# TIME BASED DISCOVERING OF WEB USER PATTERNS TO OPTIMIZE WEB SITES AND HYPERLINKS

**[1]EMRAH DONMEZ, [2]ALPER OZCAN**

[1]Computer Engineering İnönü University Malatya, Turkey
[2]Computer Engineering Softtech Research & Development Center İstanbul, Turkey
E-mail: [1]emrahdonmez@msn.com, [2]alper.ozcan@softtech.com.tr

**Abstract-** The Web composes of millions of pages and all these pages contain billions of information about the world. Web users surf on these pages randomly (randomly means that there is no specific navigation) or consciously (it means users look for a specific information). In this study; an analyzing system is offered, whether randomly or consciously surfing, to find that have users drawn a similar surfing pattern at similar days and optimize web sites and hyperlinks with this information. In this way, the pattern of surfing days of users can be mapped by comparing his other similar surfing days. This can be achieved by mining the logged Web usage data of users. Analyzing is of course not based on one or two weeks. This analyzing will be maintained until fluctuation of the Web usage in logged data reflects a coherent pattern on similar days to provide realistic results.

**Keywords-** cloud computing, data security, data privacy.

## I. INTRODUCTION

Information retrieval from Web usage and in this way creating the usage pattern of Web user(s) is a significant research issue in aspect of security, e-commerce and advertising etc. Pattern extraction from Web usage of a user or users has provided a number of useful information to Web site providers, search engines, and others who have looked for statistical information about users. Interests of Web users on a search engine or on a site can be measured as a form of mathematical structures that provide a number of analyzable information. Namely, we can measure what are the Web users looking for. Next step, the links which are clicked by the Web users could be predictable in a range of time. This range of time is a range like that sufficient analysis throughout Web usage of these users is implemented until the pattern of Web usage demonstrates a coherent instance. Therefore, we can foresee next clicking of users to speed-up surfing or reduce the time of browsing. This speed-up can be done by reducing navigation steps of visitors via refining of irrelevant steps touching on later sections. Ads (Advertises) on Web sites especially in commercial sites can be showed more properly in aspect of what users look for. Ads are most used promotional tools in Web sites. And if we think about the high profit coming by advertising, it can be said that making advertising according to interest of Web users by analyzed pattern of their Web usage is obtained more and more prominence for Web sites, especially commercial ones. Web usage mining is the process of extracting useful information from server logs i.e. user's history. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web usage mining touch upon the automatic exploration and clickstream analysis, transaction analysis and investigation of other associated data collected or generated in surfing pattern of users as a consequence of interactions with the Web sites through the Internet [9, 10, and 11]. The aim is detecting, modelling, and analyzing the surfing patterns and profiles of users interacting with Web sites. The identified surfing patterns are generally represented as collections of Web pages, objects, and resources that are frequently accessed/used by groups of users with similar needs or interests.

**This system has three main goals;**

1. Determining the Web usage pattern of users at similar days. Thence, the relevant pages that can be demonstrated to the Web users can be predictable. A user starts to search that what he wants on the internet, and then he encounters a lot of irrelevant pages. By using this method more relevant pages than others can be shown to the Web users. For example; users are searching for shopping word especially on Saturdays, if such a system analyses the Web usage of this users, it can be said that system should demonstrate shopping pages by depending on this information at relevant times of the days and weeks.
2. Web sites can be advertised with relevant ads according to days of the weeks, since the system has already offered that what is searched mostly by users. In this way, a Web site will be advertised with respect to user's interest. For example; while an ad is telecasted for five minutes on Monday, it can be telecasted for fifteen minutes on Saturday. This advertising timetable is strongly depends on that what the users want on these days mostly.
3. The designed system associates the Web pages that commonly visited with other pages according to the analyzed data logged on the systems. This ensures that the Web sites which are mostly visited together (whether they relevant or irrelevant) can be emerged. For example; users visiting shopping

Web sites, visits other pages such as news, community, forums or etc. on a certain day commonly. This correlation between the Web sites provides useful information about Web usage pattern of users.

