Web Usage Behavior Pattern Analysis and Recommender System

By

Farhana Seemi

2013-NUST-MS-IT-730

Supervisor

Dr. Hamid Mukhtar
Department of Computing

Masters of Science in Information Technology (MS IT)
In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(December 2015)

Acknowledgment

First and foremost praises be to Allah, the Praise be to ALLAH Almighty for He bestowed upon us intelligence, gave us the mental strength to demystify the secrets of the universe, conferred to us the will and the strength to explore and the power to aim for and achieve our goals.

I would like to express my deep and sincere gratitude to my supervisor, Dr. Hamid Mukhtar for all his kind help, guidance, suggestions and support throughout this project. His wide knowledge and his logical way of thinking have been of great value for me. His understanding, encouraging and personal guidance have provided a good basis for my thesis.

Besides my advisor, I would like to thank my thesis committee members for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

Last but not the least, I would like to thank my family: my parents and to my brothers and sister for supporting me spiritually throughout writing this thesis and my life in general.

Abstract

Almost anything we do online leaves traces of our activities on the Internet. These footprints offer an opportunity to study various aspects of human behavior**.** Aim of this research is to analyze web usage behavior patterns to promote self-awareness that helps bring positive changes in individual's performance. Time spent at browsing particular web page is a key metric to estimate the website usage. In this thesis, we describe the framework that first collects and processes the data including quantitative data such as time spent, number of visits and qualitative data i.e. web site category. Second, we describe the web usage behavior modeling to extract valuable and interesting temporal and categorical patterns regarding user interests, peak browsing time of day, most visited websites, related websites groups according to website categories, frequent tab switching, and session's statistics i.e. number of sessions per day, number of tabs per session etc. To discover the valuable behavior patterns from the individual's browsing data, different web usage mining techniques have been used including statistical analysis, associative rule mining and clustering. Finally, we demonstrate interactive visualizations for the analysis and monitoring of web browsing behavior patterns with the goal of providing the individual with detailed understanding of behavior, provide recommendations and present the social comparison framework to promote competition and motivation among individuals that can bring out the ambition and push one that's good for person's personal and professional growth. The experimental results demonstrate interesting correlations of web categories usage with different times of the day, project deadlines and user productivity. Questionnaire is drawn to evaluate the proposed system 'BBA'. According to survey results, 90% of users found 'BBA' very effective that made them conscious and aware of their web usage behavior.

Table of Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

Quantified self-movement[1] incorporates digital technology to acquire data on various aspects of an individual's life with an aim to improve self-awareness and human performance. People want to be self-aware and self-knowledgeable in order to improve their performance and outcomes. Today, technology logs almost everything we do with the aim to measure all aspects of our daily lives. 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.

People can become more self-aware by comparing themselves to others in their social circle to evaluate their performance, strengths and weaknesses as well as their position in society. Social Comparison will motivate a person to try harder and to perform at higher level. Based on many psychological studies, humans gain motivation through comparison with others. Corcoran, K., Crusius, J. and Mussweiler [10] have defined social comparison as a ubiquitous process which influences how people evaluate themselves, what they are motivated to do, and how they behave. Chen, P. & Garcia [14] provide the perspective that social comparisons can drive competition among users and also show how social comparison correlates with competition. According to the study [18], Individuals may seek self enhancement and improve their self-esteem by comparing themselves.

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 such as preprocessing, pattern discovery and pattern analysis. Pattern discovery includes following techniques to extract the pattern i.e. statistical analysis, sequential pattern mining, path analysis, association rule mining, classification, and clustering [13]. To quickly understand and analyze the behavior patterns, there are a large number of information visualization techniques that have been developed over the last few years that can deal with wide range of data [12].

[1] http://quantifiedself.com/

Online recommendation and prediction is one of the web usage mining applications. Recommendations help users to quickly find the information they want or find interesting. Recommendations are dynamically determined either based on manually specified rules or automatically determined by different recommendation algorithms. Collaborative filtering is one of the most successful and widely implemented recommendation technologies [25]. It predicts the potential interests of an individual by taking into account the opinions of users in his social circle with similar taste.

This chapter gives the basic idea of the concepts involved in this research. It also presents the background and motivation for this study. Moreover, it provides the hypothesis, gives an idea of expected results, and methodology to get and evaluate the results. Finally, it presents the structure of this thesis document.

1.1 Problem Context

Life has become so much fast and busy these days that even we don't have time to pay attention to our true selves. Disease of being busy is spiritually destructive to our health and wellbeing leading us towards stress, depression and anxiety. Many people waste time on activities that keeps them busy but not productive. They spend most of their time of day in surfing web without even noticing that how much time they have wasted and how badly their wrong behavior can affect their performance and productivity. According to the research in 2014 [33], internet is capturing more and more of our time each day. Daily average of web usage has increased to 6.15 hours and time spent on social networking is also growing day by day.

In order to monitor how individuals actually spend their time online whether productive or not, there is need for an automated time management application that can track their each and every online activity and help them in discovering their good and bad behavior so that they can make changes when necessary. These self-tracking applications bring self-awareness among individuals, help in making valuable decisions, improve their judgment and bring positive changes in their behavior and life.

1.2 Problem Statement

Cheaper availability of internet has made an incredible increase in the web usage. Almost everyone has access to the internet. Now it becomes easy to facilitate individuals with self-

reflection through feedback on regular basis to let them be aware about their behaviors. Aim of our research is to promote awareness to individuals about their behavior that help them manage their time spent on internet more efficiently.

"***To analyze web usage behavior patterns using interactive visualization techniques to promote self-reflection and provide recommendations on the basis of individual's comparison among people in his social circle that can help bring changes in the individual's behavior."***

1.3 Objectives and Scope

Following are the objectives of this thesis:

1. Development of framework for gathering and processing of web usage data.
2. Web usage Behavior modeling for the extraction of interesting temporal and categorical patterns.
3. Development and demonstration of interactive visualizations to analyze and monitor the extracted patterns.
4. Development of framework that allows individuals to compare their temporal and categorical web usage statistics among friends and colleagues.
5. Propose recommendations on the basis of individual's personality, interests and similarity among social circle.

Scope of our thesis is to promote self-reflection by providing individual the deeper insight about their web usage behavior via interactive visualization techniques. To achieve this goal, a browser extension has been implemented. Initially testing and evaluation of this extension has been done at the smaller scale. In future, we intend to evaluate it at large scale.

1.4 Proposed Framework

In this thesis, we propose the framework that collects and process web usage data, extract interesting behavior patterns from formulated data, demonstrate interactive visualizations to better analyze the extracted patterns and allow individuals to compare themselves among their social circle. Initially, qualitative and quantitative web usage data features are identified such as dwell time, number of hits, category, idle time and time of occurrence. Browser add-on log these data features on the trigger of different browser events such as creating of the tab/window, updating the tab/window, closing tab/window, status of window changes etc. Add-on transfers the web usage data when previous browsing activity date doesn't match

with the current one. Tabs switching behavior data is also logged and transferred to server after some predefined interval during the session in case if session time span exceeds the defined limit.

Behavior patterns are extracted from the logged data including user interests, frequent categories, user's personality traits, and peak browsing time via web usage mining techniques. To analyze and monitor these patterns, interactive visualizations are developed that facilitates the individual with the deep understanding of behavior. This framework allows individual to add people in their social circle and compare their categorical and temporal behavior trends and statistics among them. It also provides recommendations on the basis of individual's personality and interest comparison among his social circle.

1.5 Contribution

Following are the contributions of this thesis:

1.5.1 Browser Extension

Browser Extension (BBA) has been developed that runs in background and track each and every browsing activity of an individual. This extension also provides web interface where individual's daily activities can be visualized. Individual can see his detailed daily, weekly and monthly usage statistics and significant behavior patterns in the form of interactive visualization charts. Usage comparison among friends and mobile activities usage can also be visualized here in this extension.

1.5.2 Data Set

Data set has been created having web usage behavior data of 18 users. Web usage data includes information such as url, time spent, number of visits, category, time stamp, session, tab, switchTo_tab, transition type, computer usage, idle time and browser usage.

1.6 Structure of the thesis

In Chapter 2 we will discuss related work done so for. In chapter 3 we will discuss the methodology, system architecture and brief introduction of its sub-modules. In chapter 4, we will discuss about the features of chrome extension BBA. Chapter 5 gives detail discussion regarding the experiments and results of experiments. Lastly, chapter 6 gives final conclusions and future works can be done.

Chapter 2

Literature Survey

2.1 Related Work

DOBBS [5] uses a browser add-on that allows researchers to log browsing behavior of online users, capture relevant different window, session and browser events in anonymous and privacy-preserving manner and send those events to the server. In Dobbs, event is the unit of information. This paper describes all the logged events including window events, session events and browsing events. Window events includes events e.g. the opening and closing of a browser window or tabs and changing in the state of browser window. Session events include all the events that occur during the time frame. Browsing events comprise the events that are associated with navigating between web pages e.g., how a user switched between different open tabs. This paper has also presented results using visualizations to provide deeper insight in understanding behavior. DOBBS is an open and unsupervised environment. Once a user has installed the add-on, there is no interference from any controlling entity. Users can consciously manipulate the resulting logging data by behaving in a specific manner, e.g., by always leaving the same web page open when leaving the desk for a longer time. Motivating user to participate is very challenging here because users do not directly get benefit from the add-on; it provides no added value to them.

Passive browsing is the time of idleness or inactivity during a user's browsing sessions. Parallel browsing is opening of multiple tabs within one browser window and switching among them. Christian von der Weth and Manfred Hauswirth [3] have analyzed in their study the impact of parallel and passive browsing on the calculation of user's time spent at web page and introduced the new metrics, focused ratio and activity ratio, to quantify the popularity of websites that how engaging and interesting a website is. Sharad Goel, Jake M. Hofman, and M. Irmak Sirer [8] investigated different patterns of internet usage across user demographic including sex, race, education, and income with respect to the five most visited website categories i.e. social media, e-mail, games, portals and search as shown in figure 2.1. This study also has shown that different demographic attributes can be inferred using

browsing histories to facilitate personalization of content. Demographic groups spend the most of their time on the same popular activities (e.g., social media and e-mail).

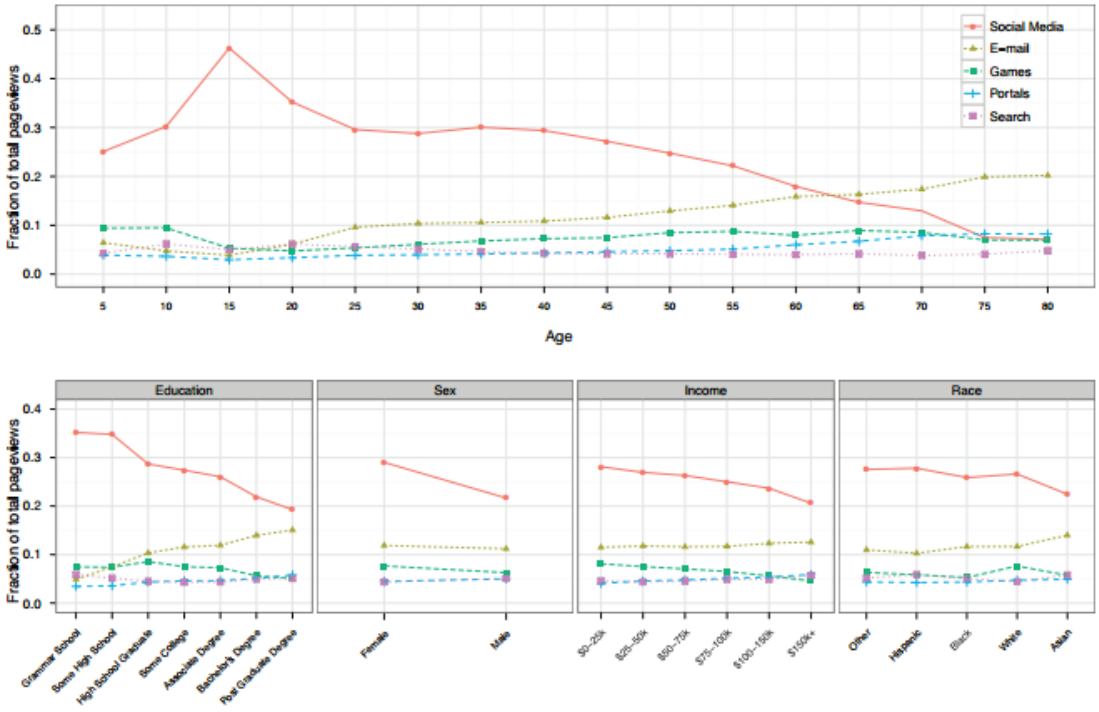


Figure 2.1 Internet usage patterns across different demographic attributes [8]

Ravi Kumar and A. Tomkins [11] provide taxonomy of page views consisting of categories content, communication and search as shown in figure 2.2 (a). They have presented a quantitative analysis of the mechanics of online behavior. Figure 2.2 (b) gives the distribution over the gaps between sessions of the same user, 70% of sessions start within twelve hours of the previous session, and only 13% of sessions occurs after a gap of a day or more. They described measures to find popular websites. They categorize the inter-arrival time between page views within a session. They studied that how users navigate between pages and examined link path within and across different types of page.

Main category	Sub-category	Fraction
Content	Games	6.2
	Multimedia	5.4
	Portal	5.4
	Head listings	3.4
	News	3.4
	Other vertical	28.1
	Total	**52.0**
Communication	Social	24.3
	Mail	9.4
	Forum	1.4
	Blog	0.4
	Total	**35.5**
Search	Main search	6.2
	Multimedia search	1.4
	Item search	1.4
	Total	**9.0**
Unknown	Total	**3.4**

Figure 2.2 (a) Page views broken by page type [11]

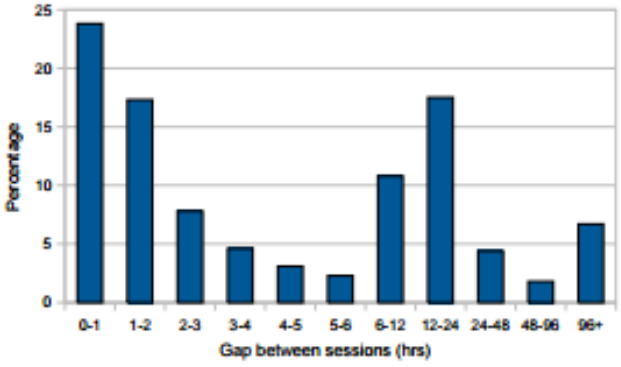


Figure 2.2 (b) Gap between sessions of the same user [11]

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 such as preprocessing, pattern discovery and pattern analysis. Pattern discovery includes following techniques to extract the pattern i.e. statistical analysis, sequential pattern mining, path analysis, association rule mining, classification, and clustering [13]. Different web usage mining techniques has been discussed in paper [1] that can be used to extract patterns from Web log files. Discovered patterns will be used for pattern analysis that helps in understanding the user behaviors. According to [2], Density based clustering algorithm has been used to discover navigation patterns. K Nearest Neighbor with inverted index has been suggested for efficient prediction.

Khovanskaya et al. [27] have presented an interface that display personal web browsing data and reveal different strategies that deliberately display sensitive, purposeful malfunction summaries in unconventional ways to raise self-awareness about data mining. They have defined a cut as subset of collected data and visualize those cuts using a variety of visualizations. They developed visualizations using different approaches to present the data from a cut because visualization that covers an interesting routine in one cut may lack detail needed to get value from another cut. Different cuts other than temporal that can also be used identify meaningful findings in data have been discussed in paper [4]. Life Flow [9] is a visualization tool that can easily analyze the log file full with diverse user activities. It provides support to analyze event sequences. It sorts the sequences by frequency and reveals the dominant activities. It also aligns the activities before and after selected event that help to see the frequent activities before and after the events.

