A Survey on Web Usage Mining Techniques

Tahira Tabassum[1], Shweta Saxena[2]
M. Tech. Scholar[1]
Assistant Professor[2]
Computer Science and Engineering Dept.
Oriental College of Technology, Bhopal, M.P., India

Abstract

As the size of web increases along with number of users, it is very much essential for the website owners to better understand their customers so that they can provide better service, and also enhance the quality of the website. To achieve this they depend on the web access log files. Users' accesses are recorded in web logs. Because of the tremendous usage of web, the web log files are growing at a faster rate and the size is becoming huge. The web access log files can be mined to extract interesting pattern so that the user behaviour can be understood. Web data mining is the application of data mining techniques in web data. Web Usage Mining (WUM) applies mining techniques in log data to extract the behaviour of users which is used in various applications like personalized services, adaptive web sites, customer profiling, pre-fetching, creating attractive web sites etc. WUM consists of three phases pre-processing, pattern discovery and pattern analysis. For discovering patterns sessions are to be constructed efficiently. This research paper gives analysis of web usage methods including preprocessing stages used in web mining.

Keywords: *Web mining, Pattern analysis, Web mining classification.*

1. Introduction

In this world of Information Technology, accessing information is the most frequent task. Every day we have to go through several kind of information that we need and what we do? Just browse the web and the desired information is with us on a single click. Today, internet is playing such a vital role in our everyday life that it is very difficult to survive without it. The World Wide Web (WWW) has influenced a lot to both users (visitors) as well as the web site owners. The web site owners are able to reach to all the targeted audience nationally and internationally. They are open to their customer 24X7. On the other side visitors are also availing those facilities [1].

Data in Web Usage Mining (WUM), can be obtained in server logs, browser logs, proxy logs, or collected from an organization's database. These data collections vary in terms of the location of the data source, the kinds of data available, the segment of population from which the data was obtained, and techniques of implementation [1].

WUM is a division of Web Mining, which, sequentially, is a component of Data Mining. The process of mining significant and valuable information from vast database is called Data Mining . WUM mines the usage features of the users of Web Applications. This obtained data can then be applied in a various ways such as, checking of fake elements etc [2].

WUM is considered as a component of the Business Intelligence in an organization [3]. It is applied for deciding business approaches via the competent use of Web Applications. It is very vital for the Customer Relationship Management (CRM) since it can guarantee customer fulfilment till the interface between the customer and the organization is concerned [4].

There are many kinds of data that can be used in Web Mining.

1. Content: The visible data in the Web pages or the data which was intended to be provided to the users. This greatly includes text and graphics (images).

2. Structure: The organization of the website is illustrated by this data. It is partitioned into two categories. Intra-page structure data consist of the arrangement of several Hyper Text Markup Language (HTML) or Extended Markup Language (XML) tags within a given page. The key type of inter-page structure information is the hyper-links used for site navigation [13].

3. Usage: Data that illustrates the usage patterns of Web pages, such as IP addresses, page references and the date and time of accesses and other information based on the log format [4].

2. Web Mining Processes

2.1 Preprocessing

Data preprocessing illustrates any sort of processing executed on raw data to organize it for another processing process [5]. Data preprocessing alters the data into a format that will be more efficiently processed for the convenient of the user. Preprocessing steps used in Web Mining are [6]:

1. **Usage Pre-Processing:** Pre-Processing involving Usage patterns of users.
2. **Content Pre-Processing:** Pre-Processing of content accessed.
3. **Structure Pre-Processing:** Pre-Processing involving structure of the website.

2.2 Pattern Discovery

Web mining can be utilized to expose patterns in server logs but is frequently executed only on samples of data. The mining procedure will be unproductive if the models are not a significant illustration of the larger body of data [7]. The following are the pattern discovery methods [20]-

1. Statistical Analysis
2. Association Rules
3. Clustering
4. Classification
5. Sequential Patterns
6. Dependency Modelling.

2.3 Pattern Analysis

This is the ultimate step in the web mining process [13]. After the completion of the preprocessing and pattern discovery, the collected usage patterns are examined to filter insignificant information and obtain the valuable information. The techniques like Structured Query Language (SQL) processing and Online Analytical Processing (OLAP) can be used. There are several approaches present in the literature for web mining [8].