## II. RELATED WORKS

There are a remarkable number of studies conducted with related to pattern extraction of Web users. We have studied on a comprehensive related works which can enhance the significance of the-state-of –the-art.
Costantinos et al. [1] consider the problem of Web page usage prediction in a Web site by modeling users' navigation history and Web page content with weighted suffix trees. They proposed a method which has the advantage that it demands a constant amount of computational effort per one user's action and consumes a relatively small amount of extra memory space. Resul and Ibrahim [4] studied a method for improving the impressiveness of a Web site by using path analysis method. The aim of their study is to help the Web designer and Web administrator to improve the impressiveness of a Web site by determining occurred link connections on the Web site. Dongshan and Junyi [5] studied efficient data mining for Web navigation patterns. The concept of preference is proposed on the analysis of the present algorithms for mining user navigation patterns. The higher preference, the more prefer to choose the selection. Yueh-Min et al. [6] proposed a Navigational Pattern Tree structure for storing the Web accessing information. To provide real-time recommendations efficiently, they developed a Navigational Pattern mining (NP-miner) algorithm for discovering frequent sequential patterns on the proposed Navigational Pattern Tree.

## III. PROBLEM STATEMENT

Web is a huge information area, whether relevant or irrelevant; we cannot distinguish information that is necessity, with only classical data mining techniques. At the same time, Web usage pattern of users cannot be revealed without additional techniques. Namely, for example; let the term automobile is searched by a user or a group of users in a certain time period, and on the other hand, let the term car, auto or motor car is searched by a user or users. If we try to cluster the term automobile which information; automobile, car, auto or motor car is taken to segment this cluster from other irrelevant ones. Of course we have to look at other information used with above searched terms to good segmentation, "automobile price", "auto cleaner", which one related to automobile known as car or auto or motor car, which one should be segmented and how. To overcome such problems, semantic Web techniques are used to process Web user log data to segment pattern of a user or a group of users with respect to data relation. These semantic techniques are especially specialized to perform segmentation of critical data in dynamic web sites to create an interactive experience with pages by processing previously extracted data.

## IV. WEB USAGE PATTERN ANALYSIS

A. Analysis of Session and User
The statistical analysis of pre-refined log data constitutes the most widely form of analysis. Therefore, predefined units; such as, sessions, users, domains or time aggregate data. Standard statistical methods can be utilized on this data to obtain knowledge related to user behavior.
Integrated usage data can be analyzed by using Online Analytical Processing (OLAP). OLAP provides a more built framework for analysis that can be appeared with a greater scale of flexibility. OLAP can categorize data in a more operational way with its structured functions. The data resource for OLAP analysis is usually a multi-dimensional data repository that congregates usage, content, and e-commerce (online shopping etc.) data at distinct levels of accumulation for each dimension.
B. Cluster Analysis and User (Visitor) Segmentation
Clustering is a data mining technique grouping together a set of items which have simulant characteristics. In the usage domain, there are two types of intriguing clusters that can be discovered: user clusters and page clusters. Each of these clusters can be used separately or together as a hybrid system.
Clustering of users/visitors tends to generate groups of users exhibiting similar browsing patterns. Supplemental analysis of user groups based on their demographic attributes (e.g., age, gender, income level, etc.) may give rise to the exploration of valuable commercial or business intelligence. Usage-based clustering has also been utilized to generate Web-based "user communities" demonstrating similar interests of groups of users [12], and to learn user models that can be utilized to ensure dynamic recommendations in Web customization application [13].
Given a transaction cluster tc, we can build the complete profile $pr_{tc}$ as a set of pageview-weight couples by computing the centroid of tc:

$$pr_{tc} = \{(p, weight(p, pr_{tc})) | weight(p, pr_{tc}) \geq \mu\},$$

Where:

- the importance weight, weight(p, $pr_{tc}$ ), of the page p within the aggregate profile $pr_{tc}$ is given by

$$weight(p, pr_{tc}) = \frac{1}{|tc|} \sum_{s \in tc} w(p, v);$$

- | tc | is the number of transactions in cluster tc;
- w(p, v) is the weight of page p in transaction vector v of cluster tc; and

- the threshold μ is applied to focus solely on those pages in the cluster emerging in a sufficient number of vectors in that cluster.

C. Analysis of Association and Correlation

Analysis by using association rule discovery and statistical correlation can discover groups of items or pages that are commonly accessed or acquired together. This activates Web sites to organize the site content more effectively, or to ensure efficacious cross-sale product recommendations in turn.

Most common approaches to association exploration are depended on the Apriori algorithm. This algorithm explores groups of items occurring frequently together in a great number of transactions (i.e., meeting a user defined minimum support threshold). Such groups of items are designated as "frequent itemsets". Association rules fulfilling a minimum confidence threshold are then procreated from the frequent itemsets.