Social interaction plays an important role in behavioral model. Behavior of a person is determined and controlled significantly by others around [17]. Mukhtar et al. [7] have proposed a framework that enables social interactions between the patients, doctors, and other users in their online social community through a web portal as well as through their smart phones. Figure 2.3 shows the various components of the framework.

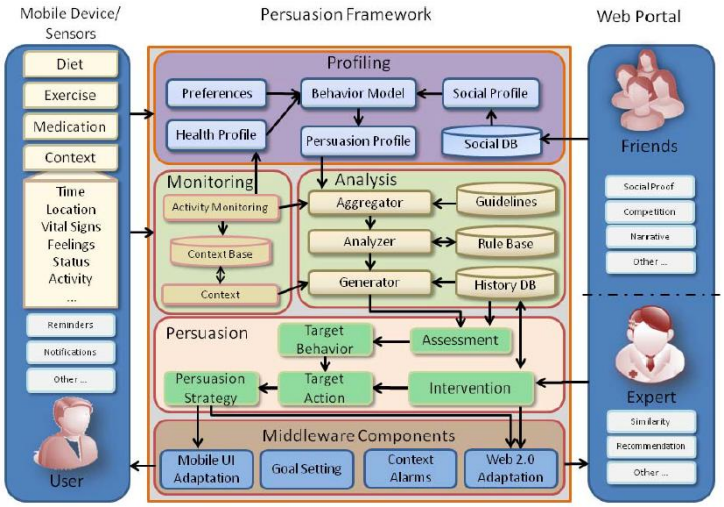


Figure 2.3 Framework for Intelligent Healthcare Self-Management [7]

Kosinski et al. [6] shows that there is a psychologically meaningful relationship between personality, website and website categories. According to this paper, extroverted users' frequent websites related to Music and Social, while Introverts prefer websites related to Comics, Literature, and Movies. Similarly, creative and liberal are attracted to blog, media, culture, astrology, eBooks and fine arts. Below is the table that shows the relationship between personality, web sites and website categories

Openness	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Liberal & Artistic	**Well Organized**	**Outgoing & Active**	**Cooperative**	**Emotional**
Arts.Animation	Reference.Education	Computers.Internet	Reference.Education	Recreation.Pets
Business.Marketing	Shopping.Electronics	Reference.Education	Computers.Internet	Recreation.Scouting
Business. Services	Shopping.Children	Science.Environment	Business.Logistics	Science.Physics
Arts.Photography	Reference.Dictionaries	Arts.Music	Health.Diseases	Sports.Hockey
modcloth.com	lww.com	clubzone.com	abebooks.com	cineplex.com
senate.gov	ecollege.com	ideeli.com	socialsecurity.gov	comparedby.us
boingboing.net	ecnext.com	thanksmucho.com	myrecipes.com	myprofilepimp.com
astrology-online.com	exct.net	discoveryeducation.com	bluemountain.com	barbie.com
gutenberg.org	education.com	list-manage.com	serialssolutions.com	yellowpages.ca
cafeastrology.com	kodak.com	trails.com	ecollege.com	biglots.com
...	...	...	...	...
gateway.com	candystand.com	lyricsty.com	localtribune.org	ncsu.edu
newegg.com	crunchyroll.com	fanfiction.net	funnyjunk.com	sheetmusicplus.com
fitnessmagazine.com	allthetests.com	behindthename.com	sciencebuddies.org	pitt.edu
ourtoolbar.com	bestuff.com	newworldencyclopedia.org	allthetests.com	highschoolsports.net
nhl.com	lyricsdepot.com	personalitypage.com	marvel.com	myrecipes.com
pier1.com	letmewatchthis.com	gaiaonline.com	supercheats.com	lww.com
Reference.Education	Health.Mental Health	Arts.Movies	Kids&Teens.Society	Arts.Photography
Arts.Television	Arts.Music	Shopping.Children	Health.Mental Health	Science.Maths
Sports.Soccer	Arts.Animation	Arts.Literature	Science.Physics	Business.Marketing
Shopping.Children	Arts.Literature	Arts.Comics	Recreation.Pets	Business.Logistics
Conservative	**Spontaneous**	**Shy & Reserved**	**Competitive**	**Calm & Relaxed**

Table 2.1 Websites and website categories with highest and lowest mean personality levels [6]

Jafari et al. [1] have presented web usage mining techniques for designing web recommender system. Problems and challenges in deploying recommender systems and solutions which address these problems have been proposed. Teng-Sheng Moh and Neha Sushil Saxena [23] have proposed a system that uses information from web usage mining, web semantics and the time spent on web pages to improve user recommendations. R Suguna and D Sharmila [24] have used Collaborative filtering approach, association rule mining and Markov model to recommend the web pages to the user. Chughtai et al. [21] have proposed a novel goal-based

filtering approach for recommender systems. k-nearestneighbor, collaborative filtering and content-based filtering techniques have been combined in this hybrid approach to increase the relevant recommendation accuracy and decrease the new-user profile (cold-start) issue. This goal based approach compare personalized profile preferences and gets the similarities between users. Rong Hu and Pearl Pu [22] presented a framework that incorporated users' personality information into the collaborative filtering framework to address the cold-start problem.

2.2 Related Browser Extensions

Different Chrome extensions are available that provide statistics regarding individual's time spent on browsing. Time stat [30] shows daily and monthly web usage statistics to the user. Rescue Time [31] provides detailed reports about the time spent on different applications, websites and categories. It allows users to set their daily goals to get them aware how productive they are.

2.2.1 Rescue Time

RescueTime makes people aware about their daily habits so they can focus and be more productive. Below figure shows the information that RescueTime displays such as time spent, productivity score and category information.

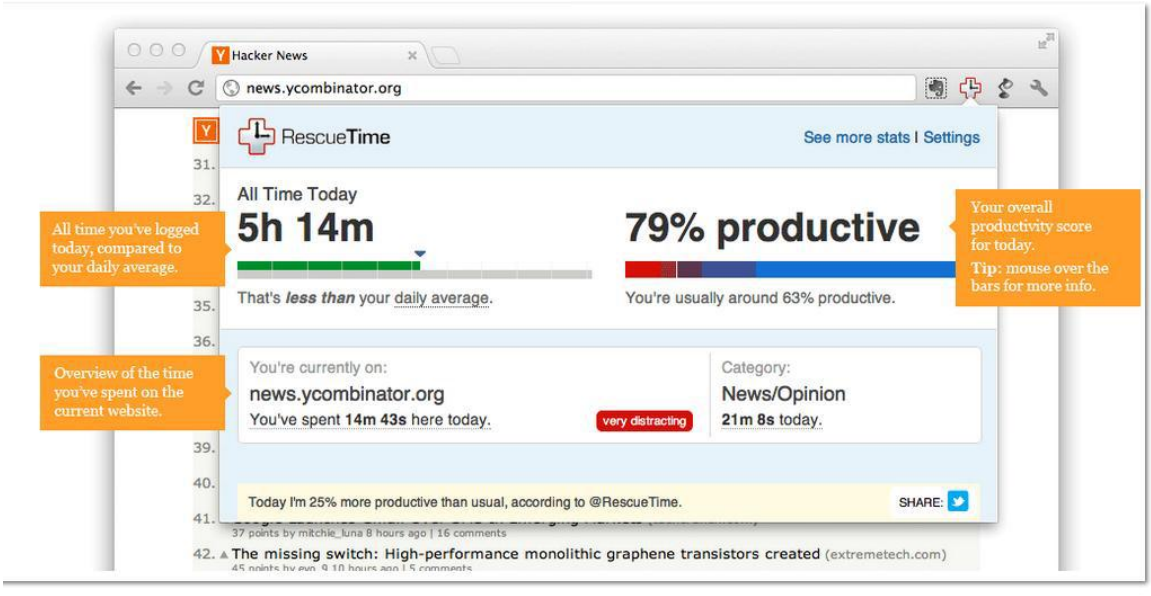


Figure 2.4 Chrome Extension –Rescue Time

Following are the features of RescueTime

1. Block out the distracting websites.

2. Show alerts to the user about the productive and distracting time.
3. Keep track if user away from the computer
4. Log daily accomplishments.
5. Display visualization related to daily usage

There are some drawbacks of rescue time that have been mentioned in the study [29]. According to the study, Reason behind the failure of rescue time is insufficiency of data collection. Comprehensive data collection is required to accurately measure qualitative data. RescueTime uses only one dimension to analyze productivity. Productivity with single dimension will lead to inaccuracy [29].

2.2.2 TimeStats

TimeStats [30] log the website stats and show to the individual daily and monthly statistics, most visited domains, busiest day and categories.

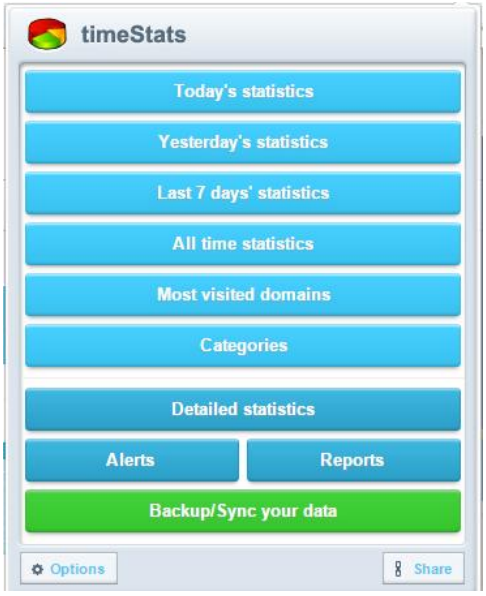


Figure 2.5 Chrome Extension - TimeStats

Accuracy in time spent calculation must be assured. Time Stats doesn't show accurate results as sometimes it happens that time spent at other applications on computer gets added to the browser usage. This occurs in case when browser window is maximized but not active and user is busy using other applications on computer.

A great deal of work has been done in web usage mining. Different browser add-ons are available that keep track of user browsing activities and provides daily reports regarding

usage statistics but motivating user to use these add-ons on regular basis is still a challenge. People will definitely stop using add-on in case of inaccurate and wrong results. There are following limitations we have found out in existing applications that irritate user i.e. lot of user intervention in an application, blocking websites, inaccuracy and wrong productivity report. Aim of our thesis is to motivate user towards self awareness and provide them comprehensive analysis and recommendation of their browsing activities and usage statistics using interactive visualization techniques. Social comparisons are also provided to motivate users and drive competition among them. Zhang et al. [28] suggest that a system for sharing some aspects of web activity publicly may be interesting and viable to users.

Chapter 3

Methodology

This chapter gives an overview of dataset, chrome extension and web data mining techniques. The first section explains the dataset that was utilized to extract the patterns. Collection of data attributes has been identified initially that is related to user web behavior and activities. Collected data has been stored in our personalize database. The section 3.2 gives the brief description of the chrome extension. Pattern extraction and visualization techniques have been described in section 3.3.

3.1 Feature Modeling

Browsing data history is maintained by all browsers that provide information that how often user requested a page but unable to capture how long the user stayed on the page. Our proposed system doesn't use the browser history logs. Its data collection module efficiently runs in the background of browser and autonomously captures a wide range of browsing information. To infer user's context and behavior, behavioral data features such as websites usage, computer usage, sessions, and tabs switching data have been identified and collected.

$$\boldsymbol{Features} = \{\boldsymbol{sessions}, \boldsymbol{tabs}, \boldsymbol{browser\ states}, \boldsymbol{websites\ usage}\}$$

Sessions and tabs data can infer the user's behavior regarding how often user switch the tab, how long the session is and how many tabs created in a session etc. Attributes for sessions and tabs are mentioned below,

$$\boldsymbol{Sessions} = \{\boldsymbol{session_{id}}, \boldsymbol{start\ time}, \boldsymbol{end\ time}\}$$

$$\boldsymbol{Tabs} = \{\boldsymbol{tab_{id}}, \boldsymbol{window_{id}}, \boldsymbol{session_{id}}, \boldsymbol{creation\ time}, \boldsymbol{close\ time}, \boldsymbol{transition\ type}, \boldsymbol{switchTo_{tabid}}\}$$

Websites usage data helps in analyzing user behavior that how much time user spent at a particular website, how often user clicks that website, what is the peak browsing time of the user. It infers user interests and mental well-being.

$$\boldsymbol{Websites\ Usage} = \{\boldsymbol{url}, \boldsymbol{timespent}, \boldsymbol{date}, \boldsymbol{time}\}$$

Browser window states are mentioned below.

$$\boldsymbol{Browser\ window\ states} = \{\boldsymbol{idle}, \boldsymbol{focus}, \boldsymbol{not\ focus}, \boldsymbol{lock}\}$$

Where 'focus' state represents that browser window is maximized, active and on top of other desktop windows. 'Idle' shows inactive time. 'Lock' is the state when computer locks down or on standby.
Idle time of browser is calculated when the browser window is not focused or if window is focused but idle or locked.

$$\boldsymbol{Browser\ Idle\ time} = \boldsymbol{time}_{\boldsymbol{not\ focused}} + \boldsymbol{time}_{\boldsymbol{focused}\ \&\ idle} + \boldsymbol{time}_{\boldsymbol{focused}\ \&\ locked}$$

Computer usage is how long user stays at computer while browser is running. Computer idle time is calculated by adding the time how long the computer stay standby, locked or idle.

$$\boldsymbol{Computer\ Idle\ time} = \boldsymbol{time}_{\boldsymbol{standby}} + \boldsymbol{time}_{\boldsymbol{idle}} + \boldsymbol{time}_{\boldsymbol{locked}}$$

These features can infer different behavior dimensions of a user as described in Table 3.1.

Features	Attributes	Behavior
Sessions	id, start time, end time	Session Time Span, Sessions per day
Tabs	id,window_id, session_id, creation time, close time, transition type,switchTo_tabid	No of clicks, time spent, tab switching time, tabs per session
Websites usage	url, timespent, date, time	User interests at particular website category at particular time of the day
Browser states	idle, focus, not focus, lock	Browser idle time, Computer usage

Table 3.1 Attributes of web browsing data that help in inferring behavior

3.2 Developing Chrome Extension

Aim of our proposed system is to promote self-awareness among individuals about their browsing behavior trends by providing web usage statistical visualizations. To achieve this goal, browser add-on has been implemented that logs the data, send the data to the server, display different browsing statistical charts, provide behavior comparisons among individual's social circle and propose the recommendations.

Chrome Extension can modify and enhance the functionality of chrome browser. It contains persistent background page that holds the main logic and runs silently in the background when browser is running. Data collection and data transfer logic has been implemented in this background page. Extensions can also contain other HTML pages that display the extension's UI. BBA user interface contains the web pages that display the user different browsing behavior trends.

3.3 Web Usage Mining

Web usage mining is the data mining technique to discover web usage behavior patterns from web data. Figure 3.1 shows the process of web usage mining. It comprises of three phases, i.e. preprocessing, pattern discovery, and pattern analysis [13]. Focus of this thesis is on pattern discovery and analysis techniques. There is variety of pattern discovery techniques including associative rule mining, sequential pattern mining, classification, and clustering, that discover the correlations among Web pages, sequential patterns over time intervals and clustering the users according to their access patterns.

Visual data mining techniques have proven to be of high value in exploratory data analysis [12]. There are large numbers of 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 data as possible in one plot. Visualization allows the user to mine and gain insight into the data and come up with new mining recommendations.

Pattern discovery techniques and visual data mining techniques have been discussed in next subsections 3.4 and 3.5.

