Journal of Computer Sciences and Applications, 2013, Vol. 1, No. 3, 39-45
Available online at http://pubs.sciepub.com/jcsa/1/3/3

DOI:10.12691/jcsa-1-3-3

Extracting Users'Navigational Behavior from Web Log Data: a Survey

Maryam Jafari[1], Farzad SoleymaniSabzchi[1,*], Shahram Jamali[2]

[1]Sama Technical and Vocational College, Islamic Azad University, Ardabil Branch, Ardabil, Iran
[2]Computer Engineering Department, University of Mohaghegh Ardabili, Ardabil, Iran
*Corresponding author: f_soleymani63@yahoo.com

Received December 26, 2012; Revised May 04, 2013; Accepted May 10, 2013

Abstract Web Usage Mining (WUM) is a kind of data mining method that can be used to discover user access patterns from Web log data. A lot of research has been done already about this area and the obtained results are used in different applications such as recommending the Web usage patterns, personalization, system improvement and business intelligence. WUM includes three phases that are called preprocessing, pattern discovery and pattern analysis. There are different techniques for WUM that have their own advantages and disadvantages. This paper presents a survey on some of the existing WUM techniques and it is shown that how WUM can be applied to Web server logs.

Keywords: *web usage mining, web log mining, pattern discovery, preprocessing, sequence mining*

1. Introduction

World Wide Web (WWW) is very popular and interactive. It has become an important source of information and services. The Web is huge, diverse and dynamic. Extraction of interesting information from Web data has become more popular and as a result Web mining has attracted lot of attention in recent time [1]. Web mining can be defined roughly as data mining using data generated by the Web [2]. It can be divided in to three categories namely Web Structure Mining (WSM), Web Content Mining (WCM) and Web Usage Mining (WUM) [3]. WSM tries to discover the link structure of the hyperlinks at the inter-document level and generates a structural summary to examine data related to the structure of a particular Website. WCM mainly focuses on the structure of inner-document to find useful information in the content of Web pages such as free text inside a Web page, semi-structured data such as HTML code, pictures, and downloadable files.

WUM attempts to discover useful knowledge from the secondary data, especially those contained in Web log files. Other sources can be browser logs, user profiles, user sessions, bookmarks, folders and scrolls. These data are obtained from the interactions of the users with the Web. Effective Website management, creating adaptive Websites, business and support services, personalization, and network traffic flow analysis efficiently use WUM for better performance. WUM focuses on the techniques that can predict user's navigational behavior.

In [3], there are three main tasks for performing WUM: Preprocessing, Pattern Discovery, and Pattern Analysis, as shown in Fig. 1.Because of the importance of pattern discovery, this paper focuses on describing this phase and presents an overview of WUM and also provides a survey of the different techniques of pattern extraction used for WUM.

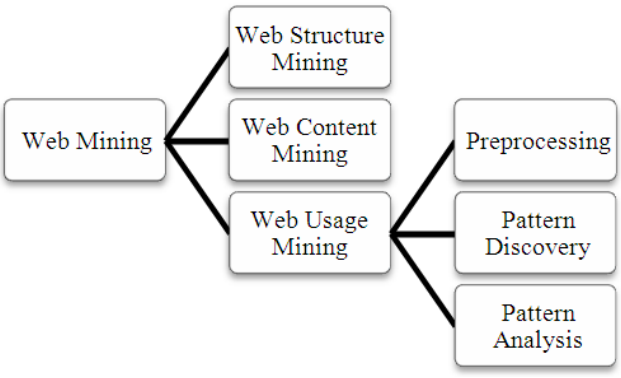

Figure 1. Taxonomy of web mining

The remainder of this paper organized as follows. In the next section related works on WUM and pattern discovery phase are reviewed. In section 3 an overview on WUM is presented and then different phases of WUM and the related techniques of these phases are examined. Finally a conclusion of this work is presented in section 4.

