Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Mining Approach in Analysing User Behaviour and Interest for Website Modification

Sadik Khan*
Research Scholar
Bhagwant University, Ajmer, India

Dr. Yashpal Singh
Associate Professor
B.I.E.T., Jhansi, India

Dr. Ajay Kumar Sachan
Professor & Director
R.I.T.S.., Bhopal, India

Abstract— Internet is grow day by day ,so online users are also increases. Online users through internet is very common, it is very useful for the websites. There is an immediate need for an autonomous and robust system that can offer decision-making support for users who are looking for the cheapest, the most familiar, or the best-quality services .This paper presents an web-mining application to website modification.

The rapid growth in the amount of information and the number of users has lead to difficulty in providing effective search services for the web users and increased web latency; resulting in decreased web performance. Web mining is becoming the tool for success for those who adopt electronic means of operation for conducting their business and modification. In contrast to the traditional multi agent systems, which use rule-based or case-based process flows to coordinate communications for system automation, such as consumer preferences, product specification, product selection, price negotiation, purchase, delivery, after-sales service and evaluation. Analyzing users behavior when shopping online like his interest on a particular category. Log file analyzing for website is a part of web mining which is very effective.

Keywords— Web Mining , Server Log files, user behaviour , Data Mining, websites modification ,Web Usage Mining.

I. INTRODUCTION

We have the opportunity to develop a web mining system for business users and data analysis from the ground up,including data collection and transformation. Designing and personalising systems for specific user groups encompasses a lot of effort with respect to analysing and understanding user behaviour. As the on-line services and Web-based information systems proliferate in all domains of activities, it became increasingly important to model user behaviour and personalising, so that these systems will appropriately address user characteristics.

Online platform ,is the interaction between supplier and the customer. Users are attracts through some promotions schemes such as discount , gift voucher , couponing etc.

Focusing on the retail e-commerce domain allowed us to provide solutions to some tough problems. Web Mining is the mining of data related to World wide Web. Web search engines are usually designed to serve user requirement without considering the user's interest**.** Web mining aims to discovering useful information or knowledge from the web hyperlinks structure , page content and usage data. With the development of internet and multimedia technology modern shopping has become a new kind of model, it breaks the multiple limitations in terms of personnel, time and space in traditional sopping more.

II. WEB MINING

To find the interesting pattern into a huge collection of data is called data mining.Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Even though it is strongly related to data mining, it is not equivalent to it. Three main axes of Web mining have been identified, according to the Web data used as input in the data mining process, namely Web structure, Web content and Web usage mining.

III. WEB STRUCTURE MINING

Web is a graph. It is a directed labeled graph whose nodes are the documents and the edges are the hyperlinks between them. Web is a huge structure, growing rapidly. This network of information lacks organization and structure, and is only held together by the hyperlinks.

A. Document Structure

In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents

B. The role of hyperlinks in web searching

In order to make navigation in this chaotic structure easier, people use search engines, trying to focus their search by querying using specific terms/keywords. At the beginning, where the amount of information contained in the Web did not yet have these big proportions, search engines used manually-built lists covering popular topics. They maintained an index, containing a list for every word, of all Web pages containing this word. This index was then used in order to answer to the users' queries. However, after a few years, when the Web evolved including millions of pages, the manual maintenance of such indices was very expensive. The automated search engines relying in keyword matching, give results including hundreds (or more) Web pages, most of them of low quality.

IV. WEB CONTENT MINING

Web content mining has to do with the retrieval of information (content) available on the Web into more structured forms as well as its indexing for easy tracking information locations. Web content may be unstructured (plain text), semistructured (HTML documents), or structured (extracted from databases into dynamic Web pages).

A. Data Preprocessing

Web content mining is strongly related to the domain of Text Mining, since in order to process and organize Web pages their content should be first appropriately processed in order to extract properties of interest. These selected properties are subsequently used to represent the documents and assist the clustering or classification processes. We discriminate four stages of data preprocessing, based on techniques used in text mmmg, namely Data Selection, Filtering, Cleaning, and Representation [9].