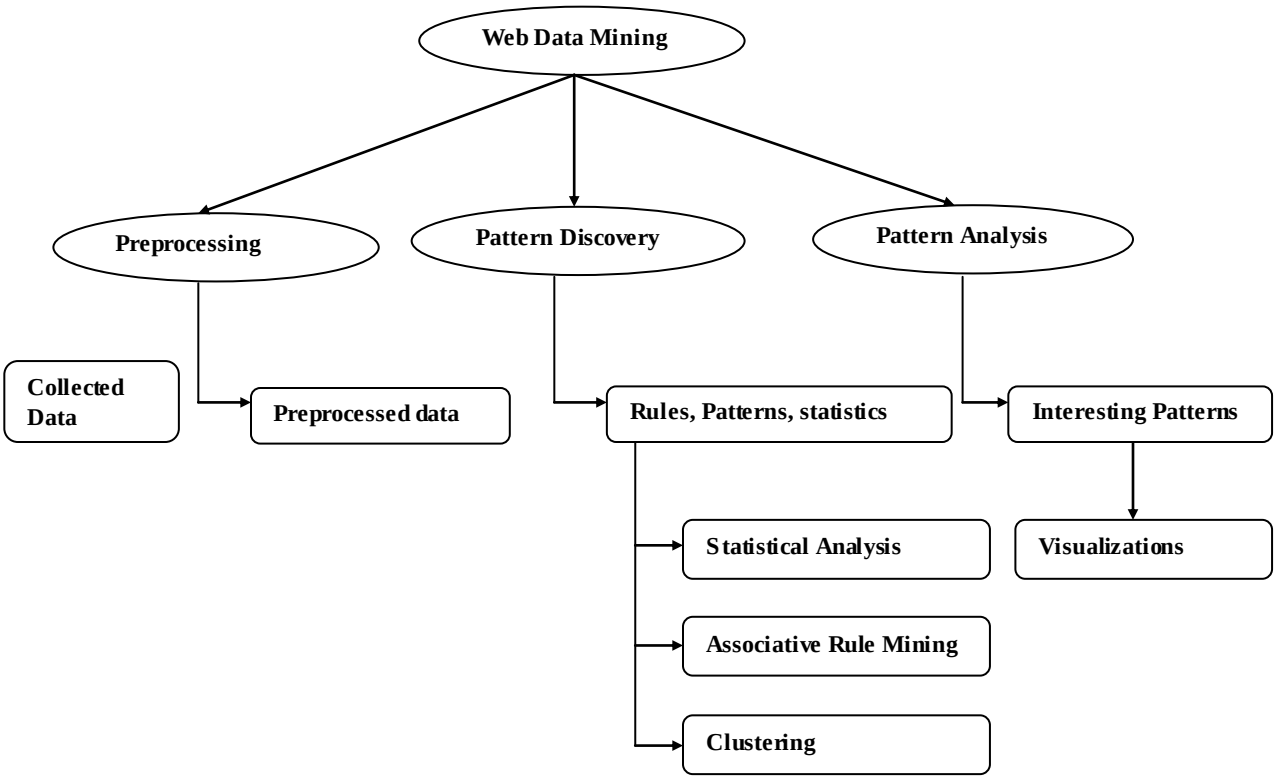


Figure 3.1 Web usage mining process

3.4 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 web usage data. Different kind of descriptive statistical analysis e.g. frequency, count, min, mean, max, median, mode etc. can be performed on the data attributes like page views, time spent at particular page, frequently accessed pages, tabs switching time, number of sessions per day , session time span ,number of tabs per session etc.

Associative Rule is used to find out the frequent items which are used together. Association or correlation rules are measured by its support, confidence and correlation.

Association rule is the correlation between the two item sets of the form A→ B where web site/ web category is referred as an item.

Support measures how frequently all items in the item set occur together in the set of sessions in data set that contain A∪B. Data set is a set of sessions containing all websites/website-categories browsed by a user during a browser session. Browser session is a continuous period of user activity in the browser. It breaks as the browser gets inactive for a period of time.

Confidence is percentage of sessions in dataset containing A that also contain B.

$$\boldsymbol{Confidence}\,(\boldsymbol{A}\rightarrow\boldsymbol{B})=\boldsymbol{P}(\boldsymbol{B}|\boldsymbol{A})=\boldsymbol{support}(\boldsymbol{A}\cup\boldsymbol{B})/\boldsymbol{support}(\boldsymbol{A})$$

Lift is a correlation measure and can be computed as

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

Mining, 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. K-means [16] is a well-known algorithm that efficiently clusters large data sets. It works well on numeric data but cannot cluster categorical data. To calculate dissimilarity between two objects, Euclidian distance formula has been used.

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Cost function of K-means is mentioned below

$$C(U) = argmin \sum_{i=1}^{k} \sum_{j=1}^{n} (\| x_j - v_i \|)^2$$

Where '$||x_j - \mu_i||$' is the Euclidean distance between x_j and μ_i. 'n' is the number of data points in i^{th} cluster. k' is the number of cluster centers.

Clustering technique (K-means) is used to find the similarity among users on the basis of different web category usage. It helps to measures the comparison on the basis of web usage for each user at different time of the day.

3.5 Visual Data Mining Techniques

Information visualization and visual data mining can help to deal with the flood of information [12]. 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 non-homogeneous data. No understanding of complex mathematical or statistical algorithms or parameters is required.

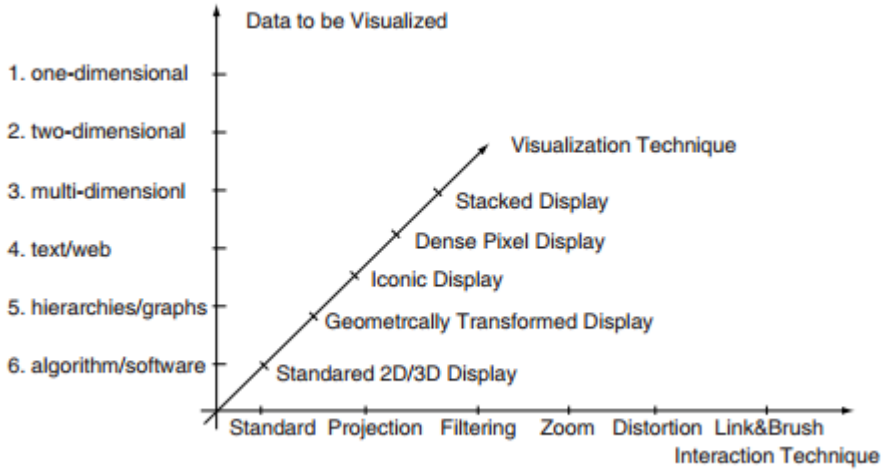


Figure 3.2 Classification of Information Visualization Techniques [12]

Figure 3.2 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 [19]. The visualization technique used may be classified as standard 2D/3D displays, such as bar charts, x-y plots, heat map, parallel coordinates[20], 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 axes 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** technique is 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.

Chapter 4

Browsing Behavior Analysis (BBA)

Browsing behavior analysis (BBA) is a chrome extension that collects and displays the browsing data, sends it to the server where individual's web browsing and mobile activities data from different devices get integrated to display the aggregate web and mobile usage statistics, allow individuals to add people in their social circle, determine and display the web usage comparison among the social circle, group the similar users on the basis of categorical web usage and provide recommendations to them.

4.1 Design Requirements

BBA addresses the following questions and provide the detailed information about:

- How much time user spends on computer and browser?
- How long user remains idle?
- How long user stays on a particular web page or category?
- What are the browsing peak time, top website and top category of the day/month?
- How often user switch between the tabs?
- How many tabs user open during a session?
- How many sessions user open during a day?
- How long users stay on a session?
- How one navigates between pages (e.g. by clicking on hyperlinks, typing url , reloading page etc)
- How many people user interact with via SMS and Calls?
- How user behavior differs from others in his social circle (calculate on the basis of categories and peak time)?
- Which group and cluster user lie in?
- What are the interests and personality of users who are similar to you?

4.2 System Architecture

Figure 4.1 shows how this system works. Data is logged on the trigger of different browser events such as creating of the tab/window, updating the tab/window, closing tab/window, status of window changes etc. Collected behavioral data with individual's identity information i.e. email id and device name is transferred to the server automatically. Add-ons stores the data on local storage for a whole day and send set of data as a bulk to the server as the browser gets active first time during the day. Behavioral data is further processed to extract significant behavior patterns such as browsing peak time, peak hour, top categories, top websites, user interests, frequent tab switching, number of sessions per day etc. These behavior patterns are then presented to the individual through interactive visualizations. Individual's friends' browsing data is retrieved from the server and display to him at client end where individual can compare his performance among his social circle.

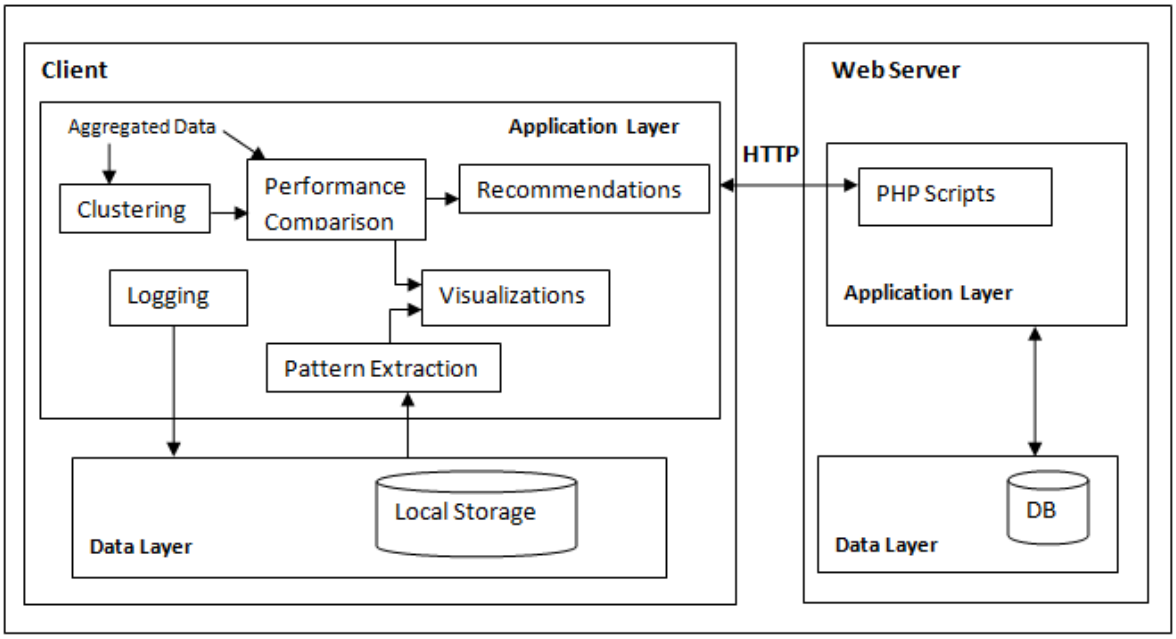


Figure 4.1 System Architecture of BBA

System Architecture has five modules. Brief descriptions of modules are described below.

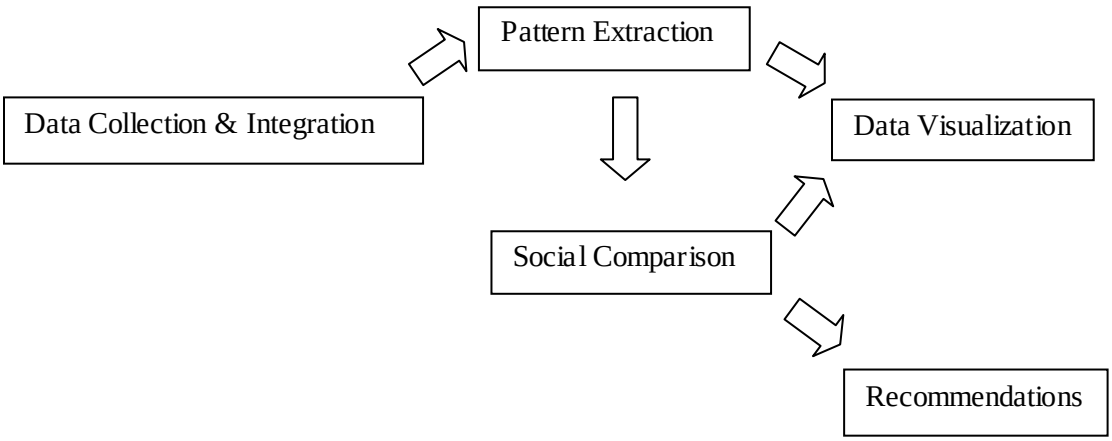


Figure 4.2 Modules of Proposed Framework

Browsing data collection and integration is described in section 4.3. Section 4.4 explains how behavior patterns are extracted. Visualizations have been demonstrated in section 4.5 and social comparison is presented in section 4.6.

4.3 Browsing Data Collection and Integration

Data logging process of BBA is described in the below section.

4.3.1 Data Logging

Behavioral data is logged as the browsing events trigger. Browsing events include, e.g., creating/updating/closing of tabs and changing of window states i.e. idle, not focused, focused, open, close. Behavioral data comprises of websites usage, sessions, tabs details and computer usage. Dwell time of each page visit is calculated on the basis of consecutive page visits with in the session. Last page dwell time is calculated at the start of the next session.

Algorithm 4.1(a) : Data Logging - Activity Calculation at Browser Events

```
chrome.tabs.onActivated.addListener(function tab
 {
          if (tab.active)
          Calculate_ DwellTime(tab);
 });

chrome.tabs.onUpdated.addListener(function(tabId, changeInfo,tab) {

     if (changeInfo.status == 'complete' && tab.active) {
          Calculate_DwellTime(tab);

}
```

Algorithm 4.1 (a) Data Logging-Activity calculation at browser events

Algorithm 4.1(b): Data Logging- Dwell Time Calculation

```
Calculate_ DwellTime (tab)
  timespent=currenttime-previoustime;
  idletime=idletime_focused+idletime_notfocused+lockedduration;
  timespent=timespent-idletime;
    for each (webusage_data ) do begin
      url← webusage_data [i].url;
      if(url==previous_selected_url)
          webusage_data [i].timespent=timespent;
          webusage_data[i].lastvisit= previous_visit_time;
          found=1;
          break; end;
      if(found==0) begin
        data={};
        data["url"]=previous_url;
        data["timespent"]=timespent;
        data["lastvisit"]=previous_visit_time;
        webusage_data.push(data); end;
      if(dateMatch) begin
        daily_total_timespent=daily_total_timespent + timespent;
        if(!hourMatch)
            UpdateDailyhours_timespent(daily_hour_timespent);
        else
          daily_hour_timespent= daily_hour_timespent+timespent;
      else
        webusage_data ←UpdateTimeSpent_day(daily_total_timespent);
        TransferData(webusage_data , email); end

  previous_visit_time=currenttime;
  previous_selected_url= getsubdomain(tab.url);
  idletime_focused=0;idletime_notfocused=0;lockedduration=0;
```

Algorithm 4.1 (b) Data Logging- Dwell time calculation

Algorithm 4.1(c): Data Logging - Idle Time Calculation

```
updateCounterInterval=1000*60;
idletime_focused=0;
lockeddur=0;
idletime_notfocused=0;
chrome.idle.onStateChanged.addListener(checkIdleTime);
window.setInterval(updateCounter, updateCounterInterval);
checkIdleTime(newState)
        if (newState=='idle')
                status=idle;
         if (newState=='locked')
                status=locked;
updateCounter()
         if(status==idle)
              idletime_focused= idletime_focused+1;
        if(status==locked)
             lockeddur= Lockeddur+1;
        if(!window.focused)
            idletime_notfocused= idletime_notfocused+1;
```

Algorithm 4.1 (c) Data Logging- Idle time calculation

4.3.2 Data Transfer

Logged data is sent via HTTP POST requests to PHP scripts residing on the BBA backend server. These PHP scripts insert the data into database. Web pages daily usage data is transferred when the browser window get active and last transfer date doesn't match with the current date. Extension continuously checks data transfer status after each 2 hours and in case of failure, data is resent again. Tabs switching data is transferred to server at the startup of next session but if session lasts for more than 2 hours, data is sent during the session to avoid any failure that can occur in sending large amount of data.

