# Browsing Behaviour Analysis using Data Mining

Hamid Mukhtar[1], Farhana Seemi[2], Hania Aslam[3], Sana Khattak[4]
[1]College of CIT, Taif University,
Taif, Saudi Arabia
[2,3]National University of Sciences and Technology (NUST),
Islamabad, 44000, Pakistan
[4]University of Engineering and Technology (UET), Peshawar,
Peshawar, 25000, Pakistan

***Abstract*—Now-a-days most of our time is spent online using some form of digital technology such as search engines, news portals, or social media websites. Our online presence makes us engaged most of the time and leads us to become oblivious of our important work, resulting in a form of procrastination that decreases our productivity significantly. Some desktop and mobile applications have recently emerged to counter the problem by introducing various means of self-tracking to reduce the wasting of time and engage in productive activities. However, these systems suffer several shortcomings in terms of being static or providing a limited view of actions using one aspect only. To promote self-awareness that helps bring positive changes in individual's performance, there is a need to present the data in a more persuasive ways, bringing interaction to it and present the same data in different ways using both temporal and categorical dimensions. We describe a framework that collects and processes the browsing data and creates a user behavior model to extract valuable and interesting temporal and categorical patterns regarding user online behavior and interests. To discover the valuable behavior patterns from the individual's browsing data, different web usage mining techniques have been used. 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 his/her behavior. We also present a small-scale study including university students, which proves the importance of our work.**

***Keywords*—Pattern discovery; visualization; behavior modeling; web usage mining; browsing**

## I. Introduction

Quantified self-movement incorporates digital technology to acquire data on various aspects of an individual's life with an aim to improve self-awareness and human performance. People want to be self-aware, 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.

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

### A. Problem Context

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

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

### B. Objectives and Scope

The main objective of our research is to develop a system for analysis of web-usage behavior patterns using interactive visualization techniques to promote self-reflection among users. Moreover, the system should be able to present the behavior from different perspectives using temporal and categorical dimensions.

Following are the objectives of our research work:

- Development of framework for gathering and processing of web usage data.

- Web-usage behavior modeling for the extraction of interesting temporal and categorical patterns.
- Development and demonstration of interactive visualizations to analyze and monitor the extracted patterns in different dimensions.

The scope of our research is limited to online browsing behavior and does not include tracking other applications used by the users on the computers. To achieve this goal, a web browser extension has been implemented. Initially, a small-scale investigation has been carried out on a group of university students, and the results have been reported here. In future, we intend to evaluate it at large scale.

## II. Literature Review

### A. Related Work

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 [1]. Different web usage mining techniques have been discussed in [4] that can be used to extract patterns from Web log files. Discovered patterns are used for pattern analysis that helps in understanding the user behaviors. According to [5], density-based clustering algorithm has been used to discover navigation patterns. K-Nearest Neighbor algorithm with inverted index has been suggested for efficient prediction. Thus, several methods from data mining are used in the area of web usage analysis.

DOBBS [6] 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. Authors in [7] 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. 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).

Ravi Kumar and A. Tomkins [8] provide taxonomy of page views consisting of categories content, communication and search. They have presented a quantitative analysis of the mechanics of online behavior. Accordingly, 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. While their contribution is generally useful for research community, it cannot be used by the users to evaluate themselves.

Khovanskaya et al. [9] have presented an interface that displays personal web browsing data and reveals 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 developed 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 [10]. Life Flow [11] 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.

Kosinski et al. [12] show 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.

### B. Related Applications

Different browser extensions are available that provide statistics regarding individual's time spent on browsing. TimeStats[1], Webtime Tracker[2], BHVis[3], and RescueTime[4] show daily and monthly web usage statistics to the user using different visualizations.

For example, RescueTime, which offers most of the functionalities like the other tools, provides detailed reports about the time spent on different applications, websites and categories. It allows users to set their daily goals to get them

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

[2] https://chrome.google.com/webstore/detail/webtime-tracker/

[3] https://chrome.google.com/webstore/detail/bhvisvisualization-of-you

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

aware how productive they are. RescueTime makes people aware about their daily habits so they can focus and be more productive. The main features of RescueTime are: block out the distracting websites, show alerts to the user about the productive and distracting time, keep track if user away from the computer, log daily accomplishments, and display visualization related to daily usage.

There are some drawbacks of rescue time (and hence the other tools) that have been mentioned in the study [13]. According to the study, the reason behind the failure of RescueTime (and similar tools) 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.

