sets show that learned edit distance resulted in improved precision and recall results. The research [20] utilizes both supervised and unsupervised machine learning techniques in the data linkage process, and introduces metrics for determining the quality of these techniques. The results have found that the machine learning techniques outperform the probabilistic techniques and provide

a lower proportion of possible matching pairs. And it has stated techniques to overcome the problem of lack of availability of training data in real world data sets.

Many of the studies have provided comparison studies on different matching techniques. But, none of the studies has analyzed and compared a comprehensive number of techniques specifically with the application to localized personal names and records. In theory, clerical review is needed to decide the final linkage status of record pairs and it is assumed that the person performing this clerical review has access to additional data which enables to resolve the linkage status [4]. In practice, however, often no additional data is available for this clerical review process. The study aims to provide accurate and applicable similarity techniques that will provide better classification of matches, non-matches and potential matches to reduce or even eliminate the amount of clerical review needed, while keeping a high linkage quality.

2.10 Quality Measures

In order to select a threshold for a particular technique, comparative evaluations must be conducted. When different techniques are compared on the same problem, it has to make sure that the quality results achieved are statistically valid and not just an artifact of the comparison procedure. Hence, the use of much used and strongly underpinned methodology of use of statistical techniques is involved.

The authors of the base research [18] of our study has discussed the issue of evaluating data linkage. They advocate the use of accuracy based on precision and recall over the use of single number measures such as accuracy or maximum F-measure, basing the fact that the single number measures assume that an optimal threshold value has been found. Also, single number measures hide the fact that one technique might perform better for lower threshold values, while another has improved performance on higher thresholds.

2.11 Summary

Through this chapter many techniques have been mentioned and discussed through the related work and also have discussed the gap which has recognized after analyzing many research papers which are related to the research. After considering all the background work, suitable design was implemented and it is presented in the next chapter.

Chapter 3 - Design

3.1 Introduction

This chapter explicates the justification of the research methodology and the proposed solutions to the research problem. It consists of three major sections as research strategy, research design and case studies.

3.2 Research Method

This study has chosen pragmatic approach as the research approach for several compelling reasons [3, 10]. It appears best suited to the research problem, with the ability to use any of the methods, techniques and procedures associated with qualitative and quantitative research. In general, the pragmatic research methods are useful in discovering the business scenarios comprising the addressed problem with the relevant statistical information. The use of variety of data sources, use of multiple perspectives to interpret the results and the use of multiple methods to study the research problem are the base for using pragmatic method of research. Qualitative approach was recognized within the research mainly to address the first research question, to determine the adverse implications of dirty data in Customer Profiles. Quantitative approach distinguished mainly to address the second question of the research, controlling and mitigating the effect that have identified within the previous phase. The dirty data component in Customer Profiles are analyzed and evaluated through techniques of similarity measure, in order to take a collective decision on customer profiles.

3.3 Research Design

The research design encompasses the solution for the research problem. Figure 3.1 represents a high-level diagram of the proposed research design.

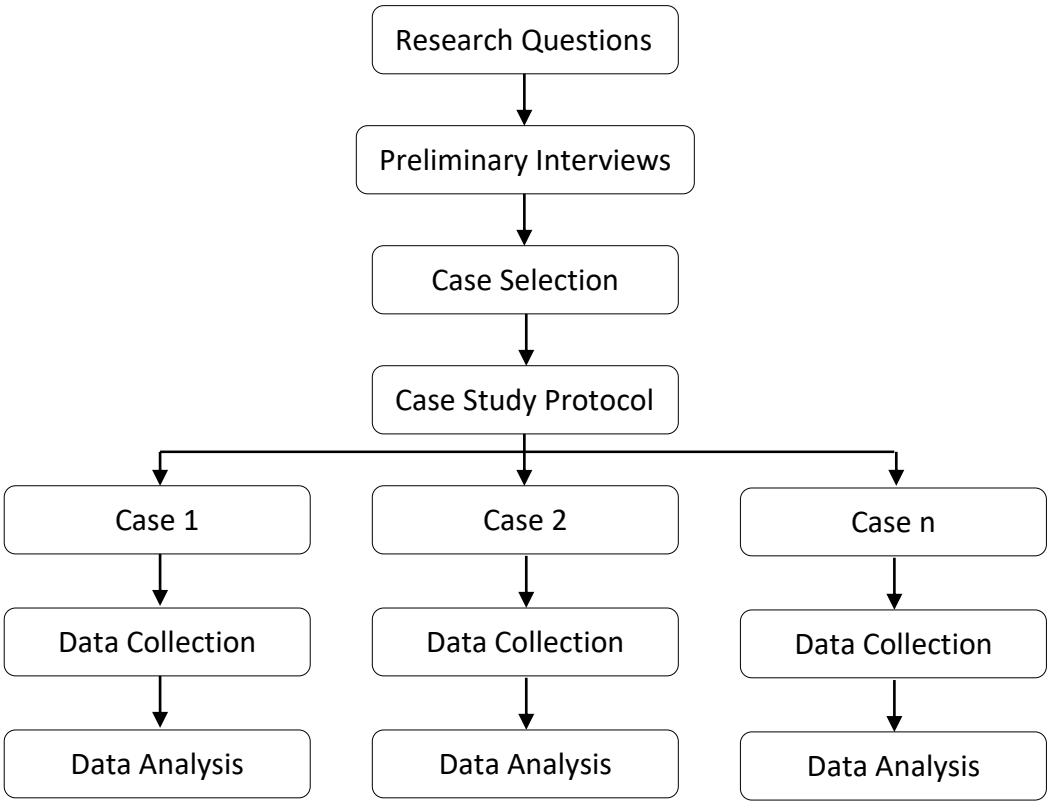

Figure 3. 1: Proposed Research Design

Design begins with formulating appropriate research questions that will shape the structure of the study. Initially, preliminary interviews were carried out to identify different areas where dirty data in customer profiles have affected the businesses in telecommunication industry. Base on a ratings of industry experts and practitioners, on the magnitude of the implications and effect of the identified areas against the organization and its decision-making process, the most affected area was chosen. Table 5.1, in Chapter 5 provides the evaluation results of the preliminary interviews and the most rated area was selected to continue to address the effects of dirty data in the telecommunication industry. Information gathered from the preliminary interviews, based on the selected area were used to construct the case studies that are being addressed by the research. The case studies are further explained in Section 3.4. The case study approach was

selected based on its usefulness and appropriateness for this particular study. The need to in-depth explore of business scenarios, to collect detailed information using variety of data collection procedures bounded by the existing business activities, in the current context has influenced the choice of methodology.

Case study protocol represents the structure of each case study that will be presented. This case study protocol is used in presenting the qualitative aspect of analyzing the effect of dirty data in Telecommunication industry. All three case studies will adhere to the same structure. Case study consists of two sections, Incident and Impact;

Incident will elaborate the exact scenario of case study occurrence at a telecommunication organization. It will further take one instance to showcase the gravity of the case study.

Impact of a case study will be showcased via a table. It will include the complications towards the cost and revenue factors separately, in order to elaborate the effect of dirty data in customer profiles.

Data collection and analysis were done for individual case studies based on the context of the problem that is being addressed. Since the research exercise pragmatic research methodology, in each case study, the first research question is being addressed using the qualitative aspects and second research question is being addressed by qualitative aspects. Data collection process carried out to comprehend the effect of dirty data on customer profiles (first research question) was performed through semi structured interviews with open ended questions. It was used to allow the participants to provide the information that is important to them but not necessarily reflected in the interview questions. A set of interview questions (Appendix B: Interview Guide) was used to guide the interview to explore effects of dirty data in the context of day to day business activities and long-term business decisions. The quantitative aspect of the research, experimentation of each case study, was carried out in different attribute levels of customer data in order to evaluate the possibilities to reduce the effect that is caused by dirty data. The design of the experiments is further elaborated in Section 3.4.

Finally, each of the case study experiments are evaluated. The matching quality was evaluated using accuracy which is based on precision and recall and defined in the equation (1):

$$\text{Accuracy} = (\text{float}\,(\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})) * 100 \qquad (1)$$

with precision and recall defined as $\text{Precision} = |\text{TP}|/(|\text{TP}| + |\text{FP}|)$ and $\text{Recall} = |\text{TP}|/(|\text{TP}| + |\text{FN}|)$ and TP being the true positives (known matched pairs classified as matches), TN true negatives (known unmatched pairs classified as non-matches), FP the false positives (unmatched pairs classified as matches) and FN the false negatives (known matched pairs classified as non-matches).

3.4 Case Studies

The case studies defined in this study focuses to identify different contexts that the effect of dirty data would affect in customer profiling in the telecommunication industry of Sri Lanka. The existence of dirty data in customer profiles are considered in the case studies at different attribute levels of customer data and it has evaluated the similarity measure techniques that can be used to reduce the effect by controlling the customer accounts with dirty data. The Customer Profiling issue elaborated in Section 5.3, Chapter 5 have considered throughout the research and techniques of identifying two or more different customer accounts under one customer profile, have been investigated to control and mitigate its effects through merging the duplicated accounts formed as a result of dirty data. Further, the effects of dirty data in the area of customer profiling have affected mainly in different sub-areas, and it is presented in Sections 5.4.1 to 5.4.3.

In order to merge different customer accounts under one customer profile, the use cases try to determine the matching quality and accuracy rate of similarity measure techniques to merge two different accounts, by applying matching of different attributes of customer data. This is done to make sure that no misleading or duplicate information is available in customer data. Each of the use cases, isolates the levels which determine the matching of two accounts considering similarity measure techniques that can be used for: full name attribute; address attribute; and email address attribute. Similarity measure techniques were evaluated and presented for each attribute of customer data collected from the telecommunication customer account dataset. Figure 3.1 depicts design of the experiments of each case study.

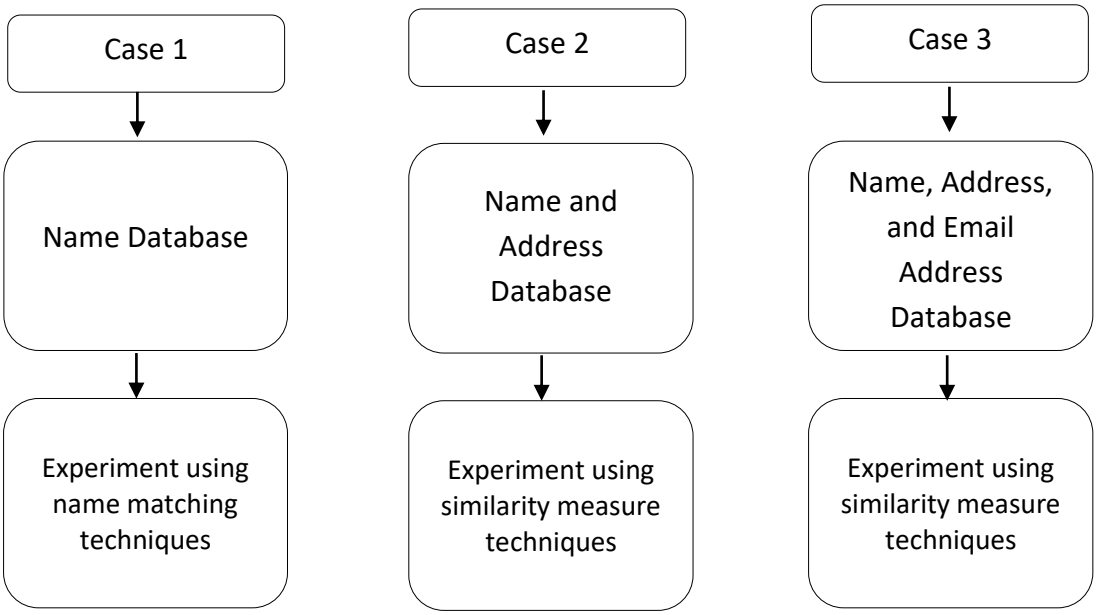

Figure 3. 2: Design of the Experiments

The first case study experiment tries to find the best similarity measure technique applicable for Sri Lankan name dataset. The purpose of the experiment is to merge two duplicated accounts of the same customer profile, by measuring the similarity of full name, when no unique identifier is available. A comparison of similarity measure techniques using a real world telecommunication dataset containing personal names is performed. Here the name matching is executed on the dataset individually without taking any other context information into consideration.

The experiment of the second case study evaluates the similarity measure techniques that can be used to measure the similarity of the address attribute of two customer accounts, in order to support the merging of duplicate accounts of the same customer profile. Here the demographic information, which is personal to each and every customer account is taken into consideration, to further confirm the state of linkage of two customer accounts.

The third case study takes the email address attribute which has been incorporated with dirty data into account and evaluates the similarity measure techniques that can be used to assess the similarity of the attribute. The purpose of using email address attribute is to confirm more on the merging of two accounts of the same customer profile, continuing the usage of more information of the customer.

The similarity measure techniques used to match the attributes of customer information are established basing the paper '*A Comparison of Name Matching: Techniques and Practical Issues*' by Peter Christen [18]. Since the paper focuses on approximate string matching techniques and aim to determine the best matching quality for different name types, the mentioned study is selected. The techniques evaluated are most commonly used in approximate string matching, which has widespread applications, from data linkage and duplicate detection, information retrieval, correction of spelling errors, approximate data joins to bio and health informatics [18]. Further, these techniques can be broadly classified into edit distance and q-gram based techniques and some of the techniques have combined phonetic encoding and pattern matching in order to improve the matching quality. Phonetic encoding techniques are language dependent, which attempt to convert a name string into a code according to how a name is pronounced (the way the name is pronounced) [18].

The following string matching techniques [18] were used in the study to measure the similarity measures of the case studies. For some of the presented techniques, different approaches to calculate the similarity exists and the techniques provide the similarity measure between 1.0 (strings are the same) and 0.0 (strings are totally different).

- **Jaro**

  Jaro algorithm is commonly used for name matching in data linkage systems. It accounts for insertions, deletions and transpositions. The algorithm calculates the number c of common characters and the number of transpositions t. A similarity measure is calculated as;

  $$simjaro\ (a1, a2)\ =\ (1/\ 3)\ \{c\ /|a1|\ +\ c\ /|a2|\ +\ (c\ -\ t)/\ c\}$$

  The time and space complexities of this algorithm are $O\ (|a1|\ +\ |a2|)$.

- **Winkler**

  The Winkler algorithm improves upon the Jaro algorithm by identifying that fewer errors typically occur at the beginning of names. Therefore, Winkler algorithm increases the Jaro similarity measure for agreeing initial characters (up to four). It is calculated as;

  $$simwink\ (a1, a2)\ =\ simjaro\ (a1, a2)\ +(\ s\ /10)\ (1.0\ -\ simjaro\ (a1, a2))$$

  with s being the number of agreeing characters at the beginning of two strings (for example, 'peter' and 'petra' have s = 3).

- **q-grams**

  Sub strings with length q in longer strings are identified as q-grams (also known as n-grams). Among many types of q-grams unigrams (q = 1), bigrams (q = 2) and trigrams (q = 3) are most frequently used.

  E.g.- 'peter' contains the bigrams 'pe', 'et', 'te' and 'er'.

  Similarity measure between two Strings is calculated using the count of common q-grams in both strings and dividing it by either

  1. The number of q-grams in the shorter string (Overlap Coefficient)
  2. The Number in the Longer string (Jaccard similarity) or
  3. The Average number of q-grams in both strings (Dice Coefficient)

  Complexities of q-grams based techniques using Time and space are $O\ (|s1| + |s2|)$.

  Padding q-grams can be also used to increase the matching quality. Padding strings must be done before q-gram comparison is performed, by adding (q−1) special characters to the start and end of the strings.

  E.g.- with bigrams, 'peter' would be padded to '⊳peter⊲' (with '⊳' symbolizing the start and '⊲' the end character), resulting in bigrams '⊳p', 'pe', 'et', 'te', 'er' and 'r⊲'.

  Because of this q-grams at the beginning of the strings and at the end of strings will not be matched to other q-grams in the middle. This will result in a larger similarity measure for two strings that have the same beginning and end but errors in the middle, but in a lower similarity measure if there are different string starts or ends between the two strings in comparison.

- **Positional q-grams**

  Positional q-grams can be used as an extension of q-grams to addition of Positional information (location of q-grams within the given string) and to match only common q-grams that are within a maximum distance from each other.

  Positional q-grams can also be padded with start and end characters similar to non-positional q-grams, and similarity measures can be calculated in the same three ways as with non-positional q-grams.