3. Problem Formulation

The most important difficulty with Web Mining in common and WUM in specific is the temperament of the data they deal with. With the exception of the quantity of the data, the data is not absolutely structured [17]. It is in a semi-structured arrangement hence it needs numerous preprocessing steps before the extraction of the essential information. Several researches have to be done on preprocessing the data and the on following problems [6].

3.1 Reducing the Paths of High visit Pages

The pages which are recurrently visited by the users can be seen as to follow a particular path. These pages can be integrated in a simply accessible branch of the Website thus resulting in reducing the navigation path length.

3.2 Eradicating or Integrating Low Visit Pages

The pages which are not regularly visited by users can be either eliminated or their content can be integrated with pages with frequent access.

3.3 Redesigning Pages to facilitate User Navigation

To assist the user to browse through the website in the best achievable way, the information acquired can be used to redesign the configuration of the Website [6].

The web usage mining algorithms are applied on the preprocessed web log data. The log files are collected from web server. But there are certain reasons due to which the actual logs are not collected [18]. A) Due to the cache present on client browser, most of the request, if it is present in the cache is not sent to web server. B) Most of the time user does not visit the home page of a website. They directly navigate to a particular page, by getting the URL from search engines. So it reduces the hit count of index page. C) 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. It is difficult to identify an inner page. D) Some web pages take query string as argument to the URL [17]. 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 web usage mining the pattern extraction algorithms are applied on the log data after they are processed. So preprocessing is very much important and must be carried out with proper care. 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 [7].

4. Related Work

Users are increasingly pursuing complex task-oriented goals on the web, such as making travel arrangements, managing finances, or planning purchases. To this end, they usually break down the tasks into a few co-dependent steps and issue multiple queries around these steps repeatedly over long periods of time. To better support users in their long-term information quests on the web, search engines keep track of their queries and clicks while searching online. Another method for

automatically identifying query groups this technique is helpful for a number of different search engine components and applications, such as query suggestions, result ranking, query alterations, sessionization, and collaborative search. This method goes beyond approaches that rely on textual similarity or time thresholds, and propose a more robust approach that leverages search query logs [9].

There has been prior work in determining whether two queries belong to the same search task. In recent work done [10] and [11] investigate the search-task identification problem. More specifically, [10] considered a search session to consist of a number of tasks (missions), and each task further consists of a number of subtasks (goals). They trained a binary classifier with features based on time, text, and query logs to determine whether two queries belong to the same task. Method given by [11] employed similar features to construct a query flow graph, where two queries linked by an edge were likely to be part of the same search mission.

First, the query-log based features in [10], [11] are extracted from co-occurrence statistics of query pairs. In this work, Authors additionally consider query pairs having common clicked URLs and we exploit both co-occurrence and click information through a combined query fusion graph. Work [10] will not be able to break ties when an incoming query is considered relevant to two existing query groups. Additionally, this approach does not involve learning and thus does not require manual labelling and retraining as more search data come in; here given Markov random walk approach essentially requires maintaining an updated query fusion graph. Finally provide users with useful query group's on-the-fly while respecting existing query groups. On the other hand, search task identification is mostly done at server side with goals such as personalization, query suggestions [11].

Some prior work also looked at the problem of how to segment a user's query streams into "sessions." In most cases, this segmentation was based on a "time-out threshold" [12-18]. Some of them, such as [14], [17], looked at the segmentation of a user's browsing activity, and not search activity. More effective scheme proposes in [18] a time-out threshold value of minutes, while others [12-16] used various threshold values. Time is not a good basis for identifying query groups, as users may be multitasking when searching online [29], thus resulting in interleaved query groups. The notion of using text similarity to identify related queries has been proposed in prior work. Algorithms of [15] and [19] used the overlap of terms of two queries to detect changes in the topics of the searches. Paper [20] studied the different refinement classes based on the keywords in queries, and attempted to predict these classes using a Bayesian classifier. Work done in [21] identified query sequences (called chains) by employing a classifier that combines a timeout threshold with textual similarity features of the queries, as well as the results returned by those queries. While text similarity may work in some cases [9], it may fail to capture cases where there is "semantic" similarity between queries (e.g., "ipod" and "apple store") but no textual similarity.