Other tools have problems of their own. For example, TimeStats does not 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.

Our approach towards online behavior mining is better when compared to these applications in various aspects. First, just like the existing tools, we provide visualizations but unlike these tools, our visualizations are more interactive, i.e., one can choose to select some data point to see more details in most of the visualizations. Second, we provide different aspects of visualization for the same pattern of usage, giving the user more opportunities to explore their behavior from different angles. This also includes a comparison of behavior over longer period. The user can also reveal his interests by viewing their activities with respect to temporal or categorical data.

## III. Methodology

In this paper, we propose a framework that collect and process web usage data, extract interesting behavior patterns from the formulated data, demonstrate interactive visualizations to better analyze the extracted patterns and allow individuals to compare themselves over time. Initially, qualitative and quantitative web usage data features are identified such as dwell time, number of hits, category, idle time, and time of occurrence. A web browser add-on logs 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. The framework has been developed as the Chrome browser extension and it transfers the web usage data to a server, securely and periodically.

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 facilitate the individual with the deep understanding of behavior.

### *A. 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. Considering this limitation, our system does not use the browser history logs. Our data collection module efficiently runs in the background of the 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.

Sessions and tabs data can infer the user's behavior regarding how often user switches the tabs, how long the session is and how many tabs are created in a session. Websites usage data helps in analyzing user behavior that how much time user spent on 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. Idle time of browser is calculated when the browser window is not focused or if window is focused but idle or 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 stays standby, locked or idle.

Table I summarizes the high-level features, the attributes related to the browser, and the intended behavior analyzed through them in our framework.

### *B. Developing Chrome Extension*

The Google Chrome web browser lets us use the functionality of the browsing through development of extensions[5]. An 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 user interface (UI). Our application's user interface contains the web pages that display the user different browsing behavior trends.

### *C. Web Usage Mining*

Web usage mining is the data mining technique to discover web usage behavior patterns from web data. Figure 1 shows the process of web usage mining. It comprises of three phases: preprocessing, pattern discovery, and pattern analysis [1]. Focus of this research is on pattern discovery and analysis techniques. There is a 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 [2]. Visualization allows the user to mine and gain insight into the data and come up with new mining recommendations. There are many visualization techniques that have been developed to explore the meaningful information from the large datasets. Goal of visual data mining is to represent as many of data points as possible in a single visualization or plot. Pattern discovery and visual data mining techniques have been discussed in next subsections.

[5] https://developer.chrome.com/extensions/getstarted

TABLE I. ATTRIBUTES OF WEB BROWSING DATA THAT HELP IN INFERRING BEHAVIOR.

| 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 |

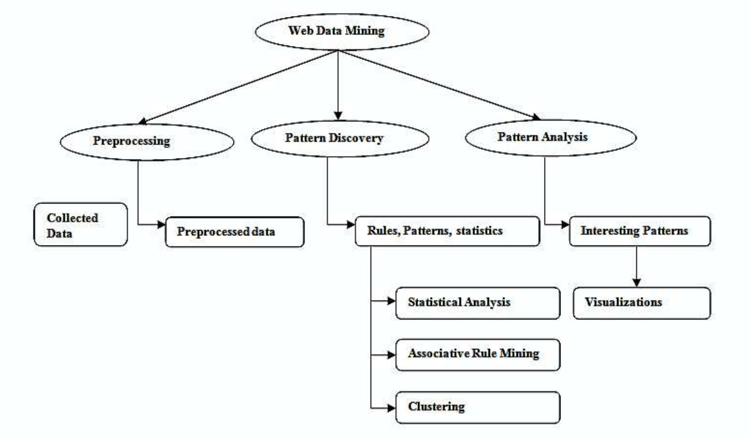


Fig. 1. Web usage mining process.

### D. 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 kinds of descriptive statistical analyses, e.g., frequency, count, min, mean, max, median, mode, etc. can be performed on the data attributes like page views, time spent at a particular page, frequently accessed pages, tabs switching time, number of sessions per day, session time span, number of tabs per session, etc.

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

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

Lift is a correlation measure and can be computed as

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

Association rules are used to find associations among web pages and web categories that frequently appear together in users' sessions. Apriori algorithm is the most classical algorithm for mining frequent item sets.