2. Related Works

Etzioni [4] proposed a new concept in 1996 that was called "Web mining". He used data mining techniques to automatic discovery and extract information from abundant data on the World Wide Web. WUM was first proposed by Chen et al. [5] and [6], Mannila and Toivonen [7] and Yan et al. [8]. Baraglia and Palmerini presented a WUM system called SUGGEST which optimizes the Web server performance by providing

useful information. This system provides an objective behavior for user navigation. Jianhan Zhu et al. [9] used the Markov chains to model user's navigational behavior. They proposed a method for building a Markov model of a Website based on previous users'behavior. Then the Markov model is used to make link predictions that help new users to navigate the Website. Jalali et al. [10] presented a system for extracting user's navigational behavior using a graph partitioning model. An undirected graph based on connectivity between each pair of theWeb pages was considered and also proposed a new formula for allocating weights toeach edge of the graph.

In [11] a method to predict the user's navigation patterns is proposed using clustering and classification from Web log data. First phase of this method focuses on separating users in Web log data, and in the second phase clustering process is used to group the users with similar preferences. Finally in the third phase the results of classification and clustering are used to predict the users' next requests. Emine Tug et al. [12] found sequential accesses from Web log files, using Genetic Algorithm (GA) that called Automatic Log Mining via Genetic (ALMG). In their work, GA based on evolutionary approach for pattern extraction was used to find best solutions for time consuming problem to discover sequential accesses from Web log data.Kim and Zhang use a genetic algorithm to learn the important factors of HTML tags which are used to re-rank the documents retrieved by standard weighting schemes for Web structure mining [13]. Picarougne et al. presented a genetic search strategy for a search engine [14]. Abraham and Ramos proposed an ant clustering algorithm to discover Web usage patterns (data clusters) and a linear genetic programming approach to analyze the visitor trends [15].

Some systems have been developed based on Web mining for automatic personalization [16,17,18]. They generally consist of two major processes: off-line mining and on-line recommendation. In the off-line mining process, all the access activities of users in a Website are recorded into the log files by the Web server. Then, some Web mining processes are applied to the server logs to mine the hidden navigation models of users. In the on-line recommendationprocess, user's requests from his current active session are recorded. By comparing these requests with the models obtained from the off-line mining, appropriate personalized recommendations are generated. Mobasher et al. [19] made an attempt to integrate both usage and content attributes of a site into a Web mining framework for Web personalization. A ''post-mining'' type approach was implemented to obtain the uniform representation for both site usage and site content profiles to facilitate the real-time personalization. However, the techniques proposed in [14] were limited to the use of clustering to separately build site usage and content profiles.

In [20], Sarukkai discussed about link prediction and path analysis for better user navigations. He proposed a Markov chain model to predict the user access pattern based on the user access logs previously collected. Chen et al. [21] introduced the concept of using the maximal forward references in order to break down user sessions into transactions for the mining of traversal patterns. A maximal forward reference is the last page requested by a user before backtracking occurs, where the user requests a page previously viewed during that particular user session.

In the following, we describe some common features of previous studies in WUM.

- The goal of papers in this field is to improve Web services and performance through the improvement of Websites, including their contents, structure, presentation, and delivery.
- They focus on the mining of server side data. Not only their data sources are almost exclusively server log files, but sometimesstructure and contents of sites are applied too.
- Considering that dealing with users on an individual basis is overwhelming for a Website, They focus on groups of users instead of individual users.

3. Overview on WUM

WUM is application of data mining techniques to discover user access patterns from Web data. Web usage data captures Web-browsing behavior of users from a Website. The task of modeling and predicting a user's navigational behavior on a Website or on a Web domain can be useful in quite many Web applications such as Web server caching that provides an interface between a single Web server and all of its users. It reduces the number of requeststhat the server must handle, and then helps load balancing, scalability and availability [22,23].Web page recommender systems that help peopleto make decisions in complex information space where the volume of information availableto them is huge [24,25].Web search engines that usually help users locate information based on the textual similarity of a query and potential documents [26,27] and Web search personalization that its goal is to tailor search results to a particular user based on that user's interests and preferences [28].