- **Skip-grams**

  Skip-grams algorithm has been developed recently to improve matching within a cross-lingual information retrieval system. It is being developed based on the goal of forming bigrams between two adjacent characters as well as skip characters (Skip-grams). The Type of skip-grams created are defined using Gram classes.

  E.g.- For a gram class gc = {0, 1} and string 'peter', the following skip-grams are created: 'pe', 'et', 'te', 'er' (0-skip grams) and 'pt', 'ee', 'tr' (1-skip grams).

- **Levenshtein or Edit Distance**

  the Levenshtein distance or Edit Distance(Distld) is a string metric for measuring the difference between two Strings. The Edit distance between two words is Calculated using the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. In the Basic form of Edit distance, the cost of each edit can be considered as 1. The Edit distance between two strings a1 and a2 can be calculated using time $O\ (|a1|\ \times\ |a2|)$and $O\ (min\ (|a1|, |a2|))$space, by using a dynamic programming algorithm. The Edit distance can be converted in to a measure of similarity of the two strings (between 0.0 and 1.0) using

  $$simld\ (a1, a2)\ =\ 1.0\ -\ distld\ (a1, a2)\ /max\ (|a1|, |a2|)$$

  distld (a1, a2) is the actual Edit Distance Function which can only have a value ≥0 depending on whether the strings are same or different. The Edit distance is symmetric (the number of additions, deletions and substitutions to get from a String a1 to a2 is the same as getting from string a2 to a2) and it always following Properties.

  1. 0 < distld (a1, a2) < max (|a1|, |a2|)
  2. abs (|a1| − |a2|) ≤ distld (a1, a2).

  Large difference in length between a given pair of strings can be quickly identified and filtered using the second Property. Different Edit Costs or even costs that depends upon characters (E.g.- a substitution from 'q' to 'g' might be given smaller costs than from 'x' to 'i' because of their visual similarity) can be allowed to the basic Edit distance using extensions. Different techniques are explored by researchers in recent years to learn the costs of edits from training data to improve the matching quality.

- **Damerau-Levenshtein Distance**

  Damerau-Levenshtein Distance is a variation of Original Levenshtein distance in which a transposition is considered as an elementary edit operation with cost 1. (in the Levenshtein distance, a transposition corresponds to two edits: one insert and one delete or two substitutions). The simdld (dld - Damerau-Levenshtein Distance) measure is calculated using the same function as simld.

- **Bag distance**

  Bag Distance is a recently propose algorithm as a cheap approximation to Edit (Levenshtein) Distance. Multiset of the characters in a string (for example, multiset ms('peter') = {'e', 'e', 'p', 'r', 't'}) is defined as a bag, and the bag distance between two strings (a1, a2) is calculated using following function.

  $$distbag\ (a1, a2)\ =\ max\ (|x\ -\ y|, |y\ -\ x|),$$

  with x = ms(a1), y = ms(a2) and $|\cdot|$ denoting the number of elements in a multiset.

  For example,

  distbag ('peter', 'pedro') = distbag ({'e', 'e', 'p', 'r', 't'}, {'d', 'e', 'o', 'p', 'r'}) = max (| {'e', 't'} |, | {'d', 'o'} |) = 2

  There is an Observation that in every instances distbag (a1, a2) ≤ distld (a1, a2), and thus simbag ≥ simld (simbag measure, calculated similarly to simld).

  Computation Complexity of Bag Distance is $O\ (|a1|\ +\ |a2|)$, and therefore it can be used as an efficient technique to filter out candidate matches before moving to more complex and expensive edit distance techniques.

- **Smith-Waterman**

  Smith-Waterman algorithm is also based on dynamic programming approach (similar to edit distance), but allows gaps as well as character specific match scores. This was originally developed to find optimal alignments between biological sequences (E.g.- DNA, proteins). Instead of looking at the entire sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure. The five basic operations are;

  1. An exact match between two characters with score 5,
  2. An approximate match between two similar characters with score 2,

3. A mismatch between two characters (that are neither equal nor similar) with score -5,

4. A gap start penalty with score -5, and

5. A gap continuation penalty with score -1.

Allowing for gaps make the Smith-Waterman algorithm more suited for compound names that contain initials only or abbreviated names. The space complexity of the Smith-Waterman algorithm is $O\ (|s1| \times |s2|)$, while its time complexity is $O\ (min\ (|s1|, |s2|) \times |s1| \times |s2|)$.

A similarity measure can be calculated using final best score bssw, the highest value within dynamic programming score matrix.

$$simswd\ (s1, s2) = bssw\ /\ (divsw \times match\ score)$$

with match score the value when two characters' match, and divsw being a factor that can be calculated in one of three ways:

(1) $divsw = min\ (|s1|, |s2|)$,

(2) $divsw = 27\ max\ (|s1|, |s2|)$, or

(3) $divsw = 0.5 \times\ (|s1| + |s2|)$ (average string length).

These three corresponds to the Overlap coefficient, Jaccard similarity, and Dice coefficient, respectively, for q-grams.

- **Compression**

Compression based similarity calculations are identified as an effective technique for multiple uses which includes clustering of biological sequences, optical character recognition, and music. The normalized compression distance (NCD) is defined based on commonly available compression techniques, like Zlib or BZ2:

$$distncd\ (a1, a2) = \{|C(a1a2)\ | - min(|C(a1)\ |, |C(a2)\ |)\}/$$
$$max(|C(a1|), |C(a2)\ |),$$

with C being a compressor (e.g. Zlib or BZ2), $|\cdot|$ the length of a compressed string, and a1a2 the concatenation of the two input strings. Compression based similarity technique does not need any parameters (besides selecting a compression algorithm), making it unique and potentially attractive for applications where no parameter tuning is possible or desirable.

- **Longest common sub-string (LCS)**

  Longest common sub-string (LCS) algorithm repeatedly (iterations) search and removes the longest common sub-string in the two strings (a1, a2) in comparison, up to a minimum length (usually set to 2 or 3).

  E.g.- The two-name strings 'gail west' and 'vest abigail' have a longest common sub-string 'gail'. After it is removed, the two new strings are 'west' and 'vest abi'. In the second iteration, the sub-string 'est' is removed, leaving 'w' and 'v abi'. The total length of the common sub-strings is now 7. If the minimum common length would be set to 1, then the common whitespace character would be counted towards the total common sub-strings length as well.

  The similarity measure between two strings (a1, a2) can be calculated by dividing the total length of the common sub-strings by the minimum, maximum or average lengths of the two original strings. This algorithm is most suited for compound names (as shown in above example) that have swapped words. The complexity of the LCS algorithm, is $O(|a1| \times |a2|)$ time using $O\ (min\ (|a1|, |a2|))$ space.

- **Sorted-Winkler**

  This can be used to strings with multiple words (Strings with at least one whitespace or separator) to improve the matching quality by first sorting the words alphabetically before the Winkler technique is applied. Sorting swapped words will bring them into same order thus improving the matching quality.

- **Permuted-Winkler**

  Permuted Winkler is a more complex approach where Winkler comparisons are performed over all possible permutation of words, and the maximum of all calculated similarity values is returned.

3.5 Summary

This chapter mainly describe about the proposed research design. Research design encompasses of three sections. It outlined the research strategy, research design and case studies. Research design has explained the main analyzing method to be used in identifying the impact of dirty data on customer data within telecommunication industry. It has further described the design of the proposed experiments to be implemented to rectify the identified impacts.

Chapter 4 - Implementation

4.1. Introduction

This chapter will mainly explain the workings of the implementation of the experiments of case studies that were presented in Section 5.2, Chapter 5. Section 4.2 elaborates the implementation details about the main software tools that were applied for implementation. Section 4.3 represents the alignment of implementation specifics about similarity measure techniques that were used as a solution for use cases. Whereas Section 4.4 presents the implementation details about the evaluation workings of similarity measure techniques that were used for use cases.

4.2. Software Tools

The proposed experiments were mainly implemented using python 2.7, and all the matching techniques were implemented in python, using Febrl (Freely Extensible Biomedical Record Linkage) data linkage system [17]. Also, Matplotlib library and NumPy arrays were used to generate the charts presented in Section 5.2.

4.3 Implementation Details – Alignment of Similarity Measure Techniques towards the corresponding Use Cases

Within this section, it has elaborated as mentioned within Section 3.4, Chapter 3 the implementation of alignment of similarity measure techniques over the selected data sets that adhere to the use cases mentioned in Section 5.2, Chapter 5.

```
with open('TEST_DATA.csv', 'rb') as f_open:
    msg = []

    for line in f_open:
        twoNames = line.rstrip('\n\r').split(',')

        s = '%13s,%13s,' % (twoNames[0], twoNames[1])
```

As shown, after loading the corresponding data set, using the rstrip function, first new line is removed with slash tab. Through split function copy of string in which all characters that have stripped from the beginning to end of the string were append. For this function a string tuple (first string attribute, second string attribute) was used as the parameters which will generate return values of 0 and 1.

After the strings are being stripped the parameters were checked against string matching techniques as mention in Section 3.4.

```
def jaro(str1, str2, min_threshold = None):
```

As mentioned in Section 3.4 all twenty-three techniques will be called to generate the similarity measure weight.

```
for i in range(len(strings[0])):
    print('calculating... ', i)
    for j in range(len(strings[0])):
      str1 = strings[0][i]
      str2 = strings[1][j]
      start_time = time.time()
      jaro(str1,str2)
      time_used = time.time() - start_time
      start_time1 = time.time()
      winkler(str1,str2)
      time_used1 = time.time() - start_time
      start_time2 = time.time()
```

```
qgram(str1,str2,1)
time_used2 = time.time() - start_time
start_time3 = time.time()
qgram(str1,str2,2)
time_used3 = time.time() - start_time
start_time4 = time.time()
qgram(str1,str2,3)
time_used4 = time.time() - start_time
start_time5 = time.time()
posqgram(str1,str2,1)
time_used5 = time.time() - start_time
start_time6 = time.time()
posqgram(str1,str2,2)
time_used6 = time.time() - start_time
start_time7 = time.time()
posqgram(str1,str2,3)
time_used7 = time.time() - start_time
start_time8 = time.time()
sgram(str1,str2,[[0],[0,1],[1,2]])
time_used8 = time.time() - start_time
start_time9 = time.time()
editdist(str1,str2)
time_used9 = time.time() - start_time
start_time10 = time.time()
mod_editdist(str1,str2)
time_used10 = time.time() - start_time
start_time11 = time.time()
bagdist(str1,str2)
time_used11 = time.time() - start_time
start_time12 = time.time()
editex(str1,str2)
time_used12 = time.time() - start_time
start_time13 = time.time()
seqmatch(str1,str2)
time_used13 = time.time() - start_time
start_time14 = time.time()
compression(str1,str2,'bz2')
time_used14 = time.time() - start_time
start_time15 = time.time()
compression(str1,str2,'zlib')
time_used15 = time.time() - start_time
start_time16 = time.time()
compression(str1,str2,'arith')
time_used16 = time.time() - start_time
start_time17 = time.time()
lcs(str1,str2,2)
time_used17 = time.time() - start_time
start_time18 = time.time()
lcs(str1,str2,3)
time_used18 = time.time() - start_time
start_time19 = time.time()
ontolcs(str1,str2,2)
time_used19 = time.time() - start_time
start_time20 = time.time()
ontolcs(str1,str2,3)
time_used20 = time.time() - start_time
start_time21 = time.time()
permwinkler(str1,str2)
time_used21 = time.time() - start_time
start_time22 = time.time()
sortwinkler(str1,str2)
time_used22 = time.time() - start_time
```

```
start_time23 = time.time()
swdist(str1,str2)
time_used23 = time.time() - start_time
start_time24 = time.time()
syllalignmdist(str1,str2)
time_used24 = time.time() - start_time
start_time25 = time.time()
charhistogram(str1,str2)
time_used25 = time.time() - start_time
start_time26 = time.time()
twoleveljaro(str1,str2)
time_used26 = time.time() - start_time
start_time27 = time.time()
twoleveljaro(str1,str2, qgram, 0.8)
time_used27 = time.time() - start_time

s = '%13s %13s' % (str1, str2)
s += ' %.3f' % (jaro(str1,str2))
s += ' %.10f' % (time_used)
s += ' %.3f' % (winkler(str1,str2))
s += ' %.10f' % (time_used1)
s += ' %.3f' % (qgram(str1,str2,1))
s += ' %.10f' % (time_used2)
s += ' %.3f' % (qgram(str1,str2,2))
s += ' %.10f' % (time_used3)
s += ' %.3f' % (qgram(str1,str2,3))
s += ' %.10f' % (time_used4)
s += ' %.3f' % (posqgram(str1,str2,1))
s += ' %.10f' % (time_used5)
s += ' %.3f' % (posqgram(str1,str2,2))
s += ' %.10f' % (time_used6)
s += ' %.3f' % (posqgram(str1,str2,3))
s += ' %.10f' % (time_used7)
s += ' %.3f' % (sgram(str1,str2,[[0],[0,1],[1,2]]))
s += ' %.10f' % (time_used8)
s += ' %.3f' % (editdist(str1,str2))
s += ' %.10f' % (time_used9)
s += ' %.3f' % (mod_editdist(str1,str2))
s += ' %.10f' % (time_used10)
s += ' %.3f' % (bagdist(str1,str2))
s += ' %.10f' % (time_used11)
s += ' %.3f' % (editex(str1,str2))
s += ' %.10f' % (time_used12)
s += ' %.3f' % (seqmatch(str1,str2))
s += ' %.10f' % (time_used13)
s += ' %.3f' % (compression(str1,str2,'bz2'))
s += ' %.10f' % (time_used14)
s += ' %.3f' % (compression(str1,str2,'zlib'))
s += ' %.10f' % (time_used15)
s += ' %.3f' % (compression(str1,str2,'arith'))
s += ' %.10f' % (time_used16)
s += ' %.3f' % (lcs(str1,str2,2))
s += ' %.10f' % (time_used17)
s += ' %.3f' % (lcs(str1,str2,3))
s += ' %.10f' % (time_used18)
s += ' %.3f' % (ontolcs(str1,str2,2))
s += ' %.10f' % (time_used19)
s += ' %.3f' % (ontolcs(str1,str2,3))
s += ' %.10f' % (time_used20)
s += ' %.3f' % (permwinkler(str1,str2))
s += ' %.10f' % (time_used21)
s += ' %.3f' % (sortwinkler(str1,str2))
```

```
        s += ' %.10f' % (time_used22)
        s += ' %.3f' % (swdist(str1,str2))
        s += ' %.10f' % (time_used23)
        s += ' %.3f' % (syllalignmdist(str1,str2))
        s += ' %.10f' % (time_used24)
        s += ' %.3f' % (charhistogram(str1,str2))
        s += ' %.10f' % (time_used25)
        s += ' %.3f' % (twoleveljaro(str1,str2))
        s += ' %.10f' % (time_used26)
        s += ' %.3f' % (twoleveljaro(str1,str2, qgram, 0.8))
        s += ' %.10f' % (time_used27)
        msg.append(s)


        if (qgram(str1, str2, 2) != sgram(str1,str2,gc=[[0]])):
          msg.append('  Error: 2-gram != s-gram (with gc=[[0]])')

        if (editdist(str1, str2) > bagdist(str1,str2)):
          msg.append('  Error: BadD > EditD')

        if (lcs(str1,str2,1) < lcs(str1,str2,2)):
          msg.append('  Error: LCS1 < LCS2')

        if (lcs(str1,str2,2) < lcs(str1,str2,3)):
          msg.append('  Error: LCS2 < LCS3')

        if (editdist(str1, str2) > mod_editdist(str1,str2)):
          msg.append('  Error: EditD > Modified EditD')
```

Jaro, J-W, 1gram, 2gram, 3gram, 1pqgr, 2pqgr, 3pqgr, Sgram, eDist, meDis, bDist, Editx, SeqMa, ComBZ, ComZL, ComAC, LCS2, LCS3, OLCS2, OLCS3, P-Win S-Win, SWDis, SyADi, Histo, 2LJaro and 2LJaroA functions were called using Febrl (Freely Extensible Biomedical Record Linkage) data linkage system [17] to get the similarity measure weight. Within the code String [0] depicts the characters in the first column while string [1] denotes the characters in the second column. As shown above after calling similarity measure technique functions, computational time has been calculated to each function.

```
f = open('result3.csv','w')
print(len(msg))
count = 0
for i in msg[4:]:
  count += 1
  i = i.replace('\n','')
  i = i.replace('\r','')
  m = ','.join(i.split(' '))


  print(count)
  f.write(m)
  f.write('\n')
f.close()
```

Above function would get the overall results of the data file to be append as a csv file.