Clustering is a technique that groups together the items having similar characteristics. Web usage clusters can be discovered by grouping the users having similar browsing trends. K-means [14] 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

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

Where $\|x_j - \mu_i\|$ is the Euclidean distance between $x_j$ and $\mu_i$. $n$ is the number of data points in $ith$ cluster. $k$ is the number of cluster centers.

K-modes algorithm [14] has extended the K-means algorithm to cluster the data with categorical values using a simple matching dissimilarity measure or the hamming distance for categorical data objects, replacing means of clusters by their modes.

The dissimilarity measure between X and Y is the total mismatches of the corresponding attribute categories of two objects. Two objects are more similar if number of mismatches is smaller.

$$d(X, Y) = \sum_{j=1}^{n} \delta(x_j, y_j)$$

Where $\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$

Cost function of K-modes becomes

$$C(Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta(x_{i,j}, q_{l,j})$$

Where $Q_l = [q_{l,1}, q_{l,2}, ............, q_{l,m}] \in Q$

K-prototypes [14] simply integrate the K-means and K-modes algorithm. It is used for mixed type of data.

Dissimilarity between two mixed type objects X and Y can be measured by

$$d(X, Y) = \sum_{j=1}^{p}(x_i - y_j)^2 + \sum_{p+1}^{n} \delta(x_j, y_j)$$

### E. Visual Data Mining Techniques

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

Fig. 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

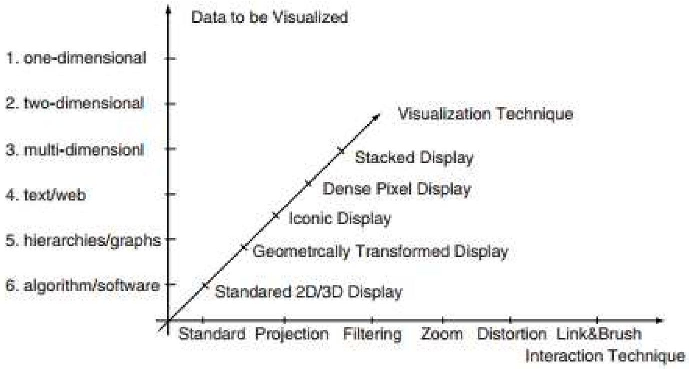


Fig. 2. Classification of Information Vis. Techniques [2].

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

- **Parallel coordinates techniques** allow exploring and analyzing the multidimensional data. Each data item is presented as a polygonal line which intersects each axis at the point equal to the value in that dimension. It maps the k-dimensional space onto the two display dimensions by using k equidistant axes which are parallel to one of the display axes.
- **Sunburst** used to visualize hierarchical data represented by concentric circles. The circle in the center represents the root node, with the hierarchy moving outward from the center.
- **Scatter Bubble chart** shows the relationship between three different variables in one plot. An additional dimension of the data is represented in the size of the bubbles.
- **Radar Chart** is a two dimensional chart that displays multivariate data over multiple quantitative variables represented on axes starting from the same point.
- **Chord diagram** shows the connection among different entities. The chords between the arcs visualize the switching behavior of the respondents between entities in both directions.
- **Heat map** is a two-dimensional representation of data in tabular format with user defined color ranges e.g. low, high and average. It provides an immediate visual summary of information.
- **Stacked Bar Chart** Bar charts are used to show two dimensional data and can be used for more complex comparisons of data with the stacked bar charts. Stacked bar chart stacks bar that represent different group on top of each other.
- **Interaction and Distortion Techniques** allow the user to dynamically change the visualization according to exploration objectives and provide the data with low level details while preserving the high level details for example interactive zooming present more details on higher zoom levels.

## IV. Browsing Behavior Analysis

Our developed chrome extension collects and displays the browsing data, sends it to the server where individual's web browsing activities data from different devices are integrated to display the aggregate web and mobile usage statistics.

### A. *Design Requirements*

Our framework addresses the following questions and provide the detailed information about:

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

The major functionality of the system is as described next.

### B. *Browsing Data Collection and Integration*

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 based on consecutive page visits with in the session. Last page dwell time is calculated at the start of the next session. Logged data is sent via HTTP POST requests to PHP scripts residing on the 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. At the server end, data sent from different devices (machines where chrome extension is installed) get integrated based on user's email.

### C. Behavior Extraction

*1) Websites Categorization:* Web URLs are grouped into various categories, such as social networking, research and development, news media, career and education, etc. Website categorization APIs [16] [17] have been used to automatically retrieve category and subcategory for the web site via HTTP request.

