ANALYZING WEB LOGS OF AN ASTROLOGICAL WEBSITE USING KEY INFLUENCERS

Neha Goel, Sonia Gupta, C.K. Jha

International Research Journal
p-ISSN 2202-2821 e-ISSN 1839-6518 (Australian ISSN Agency)
Vol. 05 No. 01 2015
82800501201504

ANALYZING WEB LOGS OF AN ASTROLOGICAL WEBSITE USING KEY INFLUENCERS

NEHA GOEL, SONIA GUPTA, C.K. JHA

Abstract---With the growth of Internet, websites have become a dynamic tool in the market for every business to both acquire and service their customers. Online presence through websites has given them a global and wider reach. Good design and well management are the two key aspects of the effective website to catch the attention of visitors. Nowadays, the analysis and behavior of website visitors can be used to convert them into customers and serve them in a better way. This can be carried out on the weblogs which are generated as a result of user's access to a website. This paper aims to study the log files of an astrological website. An astrological website has not gained much attention in the internet community as people prefer going to the astrologers in person to resolve their problems. The log files of an astrology website were taken, preprocessed and analyzed using Analyze Key Influencers technique a feature of Microsoft SQL Server Data Mining Add-ins for Microsoft Office 2007. The information obtained can be used to enhance the effectiveness of the website.

KEY WORDS: CUSTOMER, PREPROCESSING, WEB LOGS, WEB LOG MINING, WEB USAGE MINING

INTRODUCTION

Currently websites are known to be the most effective marketing strategy for all business environments. It provides a wider reach to the customers as they can access the website 24x7 hence providing them flexibility and ease. World Wide Web (WWW) has evolved into a gold mine that generates huge data every day. To utilize this data completely a discipline called Web mining is becoming increasingly important these days. It applies data mining techniques and extracts patterns from web data. Web Mining can be classified into three types of mining namely Content which deals with the Content data present on the web), Structure (which examines the Hyperlink structure of web) and Usage dealing with the data stored in Web access logs of server) Mining. The major hurdle that comes up in front of website owners is the difficulty in attracting the visitors, turning these visitors to customers and retaining them. An alternative to this problem could be designing attractive web pages. People who access the internet are increasing and thus leading to significant rise in the amount of usage data which can be collected from the websites in the form of web logs. Whenever the user hits a particular Uniform Resource Locator (URL) it gets recorded in the web server associated with it. It is generated automatically and contains information about people accessing the website ranging from the client's IP (Internet Protocol) address, the web pages that he surfed to the time taken by him on a particular web page on the website.

Web Usage Mining accesses such information present in the web logs and analyzes it to find interesting patterns [1]. It consists of three major steps – Data Preprocessing, Pattern Mining and Pattern Analysis [2]. Web logs data is highly diverse, voluminous [3] and contains noisy and junk values in the form of .gif, .png, multimedia files, error files which needs to be removed in the preprocessing stage. Pattern Mining includes application of data mining techniques on cleansed data. After application of techniques such as Clustering, Association Rules etc. interesting rules and patterns can be discerned. The patterns extracted from this stage are utilized in the Pattern Analysis stage wherein after eliminating the irrelevant patterns interpretations are drawn by utilizing visualization techniques such as graphing pattern.

Web Usage Mining is being popularly used in areas like E-commerce, Tours & Travel, Matrimony, E-learning etc. Astrologers and astrology are in great demand these days. There are lots of websites also which are offering various astrological services. But now the customers have become increasingly savvy when it comes to the websites. There expectations are set by the experience they are getting from other similar websites. So, there is a severe competition among these websites. Every website wants to attract the

visitors to convert them into customers. Therefore, the information related to the consumer behavior is very precious to the website owners. In order to understand and study the behavior of customer the web access logs of an astrological website were taken and analyzed. The data was analyzed using Microsoft SQL Server Data Mining Add-ins for Microsoft Office 2007. These Add-ins helps in obtaining trends and patterns from complex data with the help of visualization techniques and generates attractive, easy to understand and meaningful representations.