WUM has other several applications [29] such as: business intelligence, e-Learning, e-Business, e-Commerce, e-Newspapers, e-Government and Digital Libraries. Most of the WUM techniques are based on association rules, sequential patterns and clustering [30]. WUM involves of three main phases that are described as follows.

3.1. Preprocessing

The information available in the Web is heterogeneous and unstructured. Therefore, data preprocessing is predominantly significant phase in WUM. The goal of preprocessing is to transform the raw collected data into a set of user profiles [31]. This phase is often the most time-consuming and computationally intensive step in WUM, but it is necessary to have a successful analysis of Web usage patterns.

Every log entry of Web server log contains the traversal time from one page to another, the IP address or domain name, time and type of request (GET and POST, etc.), address of the page being accessed and other data [32]. Preprocessing removes many entries from the data files that are considered uninteresting data for pattern discovery.

Various research works are carried in this area for grouping sessions and transactions, which is used to discover user's navigation patterns. In brief, the whole

process deals with the conversion of raw Web server logs into a formatted user session file in order to perform effective pattern discovery and analysis phases. Generally,data preprocessing has four main tasks that are called data cleaning, user identification, session identification and path completion, as shown in Figure 2.

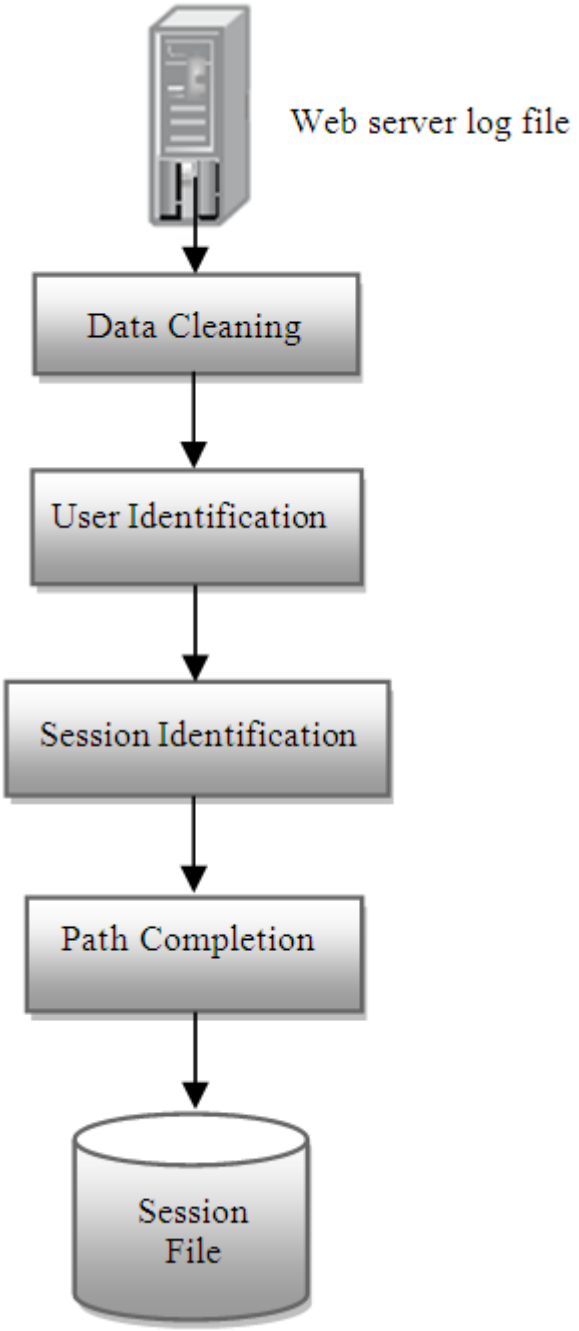

Figure 2. Steps in data preprocessing for WUM

3.1.1. Data Cleaning

In this task the server log is examined to remove the irrelevant and redundant items for the mining process. There are three kinds of irrelevant or redundant data needed to clean:

(1) Accessorial resources embedded in HTML file: A user's request to view a favorite page often records in several log file entries since file requests that the user did not explicitly request such as graphic files and scripts add entries in Web log file. Therefore, all log entries with an extension such as gif, jpeg, GIF, JPEG, jpg, JPG, css, cgi and map in their filename should be removed.