4.4 Implementation Details – Evaluation of Similarity Measure Techniques corresponding to the Dataset

Importing the matplotlib library, the evaluation points were graphed on a chart using the precision, recall and accuracy functions. The main input that was appended to this implementation was the result that was output through Section 4.3.

```
def sensitivity(tp, fp, tn, fn):
    sensitivity = (float(tp) / (tp + fn)) * 100
    return sensitivity


def precision(tp, fp, tn, fn):
    precision = (float(tp) / (tp + fp)) * 100
    return precision


def accuracy(tp, fp, tn, fn):
    accuracy = (float(tp + tn) / (tp + fp + tn + fn)) * 100
```

As shown above the similarity weights generated in Section 4.2, precision and recall values are generated.

```
x = np.array([0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0])
        y = np.array(precisionList)
        b = np.array(sensitivityList)

        plt.plot(x, y, 'r*')
        plt.plot(x, b, 'b+')


        plt.xlabel('x - axis')

        plt.title('graph! method ' + str(changeList))
        plt.show()
```

As shown above, np.array function will append the threshold values from 0 to 1, to compute the accuracy values for similarity matching techniques. These accuracy levels will generate the most accurate similarity technique to the considered data set. Function Plt.plot will plot the graph and the function plt.show will display the graph.

4.5 Summary

In this chapter, the software tools utilized to implement the proposed experiments were elaborated. Then the functionalities that are used to get the similarity weighted values through similarity measure techniques are being elucidated. Finally, evaluation functions are being clarified to show case the accuracy of the similarity measure techniques. The next chapter will depict the results generated from the implementations that are listed within this chapter.

Chapter 5 - Results and Evaluation

5.1 Introduction

This chapter elaborates the results of the analysis that has been done and the evaluation of the results.

5.2 Analysis of Preliminary Interviews

The preliminary interviews were carried out to narrow down the field into the areas with great specificity and detail. It is being identified that the effect of dirty data on customer profiles has been escalated to many areas in businesses directly and indirectly. And the aim of the preliminary interviews was to identify the area/s in businesses in telecommunication industry that dirty data would affect the most in order to develop the case studies. The purpose of the preliminary interviews was to narrow down the research to the most significant case/s among the cases identified.

The preliminary interviews carried out involved nearly 10 individuals. The participants in the study were comprised of department heads and industry experts of Dialog Axiata PLC, Sri Lanka Telecom and Mobitel, leading telecommunication companies in Sri Lanka. Our participants had seven to eleven years of experience on the field on average and were engaged in similar data analytics projects. All participants described themselves as having special education experience as part of their current roles and past roles as persons who have realized the existence of dirty data on customer profiles.

After conducting the preliminary interviews, below areas were able to recognized as areas affected by dirty data.

1st area - Customer Profiling issues

This is the main case that was pointed out by many experts who participated for interviews. The issue of customer profiling mainly occurs due to duplication of same person or an individual's profile more than once within the database. The root cause for this phenomenon is attributed by the existence of many primary key fields to create a customer account. For example, in telecommunication industry, unique identification of customer accounts is done by many primary keys such as National Identification number, passport number, driving license number. This is allowed to make the account creation process more convenient in a real business world scenario. Due to this, it has caused the creation of many accounts under the same customer name as there's no unique identifier to merge them.

2nd area - Geo Locations Mapping issues

This case study mainly highlights the Geo marketing obstacle. Geo marketing mainly helps to integrate geographical locations for purposes such as marketing, distribution, territorial planning, and site selection. When the corresponding customers are not distinguished accurately overall purpose of Geo marketing is failed.

3rd area - Consequences of not maintaining proper Hierarchy with regard to Company Profiles and SBUs

This problem can be mainly found in corporate accounts. Dominant cause for the above-mentioned issue is mainly appear due to incorrect linkage between profile records of different companies of the same group. Organizations often find difficulty in identifying the relationships of each corporate customer linked into the hierarchy.

4th area - NIC Mapped issues

NIC mapped issue arises due to the availability of many primary keys, to create customer accounts, since some customers tend to register through driving license and passport. Organization face the difficulty in identifying the same customer and all his connections, when the customer has received couple of connections via different primary keys.

5th area - Single View of customer

Above mentioned matter mainly arises due to inaccurate linking of customer records with an individual's profile. When the overview of the customer is considered when there are different accounts of the same profile, it is problematic to identify the customer in his full picture, which matter to make segmentation, and packaging decisions.

6th area – Customer Lifecycle Management failures

CLM failures mainly result due to a collection of problematic circumstances. There is no specific root cause for this matter. Inaccurate customer profiling, geo marketing failures, inappropriate segregation of SBUs are some factors which create the edge to make the Customer Lifecycle Management Failed.

7th area - Governance policy

Data governance issues mainly occurs at the touch point of the data entry level. Many policies and regulations get violated at the point of data entry which initiate unnecessary circumstances. Data governance matter has been considered as one of the significant concerns to be addressed.

For the purpose of selecting the most significant case from the above areas, the interviewees were provided with a questionnaire (Appendix A: Questionnaire) to rank the identified from the scale 0 to 5.

Table 5.1 presents the analysis of the results of the rating provided to the interviewees, using a scale of 0=Not at all important, to 5=Very important.

Interviewee	Job Role	1st area	2nd area	3rd area	4th area	5th area	6th area	7th area
Shanka Rabel	Senior Director - Solutions, Virtusa (Pvt) ltd	5	3	4	4	5	1	2
Chaminda Ranasinghe	Chief Executive Officer at IdeaHub (Pvt) Ltd	4	3	5	4	5	2	3
Sandra De Zoysa	Group Chief Customer Officer	5	3	4	4	5	2	3
Eranda Adikari	Senior Manager - BI Analytics at Dialog Telekom	5	3	5	2	4	2	3
Sajeewani De Zoysa	PA to Group Chief Customer Officer / GCCO's Office	4	3	3	2	3	3	2
Waruni Algama	Head - CE, Group Loyalty & CRM at Dialog Axiata PLC	5	3	3	3	3	3	1
Maneesha Jinadasa	Deputy Chief Officer at Sri Lanka Telecom	5	4	3	3	3	2	2
Rohitha Somaratne	General Manager IT Strategy & Planning at Sri Lanka Telecom	3	3	5	3	3	2	2
Pubudu Chinthaka	Assistant Manager at Dialog Axiata PLC	3	3	3	3	5	2	2
Dhanya Herath Gunaratne	General Manager Operations at MAS Legato	3	4	5	3	3	1	2
Total		**42**	**32**	**39**	**31**	**39**	**20**	**22**

Table 5. 1 : Analysis of areas of Dirty Data

According to the results obtained from Table 5.1, the Customer Profiling issue shows a significant importance over other areas since 7 out of 10 industry experts and practitioners have rated it as *Very Important* in the questionnaire. Thus, area of Customer Profiling is considered as the most critical issue to be addressed related to dirty data in Sri Lankan Telecommunication industry. Therefore, the research aims to pursue its research questions under Customer Profiling issue.

5.3 Effect of Dirty Data on Customer Profiles

In this section the results of a series of interviews and group discussions that were conducted in order to gather information about the effect of dirty data on customer profiles are being discussed. The results gathered on each use case are presented along with the incident and impact which is constructed through case studies.

The interviews and the group discussions were held targeting to gather information pertaining to the first research question. Dialog Axiata PLC, Sri Lanka Telecom and Mobitel were taken as the case companies and around 25 personnel were contacted and interviewed. The participants comprised of the department heads and industry experts in the fields of data analysis, big data, customer relationship management and strategic level decision makers of the case companies. And the interviews lasted approximately 30 to 70 minutes and the interviews are being recorded in the audio format with the participants' approval. Also, handwritten notes were taken during the interview to track the key points and to highlight areas of particular importance.

Interviews contributed differed amount of information to the case studies and some participants talked at length on one or two case studies; some participants made nearly equal contributions across all three case studies. Thus, all participants' voices and views are represented abstractly.

Semi structured interviews with open ended questions were used for the interview procedure. It was used to allow the participants to provide the information that is important to them but not necessarily reflected in the interview questions. A set of interview questions (Appendix B: Interview Guide) was used to guide the interview to explore effects of dirty data in the context of day to day business activities and long-term business decisions.
As the first step of interview process, purpose of the study, research procedures were reminded and any participant questions or concerns about the study or procedures were asked. Also, information about the authors were provided to establish mutual understanding and trust.

After analyzing the overall interviews conducted by domain experts, following sub-issues were listed as the main repercussions of customer profiling issue.

1. Customer profiling issue formed as a result of dirty data in personal names.
2. Billing issue formed as a result of dirty data in addresses of customer profiles.
3. CRM and marketing issues formed as a result of dirty data in customer profiles.
4. Customer segmentation issue caused as a result of duplication of customer profiles.

Section 5.3.1 – 5.3.4 and Section 5.4 have elaborated the findings of the above cases separately.

5.3.1 Case Study 01 - Customer Profiling issue formed as a result of Dirty Data in Personal Names

Description: Above case study mainly tries to address the issue of customer profiling which has aroused through mismatches of personal names. Customer profiling can be simply elaborated as an individual customer having one or more accounts as a customer in an organization. The case study tries to merge different accounts based on the personal names including the first name, middle name and last name, by considering similarity measure as the merging approach.

Incident and impact sections will further showcase the scenario and the effect of the above case study as classified in Section 3.3, Chapter 3.

Incident

Below scenario is related to the first case study, and it will represent and replicate the exact situation within telecommunication industry.

Customer X has taken five service connections from a telecommunication company called Y, in different instances of a timeline. Unfortunately, all these services have not classified under the same customer profile. One of the main reasons for this situation has occurred due to inconsistencies and errors within the data entry level and because of the use of several primary key fields such as NIC, passport number and driving licence number, to identify a single customer. Due to these, it is not always possible to identify all customer accounts under one profile. Because of the occurrence of these situations, customer X has been classified under two/several customer accounts. As an example the services bought by customer X has been divided as two service connections under one account and other three service connections under another customer account. For further elaboration using a specific name, if the customer X‘s name is taken as Malith Galgomuwa, for two times it has recorded as Malith Galgomuwa but another three times it has recorded as Malith Golgomowa. But all the five connections should be recorded as Malith Galgomuwa in one individual profile. Due to above scenario following impact was discovered under the categories of revenues and cost.

Customer value creation

In terms of revenue aspect when this issue is taken into consideration, it is important to understand that one of the primary objective of organization has been classified as the value creation. Unfortunately, with accordance to the case study, customer value creation is the main area that has affected due to the customer profiling issue. Customer value creation can be identified as the qualitative measurement of how far a customer is valuable to a company. When the value of the customer is not recognized properly, organizations faced the difficulty in segregation of value as low level, middle level and high level valued customers. Further this has led in to the difficulty of understanding consumption patterns ultimately which has impacted on crucial decisional matters within the organization.

Net worth of the customer

Net worth of the customer can be identified as a quantitative approach of valuing the customer. Due to the existence of customer profiling issue monetary value assigned to a customer will not be able to calculate. This has highly affected towards the strategic decision making in an organization. When wrong decisions are taken many futuristic projects and targets has become inaccurate.

Single view of a customer

Within a telecommunication industry it is significant to understand the customers separately from each other in an individual manner. Most of the packaging, loyalty benefit allocations are done with accordance to the customer profiles. When the company cannot view each customer as an individual, most of the CRM operations has not be able to fulfill. When customer is not recognized by the company properly, customer is being allocated to unwanted privileges or not get offered by the right privileges. These privileges will not match with the customer. It will create an additional cost to be bore by the company. In the perspective of the customer, additional benefits shared, make customer much more delighted but in the perspective of the company, it has allocated an additional cost to the customer. Cost components related to customer will gradually increase.

Customer Intelligence

Presently most of the companies in order to cater the customers, focus on the customer intelligence. Through customer intelligence customer buying patterns, consumption patterns and behavioral patterns are being identified. Further these gatherings will be used to cater futuristic projects of the organizations. When misleading information is being entered, the overall output of customer intelligence projects gets failed. Future investments have forgone due to the additional cost incurred on unnecessary projects.

5.3.2 Case Study 02 - Billing issue formed as a result of Dirty Data in Addresses of Customer Profiles

Description: Above case study mainly tries to address the billing issue. Billing issue can be explained as the problem of having multiple number of addresses to be dealt with a customer. This scenario would create additional cost component towards the organization. This case study will also enhance the Case study 01. Within this case study it will enhance the accuracy of merging several customer accounts as one customer profile by considering address attribute apart from name attribute.

Incident and impact sections will further showcase the scenario and the effect of the above case study as classified in Section 3.3, Chapter 3.

Incident

Second use case is much similar to first use case where, it is an enhancement of the first use case. Through the first use case customers' names have been taken as the attribute, in merging duplicated accounts to one customer profile. Through this method another problematic matter comes into picture. It is significant to understand that there can be occasions where same name can be existed within the database but those two accounts can be completely different people. Hence it is not practical to match people only through the similarity measure of names. Address can be taken into considerations in such occasions to provide an overall better output. It has been further elaborated from following example,

X, a company which provides Telecommunication services and Z is a company which has taken many services from the X Company. Z Company has many subsidiaries under it. Each subsidiary has been given the privilege to use the services provided by company X. At the end of the month, Z Company pays the bill of the group. X Company is unable to identify their customers separately which are classified under subsidiaries due to this procedure. Due to existence of customers with same name (group name) it has made difficult to identify accounts and to create single profile per each subsidiary. As a result of that, when there is more than one customer profile in a group, final bills are produced for the number of customer profiles available but not a concatenated

group bill. Therefore, X Company has to bare up unnecessary billing cost due to the occurrence of this situation.

When an individual customer who has taken several service connections from a telecommunication company is being considered. And if the name of the customer is provided as Malith Perera and Malith Gayan Perera, in the service accounts, with no unique identifier or clear merging status, without considering context information, like address, the accounts would not be able to take into one customer profile. Therefore, two bills will be generated and posted to one individual.

Following Revenue and cost component changes have been discovered as a result of the above situation.

Bill bundling

Bill bundling is one of the main cost element incurred by telecommunication companies since it has a huge customer base. When a certain customer has taken many services from the company all the bills will be bundled under one customer name to be sent to the address listed. Existence of dirty data would create inefficiencies in the process, and will cause extra expenditure over billing in the company.

Bill returning

When the bills are bundled and sent to corresponding addresses, if there are any in accuracies in addresses, those bills will get return back to the company and the overall return cost should be covered by the company itself. As most of the telecommunication companies are in the level of maturity in terms of business development life cycle, it is significant to control overall expenditure levels to gain competitive advantage. Unfortunately, bill returning and bill bundling costs as become one of the main costing areas.

Revenue targets

When additional cost units' increase, revenue targets will not be able to achieve by the company. This will effect on the overall performance of the company. Due to this situation revenue will be

reduced gradually, the profit which is supposed to be invest within the other domains of the company, will get absorbed by the cost.

Proper management of promotion materials

Most of promotional materials are being sent to homes of the customers of telecommunication companies to create awareness of new products and connections available. When there are faults in addresses these promotional materials will be sent into mistaken places which will create a mismanagement of promotional activities. The overall purpose of promotions will be lost. Further, when the same customer receives irrelevant and many copies of the same promotional material, customer faces dissatisfaction.

Revenue Leakage

Due to the above mentioned reasons, under effect of Section 5.3.2, there will be more cost factors than the revenue generating paths. These cost factors lead into competitive disadvantage and when the competitive advantage is forgone, company will face unfortunate circumstances to stabilize within the maturity level of the market curve.

Only touch point between company and the customer

Billing is identified as the main method of communicating with the customer or the main touch point between customer and the company. Hence it is important to accurately manage this process, to have a healthy relationship with the customer.

5.3.3 Case Study 03 - CRM and marketing issues formed as a result of Dirty Data in Customer Profiles

Description: This case study will mainly focus the impact on CRM and marketing issues formed as a result of dirty data in customer profiles.

Incident and impact sections will further showcase the scenario and the effect of the above case study as classified in Section 3.3, Chapter 3. And this provides an overall presentation of the effect of dirty data on customer profiles, after considering Section 5.3.1 and 5.3.2.

