S	
E1	
E2	
For Office Use Only	

Masters Project Final Report

(MCS)

2019

Project Title	Identify User behavior on digital devices by using digital footprints
Student Name	N.NITHIANANTHAN
Registration No. & Index No.	2014/mcs/054
Supervisor's Name	Dr. M D R N Dayaratne

For Office Use ONLY

Identify User behavior on digital devices by using digital footprints

A dissertation submitted for the Degree of Master of Computer Science

N.NITHIANANTHAN

University of Colombo School of Computing

2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: N NITHIANANTHAN

Registration Number: 2014/mcs/054

Index Number: 14440549

_____

Signature: Date:

This is to certify that this thesis is based on the work of **Ms. N. Nithiananthan** under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. M D R N Dayaratne

_____

Signature: Date:

Acknowledgements

This thesis is the result of me being fortunate to have the unconditional assistance of several people who have been extremely supportive in various ways. First and foremost I would like offer my humble gratitude to Dr. M D R N Dayaratne, my supervisor, for their tremendous encouragement, support and the guidance given throughout this research.

I would like to sincerely thank all the lecturers at the University of Colombo School of Computing for their valuable advices and comments given at various stages of this research. Without your support, I could not have completed this research with success.

I like to thank all my dear friends who were there around me, with best of their encouragement, suggestions and support throughout this research.

Finally, I would like to express my heartfelt thanks towards my family for their support and encouragement through the many days and nights dedicated to the completion of this research.

Abstract

Now-a-days most of our time is spent using some form of digital technology such as search engines, news portals, social media websites or applications. Our presence on digital devices makes us engaged most of the time and leads us to become oblivious of our important work, resulting in a form of procrastination that decreases our productivity significantly. Some desktop and mobile applications have recently emerged to counter the problem by introducing various means of self-tracking to reduce the wasting of time and engage in productive activities. However, already available Soft wares/Systems suffer several limitations in terms of being static or providing a limited view of actions using one aspect only. To promote self-awareness that helps bring positive changes in individual's performance, there is a need to present the data in a more convincing ways, bringing interaction to it and present the same data in different ways using both temporal and categorical dimensions.

Framework is described to collects and processes the digital foot prints and creates a user behavior model to extract valuable and interesting temporal and categorical patterns regarding user behavior and interests. To discover the valuable behavior patterns from the individual's data, different web usage mining techniques have been used. Finally, demonstrate interactive visualizations for the analysis and monitoring of behavior patterns with the goal of providing the individual with detailed understanding of his/her behavior.

Table of Contents

Chapter 1

Introduction

Quantified self-movement incorporates digital technology to acquire data on various aspects of an individual's life with an aim to improve self-awareness, make informed decisions and human performance. People want to be self-aware, self-knowledgeable in order to improve their performance and outcomes. Today, technology logs almost everything with the aim to measure all aspects of our daily lives and improve the efficiency of the life. While using digital services, individuals leave behind traces of their activities that offer an opportunity to gain insights about themselves, their interests and their behavior.

Web usage mining is the major research area in data mining that facilitates to predict the individuals browsing behaviors and infer their interests by analyzing the behavior patterns. It consists of three phases: preprocessing, pattern discovery and pattern analysis. Preprocessing is required to convert the raw data into a meaningful form useful for efficient processing. Pattern discovery includes techniques to extract the pattern and encompasses statistical analysis, sequential pattern mining, path analysis, association rule mining, classification, and clustering [1]. For analysis of patterns, visualization allows to understand and analyze the patterns in an intuitive way. There are many information visualization techniques that have been developed over the last few years that can deal with wide range of data [2].

1.1 Problem Description

Life has become so much fast and busy these days that even we do not have time to pay attention to our true selves. The disease of being busy is spiritually destructive to our health and well-being leading us towards stress, depression, and anxiety. Many people waste time on activities that keep them busy but not productive. They spend most of their time in surfing the Web without even noticing how much time has been wasted and how badly this behavior can affect their performance and productivity. According to the research in 2017 [3], the Internet is capturing more and more of our time each day. Daily average of Internet usage has increased to 6.15 hours and time spent on social networking is also growing day by day. (Internet users aged 16-64)

In order to monitor how individuals spend their time online, productively, there is need for an automated time management application that can track their online activities and help them in discovering their good and bad behavior so that they can make changes when necessary. Thus, several self-tracking applications have been developed that bring self-awareness among individuals, help in making valuable decisions, improve their judgment and bring positive changes in their behavior and life. However, considering the limitations of existing applications (discussed in the next section) and the need for improved means for self-awareness, we present our research approach and findings from the digital footprints(Computer, Smart phones).

1.2 Objectives of the project

The main objective of our research is to develop a system for analysis of web-usage behavior patterns using interactive visualization techniques to promote self-reflection among users. Moreover, the system should be able to present the behavior from different perspectives using temporal (Several temporal dimensions can be defined, such as valid time, describing when the fact or the information is true in the real word, and transaction time) and categorical dimensions (categorical data derive from observations made of qualitative data that are summarized as counts or cross tabulations, or from observations of quantitative data grouped within given intervals).

Following are the objectives of our research work:

- Development of dashboard for gathering and processing of digital footprints of the user.
- Digital footprints behavior modeling for the extraction of interesting temporal and categorical patterns.
- Development and demonstration of interactive visualizations to analyze and monitor the extracted patterns in different dimensions.

1.3 Scope of the Project

The scope of our research is limited to only computer, android smartphones and does not include any other devices. To achieve this goal, Collect data only from user's computer and smartphones using already available third party tools.

Initially, a small scale investigation has been carried out on an individual user, and the results have been reported here. In future, we intend to evaluate it at large scale.

Chapter 2

Literature Review

Quantified self is the concept of individuals who use technology to collect, store and analyze their own life data to improve quality of life. The technologies, such as smartphone applications, wearable sensors and GPS devices, allow individuals to track any aspects of their daily activities. A variety of areas may be tracked and analyzed, for instance, weight, energy level, mood, time usage, sleep quality, health, performance, athletics and learning strategies. The data gathered by quantified self-trackers not only allows them to learn more about themselves, but it can also help them improve themselves, by making behavior changing decisions.[4]

The term "quantified self" was coined by Wired Magazine editors GaryWolf and Kevin Kelly in 2007. Wolf and Kelly formed a community for users and makers of self-tracking tools, in 2008, called The Quantified Self. It is now a worldwide community with close to thirty thousand members spread across 39 countries. The official website is the base for the community where users collaborate to share self-knowledge, tools and interests in the area of self-tracking. The community also hosts meetings all over the world, where quantified self-practitioners and other interested come together to share and discuss projects, tools, techniques and experiences about the subject. The Quantified Self community has established its own motto, "self-knowledge through numbers".

The implication of the motto would be taking the aspect of simply tracking the raw data to find patterns and ways to improve daily lives from it [14].