(2) Robots' requests: Search engines such as Google periodically use Web robots (also called spiders) to navigate on Web and do accurate searches on Websites update their search indexes [33]. Therefore, it is required to try to rid the Web log file of these types of automatic access behavior.

(3) Error requests: Erroneous files are useless for WUM and can be removed by examining the HTTP status codes. For example, if the status code is 404 it means that the requested resource is not existence, so this log entry can be removed. Finally, log entries with status codes between 200 and 299 that give successful response are kept and entries which have other status codes are removed.

3.1.2. User Identification

This step focuses on separating the Web users from others. User Identification means identifying Unique users considering their IP address.

Following heuristics are used to identify unique users:

(1) If there is a new IP address, then there is a new user.

(2) For more logs, if the IP address is the same, but the operating system or browsing software are different, a reasonable assumption is that each different agent type for an IP address represents a different user.

Existence of local caches, corporate firewalls and proxy servers greatly complicate user identification task. The WUM methods that rely on user cooperation are the easiest ways to deal with this problem. However, it's difficult because of security and privacy.

3.1.3. Session Identification

Visited pages in a user's navigation browsing must be divided into individual sessions. A session means a set of Web pages viewed by a particular user for a particular purpose. At present, the methods to identify user session include timeout mechanism and maximal forward reference mainly [34].

The following rules are used to identify a session:

(1) For any new IP address in Web log file, a new user and also a new session will be created.

(2) In one user session, if the refer page in an entry of Web log file is null, a new session will beconsidered.

(3) If the time between page requests is morethan 25.5 or 30 minutes, it is assumed that theuser is starting a new session.

3.1.4. Path Completion

Many of important page accesses are missed in the Web log file due to the existence of local cache and proxy server. The task of path completion is to fill in these missing page references and makes certain, where the request came from and what all pages are involved in the path from the start till the end.

3.2. Pattern Discovery

Pattern discovery is a phase which extracts the user behavioral patterns from the formatted data. For this reason the data have to be converted in the preprocessing phase such that the output of the conversion can be used as the input of this phase. In pattern discovery phase,

several data mining techniques are applied to obtain hidden patterns reflecting the typical behavior of users.

Some important techniques for this phase are: path analysis, standard statistical analysis, clustering algorithms, association rules, classification algorithms, and sequential patterns. In the following, some of these techniques are described.

3.2.1. Statistical Analysis

Statistical analysis is the most common form of analysis to extract knowledge about visitors' behavior. By analyzing the obtained session file from Web log, useful statistical information such as frequency, mean, median, etc. can be resulted. This statistical information is used to produce a periodic report from the site such as information about users' popular pages, average visit time of a page, average time of users' browsing through a site, average length of a navigational path through a site, common entry and exit pages and high-traffic days of site.

According to these reports, it is clear that used statistical technique for pattern discovery perform a sketchy analysis on preprocessed data but obtained knowledge can be useful. For instance detecting entry points which are not home page or finding the most common invalid URL lead to enhance system performance and security and also facilitate the site topology modification task. The useful statistical information discovered from Web logs is shown in Table 1.

Table 1. Important Statistical Information

Statistics	Detailed Information
Website Activity Statistics	Total number of visits Main number of hits Successful/failed/redirected/ hits Average view time Average length of a path through a site
Troubleshooting/ Diagnostic Statistics	Server errors Page not found errors
Server Statistics	Top pages visited Top entry/exit pages

3.2.2. Sequential Patterns

The technique of sequential pattern discovery is to find inter-session patterns such that the presence of a set of pages is followed by another page in the time-stamp ordered session set. This mining is trying to find the relationships between sequential visits, to find if there exists any specific time order of the occurrences. The goal of this technique is to discover time ordered sequences of URLs followed by past users, in order to predict future pages. This prediction helps Web marketers to target advertising aimed at groups of users based on these patterns. An example of Web server access logs analysis by using the Web mining system can show temporal relationships discovering among data items such as the following:

(1) 30% of clients who visited/company/products/, had done a search in Yahoo,within the past week on keyword data mining; or

(2) 60% of clients, who placed an online order in /computer/products/webminer.html, also placedan online order in /computer/products/iis.html within 10 days.