Identification of customer

When the customer is identified erroneously, overall operational and strategic decision making processes will get affected. This highlights the factor of identifying and understanding customer accurately without any assumptions or redundancies. Total worth of the customer in monetary and economic aspects, customer patterns, customer behavior, and customer retention can be analyzed, adding value to each and every stages of customer lifecycle.

Overall experience of the customer

Customer experience management, is a top prioritized aim in current business world and personalized customer services trump the organization to be the differentiator among other companies. Without a customer being identified as a single profile, the feedback and complaints, would not get concatenated to provide a seamless service.

Accuracy level of decisions

Decisions taken on inaccurate information and initiatives taken on inaccurate decisions will lead into poor performance and success. As customer profiles are managed and analyzed by operational level, a mistake, error, that has not been corrected in the operational level, gets carried away to higher management levels, causing defective decision making.

Expansion of markets

Identification of new trends, initiatives and products and services can be analyzed and predicted through the identification of the customer. And personalized customer experience would retain and attract customers creating new market segments and boundaries. An organization will be

able to map new services and products to the most ideal customer base after the proper identification of the customer. Organizations will accomplish to reach new territories.

Customer holistic view

The ability of identifying customer holistically will open new dimensions in strategic decision making. It will show case the value of customer by enriching demographic and psychographic values. Which will lead into exceptional customer service. Marketers will be able create better promotional campaigns after analyzing every aspect of the customer.

Customer personas

Customer personas or marketing personas can be identified as generalized presentation of the ideal customers. These will help to understand the customers better using story cards. This method will help to manage the customer content, product development, and special needs of customers. Customer personas are results of final outcome of customer profiling. Therefore, correct identification of customer profiles leads to correct customer personas.

Service level agreements

This category mainly focuses on the customer level service agreements which is produced after considering all the services used by the customer. These service level agreements will be up to necessary standard as it will be created according to accurate customer profiles.

5.3.4 Case Study 04 - Customer Segmentation issue caused as a result of Duplication of Customer Profiles

Description: Above case study mainly tries to address the improper segmentation. Improper segmentation issue arises due to unsuitable grouping of customer profiles into segments. Erroneous data in the attributes of customer information and unconnected information of customer from different accounts will lead to inappropriate segmentation. Family and household of a customer, net worth of a customer, customer life span, loyalty customer segments could only be identified under different segments with correct customer information.

Incident and impact sections will further showcase the scenario and the effect of the above case study as classified in Section 3.3, Chapter 3.

Incident

The Fourth and the final use case focuses on the segmentation of customers under proper segmentations. Proper Customer segmentation is considered highly valuable for many strategic and revenue generating decisions. A proper customer segmentation will help a company to focus much merrily on company objectives such as profit generation, customer personalization and also it is identified as one of the essential prospect of deciding competitive advantage over other companies. Also, customer segmentation directly has an impact on market expansion, customer retention and profitability. Hence a company would put an additional effort on proper segmentation of customers for greater benefit of the customer and the company.

X company considered as a telecommunication company where it has segmented its customers into two main segments known as household and family group. Most of the time the main income earner of the family would get the service from the X company, on behalf of the rest of the family members, and in this case family members are privileged to use maximum benefits that are offered by X company, since the net worth of the buyer of the services is really high. Unfortunately, in the perspective of the X company all the services and benefits that were used by all family members will be taken under an individual. Therefore, telecommunication companies face the difficulty to identify individual customers who actually use the service and tend to their needs. The consequences of this sub-issue are further elaborated below.

Valuation

With the identification of proper family, household, and other segments, the value of each segments could be identified separately. These segregations are done by mainly using customer profiles. If the customer profiles are inaccurate then the overall segmentation will be incorrect and improper decisions will be made.

Loyalty issues

Improper segmentation will leave some customers out of their deserved privileges and loyalty benefits. Hence if the loyalty is not properly distributed then it would have a direct impact over customer satisfaction and retention.

Family level segregation

Family level segmentation can be identified as a macro level segmentation, when overall segmentations of telecommunication industry is considered. Telecommunication industries mainly segregate their revenue models via segmentation. Within a family level segmentation, it will try to identify all the family members who have taken a service from the organization into a one cluster. When the organization is unable to connect the dots of relationships between family members' overall family level segmentation will be inaccurate. This will lead into expensive preposition. Whereas misinterpretations customer group will cause cost increments which will gradually lead in to revenue falls.

Innovation of unsuitable service packages

If the segments are improper it will result in misrepresentation of market groups. Companies might generate service or product lines by focusing these misrepresentations. These type of circumstances would lead into the complete failure of service lines.

5.4 Experiments

In this section the results of a series of comparison experiments conducted to evaluate similarity measure techniques that can be used to measure the similarity of two customer accounts propagated as a result of dirty data are presented. The objective is to find the optimal technique/s that can be used to determine high linkage accuracy in different attributes of customer information, that would be supportive for the clerical review process.

The test data comprised of the full name, address, email address, and gender attributes extracted from the customer account dataset provided by one of the telecommunication companies in Sri Lanka. Due to privacy issues, the company and the dataset given cannot be disclosed. The dataset comprised of 1108 tuples of customer accounts, that needed to be classified as accounts of the same profile and accounts of different profiles. Duplicate accounts, exact true name pairs (indicating accounts of the same person with exact information but different key identifiers) were already removed from the dataset. The dataset only comprised of account pairs with differences in full name, address and email address attributes, which were used in the similarity comparison. The nature of the differences of the attributes, has occurred most likely due to data entry showing following types of error possibilities;

- Manual keyboard based entry errors, resulting wrongly typed neighboring keys (for example ‘n’ and ‘m’ or ‘e’ and ‘r’).
- Omission or addition of letters while typing (addition of ‘l’ mistakenly twice).
- Remembering and manual entry of words, replacing the word with most common spellings, not the correct spelling (‘Ratnayake’ typed as ‘Rathnayaka’).
- Customers deliberately providing modified or part of their names (‘Kasun Supun Perera’ reporting the name as ‘Kasun Perera’ in the second account creation).

Since valid variations should not be disregarded in telecommunication industry, as they indicate an active customer account, a proper distinguish between the variations of accounts should be analyzed and a clerical reviewing process should be performed definitely. Therefore, similarity measure techniques are used to provide better analysis of huge sets of customer accounts that contain valid variations which needed to be classified as matches or non-matches.

5.4.1 Case Study 01 - Customer Profiling issue formed as a result of Dirty Data in Personal Names

Name matching can be defined as the process of determining whether two name strings are instances of the same name [18]. Since the customer profiling issue mainly occurs due to the inability to identify different accounts of a single customer under one profile (Section 5.3, Chapter 5), exact name comparisons will not result in good matching quality [18] as name variations and errors are quite common.

This case study experiments to determine an approximate measure/s of how similar the names of duplicated customer accounts can be to merge them to a single customer profile and how to decide the linkage status of possible matches. Similarity measure techniques are evaluated to determine which techniques achieve the best matching quality and suitability for a Sri Lankan name dataset.

The case views matching two name strings as an isolated problem within the localized personal names without considering any context information like address, date of birth and various other details into account and has taken the full name of the customer accounts joining first name, middle name and surname. The aim of this experiment is to identify which matching technique/s achieve the best matching quality for different localized names of a Telecommunication dataset, in order to merge two different instances of the same name string under one profile.

Matching Results

Here present the results of the tests run applying all pattern matching techniques presented in Section 3.4 with their various ways of calculating similarity measures. The similarity measure between 1.0 (two name strings are identical) and 0.0 (two name strings are totally different) is used to present the similarity between two compared name strings in order to decide whether to merge two name strings under the same customer profile or not.

Table 5.2 shows the average similarity measures achieved for each of the presented techniques on the name data set. The similarity measures lie between 0.0, showcasing no matching at all and 1.0, showcasing exact matching.

Matching Technique	Similarity Measure
Jaro	0.6113952014
J-W	0.6531009507
1gram	0.5716487098
2gram	0.2541095518
3gram	0.1773472159
1pqgr	0.3620262562
2pqgr	0.1552847442
3pqgr	0.1188682662
sgram	0.236084201
eDist	0.3249751019
meDis	0.3254640109
bDist	0.5368990493
Editx	0.3744653689
Seqma	0.4072503395
ComBZ	0.7466165686
Comzl	0.3616143051
ComAC	0.1028465369
LCS2	0.316893617
LCS3	0.2050909914
OLCS2	0.4420941603
OLCS3	0.313875962
P-win	0.6670140335
S-Win	0.6110239928
Swdis	0.2130448167
SYADI	0.03328474423
Histo	0.6762028067
2Ljaro	0.03003259393
2Ljaro	0.02942326845

Table 5. 2 : Average Similarity Measures (Name attribute)

For the similarity measures a threshold can be varied between 0.0 and 1.0 that influences the classification of matching pairs and non-matching pairs (name pairs with a similarity value above the threshold are classified matches, and pairs with similarity value below as non-matches). The matching quality was evaluated using accuracy which is based on precision and recall and defined in the equation (1) in Section 3.3, Chapter 3. The following figures shows the best results achieved for each of the presented techniques based on the precision and recall. X-axis denotes the threshold values and y-axis denotes the similarity values. And red stars depict precision whereas blue pluses depict recall.