Quantified self approaches to collecting data about an individual could support reflection amongst individuals about their own learning [5]. Certain specific behaviors or psychological states relevant to learning (e.g., time on one task, number of words written per hour, emotional state) could be tracked and connected to specific learning outcomes (e.g., perseverance and score on maths homework). Such information could inform more effective study practices (e.g., knowing what times of day is most productive, increasing time on task) and give useful feedback on learning strategies (e.g., levels of resilience in difficult learning tasks); supporting meta-cognition. Such

activities could be further broadened to connect more specific data about cognition and learning with wider data sets about the individual that are relevant to learning. This might involve, for example, adding additional information about location, content of websites explored, information about nutrition or exercise, peer or school nominated data, or data about friends or peers. Such data can be collected automatically and/or through self-report, and similarly an individual could become significantly involved in the analysis of their data, or the 'results' about their learning process and performance could be automatically fed back to the user at appropriate time points. Individuals could use this information over time on an individual or group level.

The Quantified Self refers both to the cultural phenomenon of self-tracking with technology and to a community of users and makers of self-tracking tools who share an interest in "self-knowledge through numbers."[20]

2.1 Quantified Self through Time

Even though quantified self has not received much attention until recently, it is not a new phenomena. This section presents some projects in the area of quantified self.

2.1.1 Early projects

Self-tracking has a long history, especially in the medical sciences, resulting in significant discoveries. It has been documented that self-experimentations go as far back as the 16th century, where a man named Sanctorius of Padua weighed himself for 30 years before and after meals to study the energy expanded by a living organism. Anton Strock, in 1760, drank hemlock to determine its therapeutic effects. To study how digitalis affected his vision, Purkinje ate foxglove. Lazear died from his self-experiment with yellow fever. An anonymous scientist weighed the hair he shaved from his face each day as an index of testosterone production [4].

Allen Neuringer [4] tells in his paper Behaviorism from 1981 that a student of his studied how her food intake was related to her need for sleep. She slept fewer hours when eating 1,000 calories per day, than when she ate around 2,000 to 2,500 calories. This observation showed that she was more alert and had easier to study during the low calorie phase.

Even Benjamin Franklin was a self-tracker. He tracked 13 personal virtues in a daily journal to push himself toward moral perfection [15].

2.1.2 Newer projects

Stephen Wolfram, the founder of the computational knowledge engine Wolfram Alpha, has been collecting data on his life for a long time. He has been recording every keystroke he has made, incoming and outgoing emails, meetings and events, personal calls and steps taken. Wolfram revealed his sleep pattern by plotting his email timestamps by hour of the day for about 13 years. The plot, shown in Figure 1, shows a big gap on each day that reveals his waking hours. [14] Jon Cousin measured his mood daily and discovered improvements when sharing his current mood, in form of scores, with friends by email. He later founded a website called Moodscope, where users can rate and share their mood [15]. Seth Roberts tracked the long-term effect of butter on arithmetic speed. He compared a period where he ate about 60 grams of butter per day to a period where he ate almost no butter and did a daily arithmetic test. The tracked data supported the idea that butter

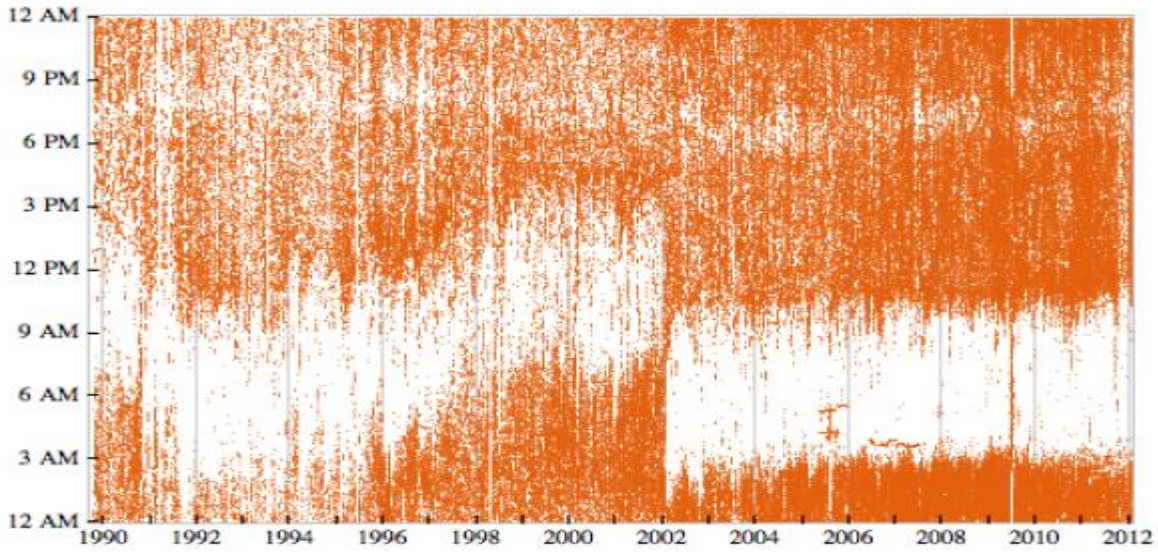


Figure 1 – Waking Hours comparision

produced an improvement in arithmetic speed [15].

To quote biohacker David Asprey:

“I don’t know if we’ll be calling any of this the quantified self in 50 years, we might just call it being human.” [14]
Lauren Manning tracked every type of food she consumed for a whole year and visualized it with a variety of over 40 graphics [14].

Robin Barooah stopped drinking coffee by making a large cup and removed 20 milliliters weekly. He tracked his hours of concentration per day and discovered that it increased after he stopped drinking coffee [15].

Bo Adler wore, even to the gym, a blood-pressure cuff, pulse oximeter and accelerometer all day long, along with a computer on a harness to collect the data. Adler was trying to figure out patterns for his sleep apnea [1].

2.2 The Rise of Quantified Self

There are a lot of contributing factors for the rise of the quantified self-trend. A main factor would be the easier and more flexible ways to collect data. Today you do not have to rely on your own memory or having to use analog tracking such as pen and paper or spreadsheets. The advancement in the areas of mobile devices, cheap small sensors, the Internet of Things, big data, the cloud and data visualization have led to the creation of the quantified self-tools of today.

2.2.1 The Internet of Things

Things that are connected to the internet are referred to as a part of the Internet of
Things phenomena. The most familiar internet-connected devices are computers such as laptops, servers, smartphones, and tablets, but the Internet of Things concept is much broader. In particular, everyday objects that have not previously seemed electronic at all are starting to be online with embedded sensors and microprocessors, communicating with each other and the internet. One of the biggest drivers of the Internet of Things is the increasing number of low-cost sensors available for many different kinds of functionality [15].

2.2.2 Mobile Devices

Smartphones have contributed to the rise from analog tracking to the first generation of quantified self-applications and wearable technology. Today we have even smarter devices with advanced embedded sensors and better applications for tracking and interpreting data. More and more tools suitable for the quantified self are being developed. According to IMS Research [14], 14 million wearable devices were sold in 2011 and in 2016, this number is expected to reach 171 million.

2.2.3 The Cloud

The cloud is basically a metaphor for the use of computers over the internet. An advantage of the cloud is the ability of using software, platforms and infrastructures as a service on the internet, without having to install anything. Basically, every internet-connected device can take use of cloud-based systems. In the case of quantified self, the cloud storage, accessibility and computing engines would be the main advantages.