*2) Browsing Times of the Day:* We have considered six times of the day i.e. Early Morning, Morning, afternoon, evening, night, midnight.

Where

$$4_{AM} \geq EarlyMorning \leq 8_{AM}$$

$$8_{AM} \geq Morning \leq 12_{AM}$$

$$12_{PM} \geq AfterNoon \leq 4_{PM}$$

$$4_{PM} \geq Evening \leq 8_{PM}$$

$$8_{PM} \geq Night \leq 12_{AM}$$

$$12_{AM} \geq MidNight \leq 4_{AM}$$

*3) Frequent Categories/Websites and their Correlation:* Apriori algorithm has been used to get the frequent categories. It extracts the categories that frequently used together. We have supposed that an item set is frequent if it appears in at least 40% of the total sessions. For example, 20 is the support threshold for 50 sessions. First step is to count the number of occurrences of each category separately by scanning all the sessions. Next step is to generate the pairs of frequent items. Pairs that meet the support threshold are frequent.

Associative rule mining is a technique for discovering interesting relations between categories. In order to select interesting rules, minimum support and confidence constraints are used. For example, rule is $Social \implies SoftwareDevelopment$. Its confidence is $Support(Social \cup SoftwareDevelopment)/Support(Social) = 0.5/0.5 = 1$ which means software development occurs in all the sessions containing social. To find the correlation among the categories, we use:
$Lift(Social \implies SoftwareDevelopment) = P(Social \cup SoftwareDevelopment)/P(Social)P(SoftwareDevelopment) = 0.5/((0.5 * 1)) = 1$ It shows that social websites and social development are used together. Recommendation can be proposed here by analyzing whether social networking affecting the productivity of user or not.

*4) Predicting User Interests:* Website's visits frequency and duration are two major metrics of a user interest in a website [18]. We consider these metrics to estimate the user interest. Duration is measured based on dwell time normalized by maximum dwell time. Frequency is measured based on number of visits of category normalized by maximum number of visits. Harmonic mean is used to mitigate the impact of large outliers and aggravate the impact of small ones. Together, they are used to find the areas of interest for any user.

### D. Data Visualization

*1) Daily Usage Visualization:* In Fig. 3 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 3 provides an insight about the daily usage of internet sites by the users. As can be inferred, among other observations, that the user uses social networking websites almost on daily basis but on one Sunday the usage duration was very high. Similarly, the user browses (watches) the TV and Videos category on almost daily basis around evening or late night. User can get the details of any of the activity (bubble) by placing mouse over the bubble.

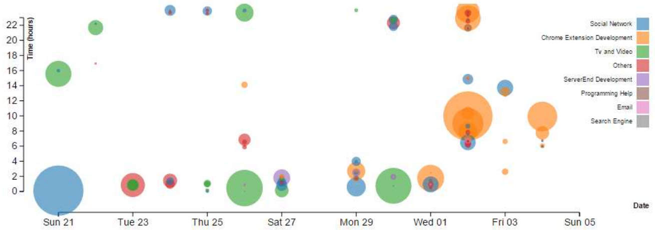


Fig. 3. Daily usage visualization

*2) Browsing Usage at different times of day:* In Fig. 4, browsing usage at different times of the day can be visualized. Distinct color has been assigned to each part of the day. Date and duration have 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 figure shows that user had approximately the same pattern from Sat 27 June 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. This analysis can lead to inferring about the temporary project ow work activity during some specified time.

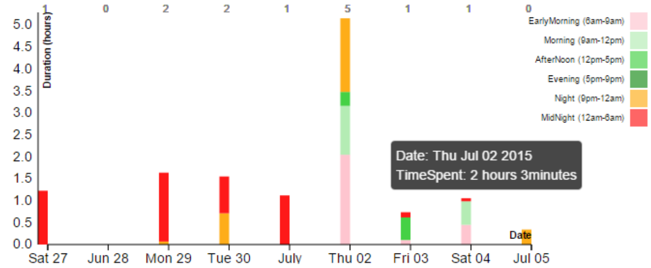
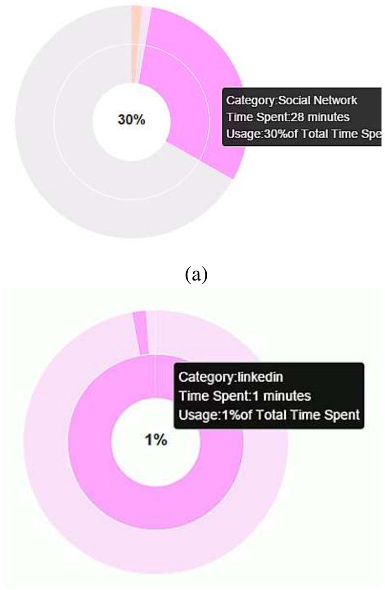


Fig. 4. Web Usage at different time of day.