The problem of online query grouping is also related to query clustering [31], [22], [30], [31], [23]. The authors in [13] found query clusters to be used as possible questions for a FAQ feature in an Encarta reference website by relying on both text and click features. In [30] and [31], commonly clicked URLs on query-click bipartite graph are used to cluster queries. The [22] defined clusters as bicliques in the click graph. Unlike online query grouping, the queries to be clustered are provided in advance, and might come from many different users. The query clustering process is also a batch process that can be accomplished offline. While these prior works make use of click graphs, this approach is much richer in that we use the click graph in combination with the reformulation graph, and we also consider indirect relationships between queries connected beyond one hop in the click graph. This problem is also related to document clustering [24], [25], with the major difference being the focus on clustering queries (only a few words) as compared to clustering documents for which term distributions can be estimated well. Graphs based on query and click logs [26] have also been used in previous work for different applications such as query suggestions [11], query expansion [27], ranking [28], and keyword generation [32]. In several cases, variations of random walks have been applied on the graph in order to identify the most important nodes. Scheme given in [28] explain, a Markov random walk was applied on the click graph to improve ranking. In Fuxman et al. [32], a random walk was applied on the click-through graph to determine useful keywords; while in [27], a random walk was applied for query suggestion/expansion with the node having the highest stationary probability being the best candidate for suggestion.

As take advantage of the stationary probabilities computed from the graph as a descriptive vector (image) for each query in order to determine similarity among query groups. The paper [33] proposed a user- and query-dependent solution for ranking query results for web databases. Author

formally defined the similarity models (user, query, and combined) and presented experimental results over two web databases to corroborate this work analysis. Also demonstrated the practicality of implementation for real-life databases. Further, discussed the problem of establishing a workload, and presented a learning method for inferring individual ranking functions.

5. Web Usage Mining

Web usage mining also known as web log mining is the application of data mining techniques on large web log repositories to discover useful knowledge about user's behavioural patterns and website usage statistics that can be used for various website design tasks [5]. The main source of data for web usage mining consists of textual logs collected by numerous web servers all around the world. There are four stages in web usage mining [3].

Data Collection: users log data is collected from various sources like server side, client side, proxy servers and so on.

Preprocessing: Performs a series of processing of web log file covering data cleaning, user identification, session identification, path completion and transaction identification.

Pattern discovery: Application of various data mining techniques to processed data like statistical analysis, association, clustering, pattern matching and so on.

Pattern analysis: once patterns were discovered from web logs, uninteresting rules are filtered out. Analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

All the four stages are depicted through the following figure 1.

All manuscripts must be in English. These guidelines include complete descriptions of the fonts, spacing, and related information for producing your proceedings manuscripts.

Figure 1. Steps of Web Mining [5].

Web mining is the use of data mining techniques [2] to automatically discover and extract information from Web documents/services. Web mining is categorized into 3 types.

1. Content Mining as Examines the content of web pages as well as results of web Searching.

2. Structure Mining as Exploiting Hyperlink Structure.

3. Usage mining as analyzing user web navigation.

Web usage mining is a process of picking up information from user how to use web sites [6]. Web content mining is a process of picking up information from texts, images and other contents. Web structure mining is a process of picking up information from linkages of web pages [5].

Web mining is the use of data mining techniques [2] to automatically discover and extract information from Web documents/services. Web mining is categorized into 3 types.

1. Content Mining as Examines the content of web pages as well as results of web Searching.

2. Structure Mining as Exploiting Hyperlink Structure.

3. Usage Mining as analyzing user web navigation.

Web usage mining is a process of picking up information from user how to use web sites [6]. Web content mining is a process of picking up information from texts, images and other contents. Web structure mining is a process of picking up information from linkages of web pages [5].

These three approaches attempts to extract knowledge from Web generate some useful result from that knowledge and apply the result to certain real world problems. Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web Log files [2].

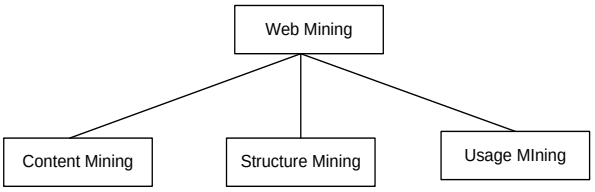

Figure 2. Types of Web Mining[5]

Web usage mining is one of the prominent research areas due to these following reasons. a) One can keep track of previously accessed pages of a user. These pages can be used to identify the typical behaviour of the user and to make prediction about desired pages. Thus personalization for a user can be achieved through web usage mining. b) Frequent access behaviour for the users can be used to identify needed links to improve the overall performance of future accesses. Pre-fetching and caching policies can be made on the basis of frequently accessed pages to improve latency time. c) Common access behaviours of the users can be used to improve the actual design of web pages and for making other modifications to a Web site. d) Usage patterns can be used for

business intelligence in order to improve sales and advertisement by providing product recommendations [6].