2.2.4 Big Data

The term big data refers to managing large amounts of data, in forms of storage and analysis. Data is the source of the quantified self - movement as practice it is mainly about collecting, storing, analyzing and presenting personal data. Quantified self users collect data on a daily basis by manually and automatically tracking aspects of their daily lives. As these data sets grow in size, users may not have the tools available on local computing resources to store, query and manipulate quantified self data sets.

Cloud-based services for quantified self data storage, sharing, and manipulation would be extremely useful. In the long-term for quantified self systems, the increase in the amount of users and data would need an implementation of a big data solution to manage the need for data and work scaling.

2.2.5 Motivators

Good motivators are essential for quantified self users. Tracking data, especially manually, on a daily basis can be boring, hard and easy to forget. By utilizing visualization and creative application user interfaces could make the experience more fun, motivating and easier.

2.3 The Practice

The self-tracking practice can be illustrated with The Stage-Based Model of Personal Informatics developed by Ian Li, Anind Dey and Jodi Forlizzi [14]. The model, shown in Figure 2, consists of five stages: preparation, collection, integration, reflection and action. This model is proposed for development of effective quantified self tools. These stages have four essential properties: barriers cascade to later stages, they are iterative, they are user-driven and/or system-driven, and they are uni-faceted or multi-faceted.

The Stage-Based Model of Personal Informatics shows a good illustration of how practitioners begin to pursue the quantified self process, in order to gain self-knowledge and self-improvement, which is usually the main purpose when dedicating to the quantified self process.

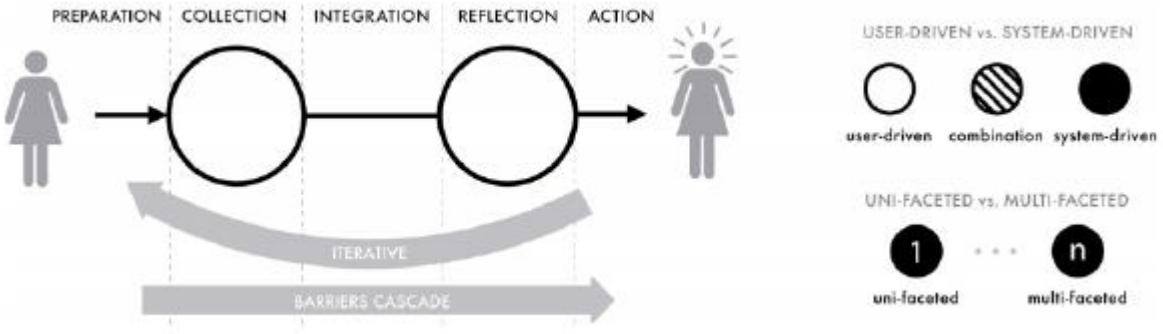


Figure 2: The Stage-Based Model of Personal Informatics Systems and its four properties

Preparation Stage

The preparation stage occurs before the collection of personal data starts. In this stage, the quantified tracker decides which data to collect and which tools to use. The data can include food consumption, sleep, mood or other collectible variables. Appropriate tools which satisfy the collector's needs should be chosen. There are two main types of tools for the practice of quantified self, which are shown in the figure 2.

– User-driven is a widely used type of quantified self applications. These applications (desktop-based, web-based or mobile-based) aid users in their data collection, by letting them enter their measurements manually. This is especially useful on data that cannot be directly measured and is often easier and more flexible to use, and cheaper to develop than hardware sensors [14].

– System-driven are the type of devices that automatically collect certain data.
Common categories of tracking for these devices are: heart rate, footsteps, acceleration and sleep cycles. Nike FuelBand, Fitbit tracker and the Pebble smartwatch are three examples of system-driven devices.

Collection Stage - The collection stage is where the practitioners collect data about themselves. The frequencies of collection can vary. It can either be continuous or episodic. A continuous collection can, for instance, be made hourly, daily or weekly. Episodic collection is made when a particular event occurs.

Integration Stage - Integration is the stage where the data produced in the previous stage is prepared for the reflection stage, by making it readable to the human eye. This is often made with some visualization of the data.

Reflection stage - The reflection stage is where the practitioners reflect on their collected data. The user will have to analyze the prepared data to gain self-knowledge and, if possible, draw conclusions that can lead towards self-improvement.

Action Stage - The action stage is where the practitioner choose what to do with the newly gained self-knowledge. It can be a number of things, such as doing nothing to tailoring behavior towards a goal or changing the behaviors completely.

Barriers Cascade - Every stage has a number of barriers, which are problems that can affect the current and later stages. For example, choosing an inappropriate tool for collecting data in the preparation stage may lead to faulty data, which affects the reflection stage. Another example would be problems in the collection stage, like lack of motivation to collect data on a daily basis, which may lead to sparse data.

Iterative Stages - Stages are iterative; practitioners will incorporate new data, tools and processes as they progress through the stages. For example, a user may change the type of data he or she is tracking or the type of tool for collecting data.

2.4 Available Online Third Party Tools

There are many tools available online to collect digital footprints from computer.

- **All In One Keylogger**

One of the most popular PC surveillance software currently available on the market. Offers a lot of advanced features at a pretty affordable price. The interface is translated into 12 languages, which makes it suitable for users from many countries.

Drawbacks

- Outdated Interface
- Not able to get total duration of application usage.

- **SpyPal - The Best PC & internet monitoring software**

SpyPal Keylogger is advanced PC & Internet monitoring software. It is NOT a simple keystroke logging program, often called keylogger. It intelligently records Facebook use, chats, emails, instant messengers, web sites visited, documents opened, applications executed, clipboard activity, Windows opened, microphone sound and much more. As an advanced monitoring software, SpyPal Keylogger takes pictures of your system Desktop screen periodically to provide you with graphics-based information, just like what a professional detective does in real life and captures passwords typed by web-based keyboards or on-screen keyboard. It displays exact activities, like WhatsApp, Skype, Facebook, Viber, Tinder, Hangouts, Youtube, Instagram, LINE, Telegram, Windows Store apps, computer games, internet searches, online shopping, file transfers, web-based emails like Hotmail, Gmail, Outlook, Yahoo mail and hundreds of others.

In Invisible Mode, SpyPal Keylogger is completely hidden to computer users. However, user can easily unhide the software with hotkey, or receive log reports via a pre-set email. These reports can be sent as often as according to preference, like every 60 minutes.

Main drawback: Generated reports are very complicated not user friendly.

- **Power Spy Lite**

Power Spy Lite allows you to monitor employees efficiently. It secretly logs all users on a PC and they won't know its existence. After it is easily installed on the PC, you can receive log reports on any device remotely via emails or ftp. You can check these reports as soon as you receive them or at any time convenient. You can also check logs from its log viewer on the monitored PC directly. No extra hardware is required to perform such monitoring. Power Spy Lite works on PC, laptop and tablet and it is compatible with Windows XP / Vista / 7 / 8 / 10.

- **ProcrastiTracker**

ProcrastriTracker is an open source *time tracking tool* that automatically tracks what applications and documents user used, and allows to view statistics.

- **WinLogOnView**

WinLogOnView is a simple tool for that analyses the security event log of Windows operating system, and detects the date/time that users logged on and logged off. For every time that a user log on/log off to your system, the following information is displayed: Logon ID, User Name, Domain, Computer, Logon Time, Logoff Time, Duration, and network address. WinLogOnView also allows you to easily export the logon sessions information to tab-delimited/comma-delimited/html/xml file.