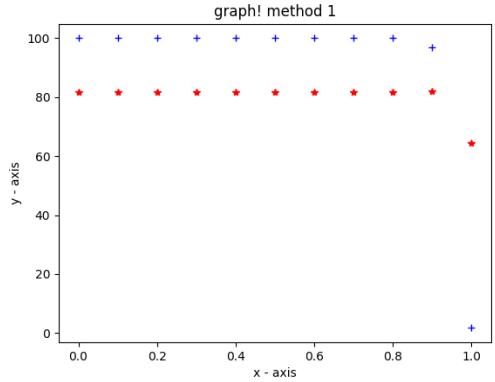

Figure 5. 1 : Jaro

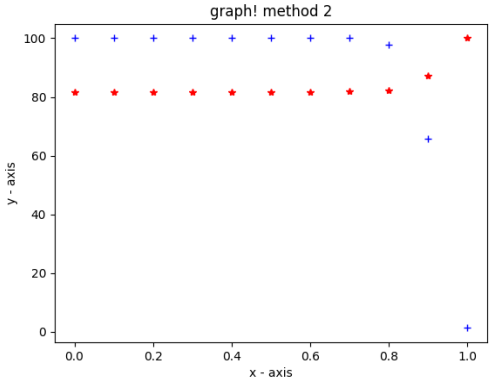

Figure 5. 2 : J-W

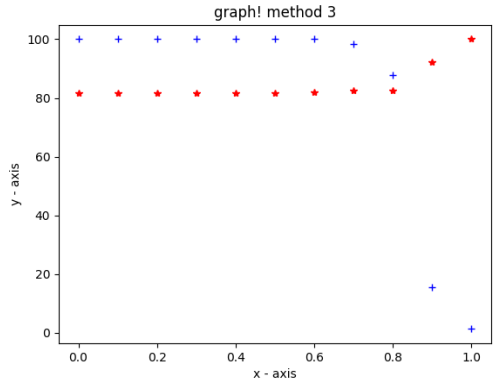

Figure 5. 3 : 1gram

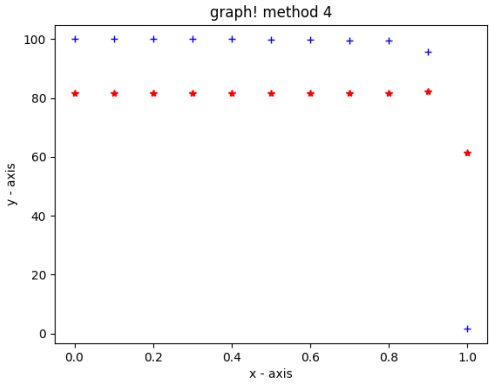

Figure 5. 4 : 2gram

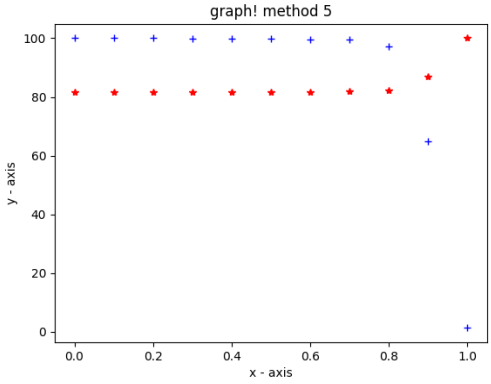

Figure 5. 5: 3gram

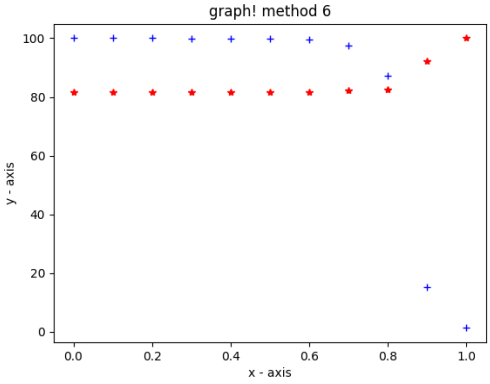

Figure 5. 6 : 1pqgr

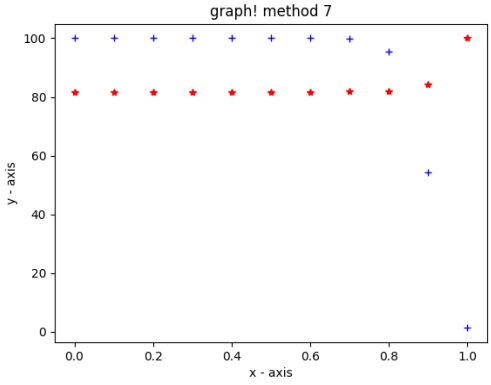

Figure 5. 7 : 2pqgr

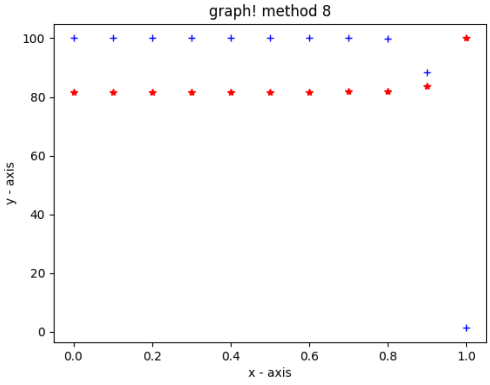

Figure 5. 8 : 3pqgr

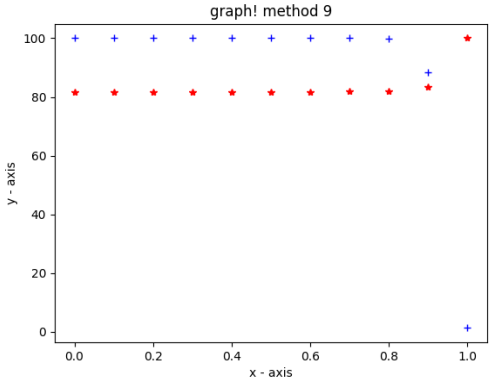

Figure 5. 9 : sgram

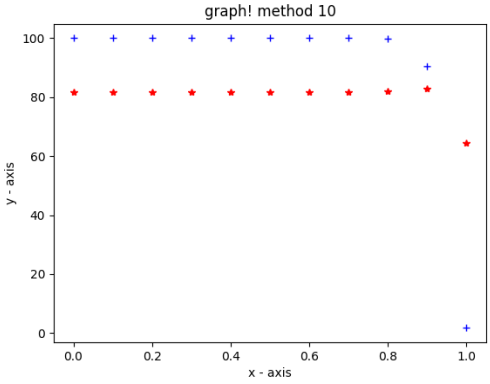

Figure 5. 10 : eDist

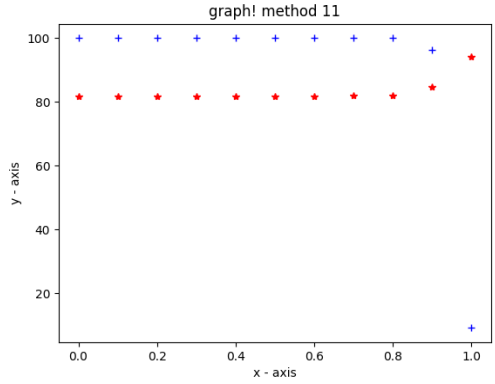

Figure 5. 11 : meDist

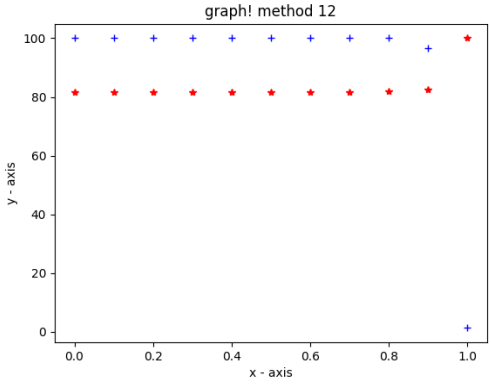

Figure 5. 12 : bDist

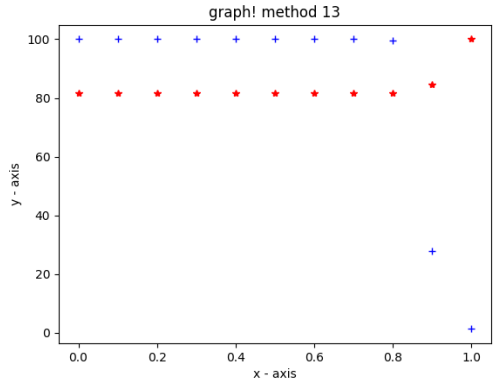

Figure 5. 13 : Editx

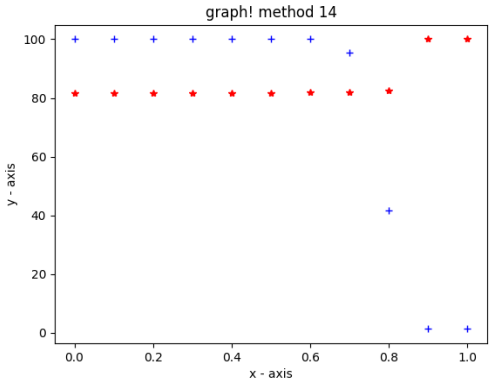

Figure 5. 14 : Segma

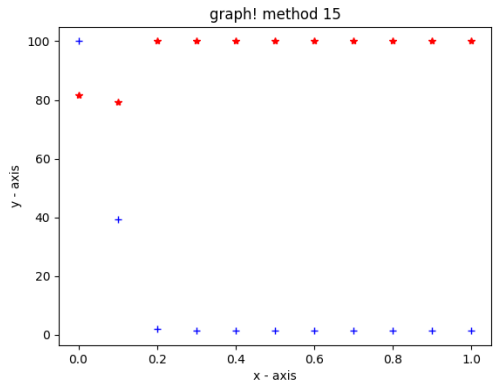

Figure 5. 15 : ComBZ

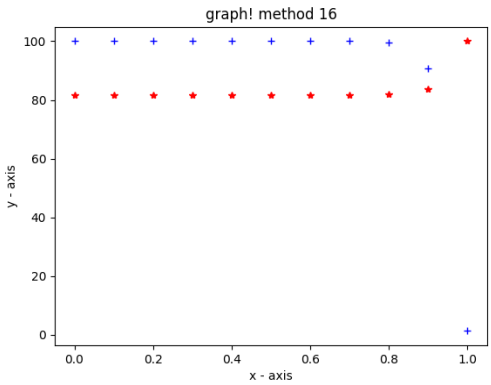

Figure 5. 16 : Comzl

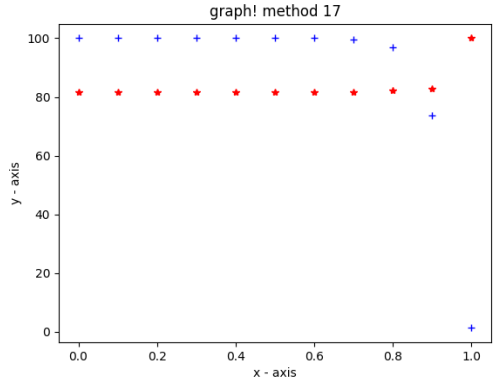

Figure 5. 17 : ComAC

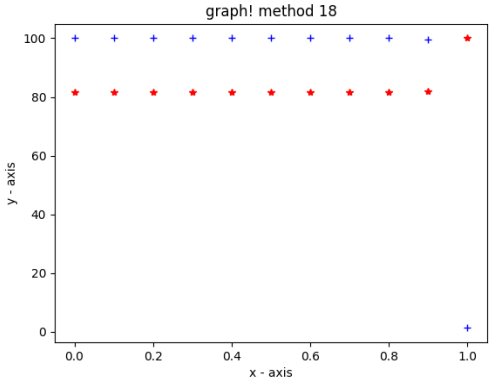

Figure 5. 18 : LCS2

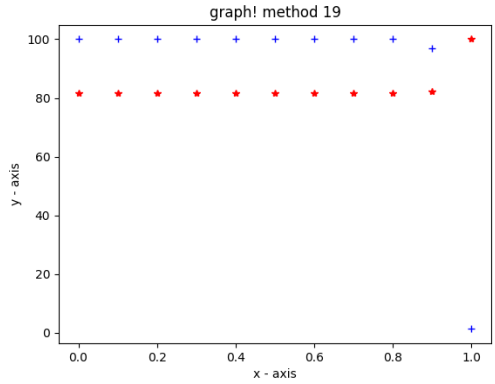

Figure 5. 19 : LCS3

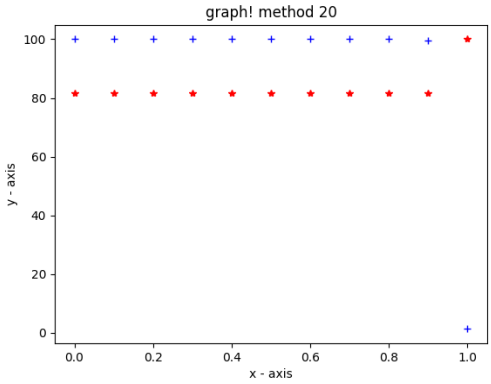

Figure 5. 20 : OLCS2

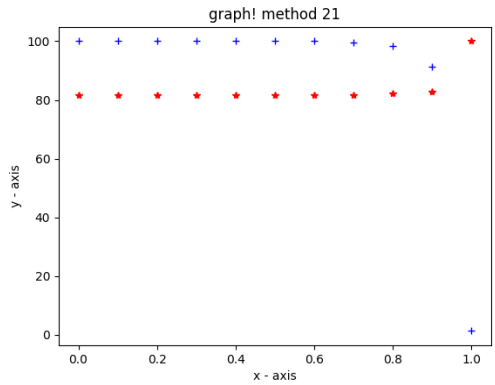

Figure 5. 21 : OLCS3

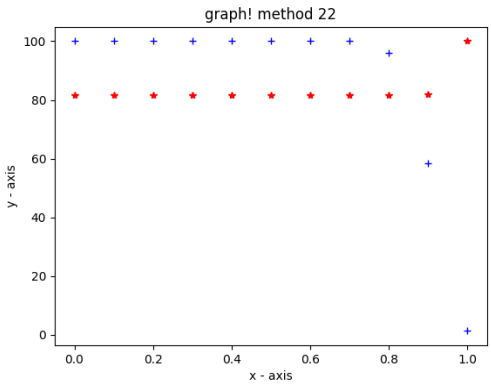

Figure 5. 22 : P-win

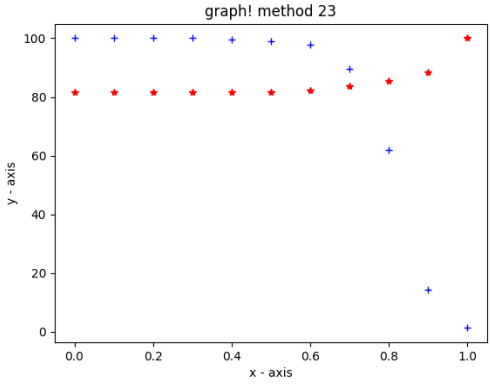

Figure 5. 23 : S-Win

Table 5.3 shows the accuracy values over all possible threshold values as it indicates the overall quality of the matching technique.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Jaro	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	80.185	19.040
J-W	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.888	81.114	64.086	19.504
1gram	81.578	81.578	81.578	81.578	81.578	81.578	81.888	81.424	74.767	30.030	19.504
2gram	81.578	81.578	81.578	81.578	81.578	81.424	81.578	81.269	81.424	79.566	18.885
3gram	81.578	81.578	81.578	81.424	81.424	81.578	81.269	81.578	80.495	63.467	19.504
1pqgr	81.578	81.578	81.578	81.424	81.578	81.578	81.424	80.804	74.303	29.876	19.504
2pqgr	81.578	81.578	81.578	81.578	81.578	81.578	81.733	81.733	79.256	54.489	19.504
3pqgr	81.578	81.578	81.578	81.578	81.578	81.578	81.733	**82.043**	81.888	76.315	19.504
sgram	81.578	81.578	81.578	81.578	81.578	81.578	81.733	**82.043**	81.888	76.315	19.504
eDist	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.888	77.089	19.040
meDis	81.578	81.578	81.578	81.578	81.578	81.578	81.733	82.043	82.043	**82.972**	25.386
bDist	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.888	80.495	19.504
Editx	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.424	36.996	19.504
Seqma	81.578	81.578	81.578	81.578	81.578	81.578	81.888	79.102	45.201	19.504	19.504
ComBZ	81.578	42.105	20.123	19.504	19.504	19.504	19.504	19.504	19.504	19.504	19.504
Comzl	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.733	78.018	19.504
ComAC	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.424	80.340	65.944	19.504
LCS2	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.733	19.504
LCS3	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.733	80.340	19.504
OLCS2	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.578	81.424	19.504
OLCS3	81.578	81.578	81.578	81.578	81.578	81.578	81.733	81.424	81.269	77.244	19.504
P-win	81.578	81.578	81.578	81.578	81.578	81.578	81.733	81.733	79.256	55.572	19.504
S-Win	81.578	81.578	81.578	81.578	81.114	80.959	80.804	77.089	60.371	28.482	19.504

Table 5. 3 : Accuracy of Matching Techniques

As it is seen in Table 5.3, Positional 3q-grams, Skip grams and Modified Edit distance techniques perform on the localized customer name data, shows the highest suitability for personal name data. The threshold to achieve best possible classification of matches and non-matches, depend on the tested data set for the mentioned three techniques are shown in Table 5.4.

Technique	Threshold with highest accuracy
Positional 3q-grams	0.7
Skip grams	0.7
Modified Edit distance	0.9

Table 5. 4 : Techniques with Highest Accuracy

5.4.2 Case Study 02 - Billing issue formed as a result of Dirty Data in Addresses of Customer Profiles

To match all records relating to the same entity, this case study focuses on the development of the name matching along with address, based on the context data collected on the tested customer profiles. The case study experiments to determine an approximate similarity measure technique/s for address attribute and to improve the number of confirmed customer accounts to be matched under one customer profile using the address attribute.

The comparison of records was done combining the results of the first use case. The name pairs that were classified as exact matches and non-matches by the matching techniques Positional 3 q-grams, Skip grams and Modified Edit distance were checked against their address attribute. And the similarities of the addresses of each customer account on the pair were computed using the string matching techniques described in Section 3.4, Chapter 3.

Matching Results

This section showcases the results of the test runs applied to all pattern matching techniques described in Section 3.4. The similarity measure of the addresses is given on the scale, 1.0 indicates an exact match and 0.0 means a non-match. And considering the results, the classification of two record pairs, considering the address attribute, under one customer profile is determined.

Table 5.5 shows the average similarity measures achieved for each of the presented techniques on the address data set. The similarity measures lie between 0.0, showcasing no matching at all and 1.0, showcasing exact matching.

Matching Technique	Similarity Measure
Jaro	0.81906
J-W	0.85774
1gram	0.804
2gram	0.70838
3gram	0.64694
1pqgr	0.70642
2pqgr	0.61666
3pqgr	0.56878
Sgram	0.68896
eDist	0.72598
meDis	0.72598
bDist	0.75754
Editx	0.7509
Seqma	0.78082
ComBZ	0.81382
Comzl	0.63564
ComAC	0.10704
LCS2	0.74398
LCS3	0.71876
OLCS2	0.82748
OLCS3	0.8053
P-win	0.8927
S-Win	0.8395
Swdis	0.67082
SYADI	0.48624
Histo	0.8972
2Ljaro	0.60236
2Ljaro	0.425

Table 5. 5 : Average Similarity Measures (Address attribute)

The matching quality of the addresses was evaluated using accuracy, based on precision and recall defined in equation (1). The following figures shows the best results achieved for each of the presented techniques based on the precision and recall. X-axis denotes the threshold values and y-axis denotes the similarity values. And red stars depict precision whereas blue pluses depict recall.