D. Analysis of Sequential and Navigational Patterns

The practice of successive pattern mining attempts to discover inter-session patterns so that the presence of a set of items is tracked by another item in a time-ordered set of user sessions or sections. By exerting this approach, Web marketers or owners can presumed future user surfing patterns which will be helpful in emplacing advertisements aimed at specific user groups.

Markov models have been proposed as the underlying modeling machinery for link prediction as well as for Web prefetching to minimize system latencies [14, 15]. Another way of efficiently exhibiting contiguous navigational trails is by appending each trail into a tree structure. A good example of this approach is the notion of aggregate tree introduced as part of the WUM (Web Utilization Miner) system [16].

E. Classification and Prediction on Web User Transactions

Classification is the process of mapping a data item into one of a few predefined classes. In the Web domain, one is interested in developing a profile of users relating to a certain class or category. This necessitates extraction and selection of features which best exemplified the properties of given class or category. Classification can be operated by utilizing supervised learning algorithms such as k-nearest neighbor classifiers, decision trees, naive Bayesian classifiers, and Support Vector Machines. It is also feasible to apply previously discovered clusters and association rules for classification of new users.

## V. DESIGN ISSUES

To implement such a system, several tools, classes and some third party software for Web mining have been used. Additionally, to perform system a data set is a requirement. For this reason; several Web log data sets have been used. Let's look at these instruments and utilized data briefly.

A. Utilized Web-Data Mining Tool(kit)s

Rapid Miner; is an environment for machine learning, data mining, text mining, predictive analytics, and business analytics. Rapid Miner ensures data mining and machine learning procedures with visualization. Carrot2 Workbench; is an Open Source Search Results Clustering Engine. It can automatically promote small collections of documents (search results but not only) into thematic categories. Google Prediction API is Google's cloud-based machine learning tools which can help analyze our data to append the following features to our applications: Customer sentiment analysis, message routing decisions, spam detection, recommendation systems, and upsell opportunity analysis etc.

B. Utilized Data

As it is emphasized above several Web log data are used to check system operation and to perform testing process on designed system to determine whether it performs well or not. Starting from this point, we used "Open Directory Project" data for creating label cluster, and log data of a Web mining resources site and a mixed Web log data that compose of 7268 entries. Eventually, we totally have focused on Carrot2 Search engine server log and query log data for main tests.

Open Directory Project Data (O.D.P. Data) is the biggest, most exhaustive intellect-edited directory of the Web. It is built and sustained by a wide, global community of volunteer editors. The Open Directory is the most vastly distributed data base of Web content categorized by humans. Mixed Log Data is obtained from w3.org site as under the GPL license. This log data have 7268 entries. Unlike the previous mentioned data set, this log data has 7268 distinct sessions. This means that there are 7268 different accessing done by users. Carrot2 Data comprises a collection of ~6.7M Web queries collected from ~250k users throughout three months. The data is sorted by anonymous user ID and sequentially collated. The goal of this collection is to provide real query log data that is based on real users. It could be used for personalization, query reformulation or other types of search research.

**Sample Server Log**

2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221
HTTP/1.1 maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)
http://dataminingresources.blogspot.com/

## VI. EXPERIMENTAL DESIGN & RESULTS

First results of performed system demonstrate that the users surfing the Web, can be classified with respect to day of the weeks. The results of offered analyzing system are based on surfing patterns of users which are extracted from server log data of a search engine (Carrot2 Search Engine Query Log Data). In order to examine analysis from different aspects several

datasets emphasized above in section 5, are used. Because of the huge data in Carrot2 dataset, we first, randomly selected 10K users from this dataset to get first results without make a comprehensive classification. Surfing patterns of selected one user which are picked from these 10K users with respect to day of the weeks are demonstrated in table 1. This table is crated according to surfing patterns of users implemented in several weeks (e.g. 10 weeks – 12 weeks). Because of huge data (6.7M Search query), unfortunately, we only give weekly surfing pattern rate of percentiles for selected one user. But, on the other hand, average surfing patterns of one user and a general average surfing pattern which is about 10K users has been given in table 2. It can be said that User1 is strongly presumably an undergraduate or graduate student on engineering area or a researcher, academician etc. His specific research area can be determined by looking his queries about engineering issues. These huge Carrot2 search engine query logs data is obtained by not only directly searching on Carrot2 search engine but also by using Carrot2 toolbar installed on browser, and address bar specialized with respect to Carrot2 search engine add-on. In order to eliminate irrelevant queries such as only composed of one single alphabet or special characters (for instance; "aaaaaaaaa", "bbbbbbxxxxx" or "-----***" etc.) data are subjected to cleaning and filtering process to create a meaningful dataset. This data pre-processing step includes general missing values, misspellings and violated attribute dependencies etc. To implement this refinement general data cleansing techniques and multiple linear regression for filtering (or noise reduction) are required. But these processes are laborious, time consuming and prone to errors itself. Because of these additional workloads, to alleviate heavy tasks we use Rapid Miner toolkit to filter data. Since such terms decrease the classification success not necessarily. Therefore, term irrelevance has to be considered for high level accuracy in classification process.