Chapter 3

Methodology

Methodology of the research begins from the selection of the experimental environment through tracing and analysis. Therefore, this section of the research describes the proposed approach followed by the experimental platform from which the data has been collected, demonstrate interactive visualizations to better analyze the extracted patterns and allow individuals to compare themselves over time. To analyze and monitor these patterns, interactive visualizations are developed that facilitate the individual with the deep understanding of behavior.

3.1 Data Collection

There is no one tool available to collect the relevant data. So decided to use Power Spy Lite, ProcrastiTracker and WinLogOnView tools and collect the relevant data.

Following data able to collect from WinLogOnView tool.

- LogonID – Random unique number
- User Name – Logged User name
- Domain – Computer work group
- Computer – Computer name
- Logon Time – Logon Time
- Logoff Time – Logoff Time
- Duration – Duration
- Network address – **127.0.0.1** is the loopback Internet protocol (IP) **address** also referred to as the "localhost." The address is used to establish an IP connection to the same machine or computer being used by the end-user. The same convention is defined for computer's that support Ipv6 addressing using the connotation of ::1.
- Logon Type – Described below

Logon Type	**Description**
2	Interactive (logon at keyboard and screen of system)

3	Network (i.e. connection to shared folder on this computer from elsewhere on network)
4	Batch (i.e. scheduled task)
5	Service (Service startup)
7	Unlock (i.e. unattended workstation with password protected screen saver)
8	NetworkCleartext (Logon with credentials sent in the clear text. Most often indicates a logon to IIS with "basic authentication")
9	NewCredentials such as with RunAs or mapping a network drive with alternate credentials. This logon type does not seem to show up in any events. If you want to track users attempting to logon with alternate credentials.
10	RemoteInteractive (Terminal Services, Remote Desktop or Remote Assistance)
11	CachedInteractive (logon with cached domain credentials such as when logging on to a laptop when away from the network)

Table 1 – Logon Type

Collected data from computer using **WinLogOnView** tool

Logon ID	User Name	Domain	Computer	Logon Time	Logoff Time	Duration	Network Address	Logon Type
0x0042e04a	User	WORKGROUP	DESKTOP-I9CABHF	4/19/2019 12:27:02 PM			127.0.0.1	Interactive (2)
0x00040f03	User	WORKGROUP	DESKTOP-I9CABHF	4/19/2019 11:59:45 AM	4/19/2019 12:26:49 PM	00:27:04	127.0.0.1	Interactive (2)
0x00042b1c	User	WORKGROUP	DESKTOP-I9CABHF	4/19/2019 8:23:04 AM	4/19/2019 9:48:31 AM	01:25:27	127.0.0.1	Interactive (2)
0x00044f0b	User	WORKGROUP	DESKTOP-I9CABHF	4/18/2019 6:54:22 AM	4/18/2019 7:15:00 AM	00:20:38	127.0.0.1	Interactive (2)
0x010e6246	User	WORKGROUP	DESKTOP-I9CABHF	4/17/2019 6:43:29 AM	4/17/2019 6:50:25 AM	00:06:56	127.0.0.1	Interactive (2)
0x00e28acc	User	WORKGROUP	DESKTOP-I9CABHF	4/16/2019 7:05:38 AM	4/16/2019 7:22:07 AM	00:16:29	127.0.0.1	Interactive (2)
0x000f69a4	User	WORKGROUP	DESKTOP-I9CABHF	4/15/2019 9:09:59 PM	4/15/2019 10:58:15 PM	01:48:16	127.0.0.1	Interactive (2)
0x00047404	User	WORKGROUP	DESKTOP-I9CABHF	4/15/2019 9:06:13 PM	4/15/2019 9:08:21 PM	00:02:08	127.0.0.1	Interactive (2)
0x004b3514	User	WORKGROUP	DESKTOP-I9CABHF	4/15/2019 8:03:20 PM	4/15/2019 9:00:33 PM	00:57:13	127.0.0.1	Interactive (2)
0x000c303e	User	WORKGROUP	DESKTOP-I9CABHF	4/13/2019 4:41:35 PM	4/13/2019 4:58:24 PM	00:16:49	127.0.0.1	Interactive (2)
0x00046d45	User	WORKGROUP	DESKTOP-I9CABHF	4/13/2019 4:27:20 PM	4/13/2019 4:29:09 PM	00:01:49	127.0.0.1	Interactive (2)
0x0041d8d8	User	WORKGROUP	DESKTOP-I9CABHF	4/11/2019 3:35:19 PM	4/11/2019 3:48:27 PM	00:13:08	127.0.0.1	Interactive (2)
0x002c91ab	User	WORKGROUP	DESKTOP-I9CABHF	4/11/2019 11:14:14 AM	4/11/2019 11:24:08 AM	00:09:54	127.0.0.1	Interactive (2)
0x00047401	User	WORKGROUP	DESKTOP-I9CABHF	4/10/2019 7:12:23 AM	4/10/2019 7:30:54 AM	00:18:31	127.0.0.1	Interactive (2)
0x0004708a	User	WORKGROUP	DESKTOP-I9CABHF	4/9/2019 1:44:53 PM	4/9/2019 1:57:52 PM	00:12:59	127.0.0.1	Interactive (2)
0x002e9548	User	WORKGROUP	DESKTOP-I9CABHF	4/9/2019 1:36:07 PM	4/9/2019 4:58:01 PM	03:21:54	127.0.0.1	Interactive (2)
0x0003fdd4	User	WORKGROUP	DESKTOP-I9CABHF	4/6/2019 4:11:52 PM	4/6/2019 4:32:32 PM	00:20:40	127.0.0.1	Interactive (2)
0x005caf90	User	WORKGROUP	DESKTOP-I9CABHF	3/25/2019 5:56:02 AM	3/25/2019 7:12:25 AM	01:16:23	127.0.0.1	Interactive (2)
0x00047f2e	User	WORKGROUP	DESKTOP-I9CABHF	3/24/2019 6:17:30 PM	3/24/2019 7:01:11 PM	00:43:41	127.0.0.1	Interactive (2)
0x0177e572	User	WORKGROUP	DESKTOP-I9CABHF	3/24/2019 11:39:12 AM	3/24/2019 4:32:34 PM	04:53:22	127.0.0.1	Interactive (2)

Table 2 – WinLogOnView Tool Data

Collected data from computer using **ProcrastiTracker** tool

Figure 3 – Collected Data format from ProcrastiTracker

Collected data from computer using **Power Spy Lite** tool

Keystrokes			
Timestamp	**User**	**Details**	**Window Caption**
4/19/2019 6:15	User	nishan866pswdevdb123{Enter}pg{Enter}2014mcs054{Tab ->}0771382684{Shift}N+{Enter}	Course: MCS3204 Individual Project (MCS) - Google Chrome
4/19/2019 6:18	User	Por	Sign in to your account - Google Chrome
4/19/2019 8:41	User	{Shift}Supervisor{Ctrl}x	Program Manager
4/19/2019 8:41	User	{Ctrl}x{Left Win}d	Downloads
4/19/2019 8:45	User	{Enter}{Back Space}={Ctrl}z{Back Space}{Del}{Back Space}	Supervisor_Thesis (Autosaved) - Microsoft Word