From these relationships, vendors can develop strategies and expand business.

3.2.3. Classification

Classification is to build automatically a model that can classify a set of pages. It is the task of mapping a page into one of several predefined classes [35]. In the Web domain, classification techniques allow one developing a profile of users which are belonging to a particular class or category and access particular server files. This requires extraction and selection of features that based on demographic information available on these users, or based on their access patterns. This technique has two steps. The first step is based on the collection of training data set and a model is constructed to describe the features of a set of data classes. In this step, data classes are predefined so it is known as supervised learning. In the second step, the constructed model is used to predict the classes of future data. For example, classification on server access logs may lead to the discovery of interesting patterns such as the following:

(1) Users from state or government agencies whovisit the site tend to be interested in the page /company/lic.html.

(2) 60% of users, who placed an online orderin /company/products /Music, were in the range of 18-25 years oldand lived in Chandigarh.

3.2.4. Clustering

Clustering is another mining technique similar to classification however unlike classification there are no predefined classes therefore, this technique is an unsupervised learning process. This technique is used to group together users or data items that have similar characteristics, so that members within the same cluster must be similar to some extent, also they should be dissimilar to those members in other clusters.

In the WUM domain, clustering techniques are mainly used to discover two kinds of interesting clusters: user clusters and page clusters. Clustering of users is to cluster users with similar preference, habits and behavioral patterns. Such knowledge is especially used for automated return mail to users falling within a certain cluster, or dynamically changing a particular site for a user, on a return visit, based on past classification of that user (provide personalized Web content to the users). On the other hand, clusters of Web pages contain pages that seem to be conceptually related according to the users' perception. The knowledge that is obtained from clustering in WUM is useful for performing market segmentation in ecommerce, designing adaptive Websites and designing recommender systems.

3.2.5. Association Rule Mining

In the context of WUM, once sessions have been identified association rules can be used to relate pages that are most often referenced together in a single server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. Since usually such transaction databases contain extremely large amounts of data, current association rule discovery techniques try to prune the search space according to support for items under consideration. Support is a measure based on the number of occurrences of user transactions within transaction logs. The typical rule mined from database is formatted as (1):

$$X \rightarrow Y\left[Support, Confidence\right] \quad (1)$$

It means the presence of item (page) X leads to the presence of item (page) Y, with [Support]% occurrence of [X,Y] in the whole database, and [Confidence]% occurrence of [Y] in set of records where [X] occurred.

$$Support = P(A \cap B) = \frac{number\ of\ sessions\ that\ contain\ A\ and\ B}{total\ number\ of\ sessions} \quad (2)$$

$$Confidence(X \rightarrow Y) = \frac{Support(X \cap Y)}{\sup port(X)} \quad (3)$$

Many algorithms can be used to mine association rules from the data available; one of the most used and famous is the Apriori algorithm proposed and detailed by Agrawal and Srikant in 1994 [36]. This algorithm, given the minimum support and confidence levels, is able to quickly give back rules from a set of data through the discovery of the so-called large item set.

For example, if one discovers that 80% of the users accessing/computer/products/printer.html and /computer/products/scanner.html also accessed, but only 30% of those who accessed/computer/products also accessed computer/products/scanner.html, then it is likely that some information in printer.html leads users to access scanner.html.

This correlation might suggest that this information should be moved to a higher level to increase access to scanner.html. This also helps in making business strategy that people who want to buy printer; they are also interested in buying scanner. So vendors can offer some discount on buying combo pack of printer and scanner. Or they can offer discount on one item for the purchase of both or they can apply buy one, get one free strategy. Discovery of such rules for organizations engaged in electronic commerce can help in the development of effective marketing strategies. Apart from being exploited for business applications, the associations can also be used for Web recommendation, personalization or improving the system's performance through predicting and pre-fetching of Web data.This type of result is for instance produced by [37] using a modification of the Apriori algorithm [38]. Reference [39] proposes and evaluates measures of interest to evaluate the association rules mined from Web usage data. Reference [40] exploits a mixed technique of association rules and fuzzy logic to extract fuzzy association rules from Web logs.