4.3.3 Data Integration

At the server end, data sent from different devices such as laptops, desktops, iPad, etc (machines where chrome extension is installed) get integrated on the basis of user's email.

4.4 Behavior Extraction

4.4.1 Pattern Discovery Algorithms

Algorithms we have used to extract behavior patterns are the following. Algorithm 4.2 extracts the web category on which user spent most of his time during the day.

Algorithm 4.2: Top Category of the day

CalculateTimeSpent_category(): Calculate time spent for all websites categories browsed in last 30 days.

GetTopCategory(date): returns the category on which user spent his maximum time.

webusage_data: Web sites browsed (based on last month data)

daily_usage: Time spent at different dates of the month.

timespent: time spent at particular date for particular web site

hm_catg: hashmap, save time spent in hashmap to quickly access the time spent at particular date and category.

Categories= websites categories saved in db.

```
CalculateTimeSpent_category()
  for each (webusage_data) do begin
     category ←GetCategory(webusage_data.url);
     daily_usage ← webusage_data. daily_usage;
     for each(daily_usage) do begin
           timespent ← daily_usage.timespent;
           date ← daily_usage.date;
       hm_catg[date][category]=hm_catg[date][category]+timespent; end; end;
GetTopCategory(date)
  for each (categories) do begin
       category←categories.category;
       timespent_category[category]=category;
       timespent_category[timespent]= hm_catg[date][ category];end;
  return Max(timespent_category);
```

Algorithm 4.2 Top Category of the Day

Algorithm 4.3 finds the peak epoch of the day on which user spent his maximum browsing time.

Algorithm 4.3: Calculate peak time of day

CalculateTimeSpent_timeofDay(): Calculate time spent at different time of day for all websites browsed in last 30 days.

Getpeaktime(date): returns the time on which user spent his maximum browsing time.

webusage_data: Web sites browsed (based on last month data).

dailyhours_timespent: Time spent at different hours of the day.

hour: hour of the day when particular website is browsed.

timespent: time spent at particular hour for particular web site.

hm: hashmap, save time spent in hashmap to quickly access the time spent at particular date and time.

Times_of_day={ Early Morning, Morning, afternoon, evening, night, midnight }

```
CalculateTimeSpent_timeofDay()
  foreach (webusage_data) do begin
     dailyhours_timespent← webusage_data. dailyhours_timespent
     foreach (dailyhours_timespent)          do begin
        timespent←dailyhours_timespent.timespent;
        hour← dailyhours_timespent.hour;
        date← dailyhours_timespent.date;
        if (hour > 5 && hour < 9)
          hm[date]["EarlyMorning"]=hm[date]["EarlyMorning"]+timespent;
        else  if (hour > 8 && hour < 12)
          hm[date]["Morning"] = hm[date]["Morning"] +  timespent;
        else if (hour > 11 && hour < 17)
          hm[date]["AfterNoon"] = hm[date]["AfterNoon"] + timespent;
        else if (hour > 16 && hour < 21)
          hm[date]["Evening"] = hm[date]["Evening"] + timespent;
        else if (hour > 20 && hour <= 23)
          hm[date]["Night"] = hm[date]["Night"] + timespent;
         else if(hour >=0 && hour <=5)
          hm[date]["MidNight"] = hm[date]["MidNight"] + timespent; end; end;

GetPeakTime(date)
  foreach (times_of_day) do begin
       timespent_timeofday[times_of_day]=times_of_day;
       timespent_timeofday[timespent]= hm[date][ times_of_day];end;
  return Max(timespent_timeofday);
```

Algorithm 4.3 Peak time of the day

Apriori algorithm is an algorithm for mining frequent itemsets for boolean association rules.

Algorithm 4.4 : Apriori Algorithm

Ck: Candidate itemset of size k

Lk : frequent itemset of size k

L1 = {frequent items};

for (k = 1; Lk !=∅; k++) **do begin**

Ck+1 = candidates generated from Lk;

for each transaction t in database **do**

increment the count of all candidates in Ck+1 that are contained in t

Lk+1 = candidates in Ck+1 with min_support **end**

return k Lk;

Algorithm 4.4 Apriori Algorithm

4.4.2 Websites Categorization

Web URLs are grouped into various categories, such as social networking, research and development, news media, career and education etc. Some of the important web categories, sub categories are mentioned in Table 4.1. Website categorization API [34] has been used to automatically retrieve category and subcategory for the particular web site via HTTP request.

Art and Entertainment	**Research and Development**	**Career and Education**	**Searching and Sharing**
TV and Video , Movies Photography , Shopping Food & Drink , Travel Games, Sports, Health	Support Development Forums	Jobs and Employment Universities/Colleges Education	Search Engine File Sharing
youtube.com, dailymotion.com playit.pk , www.imdb.com flickr.com, photobucket.com pizzahut.com,hangcheng.com bbcgoodfood.com amazon.com,aliexpress.com candycrushsaga.com emirates.com psychologytoday.com, fifa.com,espncricinfo.com	code.shutterstock.com stackoverflow.com developer.android.com www.w3schools.com www.codeproject.com androidcodeexamples.blogspot.com, play.google.com	naukri.com , rozee.pk www.mit.edu www.nust.edu.pk www.stanford.edu www.academia.edu www.code.org www.turnitin.com www.lms.nust.edu.pk	www.google.com www.bing.com www.googlescholar.com www.ask.com www.dropbox.com www.imgur.com
Email	**Social**	**Science**	**News and Media**
Email	Social Networking Sites	Environment, Social Sciences Engineering and Technology	News Papers News Channel
www.hotmail.com www.gmail.com www.live.com www.yahoo.com www.mail.seecs.edu.pk mail.google.com	www.facebook.com www.linkedin.com www.twitter.com www.instagram.com plus.google.com skype.com	www.translit.net www.treehugger.com www.cnet.com, en.wikipedia.org www.technologyreview.com www.digitaltrends.com	www.nytimes.com www.dawn.com www.dailymail.co.uk www.bbc.com www.geonews.com www.geo.tv, cnn.com

Table 4.1 Web pages, categories and sub categories

4.4.3 Frequent Categories/Websites and their Correlation

Apriori algorithm has been used to get the frequent categories. It extracts the categories that are frequently used together. Before applying this algorithm, we need to set the minimum threshold for support and confidence. The ideal minimum threshold for support is set to be 10% and for confidence is set to be 45% for web usage mining [26]. We have supposed that an item set is frequent if it appears in at least 10% of the total sessions. For example, 5 is the support threshold for 50 sessions. First step is to count up 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 considered to be frequent. Below mentioned table shows the frequent categories of a user, software development is frequent category and occurs in all the sessions. Social comes after software development that have appeared in 50% of the sessions.

Frequent Items	**No of occurrences**
{software development}	50
{social}	25
{searching and sharing}	20
{email}	15
{Art and Entertainment}	10
{software development, social}	25
{software development, searching & sharing}	20
{software development, email}	15
{software development, Art and Entertainment}	10

Table 4.2 Frequent categories and number of occurrences

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 => Software Development.

Its confidence is $\frac{Support\,(Social \cup SoftwareDevelopment)}{Support\,(Social)} = \frac{0.5}{0.5} = 1$ which means software development occurs in all the sessions containing social.

To find the correlation among the categories,

$$Lift\,(Social \rightarrow SoftwareDevelopment) = \frac{P(Social \cup SoftwareDevelopment)}{P(Social)\,P(SoftwareDevelopment)} = \frac{0.5}{(0.5*1)} = 1$$

It shows that social websites and software development are used together. Recommendation can be proposed here by analyzing whether social networking affecting the productivity of user or not.

4.4.4 Predicting User Personality

Personality of the user can be predicted by his browsing websites and categories. We have considered three personality traits here i.e. curious, workaholic, social. For example curious people browse the categories i.e. news, science and search engine. Workaholic people browse categories such as research and development, email and career and education. Social person browses social networking websites.

Curious	**Workaholic/Career Oriented**	**Social**
News and Media Science Search Engine	Research and Development, Email, Career and Education	Social Network
www.translit.net **www.treehugger.com** **www.cnet.com,** **en.wikipedia.org** **www.technologyreview.com** **www.digitaltrends.com** **www.nytimes.com** **www.dawn.com** **www.dailymail.co.uk** **www.bbc.com** **www.geonews.com** **www.geo.tv, cnn.com** **www.google.com** **www.bing.com** **www.ask.com**	**googlescholar.com** **naukri.com , rozee.pk** **www.mit.edu** **www.nust.edu.pk** **www.stanford.edu** **www.academia.edu** **www.code.org** **www.turnitin.com** **www.lms.nust.edu.pk** **code.shutterstock.com** **stackoverflow.com** **developer.android.com** **www.w3schools.com** **www.codeproject.com** **androidcodeexamples.blogspot.com,** **play.google.com**	**www.facebook.com** **www.twitter.com** **www.instagram.com** **plus.google.com** **skype.com** **linkedin.com**

Table 4.3 Relationship of personality and website categories

4.4.5 Finding Similar Users on the basis of Categorical Web Usage

K-means clustering algorithm has been used to group the similar users. Users in same cluster show that users have same browsing behavior. Vectors with six dimensions i.e. Art and Entertainment, Others, Research and Development, Career and Education, Searching and Sharing and Email are created for each user. It would be like this:

$$V = (V_{art\&development}, V_{research\&developemnt}, V_{career\&education}, V_{email}, V_{others})$$

Where magnitude of each vector component is the percentage of time spent at this dimension. Euclidian distance is calculated between centroid of each cluster for each vector and vector is assigned to the cluster having minimum distance from it.

Java script package ‘figue.js’ [32] has been used to cluster the data. The **kmeans** function takes as input the number of clusters ‘k’ and a list of input vectors ‘N’ and it returns an object with two attributes:

- **Centroids**: an Array of ‘k’ vectors containing the centroid of each cluster
- **Cluster indexes**: An Array of size ‘N’ representing for each input vector the index of the cluster.

The kmeans function will return null if:

- N < K
- The number of *different* input vectors is smaller than K

KMeans Implementation [32]

```
function kmeans (k, vectors) {
    var n = vectors.length ;
    if ( k >= n )
      return null ;

        // randomly choose k vectors among n entries
     var centroids = new Array(k) ;
     var selected_indices = new Object ;
     var cluster = 0 ;

     while (cluster < k) {
       var random_index = Math.floor(Math.random()*(n)) ;
       if (random_index in selected_indices)
           continue
         selected_indices[random_index] = 1;
         centroids[cluster] = vectors[random_index] ;
         cluster++ ;
         }

       var assignments = new Array(n) ;
       var clusterSizes = new Array(k) ;
       var repeat = true ;
       var nb_iters = 0 ;

       while (repeat) {
         // assignment step
         for (var j = 0 ; j < k ; j++)
           clusterSizes[j] = 0 ;
           for (var i = 0 ; i < n ; i++) {
             var vector = vectors[i] ;
             var mindist = Number.MAX_VALUE ;
             var best ;
             for (var j = 0 ; j < k ; j++) {
               dist = euclidianDistance (centroids[j], vector)
               if (dist < mindist) {
                 mindist = dist ;
                 best = j ;
                   }
```

```
                }
                clusterSizes[best]++ ;
                assignments[i] = best ;
                 }



// update centroids step

    var newCentroids = new Array(k) ;
    for (var j = 0 ; j < k ; j++)
      newCentroids[j] = null ;
      for (var i = 0 ; i < n ; i++) {
       cluster = assignments[i] ;
       if (newCentroids[cluster] == null)
          newCentroids[cluster] = vectors[i] ;
          else
 newCentroids[cluster] = addVectors (newCentroids[cluster] , vectors[i]);
       }

       for (var j = 0 ; j < k ; j++) {
newCentroids[j] = multiplyVectorByValue (1/clusterSizes[j],newCentroids[j])
                }
          // check convergence
          repeat = false ;
          for (var j = 0 ; j < k ; j++) {
            if (! newCentroids[j].compare (centroids[j])) {
                repeat = true ;
                break ;
             }
          }
         centroids = newCentroids ;
         nb_iters++ ;
         if (nb_iters > figue.KMEANS_MAX_ITERATIONS)
          repeat = false ;

       }
  return { 'centroids': centroids , 'assignments': assignments} ;

     }
```

Algorithm 4.5 K-Means Algorithm

4.4.6 Predicting User Interests

Website's visits frequency and duration are two major metrics of a user interest in a website, [15]. We will consider these metrics to estimate the user interest. Duration is measured based on dwell time normalized by maximum dwell time.

$$Duration_{category} = \frac{dwellTime_{category}}{\max \forall_{category \in categories} dwellTime_{category}}$$

Frequency is measured based on number of visits of category normalized by maximum number of visits.

$$Frequency_{category} = \frac{vists_{category}}{\max \forall_{category \in categories} visits_{category}}$$

Harmonic mean is used to mitigate the impact of large outliers and aggravate the impact of small ones.

$$\frac{2 * Duration_{category} * Frequency_{category}}{(Duration_{category} + Frequency_{category})}$$

4.5 Data Visualization

- ***Daily Usage Visualization***

In figure 4.3 (a) different websites browsed during last 15 days are visualized. The size of the bubble represents time spent at a particular website. Time is shown across vertical axis and date is shown along horizontal axis. Color shows the category a website belongs to. Colors of the bubbles also help the users to identify the most frequently browsed websites and websites categories. User can analyze the time spent at different websites by the size of bubble and get to know at which site and category he spent most of his time. This visualization also helps in detecting daily patterns e.g. at what time of day, a person browses which sites, does the user browse any website daily at the same time and how his browsing affecting his performance.

Figure 6a provides an insight about the categories that user browsed during last days i.e. Social Network, Chrome Extension Development, Server End Development, Tv and Video, Programming help, email and search engine. This figure shows that user spent his most of the time on the categories i.e. social network, Tv and Video and chrome Extension development and least time spent on email, programming help and search engine. It also indicates that a person spent most of his time on browsing on Thursday 2nd july.

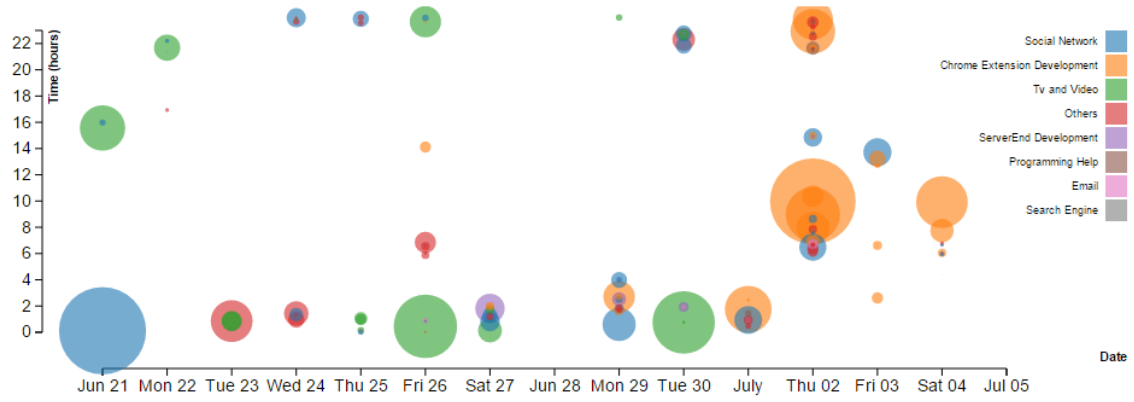


Figure 4.3(a) Daily usage visualization

User can get the details by placing mouse over the bubble. According to the visualization in figure 4.3 (b), maximum time spent on that extension that comes under the category i.e. chrome extension development.