4/19/2019 8:45	User	{Left Win}d	Inbox - nishanthi09@gmail.com - Gmail - Google Chrome
4/19/2019 8:46	User	Recoomen	Cortana
4/19/2019 8:47	User	{Ctrl}{Shift}{Del}	Desktop
4/19/2019 8:47	User	{Enter}	Delete Multiple Items

Table 3 - Collected data from computer using Power Spy Lite tool

Collect Android Smartphone usage digital data using email

Step1: Download your Android Mobile usage data:

- using the Gmail id sync user's Android Phone
- Using Google Takeout download all the data

Step 2: Process JSON data and extract features (Example as follows)

{

"Search Engines": [

{

"suggestions_url": "",

"favicon_url": "http://search.espncricinfo.com/favicon.gif",

"safe_for_autoreplace": true,

"date_created": 13039284095640706,

"url": "http://search.espncricinfo.com/ci/content/site/search.html?search\u003d{searchTerms}\u0026gblsearch\u003d",

"new_tab_url": "",

"instant_url": "",

“originating_url”: “”,

“search_terms_replacement_key”: “”,

“deprecated_show_in_default_list”: false,

“sync_guid”: “8861F571-537F-4F5F-AEC0-911BF41E5948”,

“short_name”: “espncricinfo.com”,

“keyword”: “espncricinfo.com”,

“input_encodings”: “UTF-8”,

“prepopulate_id”: 0,

“last_modified”: 13039284095640706

}, {

“suggestions_url”: “”,

“favicon_url”: “”,

“safe_for_autoreplace”: true,

“date_created”: 13048863240064198,

“url”: “http://stackoverflow.com/search?q\u003d{searchTerms}”,

“new_tab_url”: “”,

“instant_url”: “”,

“originating_url”: “http://stackoverflow.com/opensearch.xml”,

“search_terms_replacement_key”: “”,

“deprecated_show_in_default_list”: false,

“sync_guid”: “568E8DD7-EB70-4553-85FE-451B2D8632EA”,

“short_name”: “Stack Overflow”,

“keyword”: “stackoverflow.com”,

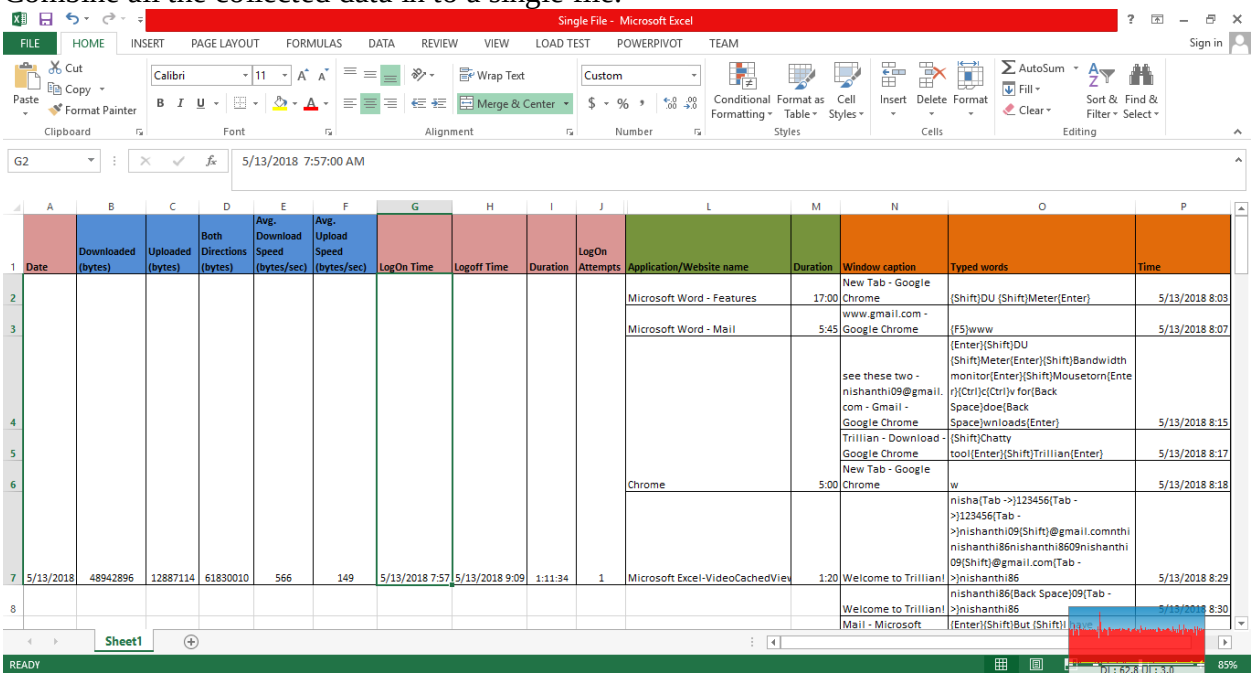
"input_encodings": "UTF-8",

"prepopulate_id": 0,

"last_modified": 13048863240064198 }

3.2 Data Preprocessing

Combine all the collected data in to a single file.



Date	Downloaded (bytes)	Uploaded (bytes)	Both Directions (bytes)	Avg. Download Speed (bytes/sec)	Avg. Upload Speed (bytes/sec)	LogOn Time	Logoff Time	Duration	LogOn Attempts	Application/Website name	Duration	Window caption	Typed words	Time
										Microsoft Word - Features	17:00	New Tab - Google Chrome	{Shift}DU {Shift}Meter{Enter}	5/13/2018 8:03
										Microsoft Word - Mail	5:45	www.gmail.com - Google Chrome	{F5}www	5/13/2018 8:07
												see these two - nishanthi09@gmail.com - Gmail - Google Chrome	{Enter}{Shift}DU {Shift}Meter{Enter}{Shift}Bandwidth monitor{Enter}{Shift}Mousetorn{Enter}{Ctrl}c{Ctrl}v for{Back Space}doe{Back Space}wnloads{Enter}	5/13/2018 8:15
												Trillian - Download - Google Chrome	{Shift}Chatty tool{Enter}{Shift}Trillian{Enter}	5/13/2018 8:17
										Chrome	5:00	New Tab - Google Chrome	w	5/13/2018 8:18
5/13/2018	48942896	12887114	61830010	566	149	5/13/2018 7:57	5/13/2018 9:09	1:11:34	1	Microsoft Excel-VideoCachedView	1:20	Welcome to Trillian!	nisha{Tab ->}123456{Tab ->}123456{Tab ->}nishanthi09{Shift}@gmail.comnthi nishanthi86nishanthi8609nishanthi 09{Shift}@gmail.com{Tab ->}nishanthi86	5/13/2018 8:29
												Welcome to Trillian!	nishanthi86{Back Space}09{Tab ->}nishanthi86	5/13/2018 8:30
												Mail - Microsoft	{Enter}{Shift}But {Shift}I have	

Figure 4 – Combine Data to single file

Then decided to extract following attributes, visualize on the dashboard and finally find correlations between the digital data.