**TABLE 1. AVERAGE RATE OF WEB SURFING PATTERN FOR A USER**

| User1 | |
|---|---|
| Monday | 53.4% Engineering topics, 21.3% Social topics (i.e. facebook, youtube, delicious, video yahoo…), 9.9% News, 10.2% Mails, 5.4% Others (products; books, clothing etc.) |
| Tuesday | 58.1% Engineering topics, 18.8% Social topics (i.e. facebook, youtube, delicious, video yahoo…), 10.4% News, 9.7% Mails, 3% Others (products; books, clothing etc.) |
| Wednesday | 60.8% Engineering topics, 17.5% Social topics (i.e. facebook, youtube, delicious, video yahoo…), 7.9% News, 8.7% Mails, 5.1% Others (products; books, clothing etc.) |
| Thursday | 61.5% Engineering topics, 15.6% Social topics (i.e. facebook, youtube, delicious, video yahoo…), 8.2% News, 12.5% Mails, 2.2% Others (products; books, clothing etc.) |
| Friday | 38.7% Engineering topics, 26.9% Social topics (i.e. facebook, youtube, delicious, video yahoo…), 15.2% News, 10.5% Mails, 8.7% Others (products; books, clothing etc.) |
| Saturday | 32.3% Social topics (i.e. facebook, youtube, delicious, video yahoo…), 30.5% Engineering topics, 14.4% Others (products; books, clothing etc.), 13% News, 9.8% Mails |
| Sunday | 45.2% Engineering topics, 24.6% Social topics (i.e. facebook, youtube, delicious, video yahoo…), 11.3% Others (products; books, clothing etc.), 9.9% Mails, 9% News |

In table 1, there are several things which have to be emphasized. Firstly, to classify each topic, we used k-means clustering algorithm by extracting queries from query logs. For instance; to classify topics ensured by surfing pattern of User1, topics labeled approximately as about 10 topics more than one label and each of the labeled topic is partitioned into at most 20 labeled subtopics, and each of these labeled subtopics partitioned into at most 20 labeled base-subtopics. At the end, we have obtained $10x20^2 = 4000$ total labeled topics. If a query is performed, firstly we scanned the topics. If query is not found, then the searching process continues in lower layer which includes subtopics. If again there is nothing, then process continues in the deepest layer which includes base-subtopics. Then, the each found query is clustered according to this topic tree structure. To support clustering process, we use simple hybrid natural language processing (NLP) techniques with some heuristics to discover relation between searched query and labeled topics. Each topic have evaluated as an item. After the daily clustering process that performed in 24h, is completed, the percentile rate of each cluster is computed. Total cluster size is evaluated over one hundred percent. For instance social topics represent 21.3% or a bit over 1/5 of the total topics size formed by User1 at Monday. These surfing pattern knowledge will be useful to demonstrate him relevant pages according to interest of him by topics. For instance; a commercial education Web site can recommend him e-education courses as video or other multimedia components.

Secondly, each day of the week mentioned above reflects division of total number of surfing rate to the number of same days throughout the researches. Let's donate the total number of rate on same days as;

$$\sum_s^f R_s$$

s represents the starting week of analysis, and f represents the number of weeks which starts s and continue until analysis are finalized. $R_s$ represents the surfing rate of relevant day. Then we can easily find the average rate of surfing pattern for each day of week, by using;

$$\frac{\sum_s^f R_s}{f}$$

In order to cluster each user to a clustering group there should be similarities between users according to their interests of topics. These similarities can be obtained from users interests to labeled topics. Starting from this point, we use three layered topics tree to reveal these similarities. Firstly, the histograms of each user daily interests are extracted from topic interest rates. Next step, the acquired histogram values are converted to matrix values. Then, each of the matrix are compared each other, and at last step, clustering process according to these matrix is performed. Eventually we have obtained clustered users whose interests to topics are similar in each cluster. There is one more thing which has to be

mentioned that; to similarity measure we have defined a manual threshold value.