*3) Categorical Usage:* Fig. 5 (a) shows 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 the figures, user has spent 30% of his time in social networks. By clicking on social network category, it can be seen in Fig. 5 (b) that user has browsed LinkedIn for only 1 minute in the social networks category.



(a)

(b)

Fig. 5. Categorical usage of website

*4) Weekly Usage at different hours of each day:* Visualization in Fig. 6 gives the complete view of weekly usage during the particular week. This figure shows that user does not 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 some activity other than browsing the Internet.

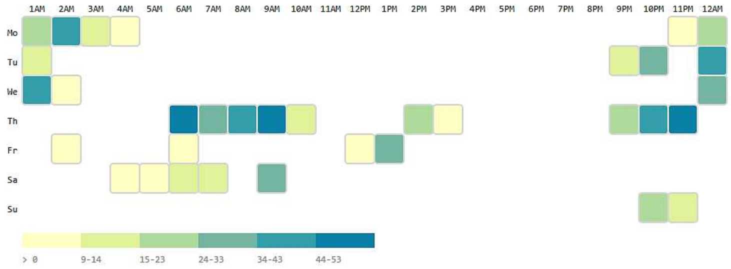


Fig. 6. Weekly usage at different hours of the day

*5) Tab Switching Visualization:* Top ten most clicked tabs during a session are visualized as shown in Fig. 7. Size of the arcs shows number of clicks. Big arc shows large number of clicks. Switching to the same web page shows the refresh or reload rate. According to this visualization, user refreshed the Facebook page many times.

*6) Frequent websites at different time of day:* As shown in Fig. 8, six clusters are formed based on different times of the day, i.e., early morning, morning, afternoon, evening, night, and midnight. The inner circles represent web page and size of circle shows how frequently this web page is visited.

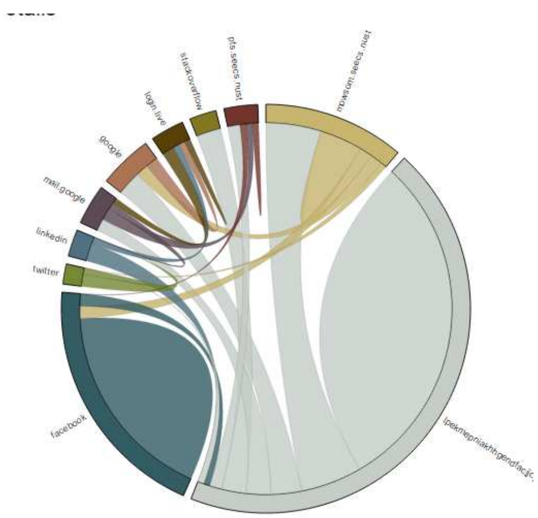


Fig. 7. Tab Switching Behavior

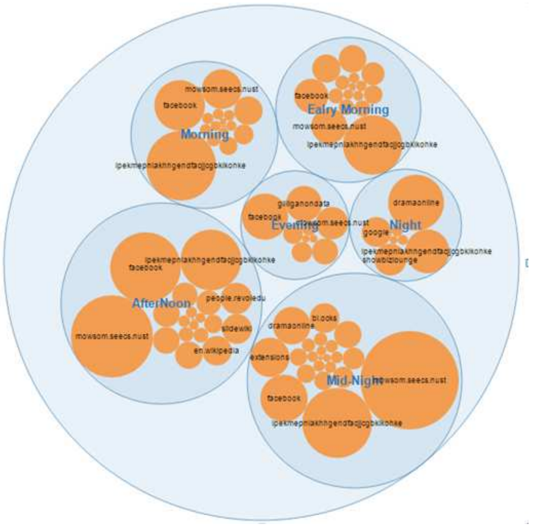


Fig. 8. Cluster of frequent websites at different time of day

## V. Results Analysis and Evaluation

### A. Experimental Evaluation

We collected two-weeks of browsing data from 15 students belonging to the computer science department of the university. They willingly added the extension to their chrome browser from the chrome web store. After installation, they registered to our system after filling the required information in the 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.