- How much time the user spends on computer and browser?
- How long the user remains idle?
- How long the user stays on a particular application/websites
- What are the browsing peak times, top website and top category of the day/month?
- How often user switches between the tabs?
- How many tabs the user opens during a session?
- How many sessions the user open during a day?
- How long the user stays on a session?
- How one navigates between pages (e.g. by clicking on hyperlinks, typing url, reloading page, etc.), and between which group of pages the user navigates?

- What are the information the user is curious about
- Total bytes uploaded and downloaded
- What is the most productive time of the day for a user
- What are the top applications and top category of the day/month?

3.3 Algorithms & Techniques

Feature Modeling

Our data collection module efficiently runs in the background of the computer and autonomously captures a wide range of information. To infer user's context and behavior, behavior data features have been identified and collected.

Data mining techniques

There is a variety of pattern discovery techniques including associative rule mining, sequential pattern mining, classification, and clustering, that discover the correlations among the collected computer data, sequential patterns over time intervals.

Visual data mining techniques have proven to be of high value in exploratory data analysis [2]. Visualization allows the user to mine and gain insight into the data and come up with new mining recommendations. There are many visualization techniques that have been developed to explore the meaningful information from the large datasets. Goal of visual data mining is to represent as many of data points as possible in a single visualization or plot. Pattern discovery and visual data mining techniques have been discussed in next subsections.

Pattern Discovery Techniques

Statistical Analysis is the science of collecting, exploring, and presenting data to discover underlying patterns and trends. Statistical techniques are most common to extract pattern from the digital footprints. Different kinds of descriptive statistical analyses, e.g., frequency, count, min, mean, max, median,
Mode 'etc. can be performed on the data attributes like page views, time spent at a particular page, frequently accessed pages, number of sessions per day, session time span, etc.

a) **Associative Rules**: are used to find out the frequent items which are used together. Association or correlation rules are measured by its support, confidence and correlation. Support is the percentage of transactions in dataset that contain A U B. Confidence is percentage of transactions in dataset containing A that also contain B.

$$Confidence(A \rightarrow B) = P(B|A) = support(A \cup B)/support(A)$$

Lift is a correlation measure and can be computed as

$$Lift(A, B) = P(A \cup B) = P(A \cup B)/P(A)P(B)$$

Association rules are used to find associations among web pages and web categories that frequently appear together in users' sessions. Apriori algorithm is the most classical algorithm for mining frequent item sets. Clustering is a technique that groups together the items having similar characteristics. Web usage clusters can be discovered by grouping the users having similar browsing trends.

Visual Data Mining Techniques

Information visualization and visual data mining can help to deal with the flood of information [2]. Presenting data in an interactive, graphical form often bring new insights and provide deeper domain knowledge. There are three steps that visual data exploration follows such as Overview, zoom and filter, and then details-on-demand. Visual data exploration can easily deal with highly noisy and nonhomogeneous data. No understanding of complex mathematical or statistical algorithms or parameters is required.

Fig. 5 shows the three dimensions such as datatype to be visualized, visualization technique and interaction technique. Any of the visualization techniques can be used with any of the interaction technique [2].

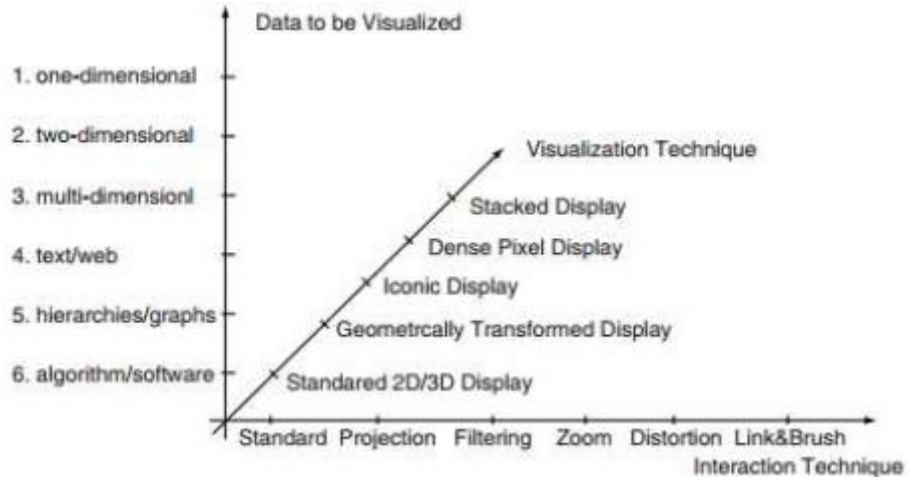


Figure 5 – Data Visualization format

The visualization technique used may be classified as standard 2D/3D displays, such as bar charts, x-y plots, heat map, parallel coordinates [15], icon-based displays, circle segments, chord diagrams, stacked displays, such as tree maps.

Parallel coordinates techniques allow exploring and analyzing the multidimensional data. Each data item is presented as a polygonal line which intersects each axis at the point equal to the value in that dimension. It maps the k-dimensional space onto the two display dimensions by using k equidistant axes which are parallel to one of the display axes.

Sunburst used to visualize hierarchical data represented by concentric circles. The circle in the center represents the root node, with the hierarchy moving outward from the center.

Scatter Bubble chart shows the relationship between three different variables in one plot. An additional dimension of the data is represented in the size of the bubbles.

Radar Chart is a two dimensional chart that displays multivariate data over multiple quantitative variables represented on axes starting from the same point.

Chord diagram shows the connection among different entities. The chords between the arcs visualize the switching behavior of the respondents between entities in both directions.

Heat map is a two-dimensional representation of data in tabular format with user defined color ranges e.g. low, high and average. It provides an immediate visual summary of information.

Stacked Bar Chart Bar charts are used to show two dimensional data and can be used for more complex comparisons of data with the stacked bar charts. Stacked bar chart stacks bar that represent different group on top of each other.

Interaction and Distortion Techniques allow the user to dynamically change the visualization according to exploration objectives and provide the data with low level details while preserving the high level details for example interactive zooming present more details on higher zoom levels.

Behavior Extraction

- **Websites Categorization**: Web URLs are grouped into various categories, such as social networking, research and development, news media, career and education, etc. Website categorization APIs [18] [19] have been used to automatically retrieve category and subcategory for the web site via HTTP request.
- **Browsing Times of the Day**: We have considered six times of the day i.e. Early Morning, Morning, afternoon, evening, night, midnight.