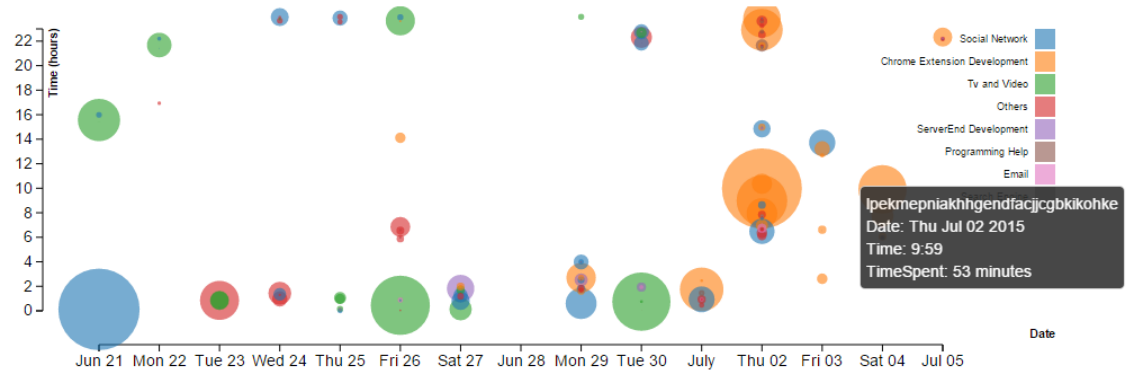


Figure 4.3 (b) Daily usage visualization tooltip

- ***Browsing Usage at different times of day***

In Figure 4.4, browsing usage at different times of the day can be visualized. Six epochs of the day have been considered here. Distinct color has been assigned to each part of the day. Date and duration has been shown along x-axis and y-axis respectively. Time spent during the particular date can be seen right above the bar. This visualization helps user in finding the peak time during a day and repetitive pattern during the last 7 days. For example, the below figure shows that user had approximately the same pattern from june 27 to 1st july as he spent most of his time in browsing during midnight. From the 2nd of the july, user's pattern is changed and peak time during this day is early morning.

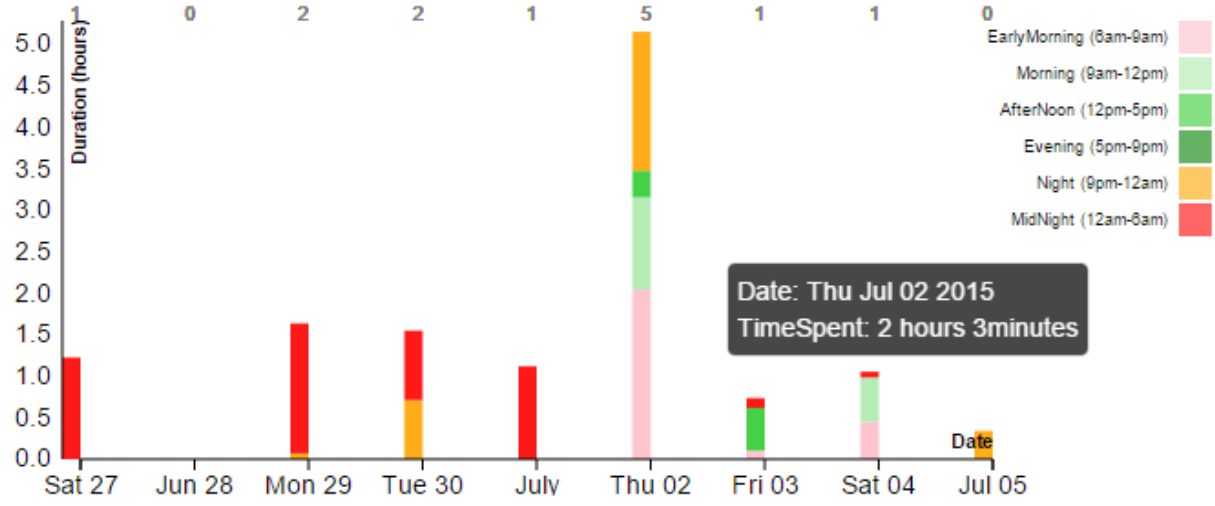


Figure 4.4 Web Usage at different time of day

Below figure 4.5 shows the time spent at different hours of the day july thu 02.Here peak hour of the day can be found i.e. 9AM.

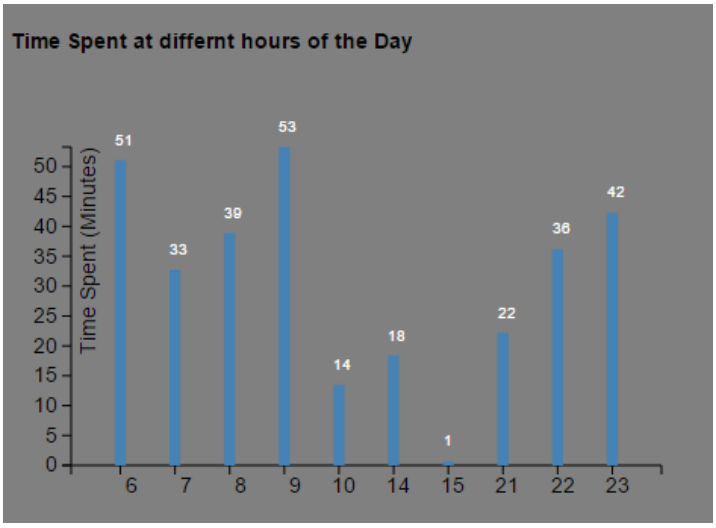


Figure 4.5 Web Usage at different hours of day

- ***Categorical Web Usage***

Figures 4.6, 4.7 and 4.8 show the time spent on categories, subcategories and websites. Inner circle represents categories, outer circle represents subcategories and by clicking on the outer circle websites can be visualized. According to below figures, user has spent most of his time on two categories i.e. Software development and social. Under the category, software development, user spent time on two of its subcategories, chrome extension development and server end development.

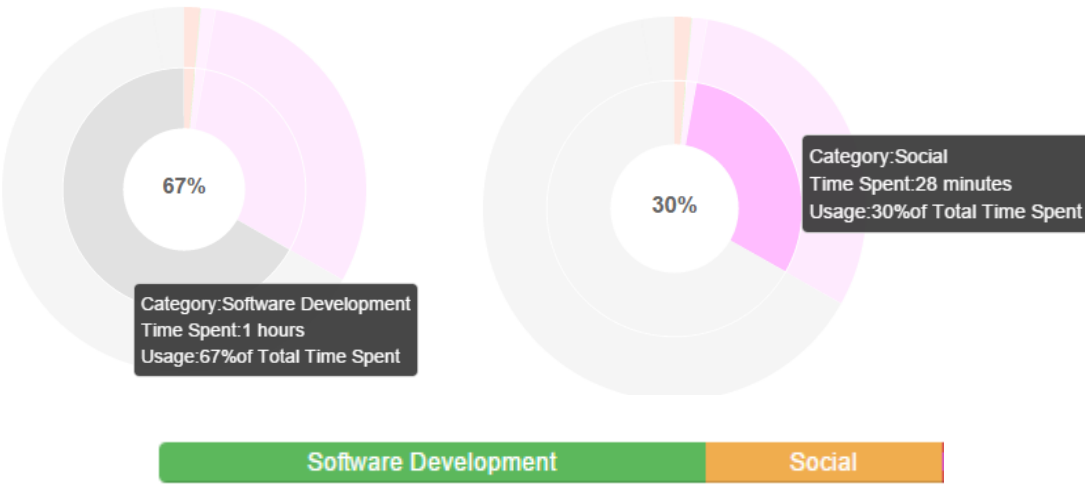


Figure 4.6 Categorical usage - Top Categories of day

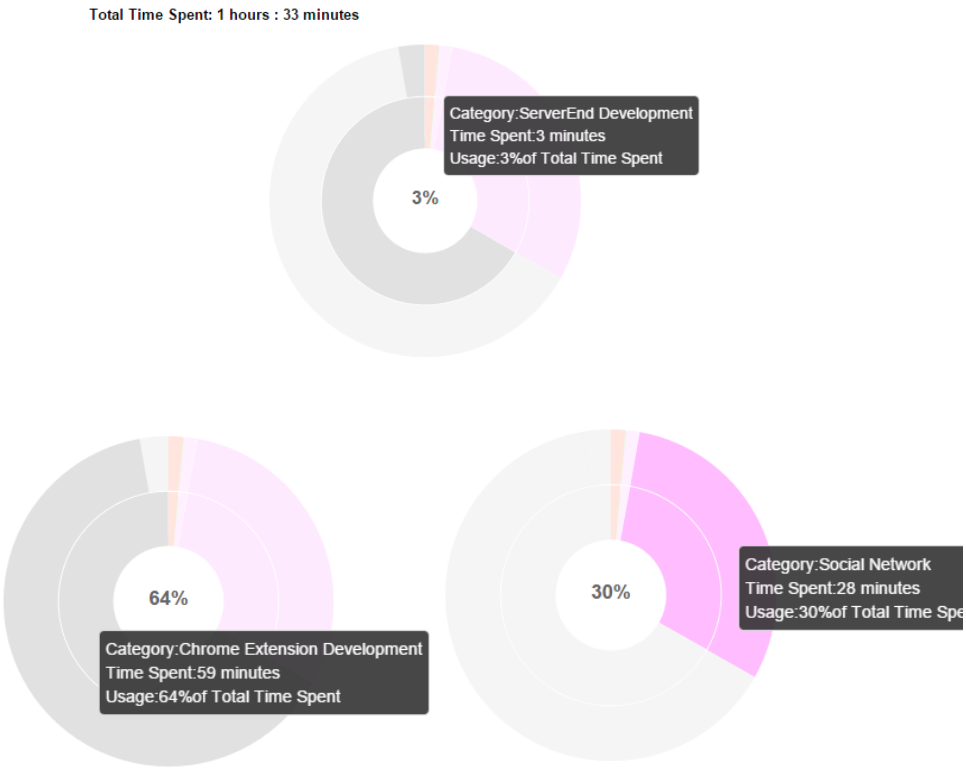


Figure 4.7 Outer layer- subcategories usage

In case of social category, there is only one subcategory under social i.e. social network where time spent on. By clicking on social network it can be seen in figure 4.8 that user has browsed two websites under this category i.e. facebook and linkedin.

Figure 4.8 Websites usage

- ***Weekly usage comparison on the basis of website categories***

In Figure 4.9, web usage for week1 and week2 are visualized respectively. These graph shows interests areas of user as most visited categories are same during these weeks. According to this visualization, it can be inferred that user was busy during week2 due to some deadline or exams and fig4 shows that his routine is not busy this week as he spent most of his time on watching movies, tv or videos.

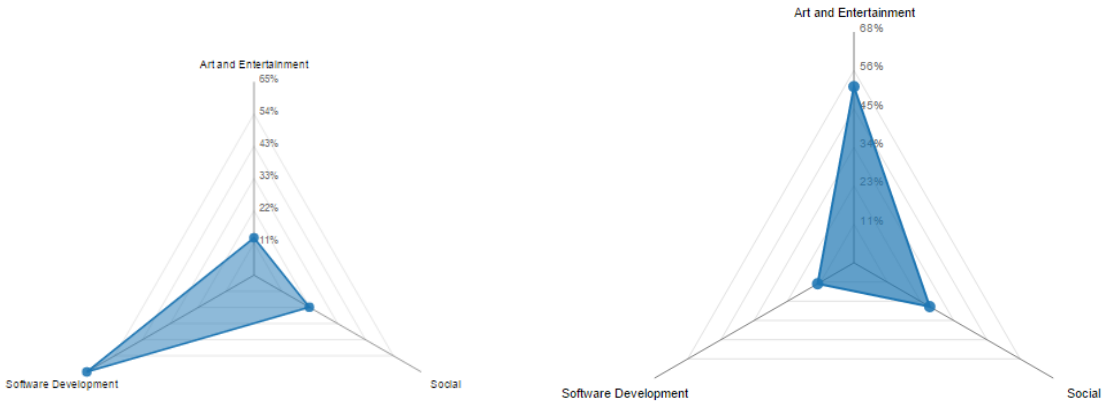


Figure 4.9 Weekly usage Comparison on the basis of website categories (time spent >10%)

- ***Weekly Usage at different hours of each day***

Visualization in figure 4.10 gives the complete view of weekly usage during the particular week. This figure shows that user doesnot browse during the time from 4PM to 8PM. From this pattern it can be predicted that during these hours he had no internet access or busy in class,office, meeting,studying or playing.

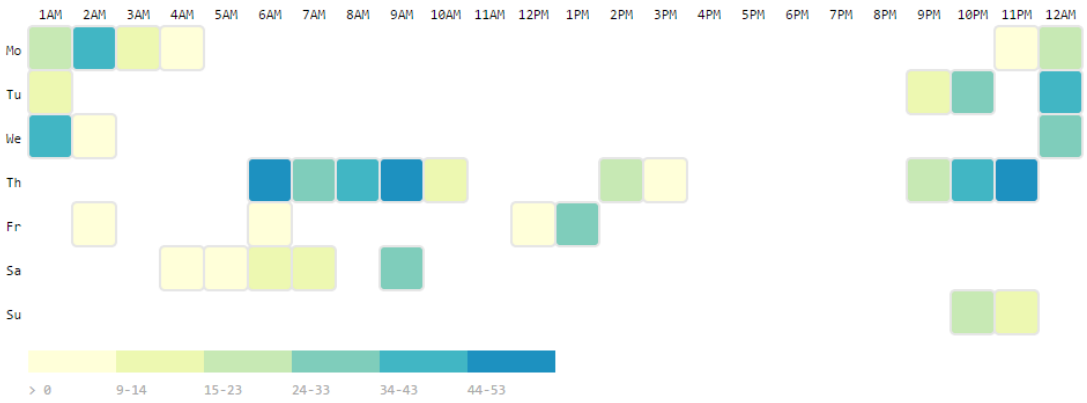


Figure 4.10 Weekly usage at different hours of day

- ***Tab Switching Visualization***

Top ten most clicked tabs during a session are visualized in this figure 4.11. Size of the arcs shows number of clicks. Big arc shows large number of clicks. Here top most clicked website is chrome extension 'lpekmepniakhhgendfacjjcgbkikohke'. Bidirectional arc shows the switching from one tab or webpage to other. Switching to the same web page shows the refresh or reload rate. According to this visualization, it can be seen that user refreshed the following pages, i.e. 'lpekmepniakhhgendfacjjcgbkikohke', 'facebook' , many times.