### B. Evaluation

We evaluated the extension by arranging an interview session and conducting survey of the 15 participants. We 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 restrict them when they see the unusual behavior and big number in statistics at particular website. Some users have privacy concerns and suggested that user's identity should be removed, and data should be transferred as anonymous user. This aspect will be considered in the future, but for the experiments it was needed for some individual tracking purposes. Although the graphs generated by the activities of different participants revealed interesting patterns and insights, we do not reproduce them here as the previous figures have explained the concepts behind each type of visualization.

### *C. Challenges*

One of the major challenges is to motivate and convince people to use this extension. Interactive visualizations have been implemented that provide users with the quick view about their behavior.

Major challenge in using the browser extension is privacy; people have privacy concerns about data collection. Some users feel hesitated in sharing their data. In order to deal with privacy, domain name of web page is just logged instead of complete URL. The users have also the option to delete all or selected data from any session; however, the interface is rigid and needs future improvement.

Accuracy cannot be assured in case if user deliberately changes his logged data by deleting some data or disabling the extension while browsing some specific websites. If user is watching some video without interacting with the computer, the state of the computer becomes idle, so extension logs this time as idle. This behavior needs to be fixed in the future as well.

## VI. Conclusion and Future Work

This research work has introduced an approach towards capturing and analyzing browsing behavior of individuals over temporal and categorical contexts. A Chrome browser extension has been developed that runs autonomously in the background and captures the browsing activities. 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 for bringing positive changes in their behavior and life.

To extract the valuable patterns from data, different pattern discovery techniques have been utilized including statistical analysis, associative rule mining, sequential pattern mining and clustering. This 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.

Additional data mining and visualization techniques will be integrated at large scale to yield more interesting, effective, and valuable insights from the behavioral data. The current extension 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, so alternative means may need to be found in the future.

## VII. Acknowledgment

This study was funded by the Deanship of Scientific Research, Taif University, KSA, through research project number 1-437-5330.

## References

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

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

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

[4] M. Jafari, F. S. Sabzchi, and A. J. Irani, "Applying web usage mining techniques to design effective web recommendation systems: A case study," *Advances in Computer Science: an International Journal*, vol. 3, no. 2, pp. 78–90, 2014.

[5] P. Mehta, S. B. Jadhav, and R. Joshi, "Web usage mining for discovery and evaluation of online navigation pattern prediction," *International Journal of Computer Applications*, vol. 91, no. 4, 2014.

[6] C. von der Weth and M. Hauswirth, "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, 2013, pp. 51–56.

[7] ——, "Analysing parallel and passive web browsing behavior and its effects on website metrics," *arXiv preprint arXiv:1402.5255*, 2014.

[8] R. Kumar and A. Tomkins, "A characterization of online browsing behavior," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 561–570.

[9] V. Khovanskaya, E. P. Baumer, D. Cosley, S. Voida, and G. Gay, "Everybody knows what you're doing: a critical design approach to personal informatics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 3403–3412.

[10] D. Epstein, F. Cordeiro, E. Bales, J. Fogarty, and S. Munson, "Taming data complexity in lifelogs: exploring visual cuts of personal informatics data," in *Proceedings of the 2014 conference on Designing interactive systems*. ACM, 2014, pp. 667–676.

[11] J.-w. Ahn, K. Wongsuphasawat, and P. Brusilovsky, "Analyzing user behavior patterns in adaptive exploratory search systems with lifeflow," 2011.

[12] M. Kosinski, D. Stillwell, P. Kohli, Y. Bachrach, and T. Graepel, "Personality and website choice," 2012.

[13] H. Zhuang, "I productive? examining the reliability of the quantified self technology," 2013. [Online]. Available: https://www.ucl.ac.uk/uclic/studying/taught-courses/distinction-projects/2013-theses/2013-Zhuang

[14] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[15] A. Inselberg and B. Dimsdale, “Parallel coordinates: a tool for visualizing multi-dimensional geometry;1990,” *San Francisco CA*, pp. 361–375, 1990.

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

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

[18] P. K. Chan, “A non-invasive learning approach to building web user profiles,” 1999.