3.3. Pattern Analysis

Pattern analysis is the last step in the overall WUM process that has two fundamental goals. The first goal is to extract the interesting rules, patterns or statistics from the output of the pattern discovery process by filtering the irrelative rules or statistics. Another aim of this analysis is to obtain some information can offer valuable insights about users' navigational behavior.For example we can understand the number of users that started from a page and proceeded through some certain pages and finally visited their goal page. Also, we can obtain some information about page popularity or some pages that contain the most information for a visitor.The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis is combining WUM tools with a knowledge query mechanism such as SQL. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

4. Challenging and Open Issues of WUM

Based on gained observations from reviewed research works, in this section we analyze some challenges and open issues in related works in the field of WUM. According to performed studies, we concluded that the major problems with WUM are the quantity of the Web usage data to be analyzed and its low quality. More and larger Websites and more visitors are the main reasons for this. When the classic algorithms of data mining applied to these no preprocessing data give unsatisfactory results in terms of behaviors of the Websites' users. Therefore, some preprocessing tasks are needed to reduce the noisy data of Web log files before applying the pattern discovery techniques to find the relationship between the log files. Apart from the volume of the data and its low quality, the data is not completely structured. It is in a semi-structured format so that it needs a lot of preprocessing and parsing before the actual extraction of the required information.

In addition, by examining various implemented techniques for mining Web usage data we realized that another challenge in WUM is user's interest issue. Most of the researches have analyzed usage logs with the purpose of developing an intelligent system that learns user features and builds a model of user. However, most of the studies did not fully consider the effects of various factors to measure user's interest, or they focused only on a user's interest without consideration of other aspects of the interestingness. For example, in most of the papers page view duration by a user is regarded. Although, page view duration is heavily depends on the size of the page. Moreover, due to large amount of data in log files most researches did not examine the long period of time of usage data. Recently, web usage mining and web content mining have been combined in order to provide a better understanding of the requirements a visitor has when entering a web site.

Another new topic in WUM is privacy. WUM depends on the collaboration of the user to allow the access of the Web log records. Due to this dependence, users should be made aware about privacy policies before they make the decision to reveal their personal data. Even if this collaboration is established between user and WUM, there are certain reasons due to which the definite logs are not collected.

Due to the cache present on client browser, most of the requests, if they are present in the cache are not sent to Web server. Most of the time, the users do not visit the home page of a Website. They directly navigate to a particular page, by finding the URL from search engines. So it reduces the hit count of index page. Generally in Web pages designed by server side scripting like PHP,

JSP or ASP.NET they use inner page. That is, one page consisting of more than one page. In that case the request for main page records two entries in access log so it is difficult to identify an inner page. Some Web pages take query string as argument to the URL. E.g. dept.php?dept=CSE, dept.php?dept=IT like this. In this case the same page i.e. dept.php is accessed but with different arguments. It is difficult to count the page access of the Web page without the argument.

In WUM the pattern extraction algorithms are applied on the log data after they are processed. So preprocessing is very much important and must be performed with proper attention. While preprocessing the Web access log the above points should be taken into consideration so that it will produce a good set of access logs for pattern extraction.

We list below the current main WUM open problems:

1. The quantity of data is continuously increasing.
2. The preprocessing step does not receive enough analysis efforts.
3. The Websites have no or little semantic definitions for their Web pages.
4. The sequential pattern mining techniques for WUM are not appropriate for dealing with the specifics of Web usage data, mainly with its huge quantity.
5. The sequential pattern mining techniques often provide short and uninteresting results.
6. The three steps of the WUM process are not coordinated to create a coherent and unique process.

