ISSN 2250-3153

Web Usage Mining has Pattern Discovery

Mr. Akshay Upadhyay, Mr. Balram Purswani

Gyan Ganga College of Technology, Jabalpur, M.P., India

Abstract— Ample amount of knowledge in respect of pattern discovery of web usage mining shall be provided in this paper. Users behavior of page browsing should be in hand with the website designers. They can even study about the visitor's activities through the web analysis and find patterns of the visitor's activities. This kind of web analysis involves not only involves the change and interpretation of the web log records to locate the hidden information or predictive pattern by the data mining and knowledge discovery technique, but also offers a great prospect coupled with the web warehousing.

Index Terms- Web Log File, Web Usage Mining, Data Preparation, Pattern Discovery

I. INTRODUCTION

WWW is a very popular and interactive medium for propagating information today. Due to the vast, varied and dynamic nature of web it raises the scalability, multimedia data and temporal issues respectively. The development of the web has given rise to large quantity of data that is freely available for user access. Management and organization of data should be done in such a way that they can be accessed by different users effectively and efficiently. That is why; the number of researchers in the field of application of data mining techniques on the web is increasing.

The Web Mining is the set of techniques of Data Mining applied to extract useful knowledge and implicit information from Web data. As more organizations rely on the Internet to conduct daily business, the study of Web mining techniques to discover useful knowledge has become increasingly important. However, with the magnitude and diversity of available information from the Internet, it is not insignificant to locate the relevant information to satisfy the requirements of people with different backgrounds. To assist Web surfers in browsing the Internet more efficiently, one of the topics that have attracted much attention is modeling the Web user's browsing patterns and making recommendations. Web mining enables one to discover web pages, text documents, multimedia files, images and other types of resources from web.

II. WEB MINING CATEGORIES

Web mining can be categorized into three different classes based on which part of the Web is to be mined. These three categories are:

1) Web content mining,
2) Web structure mining and
3) Web usage mining.

Web content mining describes the discovery of useful information from the web contents. However, what consist of the web contents could encompass a very broad range of data. Basically, the web content consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks. The web content data consist of unstructured data such as free texts, semi-structured data such as HTML documents, and a more structured data such as data in the tables or database generated HTML pages.

Web structure mining tries to discover the model underlying the link structures of the web. This model is based on the topology of the hyperlinks with or without the description of the links. This model can be used to categorize web pages and is useful to generate information such as the similarity and relationship between different web sites. Web structure mining could be used to discover authority sites for the subjects and overview sites for the subjects that point to many authorities.

Web usage mining tries to make sense of the data generated by the web surfer's session or behaviors. While the web content and structure mining utilize the real or primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web. The Web Usage mining includes the data from the web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data as the results of interactions.

III. WHY WEB USAGE MINING

Web Usage Mining is the discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities. In this paper we will give emphasize on Web usage mining. As the explosive growth of E-commerce, the law of business of companies has been changed. Now a day the web is not the place where only transaction has occurred. Millions of visitors interact with the web in daily life which generates an enormous amount of data. Web usage mining helps to know information about users' behaviors and their usage patterns, can lead to interesting results that go over descriptive tasks; such examples are dynamic content Web sites which perform mass customization and personalization by discovering clusters of users with similar access patterns and by adding navigational links and hints on the fly. Also, information mined from Web usage data allows restructuring and better management of the site, giving more effectiveness to it; also the network system can gain benefits from this discovery process.

IV. SOURCES OF DATA FOR WEB USAGE MINING

Web Usage Mining applications are based on data collected from three main sources: (i) Web servers, (ii) Proxy servers, and (iii) Web clients.

A. The Server Side

Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g. name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format e.g.: Common Log Format, Extended Log Format, and LogML. When exploiting log information from web servers, the major issue is the identification of users' sessions. Apart from web logs, users' behavior can also be tracked down on the server side by means of TCP/IP packet sniffers. Packet sniffers are rarely used in practice because of rise scalability issue on web servers with high traffic, and the impossibility to access encrypted packets like those used in secure commercial transactions a quite severe limitation when applying web usage mining to e-businesses. Probably, the best approach for tracking web usage consists of directly accessing the server application layer.

B. The Proxy Side

Many internet service providers (ISPs) give to their customer Proxy Server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers.

C. The Client Side

Usage data can be tracked also on the client side by using Java Script, java applets, or even modified browsers. These techniques avoid the problems of users' session identification and the problems caused by caching (like the use of the back button). In addition, they provide detailed information about actual user behaviors. However, these approaches rely heavily on the users' cooperation and rise many issues concerning the privacy laws, which are quite strict.

V. WEB USAGE MINING PROCESS

Web usage mining is a powerful tool to analyzing, designing and modifying a Web site structure as well as it is also useful to understanding and analyzing the site visitor's behavior in two aspects: i) The interest and information one access. ii) The way to access this information. Web usage mining activities pacify two different aspects: how designers expect to be used the site by the visitors and the way visitors effectively using the site.