B. Web document representation models

In order to reduce the complexity of the documents and make them easier to handle, during the clustering and/or classification processes, one should first choose the type of characteristics or attributes (e.g. words, phrases, or links) of the documents that are of importance, and how these should be represented. Since documents are represented in a uniform way, the similarity between two documents can then be easily calculated.

V. WEB USAGE MINING

The process of analyzing the user's browsing behavior is called Web usage mining. It can be regarded as a three-phase process, consisting of the data preparation, pattern discovery and pattern analysis phases [8]. In the first phase, Web data are preprocessed in order to identifY users, sessions, pageviews, and so on. The input data are mainly the hits registered in the Web usage logs of the site, some times combined with other information such as registered user profiles, referrer's logs, cookies, etc [1].

A. Web Server Data

The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

B. Identifying navigational patterns

The users' activity when browsing through Web sites is registered in these sites' Web logs. Considering the average number of visits to a medium-sized Web site per day, we can presume that the amount of information hidden in the site's Web logs is huge, yet meaningless if they're not appropriately processed. By processing these data, either using simple statistical methods, or by using more complicated data mining techniques, we can identify interesting trends, and patterns concerning the activity in the Web site. Site administrators can then use this information to redesign or customize the Web site according to the interests and behavior of its visitors, or improve the performance of their systems.

C. Data Preprocessing

The first issue in the preprocessing phase is data preparation. Web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. Furthermore, crawler activity can befiltered out, because such entries do not provide useful information about the site's usability. Another problem to be met has to do with caching. Accesses to cached pages are not recorded in the Web log, therefore such information is missed. Caching is heavily dependent on the client-side technologies used and therefore cannot be dealt with easily. In such cases, cached pages can usually be inferred using the referring information from the logs.

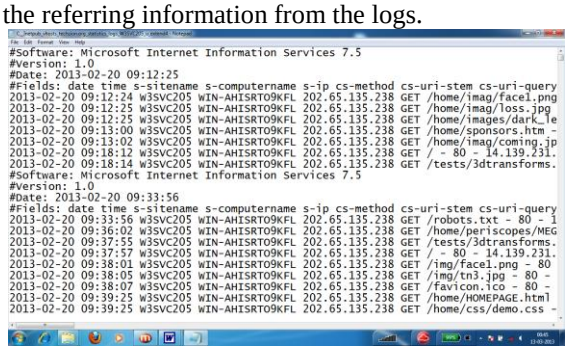

Fig1.0 Sample Log file

D. Web usage logs

Each access to a Web page is recorded in the access log of the Web server that hosts it. The entries of a Web log file consist of fields that follow a predefined format. The fields of the common log format are: remotehost rfc931 authuser date "request" status bytes Except for Web server logs, which are the main source of information, usage data can also be acquired by proxy server logs, browser logs, user profiles, registration data, cookies, mouse clicks etc.A sample log file shown in Fig.1.0 is taken by the website www.techzion.org.

The data are collected by on-line forms, by server logs and by cookie logs or java agents in a history list. This information is given as an input to the data mining component where it can be used for different purposes. The data can be used to learn the user model and the user preferences as well as the usage of the website. In the data mining component are realized data mining methods such as decision tree induction and conceptual clustering for attribute-value based and graph-structured data. Decision tree induction requires that the data have a class label. Conceptual clustering can be used to learn groups of similar data. When the groups have been discovered the data can be labeled by a group name and as such it can be used for decision tree induction to learn classification knowledge.

Based on the user model the presentation style and the content of the web site (adaptive multimedia product presentation) are controlled. Besides that the user model is used to set up specific marketing actions such as e.g. mailing actions or cross-selling actions. The results of the webusage mining are used to improve the website organization as well as for monitoring the impact rate of the advertisement of particular events. Product models and preferences are used to control the content of the website. The preferences can be learned based on the user´s navigation data. Besides that an intelligent dialogue component allows to control the dialogue with the user [9].