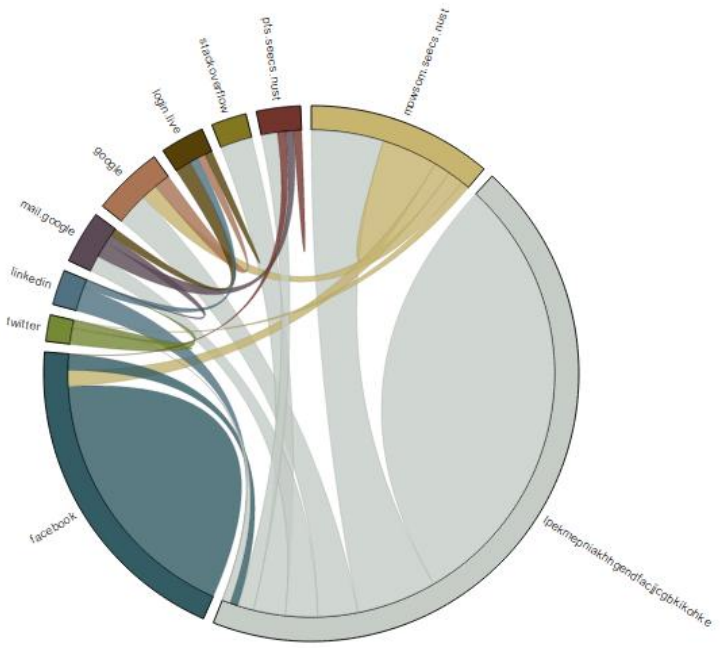


Figure 4.11 Tab switching visualization

In Figure 4.12, second chart displays information related to a specific website by placing mouse over the arc. Variation in the thickness of bi-directional arcs between different web pages indicates the frequency of switching to that webpage.

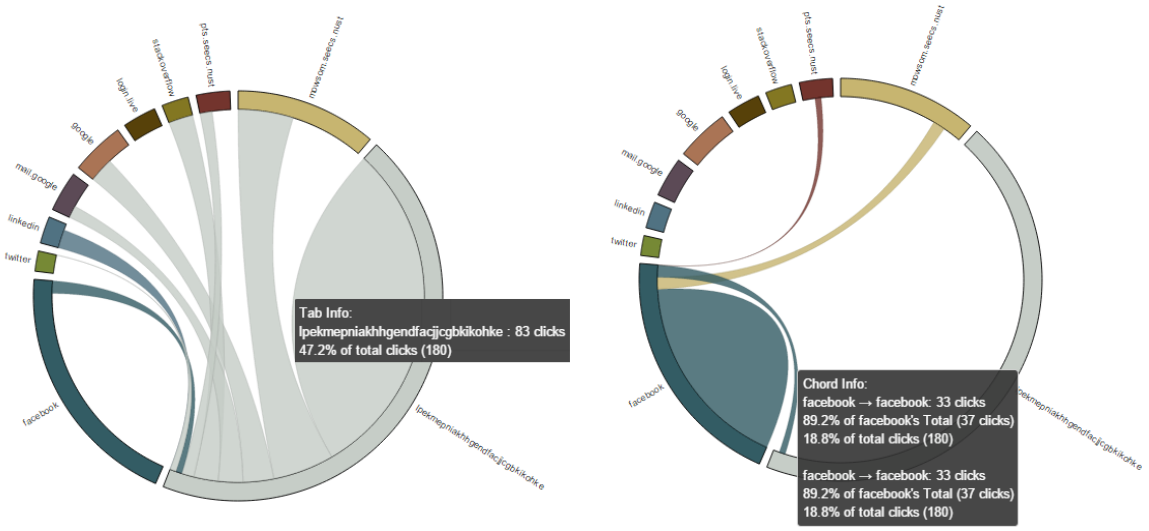


Figure 4.12 Tabs switching trends

Figure 4.13 shows the outgoing external links from any web page and these outgoing links are opened through hyperlink given on the source page.

Figure 4.13 Links transition

- ***Frequent websites at different time of day***

Six Clusters are formed based on different times of the day i.e. Early morning, Morning, AfterNoon, Evening, Night, Midnight. Brown circles represent web page and size of circle shows how frequently this web page is visited.

Figure 4.14 Cluster of frequent websites at different time of day

- ***Browser and Computer Usage***

Extension icon will get visible on the Google chrome address bar after installation. A popup below the icon becomes visible on hovering over the icon. Top arc shows the user's time

spent on social network and the below arcs displays the computer usage and browser usage over all categories. Computer usage is the cumulative time spent including browser usage (when browser in active state) and desktop applications usage when browser is inactive but running.

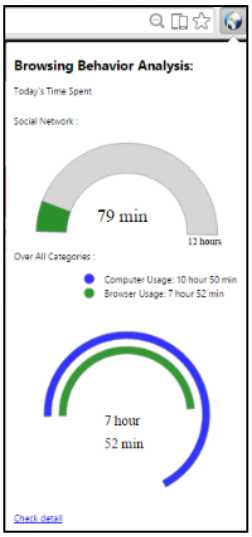


Figure 4.15 Browser Usage on social network and over all categories

4.6 Social Comparison

BBA allows individual to connect with people in their circle. Person can view and compare his categorical and temporal behavior patterns among his friends.

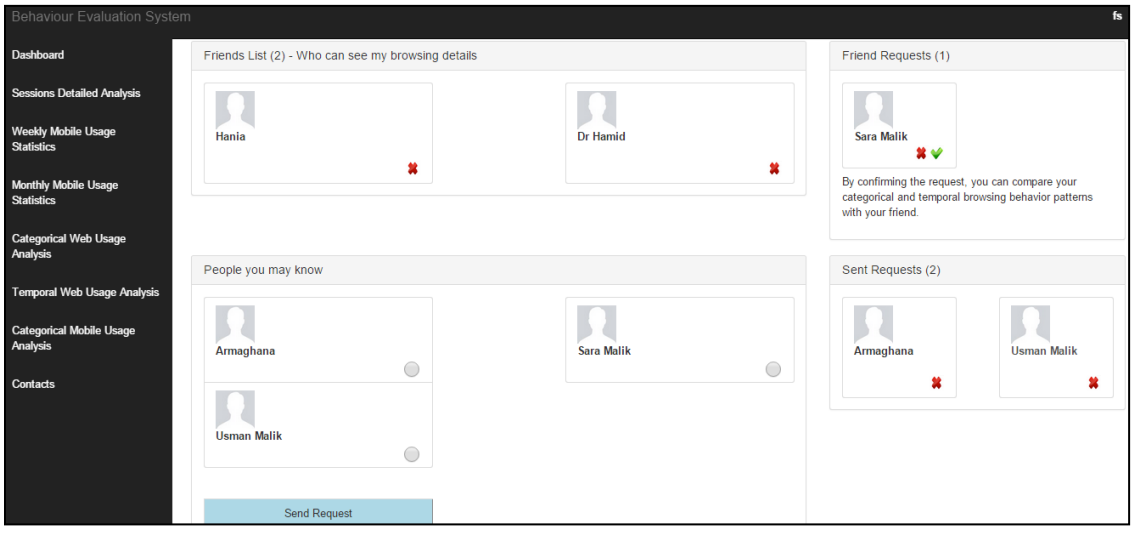


Figure 4.16 Contacts List

Different users browsing data is retrieved by the extension via HTTP Get requests to PHP scripts locating at the server. After successful retrieval of aggregated data, BBA calculates and displays the similarity among users on the basis of different web category usage using

clustering technique (K-means).It also measures the comparison on the basis of web usage for each user at different time of the day.

4.6.1 Comparison based on web categories usage

Figure 4.17 (a) shows the comparison for different user against each category. Line represents a single user's web usage trend. Color shows the group user belong to. Vertical axes values show the percentage of total time that spent at a particular category. As from the visualization, it can be seen that most of the users exhibits almost the same trend except one user. Clusters are formed to find the similar users. In case of two clusters (k=2), one user belongs to one cluster and remaining 6 belongs to other. In figure 4.17 (b) and 4.17 (c), number of clusters (k) has been increased to look the usage difference in more detail.

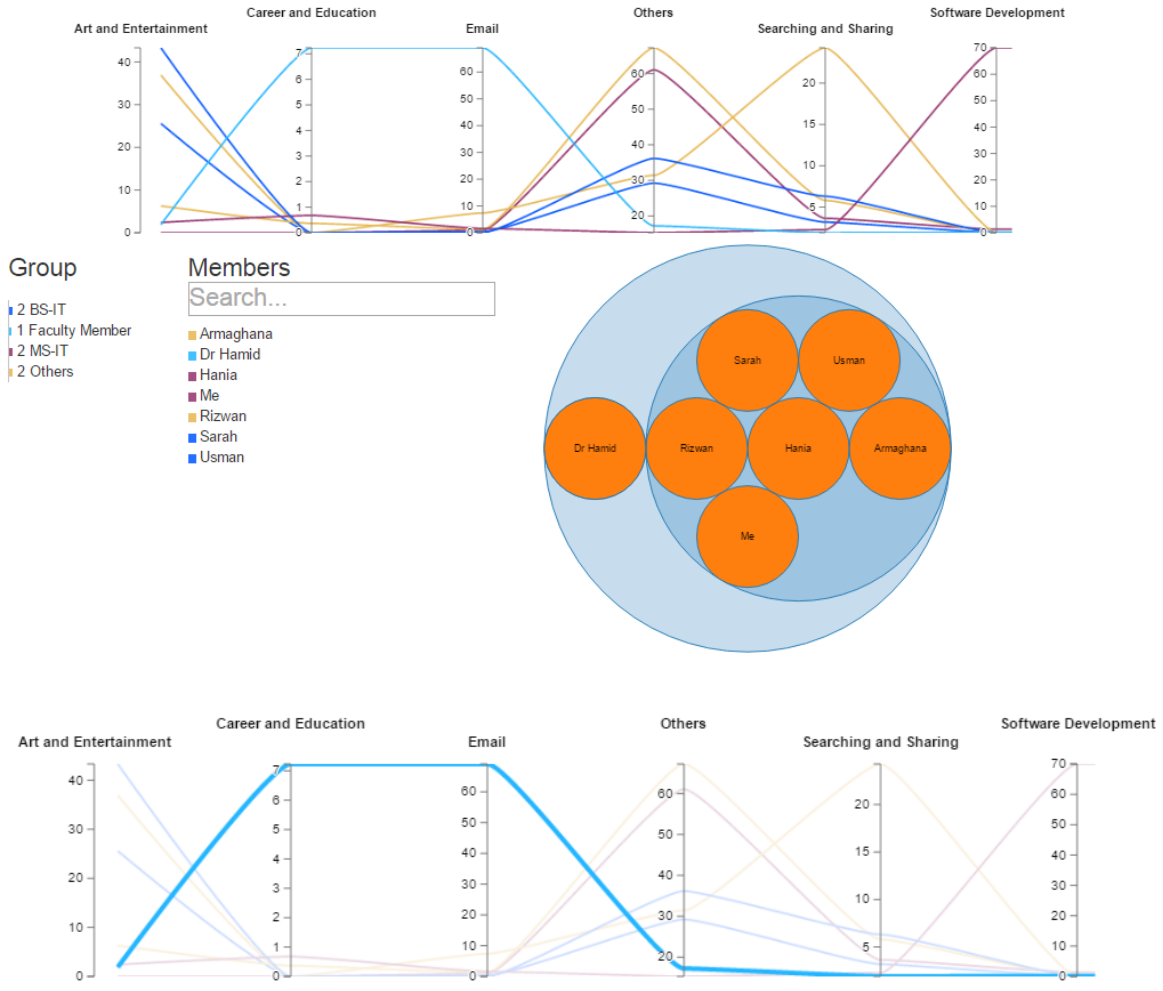


Figure 4.17 (a) Different web category usage comparison for different users (k=2)

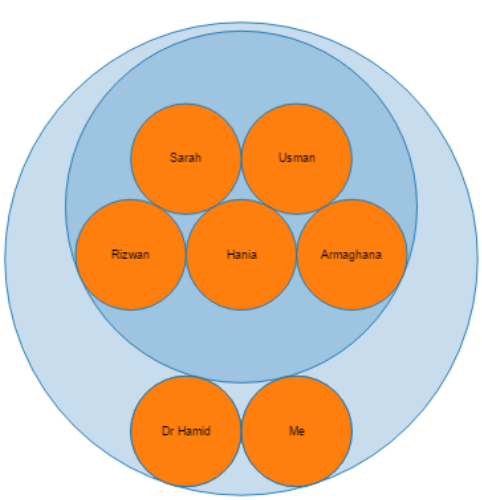


Figure 4.17 (b) Similarity among users (k=3)

Figure 4.17 (c) Similarity among users (k=4)

4.6.2 Comparison among users on the basis of peak browsing time

Figure 4.18 shows the comparison among different users at different time of day. It also represents the peak time of user that can be seen in image at right. Color shows the different times of day. In bar chart, Users and duration have been shown on horizontal and vertical axis respectively. In the figure at the right, users are ordered in descending on the basis of time spent. The color at the end of each bar shows the peak time of that particular user. For example, Peak time of top two users is 'afternoon'.

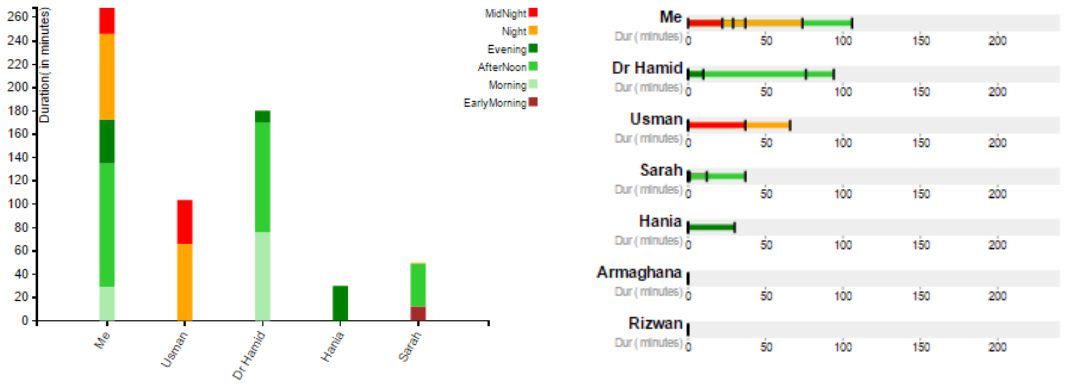


Figure 4.18 Comparison of daily web usage for different users at different time of day

4.7 Recommendations

Recommendations are proposed regarding individuals' personality and interests by identifying browsing profile similarity among their friends. Individual can see personality and interests of friends whose profiles are very similar to him. We have combined the personality and user interest information in order to overcome the new-user profile (cold-start) issue.

List of websites categories, that frequently browsed by user's most similar friends, are suggested. The user based top-N recommendation technique has been applied here. It uses a similarity based clustering model to identify the 'k' most similar users to an active user. After the k most similar users are found, their corresponding user- website categories matrices are aggregated to identify the set of categories to be recommended.

$$int_{a,j} = \sum_{U_i \in U_a^*} sim(U_a, U_i) * int_{i,j} / \sum_{U_i \in U} sim(U_a, U_i)$$

$int_{i,j}$ represents the interests of user towards a particular website category. sim is a similarity (distance) measure between users and U_a^*denotes the set of users that are most similar to U_a.

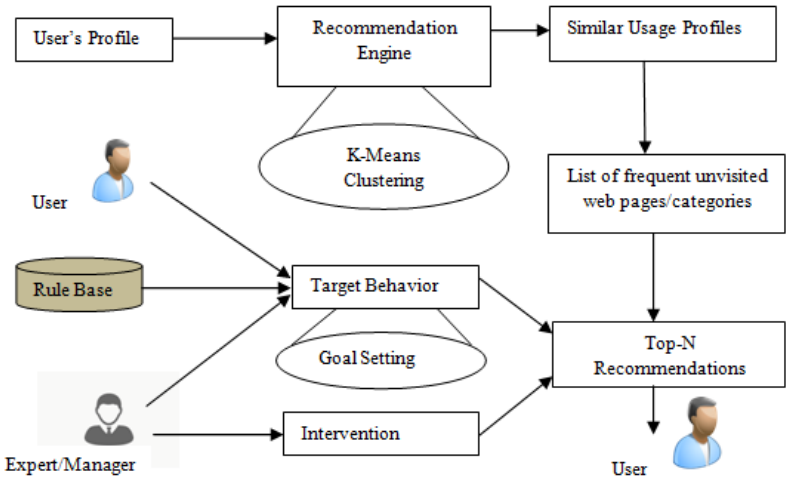


Figure 4.19 Collaborative filtering based recommender system

Recommendations are provided to users about their web usage at different categories to help them achieve their goals and improve their productivity. This system also provides suggestions regarding websites and categories visited by other users having similar usage profile. Our proposed recommender system allows users with an option to specify their target goals that they would like to achieve in their daily routine. Once users set target duration as their goal, system provides them with feedback about their current progress and recommendations are provided to help them reach their specified goal. In case if users have not specified target goals then our system recommends them according to the generalized benchmarks saved in rule base. In work environment, our system also allows manager to set project deadlines and goals for employees who work under his supervision to keep in check on their productivity and give them useful suggestions on time.

