A PERLUSTRATION ON WEB USAGE MINING

Dr. K. Prabha
Department of Computer Science
Periyar University PG Extension Centre,
Dharmapuri(TN),India
prabhaeac@gmail.com

T.Suganya
Department of Computer Science
Periyar University PG Extension Centre,
Dharmapuri(TN),India
suganyacs59@gmail.com

Abstract: Web usage mining is a type of web mining which exploits data mining techniques to discover valuable information from navigations of web uses, application of data mining techniques to web worthwhile data in order to extract usage patterns. The performance of web information retrievals and web based data warehousing are boosted with the extraction of information from the web using web mining tools. And web usage mining is one of the fastest developing areas of web mining, thus this activity that involves automatic discovery of user access patterns from one or more web servers. This paper provides an exploration about web usage mining analysis of concepts, ranking, and applications.

Keywords: Data mining, web server, web mining, ranking.

I. INTRODUCTION

Web mining is the application of data mining techniques to discover patterns from the Web. According to analyze targets, web mining can be divided into three different types, which are Web content mining, Web structure mining and Web usage mining. The extraction of hidden predictive information from large databases is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [12].

There are three general classes of information that can be discovered by web mining:

- **Class A**: Web activity, from server logs and Web browser activity tracking.
- **Class B**: Web graph, from links between pages, people and other data.
- **Class C**: Web content, for the data found on Web pages and inside of documents.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. The three categories of web mining for the following:

A. WEB CONTENT MINING

Web content mining is an automatic process that goes beyond keyword extraction. Since the content of a text document presents no machine readable semantic, some approaches have suggested restructuring the document content in is presentation that could be exploited by machines. The usual approach to exploit known structure in documents is to use wrappers to map documents to some data model. Techniques using lexicons for content interpretation are yet to come. There are two groups of web content mining strategies: Those that first group is directly mine the content of documents and the second group improves on the content search of other tools like search engines.

B.WEB STRUCTURE MINING

World Wide Web can reveal more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. This can be compared to bibliographical citations. When a paper is cited often, it ought to be important. The Page Rank and CLEVER methods take advantage of this information conveyed by the links to find pertinent web pages. By means of counters, higher levels cumulate the number of artifacts subsumed by the concepts they hold. Counters of hyperlinks, in and out documents, retrace the structure of the web artifacts summarized.

C. WEB USAGE MINING

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Many web analysis tools existed but they are limited and usually unsatisfactory. We have designed a web log data mining tool, Weblog Miner, and proposed techniques for using data mining and On-line Analytical Processing (OLAP) on treated and transformed web access files.

Applying data mining techniques on access logs unveils interesting access patterns that can be used to restructure sites in a more efficient grouping, pinpoint effective advertising locations, and target specific users for specific selling ads. Customized usage tracking analyzes individual trends [11, 2]. Its purpose is to customize web sites to users. The information displayed the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns. While it is encouraging and exciting to see the various potential applications of web log file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge one can discover from the large raw log data. Current web servers store limited information about the accesses. However, for an effective web usage mining, an important cleaning and data transformation step before analysis may be needed.

II. RANKING IN WEB USAGE MINING (WUM)

The web usage mining generally includes the following several steps: data collection, data pretreatment [6] and knowledge discovery and pattern analysis.

A. DATA COLLECTION

Data collection is the first step of web usage mining, the data authenticity and integrity will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting).

B. DATA PREPROCESSING

Some databases are insufficient, inconsistent and including noise. The data pre-treatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

1. DATA CLEANING

The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following **two kinds of records are unnecessary and should be removed**

- **The records of graphics, videos and the format information**. The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record.
- **The records with the failed HTTP status code**. By examining the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

2. USER AND SESSION IDENTIFICATION

The task of user and session identification is to find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study. The rules adopted to distinguish user sessions can be described as follows:

a) The different IP addresses distinguish different users;

b) If the IP addresses are same, the different browsers and operation systems indicate different users;

c) If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty;
d) The session identified by rule 3 may contains more than one visit by the same user at different time, the time-oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

3. PATH COMPLETION

Another critical step in data preprocessing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log. As a result, the user access paths are incompletely preserved. Web access log is to discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data preprocessing, we will improve the mined patterns' quality and save algorithm's running time. It is especially important to web log files, in respect that the structure of web log files are not the same as the data in database or data warehouse. They are not structured and complete due to various causations. So it is especially necessary to pre-process web log files in web usage mining. Through data pre-processing, web log can be transformed into another data structure, which is easy to be mined.

C. PATTERN DISCOVERY

Use statistical method to carry on the analysis and mine the pre-treated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

D. PATTERN ANALYSIS

A challenge of Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse. Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data
Or knowledge be visible. Finally, provide the characteristic service to the electronic commerce website.