Based on common characteristics of previous studies, we propose to mine Web usage data on the client side for future works. By looking into a user's Web usage data, we hope that in future works users' interests, behaviors, and preferences are used to discovering their navigation pattern. In other words, the users' Web profile could be built which would be called personal WUM, because it focuses on personal Web usage, in contrast to previous WUM which focuses on group Web usage. Also, regarding theoretical work we need adequate techniques from dynamic data mining, such as dynamic fuzzy clustering in order to update efficiently the groups of registered customers.

Some of the reasons we recommended personal WUM as a future work are as follows.

- The goal of personal WUM is to help individual users to improve their navigational behaviors. It intends to make the Web easier to traverse from a single user's point of view.
- The client side data are clean of the uncertainties therefore, client side data provide a more accurate and complete picture of a user's Web behaviors. Besides, we will be able to collect user's footprints across tens or even hundreds Websites, rather than a single Website.
- We can perform true personalization and individualism. Although we can find large groups with similar interests, it is safe to say that no two persons' needs are the same.
- Since personal WUM is done at the client side, users have full control of their data can be used for mining Web usage data. Also in the new approach of WUM, the privacy of users will be protected unlike previous WUM.

5. Conclusion

According to that nowadays discovering hidden information from large amount of Web log data collected by Web servers is very difficult, pattern discovery has become one of the most important phases in WUM. This paper presented a brief introduction to WUM and focused on methods that can be used for the task of pattern extraction from Web log files. After discovering patterns, the result will be used for pattern analysis phase. Analyzing of the Web users' navigational patterns can help understand the user behaviors and Web structure, therefore the design of Web components and Web applications will be improved.

References

[1] Cooley, R., Mobasher, B. and Srivastava, J. "Web mining: information and pattern discovery on the World Wide Web," in International Conference on Tools with Artificial Intelligence, Newport Beach, IEEE, 1997.

[2] Cooley, R., Mobasher, B. and Srivastava, J. "Data preparation for mining World Wide Web browsing patterns," in Journal of Knowledge and Information System, 1999.

[3] Srivastava, J., et al. "Web usage mining: discovery and applications of usage patterns from web data," in SIGKDD Explorations 1, 2000.

[4] Etzioni, O. "The World-wide web: Quagmire or gold mine," in Communication of the ACM, 1996, 65-68.

[5] Chen, M., Park, J. and Yu, P. "Data mining for path traversal patterns in a Web environment," in International Conference on Distributed Computing Systems, 1996, 385-392.

[6] Chen, M., Park, J. and Yu, P. "Efficient data mining for path traversal patterns," in IEEE Transactions on knowledge and data engineering, 1998, 209-221.

[7] Mannila, H. and Toivonen, H. "Discovering generalized episodes using minimal occurrences," in International Conference on Knowledge and Data Mining, 1996, 146-151.

[8] Yan, T., et al. "From user access patterns to dynamic hypertext linking," in International World Wide Web conference on Computer networks and ISDN systems, 1996, 1007-1014.

[9] Zhu, J., Hong, J and Hughes, J. "Using Markov Chains for Link Prediction in Adaptive Web Sites," in Lecture Notes in computer science, 2002, 60-73.

[10] Jalali, M., et al. "A new clustering approach based on graph partitioning for navigation patterns mining," in International Conference on Pattern Recognition, 2008, 1-4.

[11] Sujatha, V. and Punithavalli. "Improved User Navigation Pattern Prediction Technique From Web Log Data," in International Conference on Communication Technology and System Design, 2001, 92-99.

[12] Tug, E., Sakiroglu, M. and Arslan, A. "Automatic discovery of the sequential accesses from web log data files via a genetic algorithm," in Knowledge-Based Systems, 2006, 180-186.

[13] Kim, S. and Zhang, B. "Genetic mining of HTML structures for effective web-document retrieval," in Applied Intelligence 18, 2003, 243-256.

[14] Picarougne, N., et al. "Web Mining with Genetic-Based Algorithm," in NEC Research Institute CiteSeer, 2002.