Chapter 5

Results Analysis and Evaluation

5.1 Challenges

One of the major challenges is to motivate and convince people to use this extension. We have added different features to attract users such as social comparison feature has been added that is an effective strategy to motivate users and drive competition among them. Interactive visualizations have been implemented that provide users with the quick view about their behavior.

Another challenge is privacy; people may have privacy concerns about data collection. Some users feel hesitated in sharing their data publicly. In order to deal with privacy, domain name of web page is just logged instead of complete URL and individual's browsing information is only visible to the one, he is willing to show. Individuals can send request in their social circle. On confirming the request browsing data comparison across different categories becomes visible to them.

Accuracy cannot be assured in case if user deliberately changes his logged data by behaving in a specific manner e.g. disabling the extension while browsing some specific websites and if user is watching a video or movie without interacting with computer , computer becomes idle , extension will log this time as idle but in actual user is active on computer.

5.2 Experimental Data and Results

5.2.1 Data Collection

We collected four-weeks of browsing data from team of 18 users belong to the software development department of a well-known telecom organization. They willingly added BBA extension to their chrome browser by using below link.

https://chrome.google.com/webstore/detail/bba/hiifhjpiekcilpbaahdhbhnkljeiakok

User's Registration

After installation, they registered to our system after filling the required information in the form as shown in figure 5.1. All the fields are mandatory in this form. User email and device

information is used to integrate browsing data from different devices. User's mobile and browsing activities are integrated using cell number and email id.

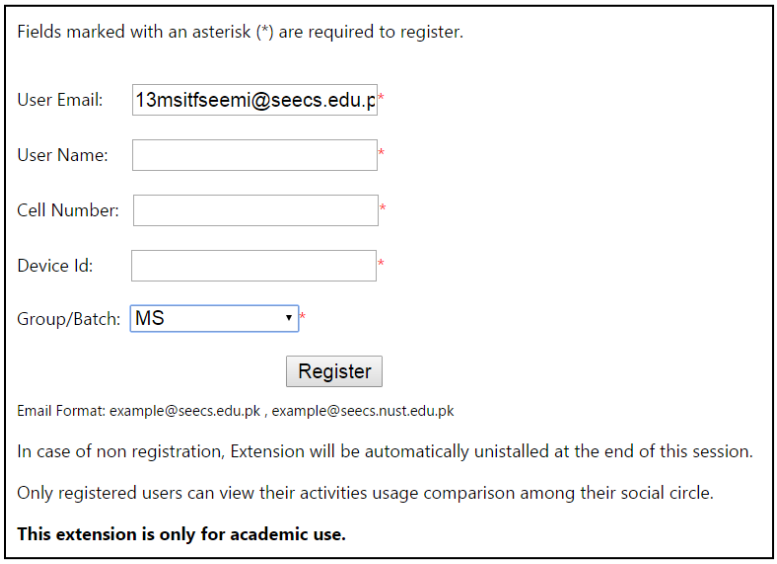


Figure 5.1 Registration Form

Data Logging

After successful registration, logging gets enabled and all browsing related activities of participant are recorded to our server. Web usage behavior data of 18 users is stored at server in privacy preserving manner. Host names of browsed websites are only logged instead of complete url to maintain user's privacy. Data set includes features such as url, time spent, number of visits, category, time stamp, session, tab, computer usage, idle time and browser usage.

5.2.2 Data Analysis and Findings

Web usage statistics and trends

Four weeks categorical usage trends for top 4 employees have been shown in figure 5.2 (at left).According to statistics, User1 (represented by blue color) spent most of his time that is approx. 60 hours, on category 'Art and Entertainment', 18 hours on 'Research and Development' , 16 hours on 'Email' and 5 hours on 'Career and Education'. User2 spent most of his time on 'Arts and Entertainment' and User4 top category is 'Research and Development'.

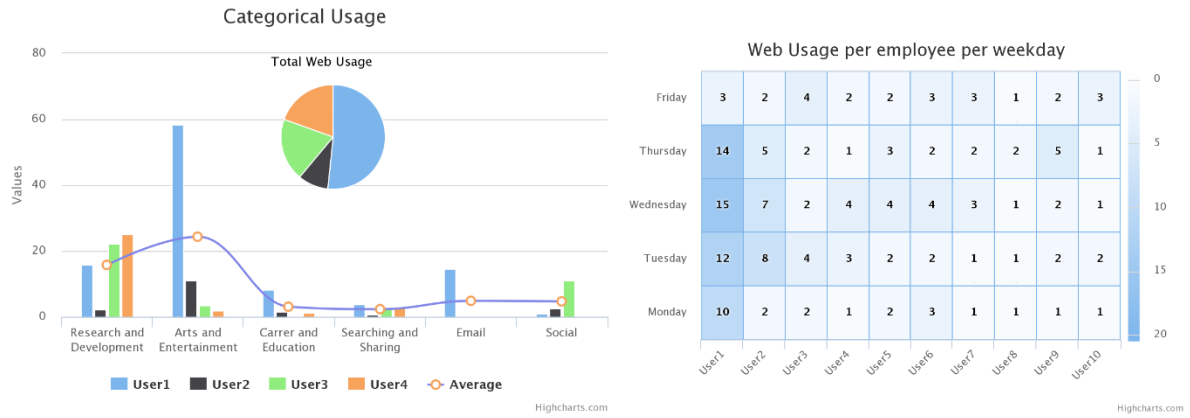


Figure 5.2 Weekly and categorical web usage trends of employees

Total web usage of user1 is far higher as compared to other users. Average daily usage for user1 is approx. 7 hours. Browsing trends of user1 shows his personality that he is curious and workaholic and his interests are on 'Art and Entertainment', Research and Development and Email. We have found that User 3 and User 4 is similar according to their categorical usage. In figure 5.2, Chart at right side shows the web usage for 10 employees at each day of a week. User1 unusually browsed this week may be due to some deadline. User2 spent 8 hours on Tuesday and 7 hours on Wednesday.

We have also collected data from two SEECS research students and one faculty member. Results are shown below in figure 5.3. User3 is faculty member, he spent mostly his time on category 'Email' and Career and Education. His trends show that he is workaholic and career oriented. User1 and User2 are students. User1 spent most of her time on research and development and social. Correlation has been found in trends of user1 that in all the browsing sessions, research and development and social category used together.

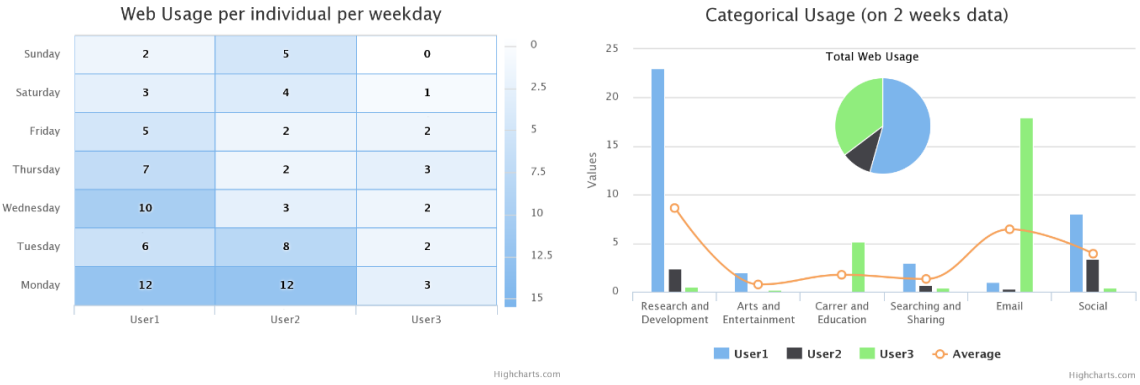


Figure 5.3 Weekly and categorical web usage trends of users at SEECS

Behavior Patterns Analysis

Collective categorical web usage behavior with respect to different epochs of the day has been shown in Fig: 5.4. This figure reveals some interesting behavior patterns that infer the dominant browsed categories and correlation among different website categories that browsed together at different time of the day. With respect to usage, Social, Email and Development appear as dominant categories. Social websites frequently used together with websites belong to 'development' and 'Email'. 'Email' and 'Social' websites browsed at every epoch of the day but their usage is evident during morning, and afternoon. 'Searching & Sharing' and 'Arts & Entertainment' related websites are mostly browsed during night and midnight. By analyzing the correlation of categorical usage with different epochs of the day, manager can guide their team members, if team's browsing activity trend affecting the deadline or performance of project. Recommendations are also automatically sent to the user by our system if his usage in specific web category exceeds or below the benchmark and affecting his performance. As shown in figure, time spent on social category during work time (morning and afternoon) is very high as compared to other categories and that behavior can be one of the reasons of missing deadlines and poor productivity.

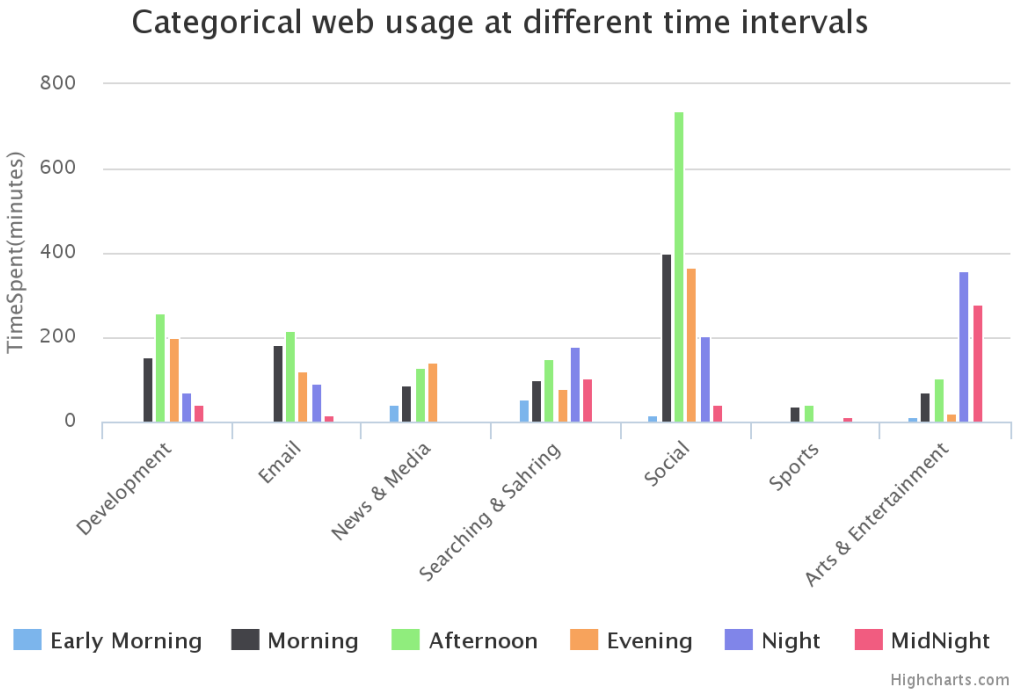


Figure 5.4 Categorical web usage trends at different times of the day

We have selected another three users having different interests. Personality can be inferred by individual's interest area. Figure 5.5 shows the comparison among these users' categorical web usage. User1 's behavior trends represent that he is workaholic who spend 55% of his time on research and development. User2 is curious; his interest rate towards searching and news is 43%. User3's personality type is inferred as social as he spends 34% of his time on surfing social media websites. Personality Information enhances collaborative filtering recommender systems performance and overcomes the cold start problem.

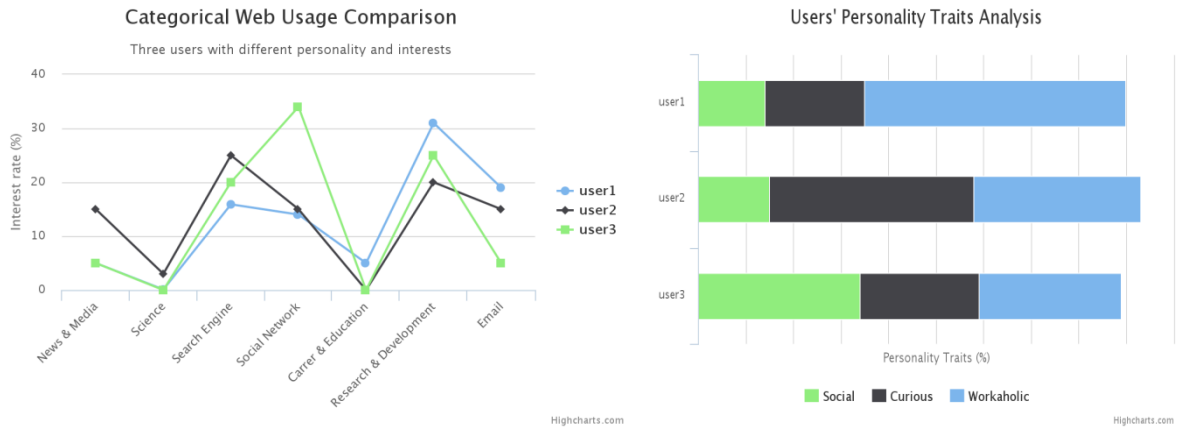


Figure 5.5 (a) Comparison among users having different interests (b) Personality traits inferred via interest rate

Figure 5.6 shows the time spent by users on different web categories against each week. In the first week, planning and requirement analysis phase of project was going on. Development tasks were assigned in 2nd and 3rd week. Development deadline was due on end of 3rd week and Testing had been done in 4th week. Benchmark had been set for each week according to the assigned tasks requirement.