6. Web Usage Mining Application

Users' behaviour is used in different applications [2] such as Personalization, e-commerce, to improve the system and to improve the system design as per their interest etc., Web personalization offers many functions such as simple user salutation to more complicate such as content delivery as per users interests. Content delivery is very important since non expert users are overwhelmed by the quantity of information available online. It is possible to anticipate the user behaviour by analyzing the current navigation patterns with patterns which were extracted from past web log. Recommendation systems are the most common application. Personalized sites are example for recommendation systems. E-Commerce applications need customer details for Customer Relationship Management [19]. Usage mining techniques are very useful to focus customer attraction, customer retention, cross sales and customer departure. System Improvement is done by understanding the web traffic behaviour by mining log data so that policies are developed for Web caching, load balancing, network transmission and data distribution. Patterns for detecting intrusion fraud, attempted break-ins are also provided by mining. Performance is improved to satisfy users. Site Modification is a process of modifying the web site and improving the quality of design and contents on knowing the interest of users. Pages are re-linked as per customer behaviour [2].

There are a number of issues in preprocessing of log data. Volume of requests in web log in a single log file is the first challenge [11]. Analyzing web user access log files helps to understand the user behaviours in web structure to improve the design of web components and web applications. Log includes entries of document traversal, file retrieval and unsuccessful web events among many others that are organized according to the date and time. It is important to eliminate the irrelevant data. So cleaning is done to speed up analysis as it reduces the number of records and increases the quality of the results in the analysis stage. Efforts in this data to find accurate sessions are likely to be the most fruitful in the creation of much effective web usage mining and personalization systems. By following data preparation steps, it is very easier to generate rules which identify directories for website improvement [19]. More research can be done in preprocessing stages to clean raw log files, and to identify users and to construct accurate sessions.

7. Conclusion

The increasing popularity of the Web has greatly attracted the Web mining technology. A vital research area in Web mining is WUM which mainly focuses on the discovery of patterns in the browsing and navigation data of Web users. The quality of a website can be analyzing by user accesses behaviour of the website. To know the user accesses pattern WUM is very efficient method. Log files are the best source to know user behaviour. But the raw log files contains unnecessary details like image access, failed entries etc, which will affect the accuracy of pattern discovery and analysis. WUM has been a potential technology for understanding behaviour of the user on the Web. There are several techniques proposed by different researchers for the web usage mining. This paper discussed about various steps used for web usage mining. This paper mainly discusses about three vial steps in WUM i.e. preprocessing, pattern discovery and pattern analysis. The mentioned research methodologies can be extended in future to create more efficient session reconstructions through graphs and mining the sessions using graph mining as quality sessions gives more accurate patterns for analysis of users.

8. References

[1] Dr. G. K. Gupta, "Introduction to Data Mining with Case Studies", PHI Publication, 2005.

[2] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, No. 2, 2000, Page 12-23.

[3] Adel T. Rahmani and B. Hoda Helmi, "EIN-WUM an AIS-based Algorithm for Web Usage Mining", Proceedings of GECCO'08, Atlanta, Georgia, USA, ACM978-1-60558-130-9/08/07, 2008, Pp. 291-292.

[4] Shailey Minocha, Nicola Millard, Lisa Dawson, "Integrating Customer Relationship Management Strategies in (B2C) E-Commerce Environments", IFIP Conference on Human-Computer Interaction-INTERACT, 2003.

[5] C. Ramya, G. Kavitha, K. S. Shreedhara, "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process", Computing Research Repository - CORR, vol. abs/1105.0, 2011.

[6] V. Chitraa, Antony Selvdoss Davamani, "A Survey on Preprocessing Methods for Web Usage Data", Computing Research Repository-CORR, Vol. abs/1004.1, 2010.

[7] Nizar R. Mabroukeh, Christie I. Ezeife, "A taxonomy of sequential pattern mining algorithms", ACM Computing Surveys - CSUR, Vol. 43, No. 1, 2010, Pp. 1-41.