Similarity matrix;

$$\cdots\cdots\begin{bmatrix} a & b & c & \cdots \\ d & e & f & \cdots \\ g & h & i & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}\cdots\cdots$$

Let's donate daily interest rate of the topics as;

$$T = \{t_1, t_2, t_3, t_4, t_5, t_6, \ldots\ldots, t_n\}$$

Similarities between two users $(U_1, U_2)$ have been calculated by using;

$$S(U_1, U_2) = \frac{\frac{min(U_1t_1, U_2t_1)}{max(U_1t_1, U_2t_1)} + \frac{min(U_1t_2, U_2t_2)}{max(U_1t_2, U_2t_2)} + \cdots + \frac{min(U_1t_n, U_2t_n)}{max(U_1t_n, U_2t_n)}}{\sum_s^x T_s + \sum_j^y T_d}$$

In this formula; $(U_1t_1)$ represents the daily rate of the topics of user1, similarly $(U_2t_1)$ body forth the daily rate of the topics of user2. $\max(U_1t_1, U_2t_1)$ stand for the maximum value of the $(U_1t_1)$ and $(U_2t_1)$, for example; let $(U_1t_1) = 21.5$ and $(U_2t_1) = 19.7$ then we select of course 21.5 and the result of the first division process are found as; $\frac{19.7}{21.5} = 0.916$. $\sum_s^x T_s$ represents the total number of the similar topic, it means that this computation gives similar topics which exist in interests of user1 and user2 at the same time. $\sum_j^y T_d$ represents the total number of the distinct topic, it means that this computation gives distinct topics which exist in one of the interests of user1 or user2; not in two at the same time. Each row and column value of matrix reflects a similarity between two users who are related a group of topics. By comparing each of these matrix values, clusters be able to be generated. In order to overcome unique users who are could not be classified, a simulation method is used. Namely, if there are unique users, then let's place them to the cluster which has nearest average similarity value to these unique users similarity. Total number of clustered users and average rate of interests according to topics are given in table 2.

**TABLE 2. AVERAGE RATE OF WEB SURFING PATTERN OF TOTAL USERS**

| Total Users Number=10K | |
|---|---|
| Monday | 40.6% Social, 18.7% Education, 10.6% News, 10.1% Sport, 9.8% Technology, 6,2% Health, 4.0% Others |
| Tuesday | 42.5% Social, 19.1% Education, 11.2% News, 10.5% Technology, 9.9% Sport, 4,9% Health, 1.9% Others |
| Wednesday | 41.3% Social, 21.4% Education, 10.9% News, 10.8% Technology, 9.5% Sport, 3,7% Health, 2.4% Others |
| Thursday | 42.2% Social, 21.7% Education, 10.4% News, 9.3% Technology, 8.7% Sport, 4,1% Health, 3.6% Others |
| Friday | 49.8% Social, 15.3% Education, 10.8% Technology, 9.8% Sport, 8.6% News, 3,6% Health, 2.1% Others |
| Saturday | 55.4% Social, 13.1% Education, 10.1% Sport, 9.3% Technology, 7.1% News, 2,5% Health, 2.5% Others |
| Sunday | 51.2% Social, 17.4% Education, 12.1% Sport, 8.3% Technology, 5.1% News, 3.1% Others, 2,8% Health |

It can be said that, almost half of the surfing pattern is consumed by social activities with respect to days. Since, we can clearly see that social surfing is increased towards the end of the week. Main reason of this situation is that people have more free time at weekend than weekdays. That explains the decreasing rate in education. News and technology demonstrate close level pattern to each other. At the end of week sport is increased by users, this is because; the sportive activities are mostly carried out at weekend. In figure 1, it has given that the average rate of surfing pattern of users with respect to average of total days that is performed throughout the three month experiments. Remind that these percentage values acquired from Carrot2 log data by classified this data according to time information after preprocessing of the related data.