Web usage mining can be divided in at least three different phases namely

1) Data preparation,
2) Pattern discovering
3) Pattern analysis and visualization.

A. Data Preparation

In Data Preparation phase the web log data must be cleaned, filtered, integrated and transformed in such a way that the irrelevant and redundant data can be removed, user session and transaction can identified.

In this paper we provide some algorithm for Data preparation process.

Data cleaning: The first step of data preparation is data cleaning or filtering. It is very important as there have many unnecessary entries in the log files. Elimination of irrelevant items can be accomplished by checking the suffix of the URL name, which tells one what format these kinds of files are. For example, the surrounded graphics can be filtered out from the Web log file, whose suffix is usually the form of "gif", "jpeg", "jpg", "GIF", "JPEG", "JPG", can be removed. In the same way the unwanted sound files can be removed.

Figure: Algorithm for Data Preparation

Algorithm: DataPreparation

1. Start
2. Check for data available in server log
3. If raw data is available goto step 4 else goto step 2
4. Cleaning data by removing gap, .jpg , .gif or sound file.
5. Execute UserIdentification.
6. Execute SessionIdentification.
7. Divide the session in transaction with a certain duration.
8. If any data available goto step 4 else goto step 9
9. exit

User Identification: Once HTTP log files have been cleaned, next step in the data preparation is the identification of the user, through heuristics. (i) By converting ip address to domain name exposed some knowledge. For example, one can estimate where visitors live by looking at the extension of each visitor's domain name, such as .ca (Canada); .au (Australia); cn (China), etc. (ii) the web server randomly assigned an Id to the web browser while it connects first time to the site. This is called cookies. The Web browser sends the same ID back to the Web server, effectively telling the Web site that a specific user has returned. Cookies help the Web site developer to easily identifying individual

visitors, which results in a greater understanding of how the site is used. Cookies also help visitors by allowing Web sites to recognize repeat visits. (iii) Cache prevents much user access to be recorded in the log file when a page hit by the user already in the cache. Cache busting is one solution of this problem.

Algorithm: UserIdentificaton
1. Start
2. Take data from cleaned HTTP log file.
3. while any data is available do
i. converting ip address to domain name by reverse DNS lookup.
ii. Sending cookies to identify user
iii. Busting cache to prevent use of cache.
iv. Referring URL.
4. Exit

Figure: Algorithm for User Identification

Session Identification: Session identification can be performed using time interval between consecutive log entries. If two accesses from the same user are separated by an interval longer than a threshold they considered as different session. Sometimes threshold considered as 30 minutes time interval. Another way to identify session is using a time out to identify the end of the session.

After data preparation the server log file data have to be prepared for pattern discovery. This data is more organized, classified which we called web warehousing.

Algorithm: SessionIdentificaton

1. Start
2. Take time of the first log entries.
3. Calculate the threshold time from the starting time.
4. if threshold >30 min session change else same session
5. exit

Figure: Algorithm for Session Identification

VI. PATTERN DISCOVERY

After data preparation phase, the pattern discovery method should be applied. This phase consists of different techniques derived from various fields such as statistics, machine learning, data mining, pattern recognition, etc. applied to the Web domain and to the available data.

The task for discovering the patterns offer some techniques as statistical analysis, association rules, sequential pattern analysis, clustering and so on. Here we will briefly describe some techniques to discover patterns from processed data.

To determine the visitor's location converting the IP address into its domain name is a good way. Looking up the extension of the domain name one may figure out the country of the visitor. Then the server administrator can assume about the most active countries visited at a particular site and he can provide the useful information relevant to that country.

Most Active Countries

	Country	Hits	Visitors
1	United States	8,791	1,788
2	France	2,908	302
3	China	810	248
4	Germany	3,016	240
5	United Kingdom	2,552	231
6	Russian Federation	1,010	166
7	Italy	1,565	142
8	Canada	1,126	119
9	Netherlands	805	76
10	Sweden	605	73
11	Australia	626	70
12	Switzerland	448	57
13	India	495	55
14	Brazil	510	52
15	Spain	441	51
16	Singapore	165	41
17	Phillippines	315	40
18	Latvia	36	35
19	South Africa	128	33

Using the path analysis technique, the information offers a valuable imminent of user navigation problems. To analyze the path the administrator can understand what pages the visitors like most or how long path they like to visit in a web site. For e.g. If 65% of visitors who accessed /sustTube/video.php by starting at / sustTube and proceeding through /sustTube /view_video.php ,or /sustTube/video.php, decided to make a decision after seeing the sample video. Since many user leave the web site after visiting four pages, the important information (for example, sample video) should contained within four pages.