The following processes can be handled with these components:

- Web-Site Administration
- Web-Site Modification
- Marketing and Selling,
- Adaptive Multimedia Product Presentation,
- Event Recognition, and
- Learning Ontology Knowledge.

In this paper we have studied a website server log(Fig1.0) file,the website name is techzion.org.In this we generate some reports which shows that which page of website users most visited.This is very important to know the pages behavior in a website. By using this analysis we can identified which page is accessed and which is not.If pages are not accessed from a long time then we can modified the website .In fig2.0 we show the popular paths through the our example website.

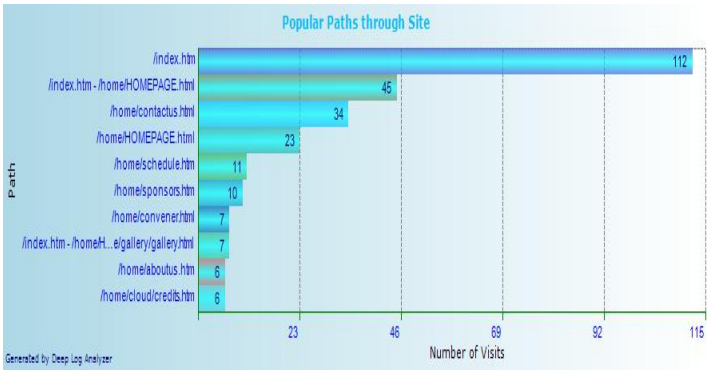

Fig 2.0 Popular Paths through sites

Another report is shows in fig3.0 the visitor history like in which date or day of the month the visitor visit the website.

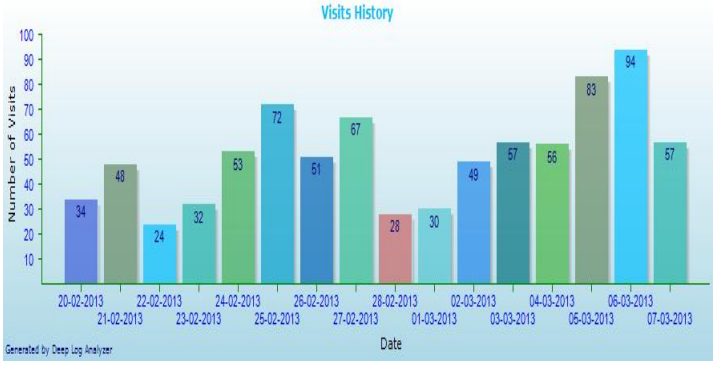

Fig 3.0 Visits History

This report shows in fig4.0 which hour the users comes in our website,if we know the exact hours then we can start the scheme which offers some discount in non peak time.In this report we analyze the total hits during the hours in a day.We can see in report there is about 830 hits in the 17 hours of the day.All the report are generated by using deep log analyzer

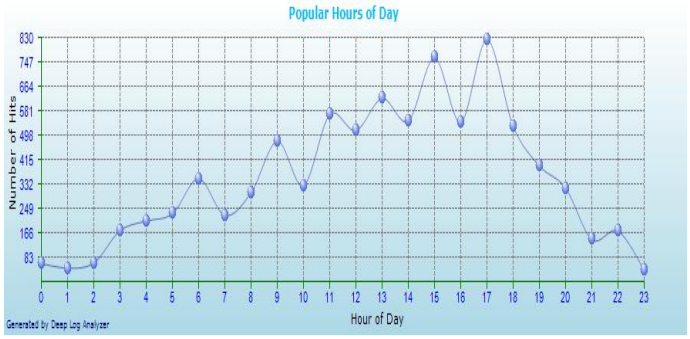

Fig4.0 Popular Hours of Day

VI. WEB USAGE MINING IN WEBSITE MODIFICATION

A website should be designed to entice the users. Web Mining analyses visitor's behavior and makes predictions on their future interaction. This can be exploited to improve website performance and to recommend products or links based on user's behavior. Visitors entering the site exhibits different behaviour.