Where

$$4_{AM} \geq EarlyMorning \leq 8_{AM}$$
$$8_{AM} \geq Morning \leq 12_{AM}$$
$$12_{PM} \geq AfterNoon \leq 4_{PM}$$
$$4_{PM} \geq Evening \leq 8_{PM}$$
$$8_{PM} \geq Night \leq 12_{AM}$$
$$12_{AM} \geq MidNight \leq 4_{AM}$$

- **Frequent Categories/Websites and their Correlation:** Apriori algorithm has been used to get the frequent categories. It extracts the categories that frequently used together. We have supposed that an item set is frequent if it appears in at least 40% of the total sessions. For example, 20 is the support threshold for 50 sessions. First step is to count the number of occurrences of each category separately by scanning all the sessions. Next step is to generate the pairs of frequent items. Pairs that meet the support threshold are frequent.
  Associative rule mining is a technique for discovering interesting relations between categories. In order to select interesting rules, minimum support and confidence constraints are used.
  For example, rule is Social $\Longrightarrow$ SoftwareDevelopment. Its confidence is
  Support (Social U SoftwareDevelopment) / Support (Social) = 0.5/0.5 = 1
  Which means software development occurs in all the sessions containing social.
  To find the correlation among the categories, we use:
  Lift (Social $\Longrightarrow$ SoftwareDevelopment) = P (Social U SoftwareDevelopment) /P (Social)
  P (SoftwareDevelopment) = 0.5/ ((0.5 *1)) = 1
  It shows that social websites and social development are used together. Recommendation can be proposed here by analyzing whether social networking affecting the productivity of user or not.

- **Predicting User Interests:** Website's visits frequency and duration are two major metrics of a user interest in a website [19]. We consider these metrics to estimate the user interest. Duration is measured based on dwell time normalized by maximum dwell time. Frequency is measured based on number of visits of category normalized by maximum number of visits. Harmonic mean is used to mitigate the impact of large outliers and aggravate the impact of small ones. Together, they are used to find the areas of interest for any user.

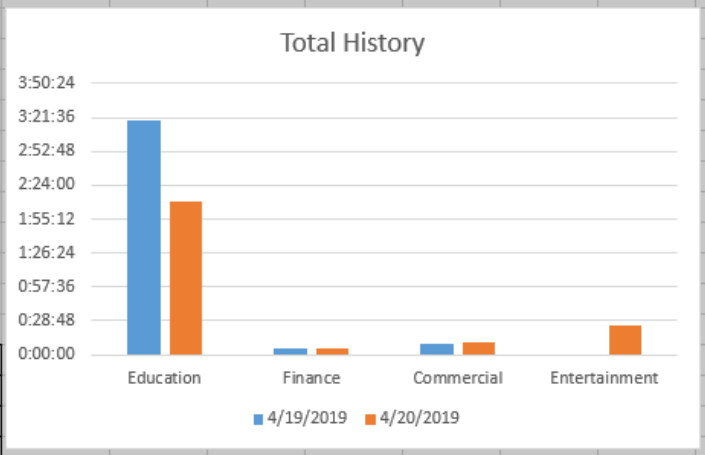
3.4 Results Analysis and Evaluation

Experimental Evaluation

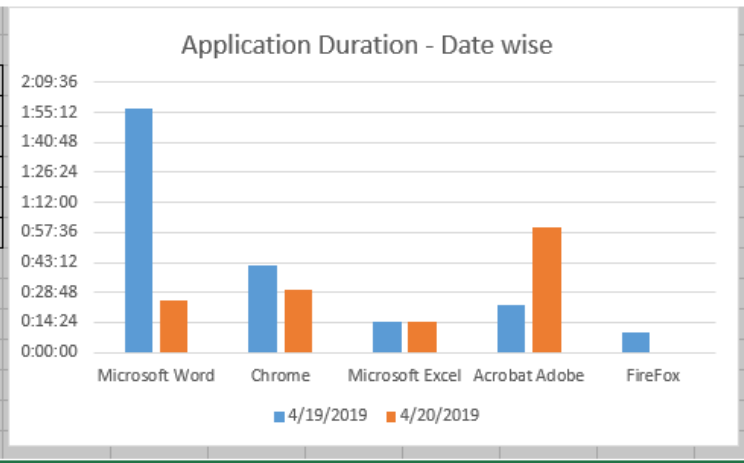
Collected two-weeks of digital footprints from user's computer as well user's mobile and browsing activities are integrated using cell number and email id.

4/19/2019	
Duration	Type
1:57:05	Education
0:24:55	Education
0:11:45	Education
0:05:00	Finance
0:14:25	Education
0:08:00	Education
0:11:25	Education
0:11:10	Education
0:09:15	Commercial

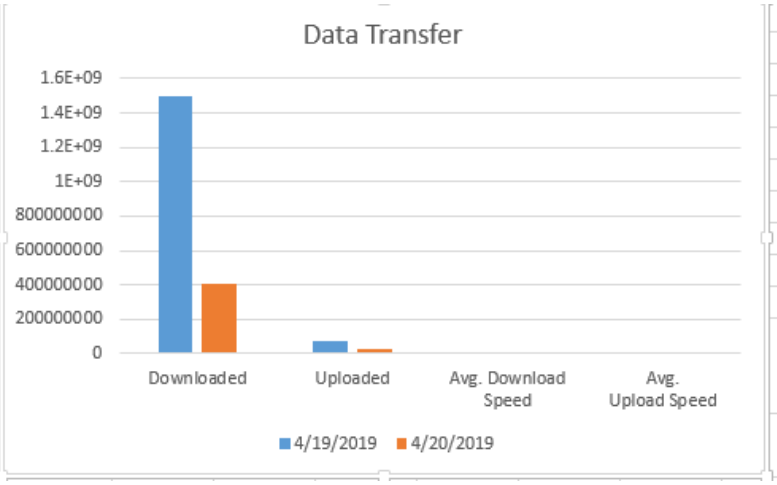
	4/19/2019	4/20/2019
Education	3:18:45	2:10:00
Finance	0:05:00	0:05:00
Commercial	0:09:15	0:10:15
Entertainment	0:00:00	0:25:00