[8] Francesco Moscato, Nicola Mazzocca, Valeria Vittorini, Giusy Di Lorenzo, Paola Mosca, Massimo Magaldi, "Workflow Pattern Analysis in Web Services",

High Performance Computing and Communications - HPCC, 2005, Pp. 395-400.
[9] Heasoo Hwang, Hady W. Lauw, Lise Getoor, and Alexandros Ntoulas, "Organizing User Search Histories", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, NO. 5, IEEE, 2012, Page 912-925.
[10] R. Jones and K.L. Klinkner, "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), 2008.
[11] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, "The Query-Flow Graph: Model and Applications," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), 2008.
[12] P. Anick, "Using Terminological Feedback for Web Search Refinement: A Log-Based Study," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, 2003.
[13] B.J. Jansen, A. Spink, C. Blakely, and S. Koshman, "Defining a Session on Web Search Engines: Research Articles," J. the Am. Soc. for Information Science and Technology, vol. 58, no. 6, pp. 862-871, 2007.
[14] L.D. Catledge and J.E. Pitkow, "Characterizing Browsing Strategies in the World-Wide Web," Computer Networks and ISDN Systems, vol. 27, no. 6, 1995, pp. 1065-1073.
[15] D. He, A. Goker, and D.J. Harper, "Combining Evidence for Automatic Web Session Identification," Information Processing and Management, vol. 38, no. 5, 2002, pp. 727-742.
[16] R. Jones and F. Diaz, "Temporal Profiles of Queries," ACM Trans. Information Systems, vol. 25, no. 3, 2007, p. 14.
[17] A.L. Montgomery and C. Faloutsos, "Identifying Web Browsing Trends and Patterns," Computer, vol. 34, no. 7, July 2001, pp. 94-95.
[18] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a Very Large Web Search Engine Query Log," SIGIR Forum, vol. 33, no. 1, 1999, pp. 6-12.
[19] H.C. Ozmutlu and F. C¸ avdur, "Application of Automatic Topic Identification on Excite Web Search Engine Data Logs," Information Processing and Management, vol. 41, no. 5, 2005, pp. 1243-1262.
[20] T. Lau and E. Horvitz, "Patterns of Search: Analyzing and Modeling Web Query Refinement," Proc. Seventh Int'l Conf. User Modeling (UM), 1999.
[21] F. Radlinski and T. Joachims, "Query Chains: Learning to Rank from Implicit Feedback," Proc. ACM Conf. Knowledge Discovery and Data Mining (KDD), 2005.
[22] J. Yi and F. Maghoul, "Query Clustering Using Click-through Graph," Proc. the 18th Int'l Conf. World Wide Web (WWW '09), 2009.
[23] E. Sadikov, J. Madhavan, L. Wang, and A. Halevy, "Clustering Query Refinements by User Intent," Proc. the 19th Int'l Conf. World Wide Web (WWW '10), 2010.
[24] T. Radecki, "Output Ranking Methodology for Document- Clustering-Based Boolean Retrieval Systems," Proc. Eighth Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, 1985, pp. 70-76.
[25] V.R. Lesser, "A Modified Two-Level Search Algorithm Using Request Clustering," Report No. ISR-11 to the Nat'l Science Foundation, Section 7, Dept. of Computer Science, Cornell Univ., 1966.
[26] R. Baeza-Yates, "Graphs from Search Engine Queries," Proc. 33rd Conf. Current Trends in Theory and Practice of Computer Science (SOFSEM), vol. 4362, pp. 1-8, 2007.
[27] K. Collins-Thompson and J. Callan, "Query Expansion Using Random Walk Models," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2005.
[28] N. Craswell and M. Szummer, "Random Walks on the Click Graph," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), 2007.
[29] Spink, M. Park, B.J. Jansen, and J. Pedersen, "Multitasking during Web Search sessions," Information Processing and Management, vol. 42, no. 1, pp. 264-275, 2006
[30] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2000.
[31] R. Baeza-Yates and A. Tiberi, "Extracting Semantic Relations from Query Logs," Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2007.
[32] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang, "Query Clustering Using User Logs," ACM Trans. in Information Systems, vol. 20, no. 1, 2002, pp. 59-81.
[33] Aditya Telang, Chengkai Li, and Sharma Chakravarthy, "One Size Does Not Fit All: Toward User- and Query-Dependent Ranking for Web Databases ", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, NO. 9, IEEE, 2012, Pages 1671-1685.