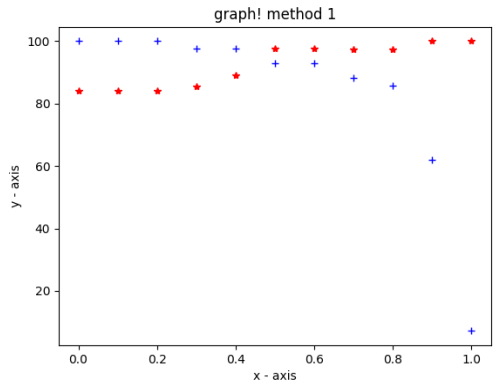

Figure 5. 24 : Jaro

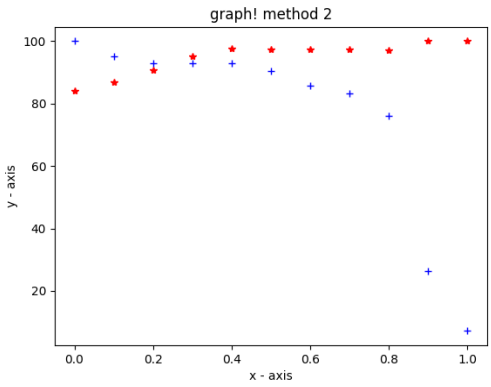

Figure 5. 25 : J-W

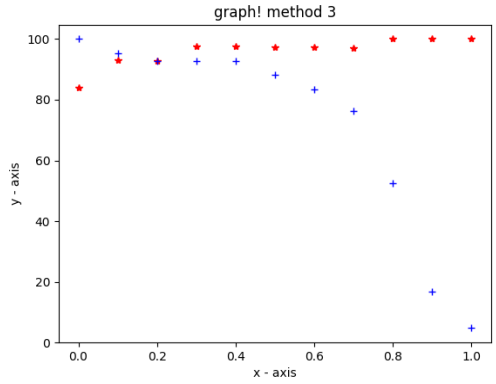

Figure 5. 26 : 1gram

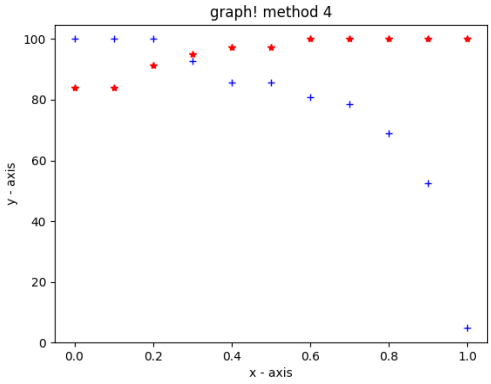

Figure 5. 27 : 2gram

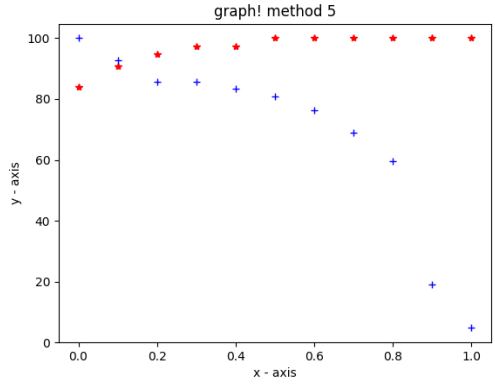

Figure 5. 28 : 3gram

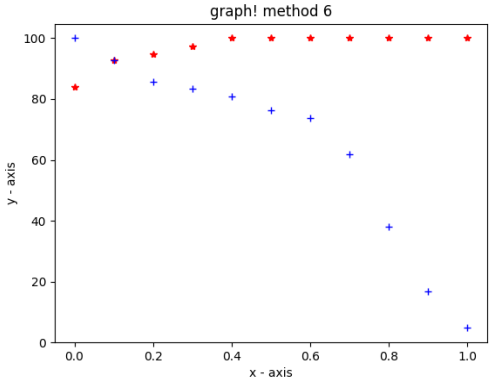

Figure 5. 29: 1pqgr

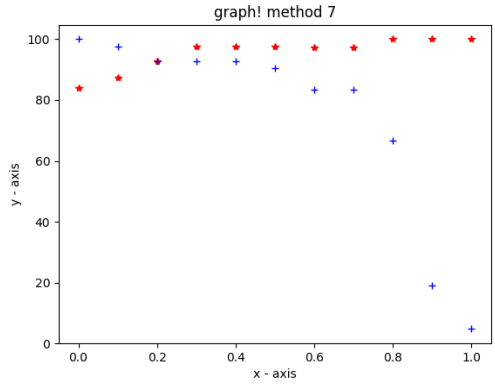

Figure 5. 30 : 2pqgr

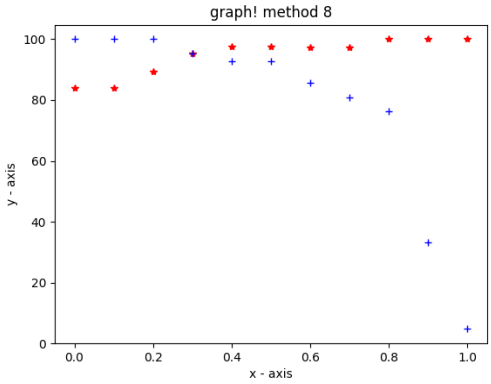

Figure 5. 31 : 3pqgr

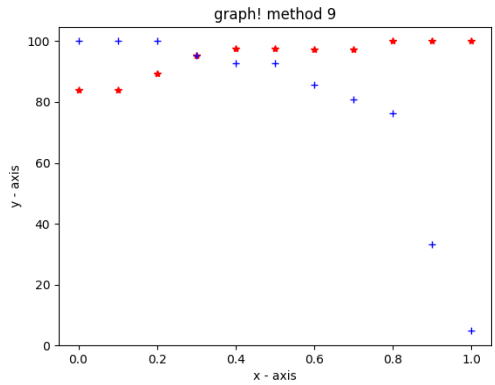

Figure 5. 32 : sgram

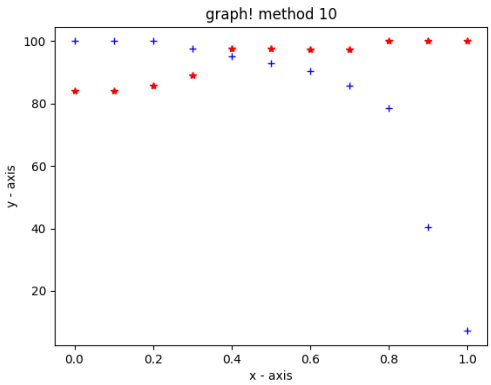

Figure 5. 33 : eDist

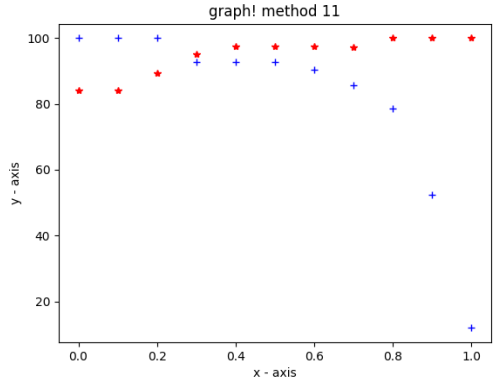

Figure 5. 34 : meDis

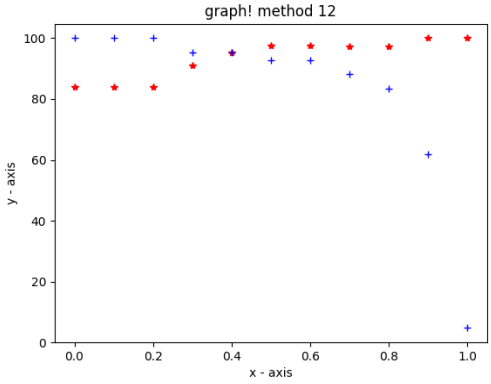

Figure 5. 35 : bDist

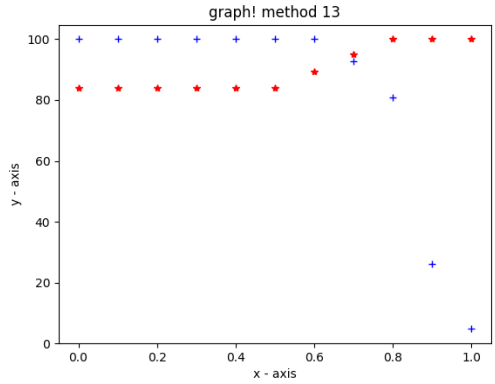

Figure 5. 36 : Editex

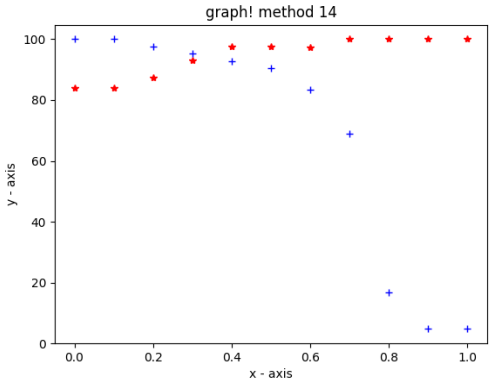

Figure 5. 37 : Seqma

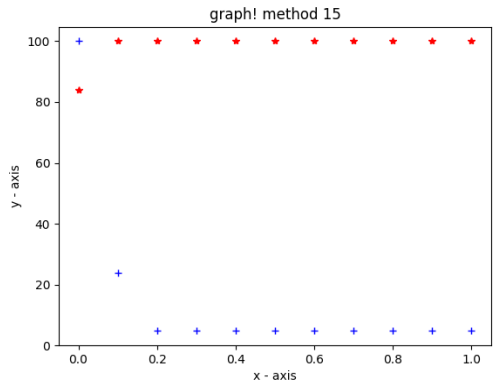

Figure 5. 38 : ComBZ

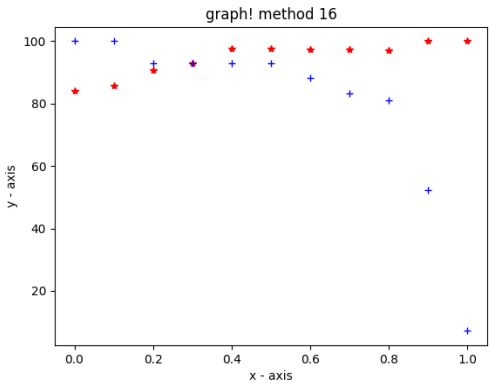

Figure 5. 39 : Comzl

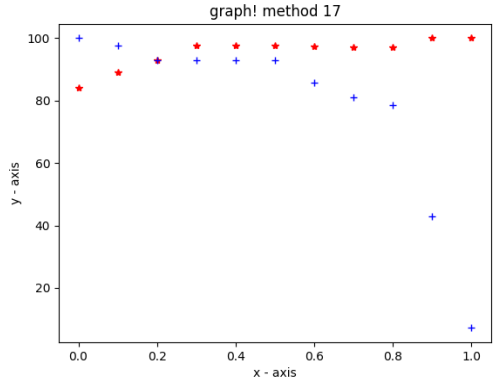

Figure 5. 40 : ComAC

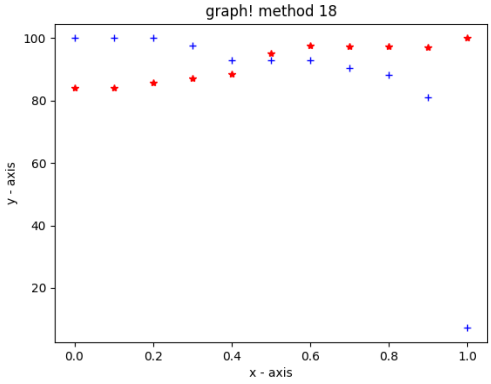

Figure 5. 41 : LCS2

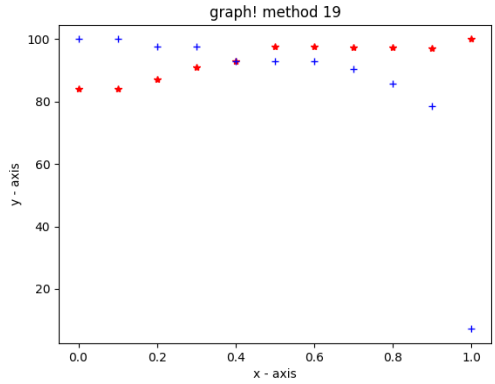

Figure 5. 42 : LCS3

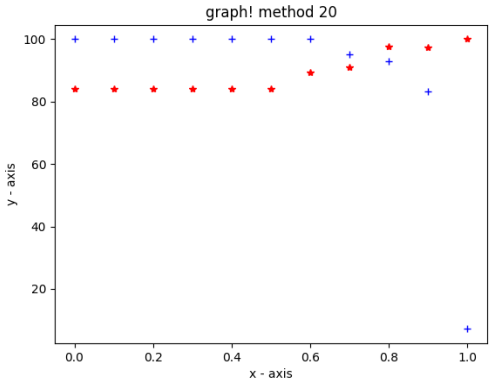

Figure 5. 43 : OLCS2

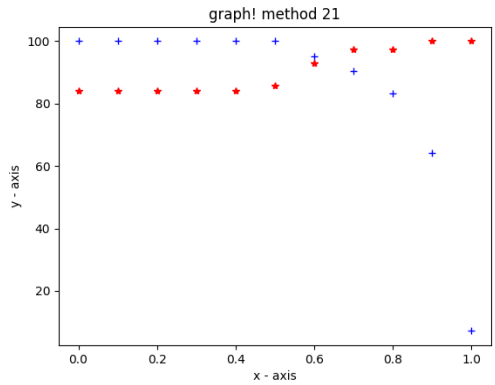

Figure 5. 44 : OLCS3

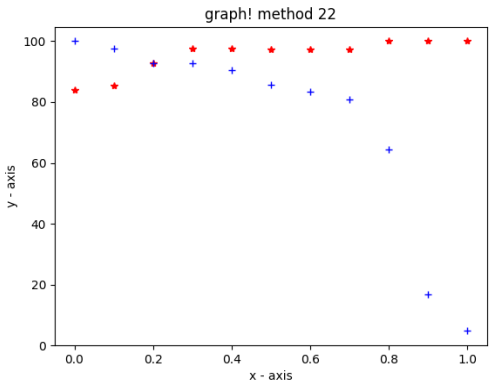

Figure 5. 45 : P-win

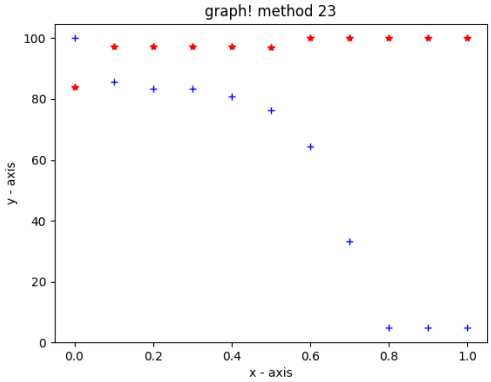

Figure 5. 46 : S-Win

Table 5.6 shows the average accuracy values over all possible threshold values of the address attribute as it indicates the overall quality of the matching technique.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Jaro	84.0	84.0	84.0	84.0	88.0	92.0	92.0	88.0	86.0	68.0	22.0
J-W	84.0	84.0	86.0	90.0	92.0	90.0	86.0	84.0	78.0	38.0	22.0
1gram	84.0	90.0	88.0	92.0	92.0	88.0	84.0	78.0	60.0	30.0	20.0
2gram	84.0	84.0	92.0	90.0	86.0	86.0	84.0	82.0	74.0	60.0	20.0
3gram	84.0	86.0	84.0	86.0	84.0	84.0	80.0	74.0	66.0	32.0	20.0
1pqgr	84.0	88.0	84.0	84.0	84.0	80.0	78.0	68.0	48.0	30.0	20.0
2pqgr	84.0	86.0	88.0	92.0	92.0	90.0	84.0	84.0	72.0	32.0	20.0
3pqgr	84.0	84.0	90.0	92.0	92.0	92.0	86.0	82.0	80.0	44.0	20.0
Sgram	84.0	84.0	90.0	92.0	92.0	92.0	86.0	82.0	80.0	44.0	20.0
eDist	84.0	84.0	86.0	88.0	94.0	92.0	90.0	86.0	82.0	50.0	22.0
meDis	84.0	84.0	90.0	90.0	92.0	92.0	90.0	86.0	82.0	60.0	26.0
bDist	84.0	84.0	86.0	88.0	**94.0**	92.0	90.0	86.0	82.0	50.0	22.0
Editx	84.0	84.0	84.0	84.0	84.0	84.0	90.0	90.0	84.0	38.0	20.0
Seqma	84.0	84.0	86.0	90.0	92.0	90.0	84.0	74.0	30.0	20.0	84.0
ComBZ	84.0	36.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
Comzl	84.0	86.0	86.0	88.0	92.0	92.0	88.0	84.0	82.0	60.0	22.0
ComAC	84.0	88.0	88.0	92.0	92.0	92.0	86.0	82.0	80.0	52.0	22.0
LCS2	84.0	84.0	86.0	86.0	84.0	90.0	92.0	90.0	88.0	82.0	22.0
LCS3	84.0	84.0	86.0	90.0	88.0	92.0	92.0	90.0	86.0	80.0	22.0
OLCS2	84.0	84.0	84.0	84.0	84.0	84.0	90.0	88.0	92.0	84.0	22.0
OLCS3	84.0	84.0	84.0	84.0	84.0	86.0	90.0	90.0	84.0	70.0	22.0
P-win	84.0	84.0	88.0	92.0	90.0	86.0	84.0	82.0	70.0	30.0	20.0
S-Win	84.0	86.0	84.0	84.0	82.0	78.0	70.0	44.0	20.0	20.0	20.0

Table 5. 6 : Accuracy of Matching Techniques

Table 5.6 depicts Edit Distance technique as the matching technique that is more suitable to measure similarity of address attribute. The threshold to achieve best possible classification of matches and non-matches, depend on the tested data set for the mentioned technique is presented in Table 5.7.

Technique	Threshold with highest accuracy
Edit Distance	0.4

Table 5. 7 : Technique with Highest Accuracy

The results of the name matching and address matching have been used to further improve the matching quality of the matched and non-matched accounts. The relationship of full name and address of all individual customer accounts has been considered here and it is categorized into four categories to measure the accuracy of the relationship; matched name pair with same address; matched name pair with different addresses; non-matched name pair with same address; and non-matched name pair with different addresses. The accounts matched by name under one profile with the same address were taken into consideration as another step to confirm the merge of the different account. Table 5.8 shows the accuracy values of the best suited name matching techniques along with the best suited address matching technique.

Name and Address Matching Technique	Accuracy
Positional 3 q-grams, Edit Distance	0.94
Skip grams, Edit Distance	0.92
Modified Edit distance, Edit Distance	0.88

Table 5. 8 : Techniques with Highest Accuracy (Name+Address)

5.4.3 Case Study 03 - Addition of other key fields apart from names and address in order to measure the changes in accuracy levels to rectify CRM and Marketing issues formed as a result of Dirty Data in Customer Profiles

In matching accounts relating to one entity of customer profile, the case study focuses on taking more customer data attributes into account. It focuses on email address attribute and aims to provide more accurate information when taking the decisions on final linkage status. The case study experiments to determine a suitable approximate similarity measure technique/s to compare email attribute and to improve the matching of two customer accounts to one or differed accounts.

The comparison of records was done combining the results of the first and second use case. The highest matching quality in combined name and address attribute techniques were shown in; Positional 3 q-grams and Edit distance; Skip grams and Edit and it was evaluated integrating email address attribute.