III. APPLICATIONS OF WEB USAGE MINING

A. E-COMMERCE

E-Commerce means two trading parties based on Internet according to certain rules or standard developing the whole traditional business activity in digital network mode. Buying and selling of products or services through Internet, E-commerce generates huge volume of interactions. This tremendous growth in the E-commerce enterprise, twisted to product surplus. It also supports e-commerce sites to retain the most profitable customers [4], improve the functionality of web based applications, provides more custom-made content to visitors. In addition, with the use of Web usage mining techniques e-commerce companies can improve products quality or sales by anticipating problems before they occur. They also provide companies with previously unknown buying patterns and behavior of their online customers. More importantly, the fast feedback the companies obtain by using Web usage mining is very helpful in increasing the company's benefit [5].

B.E-LEARNING

E-learning is a form of electronically supported learning which allows the people to learn any subject at anytime and anywhere. The simplicity in using the tools to browse the resources on the web, its easiness in deploying and maintaining resources made the web as an excellent tool for delivering courses. Web is one and only major choice to manage and maintain learning resources and has become one of the leading choice of modern advanced distance education system. As education becomes more technologically advanced, the complexity of available learning resources also increased accordingly. It is difficult to evaluate the structure of the course content and its effectiveness on the learning process. The pattern analysis capability of web usage mining has an important role in web-based learning system. They can analyze the students and instructors behavior [10] and improve the educational experience. Tracking the activities happening in the course website and mine patterns is also beneficial to improve or adapt the course contents. This allows instructors to appraise the access behavior, assess the learning activities and compare learners. The arrangement of the course contents can be enhanced by analyzing the traversal paths of the course content web pages is another advantage of Web usage mining [1].

C.E-GOVERNANCE

E-governance provides a single web portal that integrates all services that includes government, nonprofit and private-sector entities [12]. In such a type of service system which provides ready access to information, the user interface quality is an important factor. This is one of the challenging user-centric parameter since this has to provide information to extensive and various users [3].The patterns of the online behavior of the users can be discovered by using Web usage mining techniques. These patterns reveal the user interests and that can be utilized to fine tune user interfaces and suggest the most appropriate browsing paths. User requirements also are exhibiting in their navigation behavior. Analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps, web based personalized system. This technology also uses the experience of users of past sessions to provide recommendation to users of current session [10].

IV. CONCLUSION

Web usage mining is a kind of mining to server logs. Web usage mining plays an important role in realizing, enhancing the usability of the website design, the improvement of customer's relations and improving the requirement of system performance and so on. Web usage mining provides the support for the website design, providing personalization server and other business making decision, etc. Web usage mining can model user behavior and therefore to forecast their future whereabouts. Its main aim is getting useful to users for easy access information in logs to make sites perfect with effectual.

REFERENCES

[1] Bart C Palmer; Web Usage Mining: Application to an online educational digital library service; Digital Commons@USU; 2012

[2] Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", IJARCSSE, November, 2013.

[3] G .Rani; S .Chakraverty, ―Boosting Interactivity of E Governance‖, International Conference on communication Languages and Signal Processing- with Preference to 4 G Technologies‖, ICCLSP4G, January2012.

[4] http://revistaie.ase.ro/content/51/104%20-20SIVA RAMAKRISHNAN, %20 BALAKRISHNAN.pdf

[5] J. g. Liu, h. h. Huang. Web Ming for Electronic Business Application, Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies, Chengdu, China, 2003:872~876.

[6] Margaret H. Dunham "Data Mining Introductory and Advanced Topics ISBN 978-81-7758-785-2 pp 205-218

[7] Pranit Bari, P.M.Chawan "Web Usage Mining", Journal of Engineering, Computers & Applied Sciences (JEC&AS) Volume-2 no.6, June 2013, pp 37-38. ISSN NO:2319-5606.

[8] Romero C. Ventura S, Pechenizky M , Baker R. S ; Handbook of educational data mining; 2010; CRC Press.

[9] S. Chakraverty 1 , B. G. Rani, C. B. Singla and D. Anand; Experience based recommendations system for e-governance; 2012.

[10] Xiaoqing Zheng,Yiling Gu,Yinsheng Li,"Data Extraction from Web Pages Based on Structural Semantic Entropy", International World Wide Web conference committee (IW3C2),April 2012.

[11] Yong Shi, Yuqing Song and Aidong Zhang. A shrinking - based approach for multi dimensional data analysis. In the 29th VLDB conference, September 2003.

[12] Zakareya Ebrahim and Zahir Irani; ―E-government adoption: architecture and barriers‖ Emerald Business Process Management journal, vol.II, No.5 2005, pp589-611, 2005