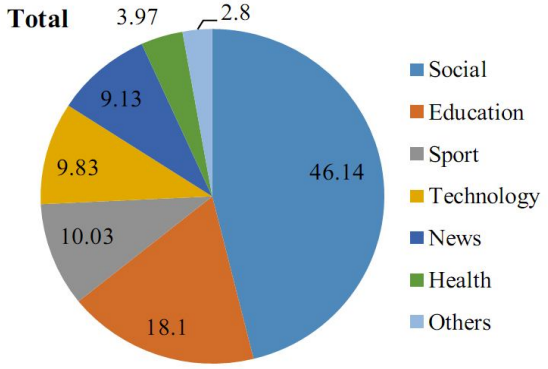


**Figure 1. Average rate of Web surfing pattern of total users according to average of total days**

Query density of users has been demonstrated in figure 2. According to this figure, it can be said that same or similar queries were performed at similar time pattern of days. Similarity of queries are represented with colors. Each color has different tones. For instance; if there are similar queries then these queries are represented different tones of the same color.

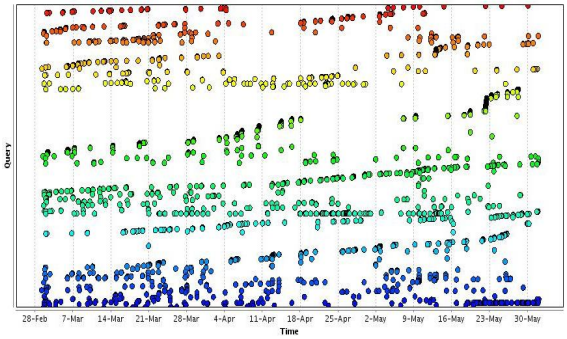


**Figure 2. Query density of Web users according to time. Coloring points show that query similarity between total users throughout defined days.**

Item-rank shows rank value of items of user search and navigation from search engine, toolbar, and browser. If the user clicked on a search result, the rank of the item on which they clicked is listed have been exhibited in figure 3. Rank process has been performed by Web users emerged between 1 and 146.

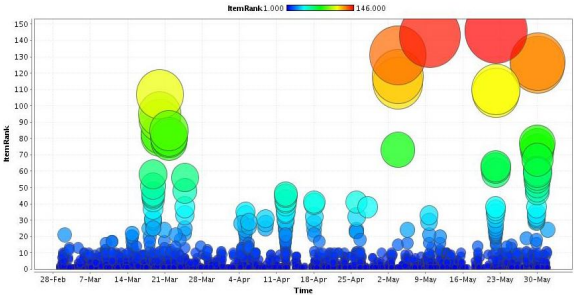


**Figure 3. Item-rank that performed by users according to time axis. Coloring points show that item-rank similarity throughout defined days.**

There are three kind of searching method performed by users in figure 5; query activities represent direct searching with search engine. Address bar represents searching activities that performed by using browser's address bar. And toolbar represent an add-on bar which provide shortcut for searching activities Figure represent the whole search engine log data set.

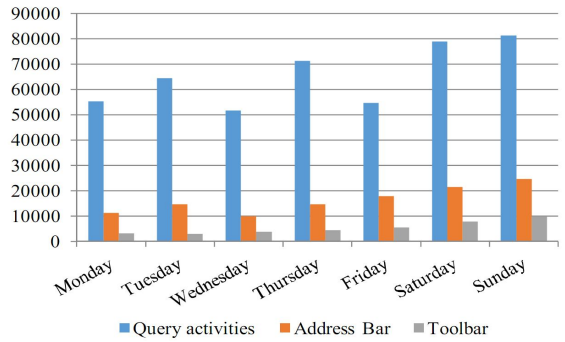


**Figure 5. Average query counts with respect to days of week..**

## VII. DISCUSSION

Early results of initial experiments show that people sometimes could demonstrate very distinct surfing patterns according to days of the week. To better understand each pattern of the users, analyzing variables can be incremented. Such as seasonal effect, day and night difference, location difference, educational difference and so on. Proposed system has been performed on three month Web log data of a search engine. Observed results demonstrate that the surfing pattern of users could be useful to design better Web sites. Namely, if we know the average surfing pattern of users, we can provide better Web service as Web site or search engine owners.

## CONCLUSION & FUTURE WORKS

Proposed system shows daily activities of users at a satisfactory level. Main advantage of proposed system is handling the existing problem with facile approach. By this way, working speed of system is kept on a good level. Web sites or search engines and hyperlinks between them can be optimized with daily surfing pattern information easily. For instance; customizable reception and transition pages could be contemplated according to each of the Web users.