Matching Results

The results of the tests run on the email attribute to all string matching techniques described in Section 3.4, Chapter 3 are represented below. The similarity measure of the email addresses is given on the scale from 0 to 1, 0 indicating a non-match and 1 indicating an exact match. Table 5.9 shows the average similarity measure achieved for each of the similarity measure technique.

Matching Technique	Similarity Measure
Jaro	0.952917
J-W	0.966917
1gram	0.956333
2gram	0.91625
3gram	0.87275
1pqgr	0.956333
2pqgr	0.91625
3pqgr	0.87275

Sgram	0.90425
eDist	0.937833
meDis	0.937833
bDist	0.942167
Editx	0.94625
Seqma	0.956333
ComBZ	0.900667
Comzl	0.817333
ComAC	0.234
LCS2	0.956333
LCS3	0.939583
OLCS2	0.981917
OLCS3	0.972583
P-win	0.966917
S-Win	0.966917
Swdis	0.895
SYADI	0.784333
Histo	0.980583
2Ljaro	0.166667
2Ljaro	0.91625

Table 5. 9 : Average Similarity Measures (Email Address attribute)

The accuracy of the results gained by evaluating the similarity measures with the email attribute, defined in equation (1), is presented by the following figures for each of the techniques presented in Section 3.4, Chapter 3. X-axis denotes the threshold values and y-axis denotes the similarity values. And red stars depict precision whereas blue pluses depict recall.

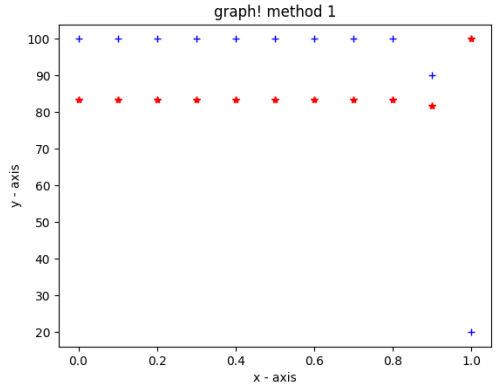

Figure 5. 47 : Jaro

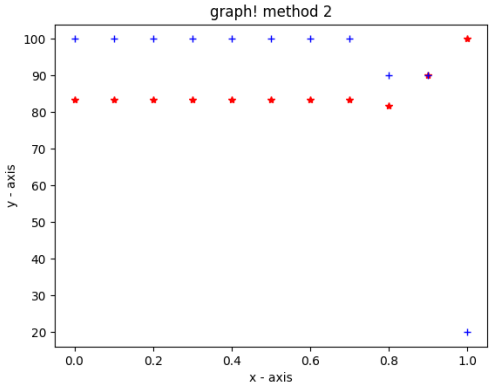

Figure 5. 48 : J-W

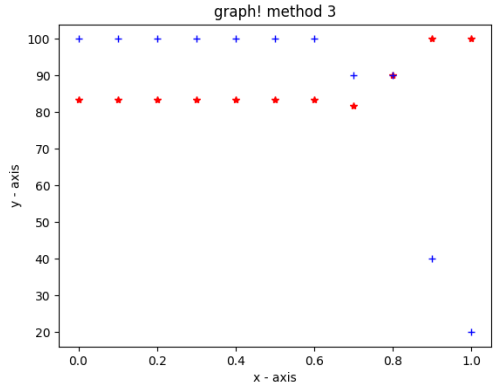

Figure 5. 49 : 1gram

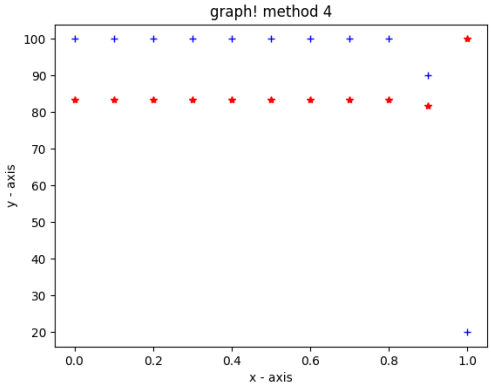

Figure 5. 50 : 2gram

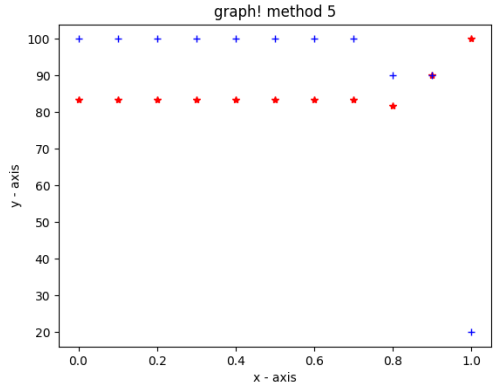

Figure 5. 51 : 3gram

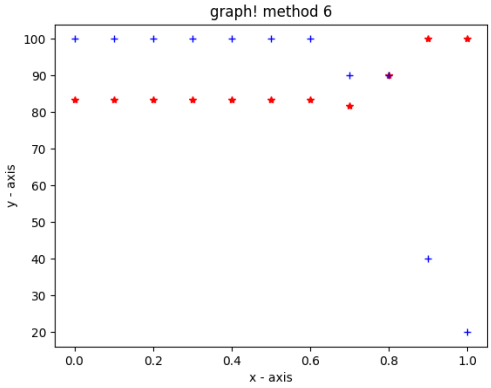

Figure 5. 52 : 1pqgr

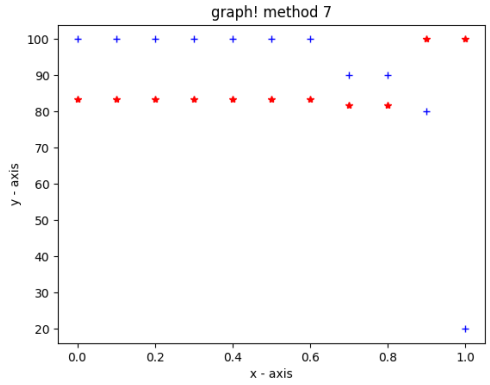

Figure 5. 53 : 2pqgr

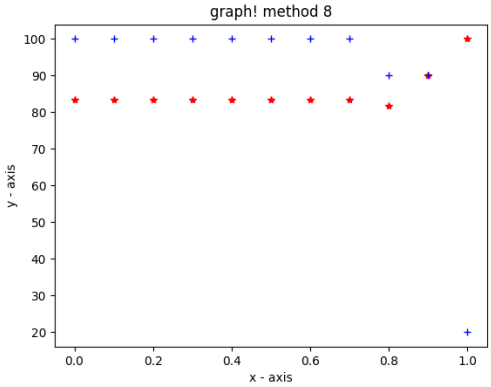

Figure 5. 54 : 3pqgr

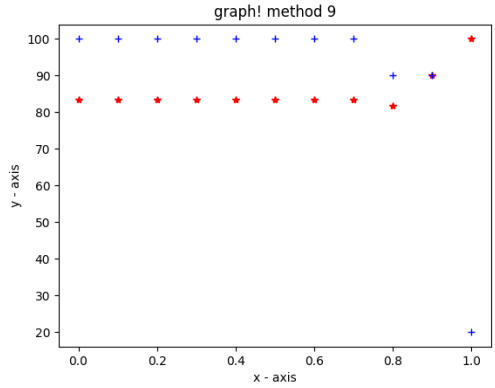

Figure 5. 55 : Sgram

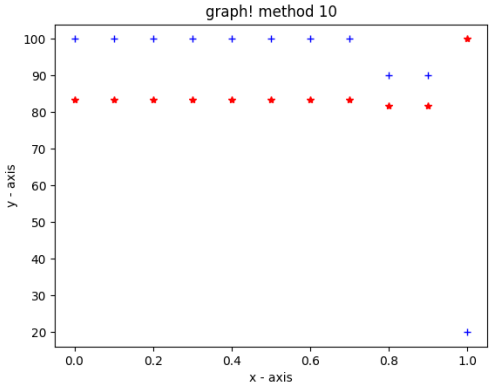

Figure 5. 56 : eDist

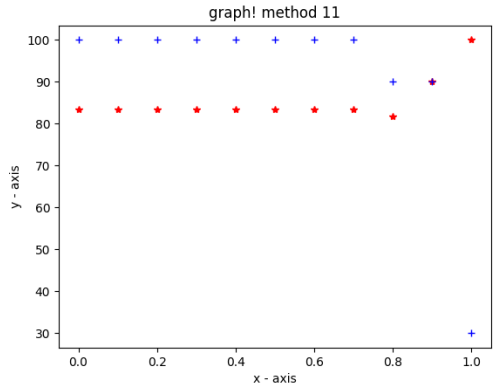

Figure 5. 57 : meDis

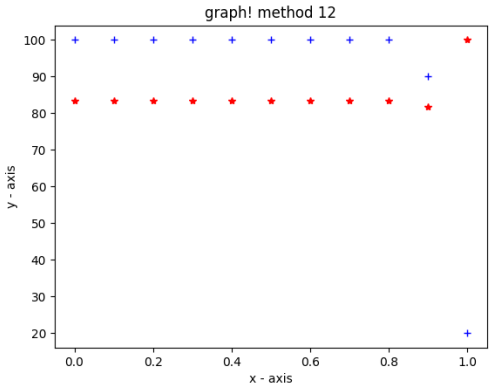

Figure 5. 58 : bDis

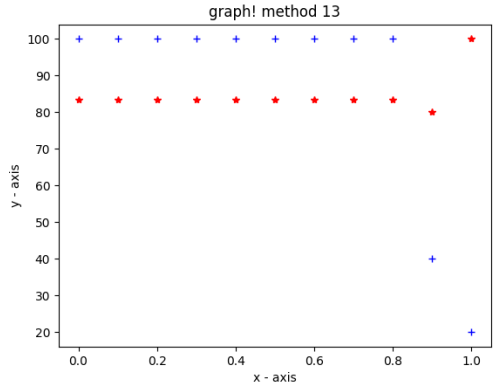

Figure 5. 59 : Editex

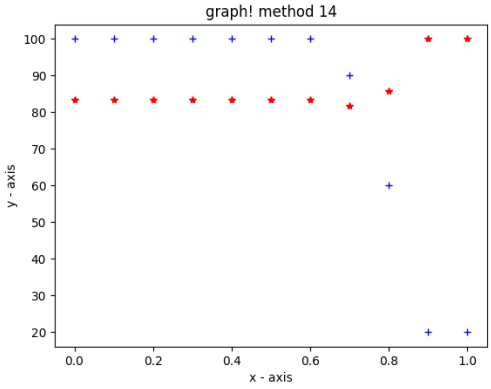

Figure 5. 60 : Seqma

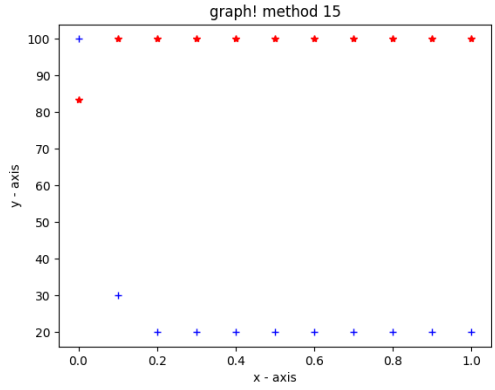

Figure 5. 61 : ComBZ

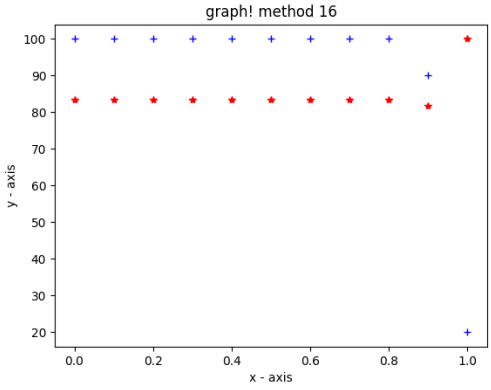

Figure 5. 62 : Comzl

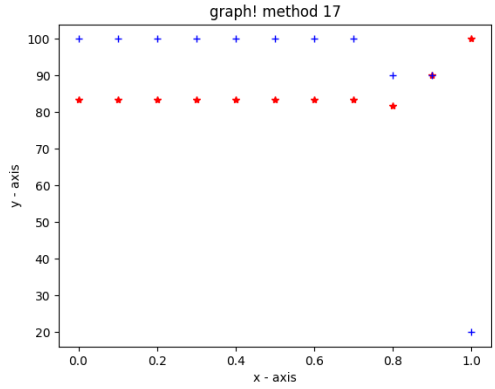

Figure 5. 63 : ComAC

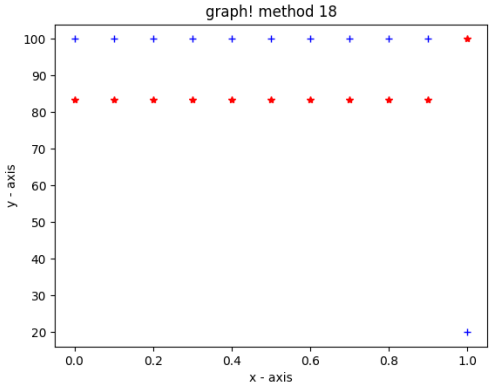

Figure 5. 64 : LCS2

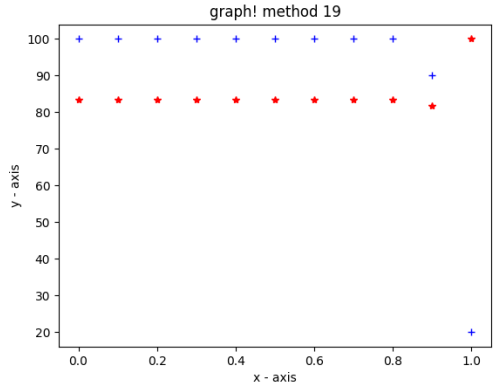

Figure 5. 65 : LCS3

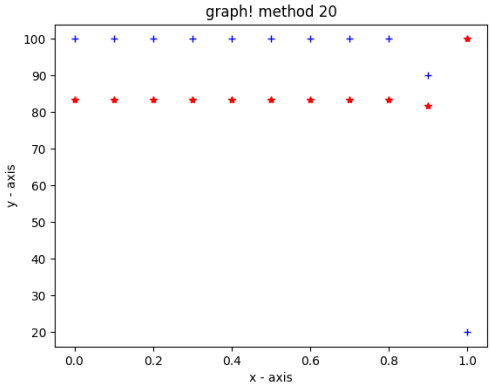

Figure 5. 66 : OLCS2

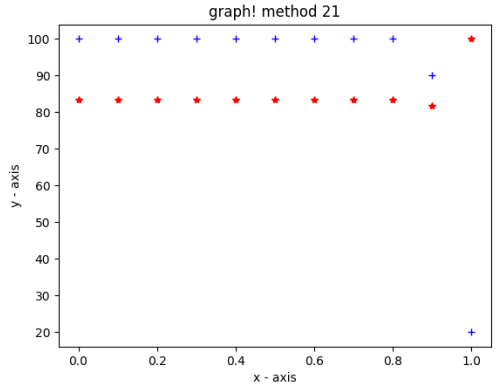

Figure 5. 67 : OLCS3

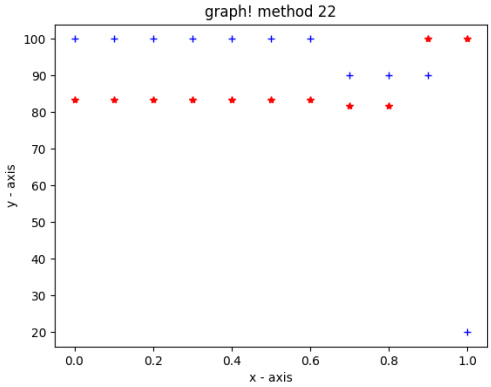

Figure 5. 68 : P-win

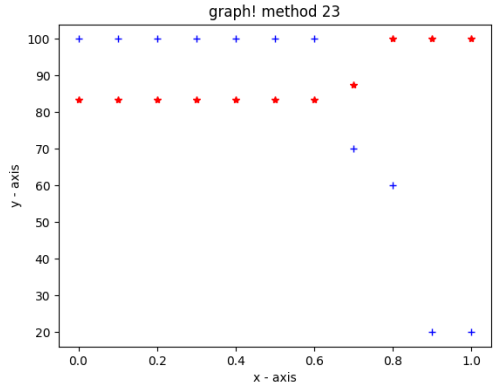

Figure 5. 69 : S-Win

Table 5.10 shows the average accuracy values over all possible threshold values of the address attribute as it indicates the overall quality of the matching technique.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Jaro	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	33.3
J-W	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	33.3
1gram	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	83.3	33.3
2gram	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	83.3	50.0	83.3
3gram	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	33.3
1pqgr	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	83.3	33.3
2pqgr	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	83.3	50.0	33.3
3pqgr	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	75.0	83.3	20.0
Sgram	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	83.3	33.3
eDist	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	83.3	33.3
meDis	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	75.0	33.3
bDist	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	83.3	41.6
Editx	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	33.3
Seqma	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	41.6	33.3
ComBZ	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	58.3	33.3	33.3
Comzl	83.3	41.6	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
ComAC	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	33.3
LCS2	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	83.3	33.3
LCS3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	33.3
OLCS2	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	33.3
OLCS3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	33.3
P-win	83.3	83.3	83.3	83.3	83.3	83.3	83.3	75.0	75.0	**91.6**	33.3
S-Win	83.3	83.3	83.3	83.3	83.3	83.3	83.3	66.6	66.6	33.3	33.3

Table 5. 10 : Accuracy of Matching Techniques

Table 5.10 depicts Permuted Winkler technique as the matching technique that is more suitable to measure similarity of email address attribute. The threshold to achieve best possible classification of matches and non-matches, depend on the tested data set for the mentioned technique is presented in Table 5.11.