[15] Abraham, A. and Ramos, V. "Web usage mining using artificial ant colony clustering and genetic programming," in IEEE Congress on Evolutionary Computation - CEC, 2003, 1384-1391.

[16] Zhou, B., Hui, S.C. and Chang, K. "An Intelligent recommender system using sequential web access patterns," in Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems, 2004, 1-3.

[17] Burke, R. "Hybrid recommender systems: survey and experiments," in User Modeling and User-Adapted Interaction, 2002, 331-370.

[18] [18] Ishikawa, H., et al. "An intelligent web recommendation system: A web usage mining approach," in ISMIS, 2002, 342-350.

[19] [19] Mobasher, B., et al., "Integrating web usage and content mining for more effective personalization," in First International Conference on Electronic Commerce and Web Technologies, 2000, 165-176.

[20] Sarukkai, R.R. "Link prediction and path analysis using Markov chains," in 9th World Wide Web conference, 1999.

[21] Chen, M.S., Park, J.S. and Yu, P.S. "Data mining for path traversal patterns in a web enviroment," in 16th International Conference on Distributed Computing Systems, 1996, 385-392.

[22] Bonchi, F., et al. "Web log data warehousing and mining for intelligent web caching," in Data and knowledge engineering, 2001.

[23] Schechter, S., Krishnan, M. and Smith, M.D. "Using path profiles to predict HTTP requests," in Seventh International Conference on World Wide Web, 1998.

[24] Dean, J. and Henzinger, M.R. "Finding related pages in the world wide web," in Eighth International Conference on World Wide Web, 1999.

[25] Chen, M., LaPaugh, A.S. and Singh, J.P. "Predicting category accesses for a user in a structured information space," in 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2002.

[26] Brin, S. and Page, L. "The anatomy of a large-scale hypertextual web search engine," in Seventh Int. Conf. on World Wide Web, 1998.

[27] Qiu, F. and Cho, J. "Automatic identification of user interest for personalized search," in 15th Int. Conf. on World Wide Web, WWW'06, 2006.

[28] Eirinaki, M. and Vazirgianis, M. "Web mining for web personalization," in ACM Trans. Internet Tehnol. (TOIT), 2003.

[29] Suneetha, K.R. and Krishnamoorthi, R. "Identifying User Behavior by Analyzing Web Server access log file," in IJCSNS International Journal of Computer Science and Network Security, 2009.

[30] Facca, F.M. and Lanzi, P.L. "Mining interesting knowledge from weblogs: a survey," in Data Knowledge Eng. 53, 2005.

[31] Dong, D. "Exploration on Web Usage Mining and its Application," in IEEE, 2009.

[32] Wang, Y. "Web Mining and Knowledge Discovery of Usage Patterns," in CS 748T Project, 2000.

[33] [33] Tanasa, D. and Trousse, B. "Advanced data preprocessing for inter sites Web usage mining," in Intelligent Systems, IEEE, 2004, 59-65.

[34] Cooley, R. and Mobasher, B. "Data Preparation for Mining World Wide Web Browsing Patterns," in Knowledge and Information Systems, 1999.

[35] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. "From data mining to knowledge discovery: An overview," in Proc. ACM KDD, 1994.

[36] Agrawal, R. and Srikant, R. "Fast algorithms for mining association rules," in Proc. of the 20th VLDB Conference, 1994.

[37] Joshi, K.P., Joshi, A. and Yesha, Y. "On using a warehouse to analyze web logs," in Distributed and Parallel Databases, 2003, 161-180.

[38] Han, J. andKamber, M. "Data Mining Concepts and Techniques,"in the Morgan Kaufmann Series in Data Management Systems, 2001.

[39] Huang, X. and Cercone, N. "Comparison of interestingness functions for learning web usage patterns," in Eleventh International Conference on Information and Knowledge Management, 2000, 617-620.

[40] Wong, S.S.C. and Pal, S. "Mining fuzzy association rules for web access case adaptation. Wong, S.S.C. and Pal, S," in Workshop on Soft Computing in Case-Based Reasoning, International Conference on Case Based Reasoning, 2001.