In the future, may be such systems will be used for determine the specific characteristics of Web users. What and when he likes in a specific season, or where he likes etc. and of course opposites of all will be predictable. By this way his time and location depended navigational characteristics can be mapped to a mathematical model. This model(s) can be a reference in designing dynamically optimized Web sites. And this system will be useful for not only the Web world but also other media such as TV, radio, advertising panels that located on the streets or etc. We called such a system as pattern recognition of Web usage.

## REFERENCES

[1] C. Dimopoulos, C. Makris, Y. Panagis, E. Theodoridis and A. Tsakalidis, "A Web page usage prediction scheme using sequence indexing and clustering techniques," Data & Knowledge Engineering, vol. 69, issue 4, pp. 371-382, April 2010.

[2] H. Liu and V. Kešelj, "Combined mining of Web server logs and Web contents for classifying user navigation patterns and predicting users' future requests," Data & Knowledge Engineering, vol. 61, issue 2, pp. 304-330, May 2007.

[3] L. Chen, S. S. Bhowmick and W. Nejdl, "COWES: Web user clustering based on evolutionary Web sessions," Data & Knowledge Engineering, vol. 68, issue 10, pp. 867-885, October 2009.

[4] R. Das and I. Turkoglu, "Creating meaningful data from Web logs for improving the impressiveness of a Web site by using path analysis method," Expert Systems with Applications, vol. 36, issue 3, part 2, pp. 6635-6644, April 2009.

[5] D. Xing and J. Shen, "Efficient data mining for Web navigation patterns," Information and Software Technology, vol. 46, issue 1, pp. 55-63, January 2004.

[6] Y.-M. Huang, Y.-H. Kuo, J.-N. Chen and Y.-L. Jeng, "NP-miner: A real-time recommendation algorithm by using Web usage mining," Knowledge-Based Systems, vol. 19, issue 4, pp. 272-286, August 2006.

[7] K. Becker and M. Vanzin, "O3R: Ontology-based mechanism for a human-centered environment targeted at the analysis of navigation patterns," Knowledge-Based Systems, vol. 23, issue 5, pp. 455-470, July 2010.

[8] B. Hay, G. Wets and K. Vanhoof, "Segmentation of visiting patterns on Web sites using a sequence alignment method," Journal of Retailing and Consumer Services, vol. 10, issue 3, pp. 145-153, May 2003.

[9] R. Cooley, B. Mobasher and J. Srivastava, "Web mining: Information and pattern discovery on the world wide web," In Proceedings of Intl. Conf. On Tools With Artificial Intelligence (ICTAI-1997), 1997.

[10] B. Mobasher, "Web usage mining," John Wang (eds.), Encyclopedia of Data Warehousing and Mining, Idea Group, pp. 449-483, 2006.

[11] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, "Web usage mining: Discovery and applications of usage patterns from Web data," ACM SIGKDD Explorations Newsletter, 1(2), pp. 12-23, 2000.

[12] G. Paliouras, C. Papatheodorou, V. Karkaletsis and C. Spyropoulos, "Discovering user communities on the Internet using unsupervised machine learning techniques," Interacting with Computers, 14(6), pp. 761-791, 2002.

[13] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, "Discovery and evaluation of aggregate usage profiles for Web personalization," Data Mining and Knowledge Discovery, 6(1), pp. 61-82, 2002.

[14] M. Deshpande and G. Karypis, "Selective Markov models for predicting Web page accesses," ACM Transactions on Internet Technology (TOIT), 4(2), pp. 163-184, 2004.

[15] R. Sarukkai, "Link prediction and path analysis using Markov chains1," Computer Networks, vol. 33(1-6), pp. 377-386, 2000.

[16] M. Spiliopoulou and L. Faulstich, "WUM: a tool for Web utilization analysis," The World Wide Web and Databases, pp. 184-203, 1999.

**Emrah Donmez** is a Ph.D. student in the Computer Engineering Department at İnonü University. His research areas are: security of cloud computing, data and Web mining with HPC systems, performance optimization in HPC systems, GPGPU acceleration and data mining, machine learning, distributed operating systems, global, grid, volunteer and hybrid computing; problems & solutions. He is a student member of IEEE.

**Alper Ozcan** is pursuing his PhD in Department of Computer Engineering at Istanbul Technical University. He received his MS degree from Istanbul Technical University. His research interests include link prediction in social networks, data mining, web mining, machine learning, artificial intelligence, software engineering. He is a student member of IEEE.

★★★