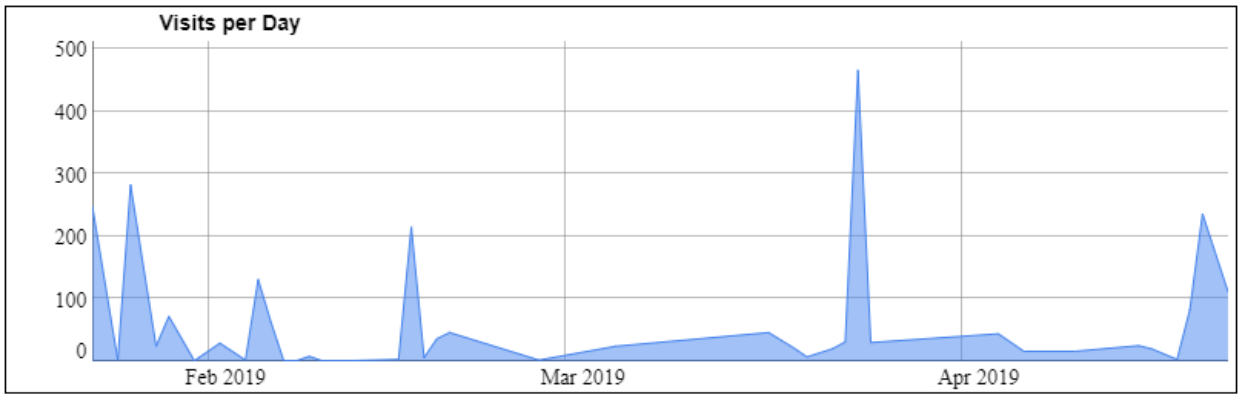


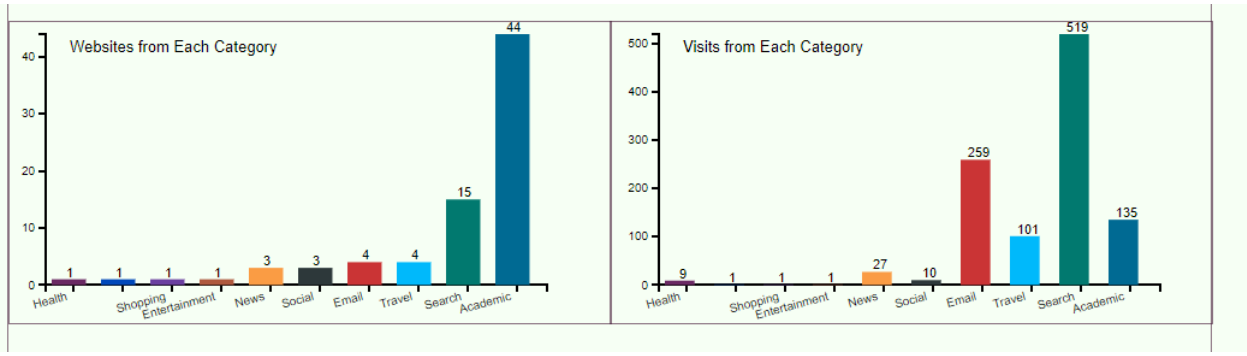
	4/19/2019	4/20/2019
Microsoft Word	1:57:05	0:25:00
Chrome	0:41:40	0:30:00
Microsoft Excel	0:14:25	0:15:00
Acrobat Adobe	0:22:35	1:00:00
FireFox	0:09:15	



	4/19/2019	4/20/2019
Downloaded	1496396627	409612377
Uploaded	71534544	23228322
Avg. Download Speed	17319	4740
Avg. Upload Speed	827	268







Daily Stats	
Average Visits Per Day:	65
Median Visits Per Day:	30
Today:	57

The following packages used for the Analysis in R

```
library(jsonlite)
library(tidyverse)
library(lubridate)
library(ggrepel)
library(viridis)
library(gganimate)
library(cowplot)
library(ggthemes)
```

```
me <- jsonlite::fromJSON("SearchEngines.json")
```

Above method used for the analysis for all the extracted JSON format data.

3.5 Evaluation

We evaluated the results by conduct survey with participants. We discuss here about some users' reviews regarding the dashboard. They have found it very interesting and were motivated that they can clearly understand their comprehensive digital footprints statistics across many dimensions. Some said that this makes them conscious and aware about their usage and restrict them when they see the unusual behavior.

Some users have privacy concerns and suggested that user's identity should be removed, and data should be transferred as anonymous user. This aspect will be considered in the future, but for the experiments it was needed for some individual tracking purposes

3.6 Challenges

One of the major challenge is to motivate and convince user to track digital footprints using third party tools. Interactive visualizations have been implemented that provide users with the quick view about their behavior. Major challenge in using the third party tools is privacy; people have privacy concerns about data collection. Some users feel hesitated in sharing their data. Accuracy cannot be assured in case if user deliberately changes his/her logged data by deleting some data or disabling the third party tools while using computer.

Chapter 4

Conclusions and Future Works

This research work has introduced an approach towards capturing and analyzing digital footprints behavior of individuals over temporal and categorical contexts. Third Party tools installed that runs autonomously in the background and captures the activities.
It allows the individuals to visualize their interesting browsing behavior patterns to gain deeper insights into their behavior by providing interactive graphical user interface to promote self-reflection and awareness among them and help in making valuable decisions for bringing positive changes in their behavior and life.

To extract the valuable patterns from data, different pattern discovery techniques have been utilized including statistical analysis, associative rule mining, sequential pattern mining and clustering. This visualizations yields some interesting results about how users browse the web such as dwell time on web pages, the time users are inactive, user's peak browsing time and hour of the day, top category of the day, frequent websites/categories and their correlation, top websites on the basis of time spent, weekly usage comparison among different categories, duration of browsing sessions, number of sessions per day, number of tabs per session, frequent transition type, cluster the frequent websites at different time of day and time spent at other desktop applications when browser is running in background but not focused.
Visual data mining techniques have been used to explore the extracted patterns as interactive visualization helps user in understanding and analyzing the wide range of data more easily and quickly.

Additional data mining and visualization techniques will be integrated at large scale to yield more interesting, effective, and valuable insights from the behavioral data. We intend to integrate our framework with persuasive feedback mechanism that will provide interventions to improve user's behavior.

References

[1] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, pp. 12–23, 2000.

[2] D. Keim et al., "Information visualization and visual data mining," Visualization and Computer Graphics, IEEE Transactions on, vol. 8, no. 1, pp. 1–8, 2002.

[3] G.W.Index, https://blog.globalwebindex.com/chart-of-the-day/dailytime-spent-on-social-networks/, 2017.

[4] Melanie Swan. The quantified self: Fundamental disruption in big data science and biological discovery. Big Data, 2:85–99, 2013.

[5] http://dl.acm.org/citation.cfm?id=2330631. [Google Scholar]

[6] The Quantified Self. The Web Site. http://quantifiedself.com/ (visited 2014-05-17).

[7] http://www.artofmanliness.com/. The Virtuous Life: Wrap Up. http://www. artofmanliness.com/2008/06/01/the-virtuous-life-wrap-up/ (visited2014-03-29).

[8] Stephen Wolfram Blog. The Personal Analytics of My Life. http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/.

[9] Seth Roberts. The Grow of Personal Science Implications For Statistics. http://blog.sethroberts.net/wp-content/uploads/2012/09/2012-09-24-The-Growth-of-Personal-Science-Implications-For-Statistics.pdf (visited 2014-03-29).

[10] Lauren Manning. A year of food consumption visualized: Flowingdata.com. http://flowingdata.com/2011/06/29/a-year-of-food-consumption-visualized/ (visited 2014-03-29).

[11] Robin Barooah. The false god of coffee: Quantifiedself.com. http://quantifiedself.com/2009/10/the-false-god-of-coffee/ (visited 2014-03-29).

[12] David Asprey. Who Really Owns Your Personal Data? http://www.details.com/culture-trends/critical-eye/201305/sharing-biodata-on-apps-and-devices?currentPage=2

[13] Melanie Swan. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. Internet of Things: Technologies and Applications, 3:217–253, 2012.

[14] IMS Research. Wearable Technology Market to Exceed $6 Billion by 2016. http://www.computerworld.com/s/article/9230095/Wearable_technology_market_to_exceed_6B_by_2016 (visited 2014-03-29).

[15] Per H¨agglund. Taking gamification to the next level. Technical report, Dept. of Comp. Sc., Ume°a University, Ume°a, Sweden, 2012.

[16] Anind Dey Ian Li and Jodi Forilizzi. A stage-based model of personal informatics systems. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1:557–566, 2010.

[17] P. K. Chan, "A non-invasive learning approach to building web user profiles," 1999.

[18] W. CategorizationAPI, https://developer.similarweb.com/, 2015.

[19] UClassify, https://www.uclassify.com/browse/uclassify/topics?input=Url,2015.

[20] https://en.wikipedia.org/wiki/Quantified_self#cite_note-4