Technique	Threshold with highest accuracy
Permuted Winkler	0.9

Table 5. 11 : Technique with Highest Accuracy

To further improve the matching quality and linkage status of two customer accounts, the results of the name matching and address matching have been used along with the email address attribute. The relationship of full name and address along with email address of all individual customer accounts has been considered here and it is categorized into four categories to measure the accuracy of the relationship; matched name and address pair with same emails; matched name and address pair with different emails; non-matched name pair and address pair with same email; and non-matched name and address pair with different emails. The accounts matched by name and address under one profile with the email address were taken into consideration as another step to confirm the merge of a same or different account. Table 5.12 shows the accuracy values of the best suited name and address matching techniques along with the best suited email matching technique.

Name, Address and Email Matching Technique	Accuracy
Positional 3 q-grams, Edit Distance, Permuted Winkler	0.98
Skip grams, Edit Distance, Permuted Winkler	0.95

Table 5. 12 : Techniques with Highest Accuracy (Name+Address+Email)

In the comparison of customer records, appropriate approximate string matching techniques were chosen for each attribute. Table 5.13 shows the attributes and the similarity measure techniques achieved with the highest accuracy level to the dataset, to compute the similarities. For any given record pair, accuracy of the sum of attribute-wise similarity techniques reflects the overall matching status of the compared customer accounts. And it is observed that the accuracy increases with the addition of more attributes.

Attribute	Similarity Measure Technique
Full Name	Positional 3 q-grams/ Skip grams/ Modified Edit distance
Address	Edit distance
Email Address	Permuted Winkler

Table 5. 13 : Similarity Measure Techniques for the Attributes

5.5 Summary

This chapter elucidates the analysis and experimentations of the case studies. It mainly comprehends of three main sections as analysis of the preliminary interviews, effect of dirty data on customer profiles and experiments. Preliminary interviews showcased the effect of dirty data in telecommunication industry and experiments extracted similarity measure techniques that are suitable for attributes of customer data. Below chapter concludes the findings of the research.

Chapter 6 - Conclusions

6.1 Introduction

This chapter includes a review of the research aims and objectives, research problem, limitations of the current work and implications for further research.

6.2 Conclusions about Research Questions (aims/objectives)

The main aim of research was to identify the effect of dirty data on customer data to understand the repercussions of its impact and to control and mitigate the effects over decision making in Telecommunication industry of Sri Lanka. After the analysis and experiments, following conclusions were extracted. It was identified that the most crucial impact lies on customer related areas. Customer profiling issue emerged as a result of dirty data infused in personal names, billing issue emerged as a result of dirty data infused in addresses of customer profiles, CRM and marketing operational failures and customer segmentation due to duplicated customer profiles were identified as sub-problems.

Experimental results on real world telecommunication customer dataset have shown that Positional 3 q-grams, Skip grams and Modified Edit distance as the similarity measure techniques suitable for personal name data, Edit distance for address data and Permuted Winkler for email address data resulting the highest suitability and accuracy. And it is observed that the accuracy of similarity matching increases with the insertion of more attributes that allows the classification of record pairs to matches or non-matches.

6.3 Discussions and Recommendations

The analysis of interviews made it realized, present methods that are used by the domain experts of Sri Lankan telecommunication Industry, do not follow a keen analysis to discover dirty data in attribute level. Present procedures within the telecommunication industry cater to existing

problems through available tools without considering its accuracy and suitability. The techniques suggested can be used to increase the confidence level of clerical review process, which can be used as a dirty data controlling and mitigating mechanism, which will verify the decision and reduce the time consumed. Further, as a controlling strategy, a portfolio managed to a single customer can be recommended. In which, all the accounts of a customer will be managed through a portfolio, where both company and customer are aware of.

When selecting a threshold that results in optimum matching quality, it was observed a dramatic drop of accuracy after a certain threshold level. And it is observed that the threshold values vary between datasets, since the similarity measure technique depend on the nature of the dataset. Therefore, setting a threshold for a particular dataset without examination is found purposeless.

6.4 Limitations

It is observed that the quality of the similarity measure techniques varies depend on the nature of the data set and nature of errors. Therefore, the results produced through the experiments are not beneficial for evaluating relative quality of techniques.

6.5 Implications for Further Research

The applicability of the proposed techniques and the combination of techniques to datasets in similar domains particularly healthcare and tourism industries in Sri Lanka can be examined. Further, how far the proposed method effects the clerical review process in telecommunication industry of Sri Lanka can be explored. Also graph method has been identified to explore the relationship between individual customer profiles and household identification linkage.

References

[1] R. Marsh, "Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management", *Journal of Database Marketing & Customer Strategy Management*, vol. 12, no. 2, pp. 105-112, 2005.

[2] Maletic, J. and Marcus, A. (2017). Data Cleansing: Beyond Integrity Analysis. [online] http://www.academia.edu. Available at: http://www.academia.edu/2722580/Data_cleansing_Beyond_integrity_analysis [Accessed 17 Jun. 2017].

[3] A. Dongre, "Data Quality and Integrity Management for Telecom Operators", *SSRN Electronic Journal*, 2014.

[4] P. Christen and K. Goiser, "Quality and Complexity Measures for Data Linkage and Deduplication", *Quality Measures in Data Mining*, pp. 127-151, 2007.

[5] W. KIM and B. CHOI, *A Taxonomy of Dirty Data*. Germany: Data Mining and Knowledge Discovery, 2003, pp. 7, 81–99.

[6] Heiko Müller, Johann-Christoph Freytag "Problems, Methods, and Challenges in Comprehensive Data Cleansing", *HUB-IB-164, Humboldt University Berlin*, 2003.

[7] T. Gschwandtner, J. Gärtner, W. Aigner and S. Miksch, "A Taxonomy of Dirty Time-Oriented Data", *Lecture Notes in Computer Science*, pp. 58-72, 2012.

[8] R. Singh and D. Singh, "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing", *IJCSI International Journal of Computer Science Issues*, vol. 7, no. 3, 2010.

[9] M. A. Her Andez and S. J. Stolfo, “Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem,” Data Min. Knowl. Discov., vol. 2, pp. 9–37, 1998.

[10] J. Vosburg and A. Kumar, “Industrial Management & Data Systems Emerald Article : Managing dirty data in organizations using ERP : lessons from a case study Managing dirty data in organizations using ERP : lessons from a case study,” 2001.

[11] V. Raman and J. Hellerstein, "Potter ’ s Wheel : An Interactive Data Cleaning System", *VLDB*, 2001.

[12] E. Rahm and D. H.H, "Data Cleaning: Problems and Current Approaches", *IEEE Techn. Bulletin on Data Engineering*, 2000.

[13] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in,” vol. 17, no. 3, pp. 37–54, 1996.

[14] Haug, Anders, Frederik Zachariassen, and Dennis Liempd. "The Costs Of Poor Data Quality". Journal of industrial engineering and management, vol. 4, no. 3,2011.

[15] T. Milo and S. Zohar, "Using Schema Matching to Simplify Heterogeneous Data Translation", *VLDB '98 Proceedings of the 24rd International Conference on Very Large Data Bases*, pp. Pages 122-133, 1998.

[[16] E. Rahm and H. Do, "Data cleaning: Problems and current approaches", *IEEE Data Eng. Bull.*, 2000.

[17] P. Christen, "Febrl -", *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, 2008.

[18] P. Christen, "A Comparison of Personal Name Matching: Techniques and Practical Issues", *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, 2006.

[19] R. GONG and T. K.Y. CHAN, "Syllable Alignment: A Novel Model for Phonetic String Search", *IEICE TRANSACTIONS on Information and Systems*, vol. 89-, no. 1, pp. .332-339, 2006.

[20] "Non adjacent Diagrams ImproveMatching of Cross Lingual Spelling Variant", *Proceedings of the 10th International Symposium,*, pp. 252-265, 2003.

Appendix A: Questionnaire

Name: Click here to enter text.

Occupation: Click here to enter text.

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☐	☐	☐
Geo location mapping issue	☐	☐	☐	☐	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☐	☐	☐
NIC mapped issues	☐	☐	☐	☐	☐	☐
Single view of customer	☐	☐	☐	☐	☐	☐
CLM failures	☐	☐	☐	☐	☐	☐
Governance policy	☐	☐	☐	☐	☐	☐

Questionnaire

Name: Sandra De Zoysa

Occupation: Group Chief Customer Officer

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☐	☐	☒
Geo location mapping issue	☐	☐	☐	☒	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☐	☐	☒
NIC mapped issues	☐	☒	☐	☐	☐	☐
Single view of customer	☐	☐	☐	☐	☒	☐
CLM failures	☐	☒	☐	☐	☐	☐
Governance policy	☐	☐	☐	☒	☐	☐

Questionnaire

Name: Dhanya Herath Gunaratne

Occupation: General Manager Operations at MAS Legato

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☒	☐	☐
Geo location mapping issue	☐	☐	☐	☐	☒	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☐	☐	☒
NIC mapped issues	☐	☐	☐	☐	☐	☒
Single view of customer	☐	☐	☒	☐	☐	☐
CLM failures	☒	☐	☐	☐	☐	☐
Governance policy	☐	☒	☐	☐	☐	☐

Questionnaire

Name: Eranda Adikari

Occupation: Senior Manager - BI Analytics at Dialog Telekom

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all 0	1	2	3	4	**Very Important** 5
Customer profiling issue	☐	☐	☐	☐	☐	☒
Geo location mapping issue	☐	☐	☐	☒	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☐	☐	☒
NIC mapped issues	☐	☒	☐	☐	☐	☐
Single view of customer	☐	☐	☐	☐	☒	☐
CLM failures	☐	☒	☐	☐	☐	☐
Governance policy	☐	☐	☐	☒	☐	☐

Questionnaire

Name: Maneesha Jinadasa

Occupation: Deputy Chief Officer at Sri Lanka Telecom

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☐	☐	☒
Geo location mapping issue	☐	☐	☐	☐	☒	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☒	☐	☐
NIC mapped issues	☐	☐	☐	☒	☐	☐
Single view of customer	☐	☐	☐	☒	☐	☐
CLM failures	☐	☒	☐	☐	☐	☐
Governance policy	☐	☒	☐	☐	☐	☐

Questionnaire

Name: Pubudu Chinthaka

Occupation: Assistant Manager at Dialog Axiata PLC

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☒	☐	☐
Geo location mapping issue	☐	☐	☐	☒	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☒	☐	☐
NIC mapped issues	☐	☐	☐	☒	☐	☐
Single view of customer	☐	☐	☐	☐	☐	☒
CLM failures	☐	☒	☐	☐	☐	☐
Governance policy	☐	☒	☐	☐	☐	☐

Questionnaire

Name: Rohitha Somaratne

Occupation: General Manager IT Strategy & Planning at Sri Lanka Telecom

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☒	☐	☐
Geo location mapping issue	☐	☐	☐	☒	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☐	☐	☒
NIC mapped issues	☐	☐	☐	☒	☐	☐
Single view of customer	☐	☐	☐	☒	☐	☐
CLM failures	☐	☒	☐	☐	☐	☐
Governance policy	☐	☒	☐	☐	☐	☐

Questionnaire

Name: Shanka Rabel
Occupation: Senior Director - Solutions, Virtusa (Pvt) ltd

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☐	☐	☒
Geo location mapping issue	☐	☐	☐	☒	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☐	☒	☐
NIC mapped issues	☐	☐	☐	☐	☒	☐
Single view of customer	☐	☐	☐	☐	☐	☒
CLM failures	☐	☒	☐	☐	☐	☐
Governance policy	☐	☐	☒	☐	☐	☐

Questionnaire

Name: Waruni Algama

Occupation: Head - CE, Group Loyalty & CRM at Dialog Axiata PLC

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☐	☐	☒
Geo location mapping issue	☐	☐	☐	☒	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☒	☐	☐
NIC mapped issues	☐	☐	☐	☒	☐	☐
Single view of customer	☐	☐	☐	☒	☐	☐
CLM failures	☐	☐	☐	☒	☐	☐
Governance policy	☐	☒	☐	☐	☐	☐

Questionnaire

Name: Chaminda Ranasinghe

Occupation: Chief Executive Officer at IdeaHub (Pvt) Ltd

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☐	☒	☐
Geo location mapping issue	☐	☐	☐	☒	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☐	☐	☒
NIC mapped issues	☐	☐	☐	☐	☒	☐
Single view of customer	☐	☐	☐	☐	☐	☒
CLM failures	☐	☒	☒	☐	☐	☐
Governance policy	☐	☐	☐	☒	☐	☐

Questionnaire

Name: Sajeewani De Zoysa

Occupation: PA to Group Chief Customer Officer / GCCO's Office

Please spare few minutes of your valuable time to answer this simple questionnaire. This is a follow up questionnaire based on the discussions we had on the Effect of Dirty Data on telecommunication industry on 6th of March 2017. The cases mentioned below are derived from the summarization of the interviews carried out with different personnel responsible for business decision making.

Please rank the following in order of importance from 0 to 5 where **0 is not important at all** and **5 very important** to the business in terms of the effect of dirty data on business.

	Not Important at all **0**	**1**	**2**	**3**	**4**	**Very Important** **5**
Customer profiling issue	☐	☐	☐	☐	☒	☐
Geo location mapping issue	☐	☐	☐	☒	☐	☐
Consequences of not maintaining proper hierarchy with regard to company profiles and SBUs	☐	☐	☐	☒	☐	☐
NIC mapped issues	☐	☐	☒	☐	☐	☐
Single view of customer	☐	☐	☐	☒	☐	☐
CLM failures	☐	☐	☐	☒	☐	☐
Governance policy	☐	☐	☒	☐	☐	☐

Appendix B: Interview Guide

1. What are the main problems face in customer relationship management?

2. What type issues are taken in relation to customers?

3. What type of dirty data found in Tele communication?

4. Do consider customer profiling as an important?

5. What type of customer profiling issues is commonly found in organization?

6. What is the impact of those incidents separately to revenue cost components and to decision making?

7. What are the current precautions that have been taken to mitigate the above mentioned situation?

8. When overall cost is considered what kind of percentage can be given to billing cost?

9. What are the main segmentations practice within the organization?

10. Why do you consider segmentation as a significant factor?

11. What are the impacts of inaccurate segmentations to the organization?

12. What type relationship does segmentation have with market expansion and retention?

13. What type of cost components are attached with segmentation?

14. What is the exact difference between household and family segmentation?

15. Is there any significance when the customers are divided by family and household?

16. With correct segmentation how business can achieve profitability?

17. What types of techniques are used to mitigate the above problems specified?

18. How far the decision making has got improved through technologies adopted?

19. Do you consider the suitability of the techniques when those are adapted to Business scenarios?

20. What type of a competitive advantage can be gained through if above mentioned situations are controlled?

21. What is a customer overview?

22. Does it have direct relationship with revenue?

23. How data quality would enhance the customer experience management?

24. When present mitigating techniques are considered, how many employees are being used to processes such as clerical review?

25. What is the loss percentage hit taken when those decisions are incorrect?