After data pre processing from the server log file we get the access page link of the visitor. If the pages link are from the same IP address so the system can decide that the user is same person but increases the page hit count. If the pages link is from different IP address so the system increase the number of visitor. When the system provides the number of hits and visitor corresponding to page link, the admin can realize which pages are most popular. In our experimental web site 'Sust Tube' we can see the number of hits and visitors count of the site.

Pages	Home	Hits	Visitors
Home Page/sustTube.php	Index.php	149	3
	Index.php	117	2
	Index.php	52	2
Video	Video.php	25	3
	Video.php	10	3
	View_video.php	11	2
	View_video.php	7	2
	View_video.php	5	2

	View_video.php	9	2
Log in	Login.php	21	3
Sign up	Sign up.php	20	3
	Sign up.php	4	2

Figure: Highest Click-Through Rates for Each Page

By the highest click the administrator can also estimate about the daily activities. We show this by a sample graph.

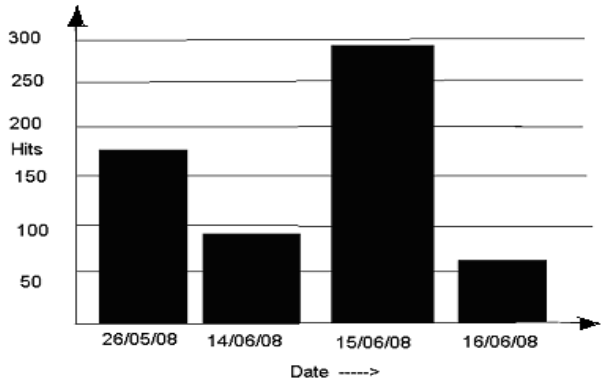
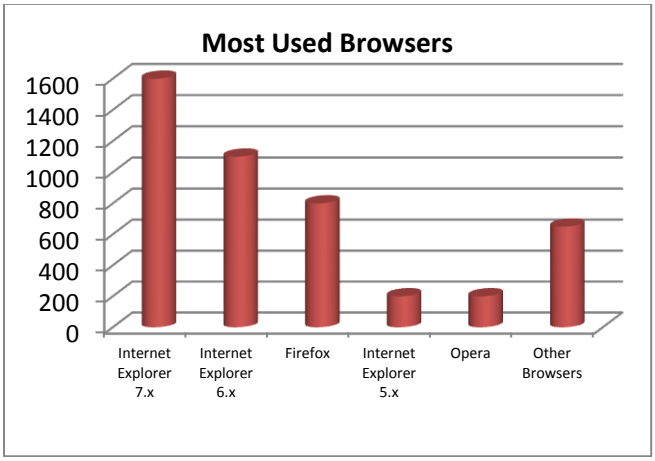


Figure: Daily Activities

From server log file's user agent portion we get the browsers name and the number of users uses a particular browser. So the system can decide from which browsers most number of visitors hit the site. It's also suitable to determine the operating system of the visitors.



VII. CONCLUSIONS AND FUTURE WORK

The web is a most important medium to conduct business and commerce. Therefore the design of web pages is very important for the system administrator and web designers. These features have great impact on the number of visitors. So the web analyzer has to analyze with the data of server log file for detecting pattern. In this paper we tried to give a clear understanding of the data preparation process and pattern discovery process. Web usage patterns and data mining can be the basis for a great deal of future research. More research needs to be done in E-Commerce, Bioinformatics, Computer Security, Web Intelligence, Intelligent Learning, Database Systems, Finance, Marketing, Healthcare and Telecommunications by using Web usage mining.

REFERENCES

[1]. R. Ivancsy, I. Vajk, "Frequent Pattern Mining in Web Log Data" Acta Polytechnica Hungarica Vol. 3, No. 1, pp. 77-90, 2006.
[2]. F. M. Facca, P. L. Lanzi , "Recent developments in Web Usage Mining Research" Artificial Intelligence and Robotics Laboratory
[3]. J. R. Boullosa, G. Xexeo, "An architecture for Web Usage Mining", 2002.
[4]. K. R. Suneetha, Dr. K. R. Kreishnamoorthy, "Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, pp. 327-332, April 2009
[5]. A. N. Mahanta "Web Mining: Application of data mining" pp. 111- 116 Proceedings of NCKM-2008.
[6]. Nikos Koutsoupias "Exploring Web Access Logs with Correspondence Analysis", 2nd Hellenic Conf. on AI, SETN-2002, 11-12 April 2002, Thessaloniki, Greece, Proceedings, Companion Volume, pp. 229-23

Mr. Akshay Upadhyay, Pursuing M.E. in Computer Science from Gyan Ganga College of Technology, soluupdh@gmail.com

Mr. Balram Purswani, M.Tech (I.T.), Asstt. Professor, Gyan Ganga College of Technology, balpurg@gmail.com