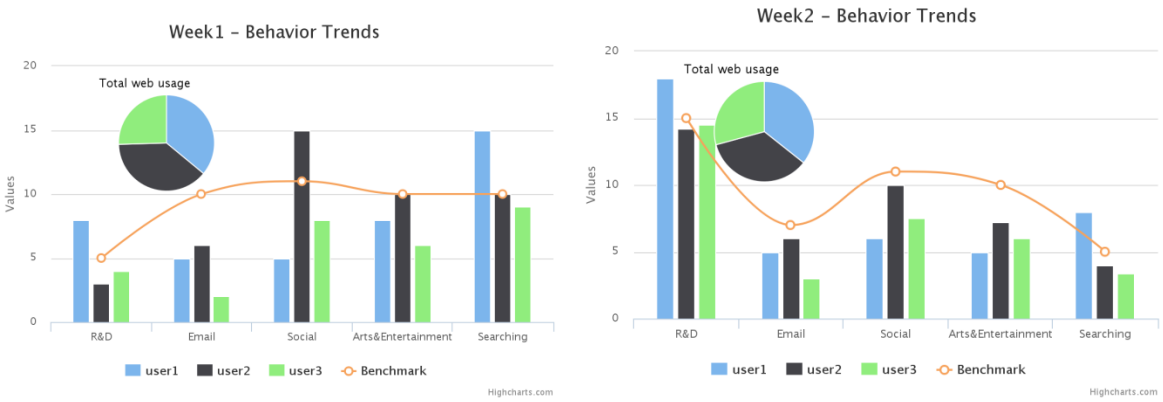


Figure 5.6 (a) Weekly trends

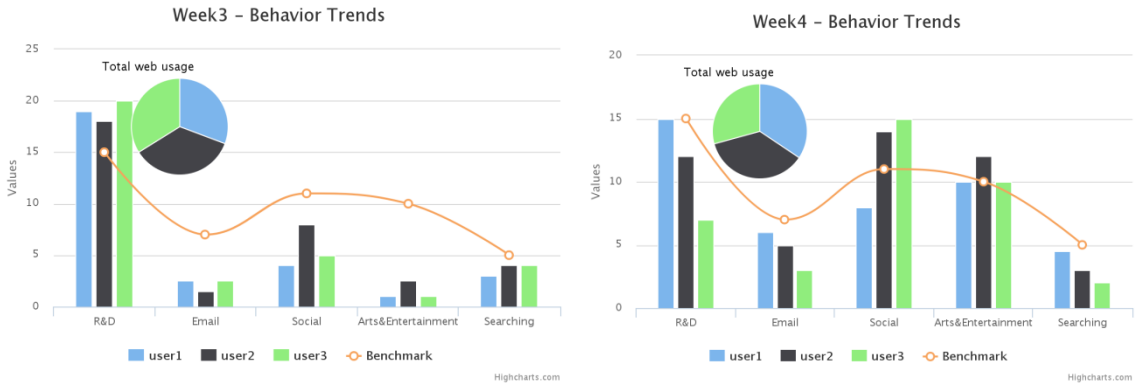


Figure 5.6 (b) Weekly trends

Figure 5.6 (a) and (b) reveals that time Spent on 'Research and Development' (R&D) and 'Email' categories increased during week2. During week3, it can be seen that time spent on 'Art and Entertainment', 'Email' and 'Social' categories dropped due to deadline of project and most of the time spent on 'R&D'. Weekly behavior trends lead us towards the interesting correlation between projects deadline and web categories usage as shown in figure 5.7. As the deadline comes closer, all the categories usage drops except 'R&D'.

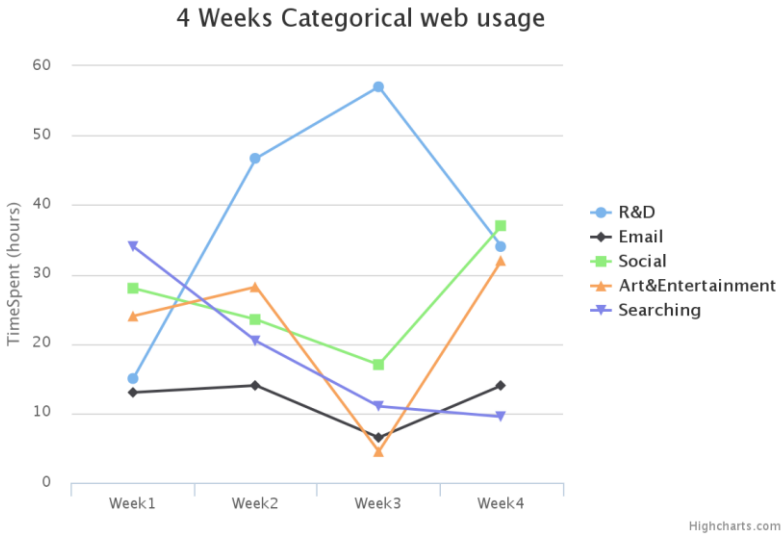


Figure 5.7 Four weeks categorical web usage

Another Correlation between user's productivity and categorical usage can be found. User1 achieved the employee of month award. His productivity was excellent. He gave his maximum to the project and spent limited time on other activities like social and entertainment as shown in figure 5.8. Productivity of user2 and user3 was good and average. Social and entertainment activities' time spent for each user drop during week3 due to

deadline of project. As weekly trends reveal that user3 spend less time on email and during week1 and week2, his performance was quite below from benchmark. Recommendations were sent to all users by our system regarding their productivity each day to further improve his performance.

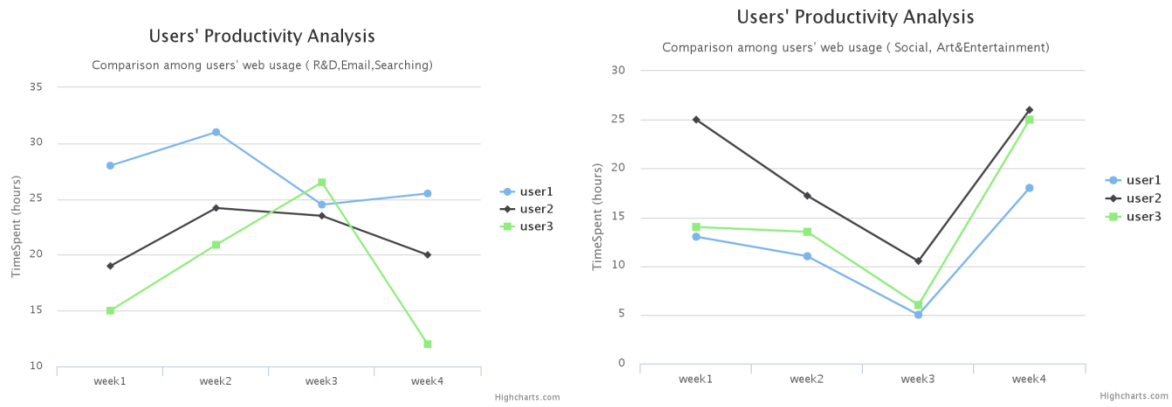
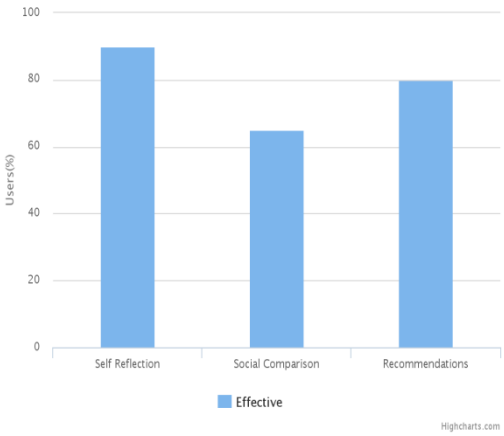


Figure 5.8 Users' productivity analysis

5.3 Evaluation

We have evaluated the extension by arranging an interview session and conducting survey over 20 participants .We will discuss here about some users' reviews regarding our extension. They have found it very interesting and were motivated that they can quickly see their comprehensive web usage statistics across many dimensions. Some said that this extension makes them conscious and aware about their usage and restricts them when they see the unusual behavior and big number in statistics at particular website. Resource manager of software development team has appreciated the social comparison feature. He said that it will promote the competitive environment within the team and well-run competition can bring out the ambition and pushing oneself that's good for person's professional growth. Some users still have privacy concerns and suggested that user's identity should not be revealed and data should be transferred as by anonymous user.



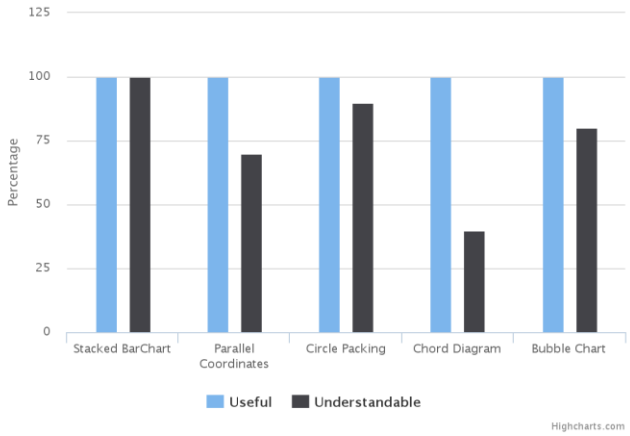


Figure 5.9(a) User's feedback regarding effectiveness of BBA Figure 5.9(b) Effectiveness of visualizations

Figure 5.9 shows results of survey. Aim of this survey is to get users feedback regarding efficiency and usefulness of our system 'BBA'. According to survey results, 90% of users rated self reflection feature effective. They said that this extension made them conscious and aware about their usage. 65% of users found social comparison feature useful. They said this feature was very interesting. It motivated them to improve their weakness by comparing with others. 35% of users didn't find it attracted as they had privacy concerns and they didn't want to share their information with others. 80% of users rated recommendation feature effective. They admitted that it brought positive changes in their behavior and improved their performance. Users found every type of visualization useful. Some of users found few of them difficult to understand.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This research work has attempted to introduce BBA, an approach towards capturing and analyzing browsing behavior of individuals over temporal, categorical and social contexts. BBA provides a chrome extension that runs autonomously in the background and captures the comprehensive browsing data set including web usage, session's events and tabs switching details on different browsing and window events. It allows the individuals to visualize their interesting browsing behavior patterns to gain deeper insights into their browsing behavior by providing interactive graphical user interface to promote self-reflection and awareness among them and help in making valuable decisions, improve their judgment and bring positive changes in their behavior and life. BBA also presents a framework for analyzing social comparison with respect to its implications for future actions or outcomes. Recommendations are also provided regarding individual's behavior in order to improve one's productivity and performance.

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 extension 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, tab switching pattern, 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.

6.2 Future Work

Testing and evaluation will be done at large scale in future to yield more interesting, effective and valuable correlations from the behavioral data. BBA is only supported on chrome browser. We aim to provide support for other browsers. We intend to integrate our framework with persuasive feedback mechanism that will provide interventions to improve user's behavior. Chrome extensions are not supported on chrome for android so we could not integrate the android phone browsing data. This system can be integrated with desktop and smartphones applications.

References

[1] Maryam Jafari, Farzad Soleymani Sabzchi, and Amir Jalili Irani. 2014. Applying Web Usage Mining Techniques to Design EffectiveWeb Recommendation Systems: A Case Study. Advances in Computer Science: an International Journal 3, 2 (2014), 78–90.

[2] Pradnya Mehta, Shailaja B Jadhav, and RB Joshi. 2014. Web Usage Mining for Discovery and Evaluation of Online Navigation Pattern Prediction. International Journal of Computer Applications 91, 4 (2014).

[3] von Christian von der Weth and Manfred Hauswirth. 2014. Analysing Parallel and Passive Web Browsing Behavior and its Effects on Website Metrics. arXiv preprint arXiv:1402.5255 (2014).

[4] Daniel Epstein, Felicia Cordeiro, Elizabeth Bales, James Fogarty, and Sean Munson. 2014. Taming data complexity in lifelogs: exploring visual cuts of personal informatics data. In Proceedings of the 2014 conference on Designing interactive systems. ACM, 667–676.

[5] Christian von derWeth and Manfred Hauswirth. 2013. DOBBS: Towards a Comprehensive Dataset to Study the Browsing Behavior of Online Users. In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, Vol. 1. IEEE, 51–56.

[6] Michal Kosinski, David Stillwell, Pushmeet Kohli, Yoram Bachrach, and Thore Graepel. 2012. Personality and website choice. (2012).

[7] Hamid Mukhtar, Arshad Ali, Djamel Bela¨ıd, and Sungyoung Lee. 2012. Persuasive healthcare self management in intelligent environments. In Intelligent Environments (IE), 2012 8th International Conference on. IEEE, 190–197.

[8] Sharad Goel, Jake M Hofman, and M Irmak Sirer. 2012. Who Does What on the Web: A Large-Scale Study of Browsing Behavior.. In ICWSM.

[9] Jae-wook Ahn, Krist Wongsuphasawat,Peter Brusilovsky: "Analyzing User Behavior Patterns in Adaptive Exploratory Search Systems with LifeFlow" , *University of Pittsburgh*, (2011)

[10] Corcoran, K., Crusius, J. and Mussweiler, T.: "Social comparison: Motives, standards, and mechanisms.", *In D. Chadee (Ed.), Theories in social psychology* (pp. 119-139). Oxford, UK: Wiley-Blackwell, (2011).

[11] R. Kumar and A. Tomkins: "A Characterization of Online Browsing Behavior.", *In Proceedings of the 19th International Conference on World Wide Web*, WWW'10, pages 561–570, New York, NY, USA, (2010).

[12] D. A. Keim : "Information visualization and visual data mining.", *IEEE Transactions on Visualization and Computer Graphics*, 7(2):100–107, 2002.

[13] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P-N. : "Web usage mining: Discovery and applications of usage patterns from web data." *SIGKDD Explorations 1*, 2 (2000.), 12–23.

[14] Chen, P. & Garcia, S. M. (manuscript) "Yin and Yang Theory of Competition: Social Comparison and Evaluation Apprehension Reciprocally Drive Competitive Motivation".

[15] Chan, P. K. , "A non-invasive learning approach to building web user profiles. In Workshop on Web usage analysis and user profiling", *Fifth International Conference on Knowledge Discovery and Data Mining*, (1999), 7–12.

[16] Zhexue Huang. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery 2, 3 (1998), 283–304.

[17] Icek Ajzen. 1991. The theory of planned behavior. Organizational behavior and human decision processes 50, 2 (1991), 179–211.

[18] Joanne V Wood. 1989. Theory and research concerning social comparisons of personal attributes. Psychological bulletin 106, 2 (1989), 231.

[19] Daniel Keim and others. 2002. Information visualization and visual data mining. Visualization and Computer Graphics, IEEE Transactions on 8, 1 (2002), 1–8.

[20] A Inselberg and B Dimsdale. 1990. Parallel coordinates: a tool for visualizing multi-dimensional geometry;1990. San Francisco CA (1990), 361–375.

[21] M Waseem Chughtai, Imran Ghani, Ali Semalat, and Seung Ryul Jeong. 2013. Goal-based Framework for cold-start problem using multi-user personalized similarities in e-Learning scenarios. In e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on. IEEE, 334–338.

[22] Rong Hu and Pearl Pu. 2011. Enhancing collaborative filtering systems with personality information. In Proceedings of the fifth ACM conference on Recommender systems. ACM, 197–204.

[23]Teng-Sheng Moh and Neha Sushil Saxena. 2010. Personalizing Web Recommendations Using Web Usage Mining and Web Semantics with Time Attribute. In Information Systems, Technology and Management. Springer, 244–254.

[24] R Suguna and D Sharmila. 2013. An Efficient Web Recommendation System using Collaborative Filtering and Pattern Discovery Algorithms. International Journal of Computer Applications 70, 3 (2013), 37–44.

[25]Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. 2000.Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce* (Minneapolis, Minnesota, United States, 2000).ACM, 352887, 158-167.

[26] Huang, Xiangji 2007. Comparison of interestingness measures for web usage mining: An empirical study. *International Journal of Information Technology & Decision Making*, 15-41.

[27]Khovanskaya, V., Baumer, E.P.S., Cosley, D., Voida, S., and Gay, G.K. 2013. Everybody Knows What You're Doing: A Critical Design Approach to Personal Informatics. *In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). Paris, France.*

[28] Zhang, Amy X., Joshua Blum, and David R. Karger. 2016. Opportunities and Challenges Around a Tool for Social and Public Web Activity Tracking.

Web Links

[29] Was I productive? Examining the reliability of the Quantified Self technology- Hongbin Zhuang- https://www.ucl.ac.uk/uclic/studying/taught-courses/distinction-projects/2013-theses/2013-Zhuang

[30] https://chrome.google.com/webstore/detail/timestats/ejifodhjoeeenihgfpjijjmpomaphmah

[31] https://www.rescuetime.com

[32] https://code.google.com/p/figue/

[33] https://www.globalwebindex.net/blog/daily-time-spent-on-social-networks-rises-to-1-72-hours

[34] https://developer.similarweb.com/website_categorization_API