All the above reports presented is clearly state that web usage mining plays a important role in websites.Another prominent issue of current online shopping model is that the system always regards resources as the core of modification, it shows the same resources to different product to achieve the aim of shopping, and ignores the differences including interests, shopping styles, shopping needs as well as shopping ability between different users. Making use of web usage mining, in one hand can be used as a reference of knowledge association, so that the link between knowledge points will be more closely and reasonable. Meanwhile through the analysis of access frequency and access time of different knowledge that the users have been accessed, we can get a more intuitive understanding about the knowledge users are interested in, as a result of this, we can recommend their interested knowledge to them for purchasing. On the other hand, when a user buy a product, and spends much time on purchasing. the system can remind the users properly, and recommend the missing. While online shopping platform in general only recommend the knowledge of users interested in.This knowledge of users behavior is very beneficial for the website modification.

VII. CONCLUSION

In this paper we provided more insights about patterns of navigation user behaviour . Using logs resulted from log analyser software which provide reports which is useful .We analysed these logs using process mining techniques.Web mining is a rapid growing research area. Web content mining is related but different from data mining and text mining. In this study, we have proposed a comprehensive and effective environment for web usage analysis in web sites offering query-based access to underlying information systems. Such sites are composed of dynamically generated pages linking data from one or more data sources. Their effectiveness in meeting their visitors' needs lies in the conformance of the query capabilities they provide to the intuition of their visitors. For web usage mining, the challenge lies in discovering navigation patterns that help in assessing and in improving this effectiveness..

ACKNOWLEDGMENT

We want to thank admin techzion.org , for providing us access to the server log data for analysing the website.

REFERENCES

[1] J.P.Callan." Passage-Level Evidence in Document Retrival.", In Proceeding of the ACM SIGIR Conference on Infromation Retrival, pages 302-309, Dulbin, Ireland, 1994.

[2] Cooley, R., Mobasher, B., and Srivastava, J. (1999) "Data Preparation for Mining World Wide Web Browsing Patterns" Knowledge and Information Systems, Vol. 1 No. 1, pp. 5-32

[3] Kohavi, R., Mason, L., Parekh, R., Zheng, Z. (2004) "Lessons and Challenges from Mining Retail E-commerce Data" Machine Learning, Vol. 57 No. 1-2, pp. 83-113

[4] A.Mendez-Torreblanca, M.Monte,"A Trend Discovery for Dynamic Web Content Mining", IEEE, Inteligence System, Vol 14, pages.20-22, 2002.

[5] M. Stolpmann, On-line Marketing Mix, Kunden finden, Kunden binden im E-Business, Galileo Press, Bonn 1999.

[6] J. Han and M. Kamber, Data Mining, Concepts and Techniques, Academic Press, San Diego, 2001

[7] Ribero-Neto, B. Laender. A. H.F. Da Silva "Topdowm ectraction of semi structured data" in string processing and information retrieval symposium, 1999 and International workshop on groupware, pages.176-183

[8] Feifei LKi, Zehua Liu, Yangfreng Huang, Wee-Keong Ng "Web Information Collection, Collaging and programming (WICCAP)"

[9] P. Cunningham,R. Bergmann, S. Schmitt, R. Traphöhner, S. Breen, and , B.Smyth, WEBSELL: Intelligent Sales Assistent for the World Wide Web, Zeitschrift Künstliche Intelligenz, 1, 2001, p. 28-32.
[10] M. Merz, E-Commerce und E-Business, dpunkt.verlag Heidelberg, 2002.
[11] P. Perner(Ed.), Data Mining in E-Commerce, Medicine, and Knowledge Management, Springer Verlag 2002, lnai 2394
[12] Reference website for analysis techzion.org
[13] Berry, J. A., Lindoff, G., Data Mining Techniques, Wiley Computer Publishing, 1997 (ISBN 0-471-17980-9).
[14] Berson, "Data Warehousing, Data-Mining & OLAP", TMH