The paper is organized as follows. Section 2 describes the work done by other researchers in this context. In Section 3 the experimental setting is described in detail. The results are discussed in Section 4. Conclusion is given in Section 5.

RELATED WORK

Web Usage Mining is the application of data mining techniques on web data or web server log files and discovering behavioral patterns. Authors Naga et.al. [4] have stated that web server log files are plain text files which store click stream data of users and comes in the following file formats 1) Common Log Format (CLF) 2) Microsoft Internet Information Services (IIS) Log file Format 3) World Wide Web (World Wide Web Consortium) Extended Log file Format and 4) NCSA Common Log file Format. They also stated that log files comprise of Access Logs, Referrer Logs, Agent Logs and Error Logs. Author Liu [5] states that log record contains lots of useful information and its analysis helps in tracing the service quality of a website. Web Usage Mining follows a three step process 1) Data Preprocessing 2) Pattern Mining 3) Pattern Analysis. Preprocessing of data includes cleaning of data, User identification, session identification and path completion. Preprocessing of web logs has been exclaimed as the crucial step as it affects a lot in result generation. Hence, this area is gaining a lot of popularity in the research community as well. Preprocessing has been studied deeply in [6, 7, 8]. Preprocessing results in significant reduction of data as it discards irrelevant data present in the form of multimedia files, error log files etc. Authors Jaideep et.al. [9] states that after the data passes through the Preprocessing stage, algorithms from Data Mining viz. Classification, Clustering, Association Rules, Statistical Analysis etc. are to be applied in the Pattern Mining phase depending upon the area in which it is to be applied. Finally, in the Pattern Analysis phase all the uninteresting patterns extracted from the previous phase are removed. Often, Visualization techniques are utilized to show these trends in data. Web Log Data has been utilized across all domains. Authors Mahendra et.al. [10] have utilized web logs for effective target marketing. On application of K means clustering algorithm on web logs and deriving useful clusters they have exclaimed that customer's behavior could be analyzed and target marketing can be done. Improving the website

structure is another area where web log data can contribute. Authors Ramakrishna et.al.[11] proposed an algorithm which finds pages in a website whose location is different from where visitors expects to find them automatically by backtracking approach. Authors Yoon et.al. [12] by learning customer preferences and product association from click stream have designed a personalized system. Author Ida [13] has also shown that this data can help significantly in improving search engine performance and moreover generate recommendations. Author Ford [14] has claimed targeted advertising to be an important application of Web Mining. In order to analyze web log data, automated web log analyzer tools are available which offer the website owners with the capability of analyzing the statistics and understanding the customer behavior. Authors Neha et.al. [15] has compared Web Log Expert (Web Analytics software) with Google Analytics. Reports that are generated by using these tools have been shown and explained in detail. Apart from the automated tools, a system that performs Web Usage Mining has been proposed in [9] and a web log mining tool has been designed by Jianli et.al.[16].Web Usage mining has always been a major topic of interest and a lot of survey has been done on the same. Chhavi [17] surveyed on Web Usage Mining and tools available and deployed in this area. Pani et.al.[18] examined the work done in Pattern Mining or Extraction and divided the algorithms studied into three broad categories viz. Association Rule Mining(ARM), Classification and Clustering. Later, research has been carried out separately in each area wherein the algorithms used by which author and which year has been surveyed.

It has been observed that web logs data is being used effectively in all areas especially e-commerce. But it was found that astrology websites are lacking the interest of the research community as well as website visitors. Hence, web log analysis should be applied on such websites and some useful inferences must be drawn which can help in increasing the popularity of these websites.

EXPERIMENTAL SETTING

Here, we are considering the web logs of an astrological website. The major objective of the proposed methodology is to acquire behavioral pattern of the visitors of an astrological website. Every click in the website is recorded in Web Server logs hence generating abundant data. This data cannot be directly used for analysis as it contains junk values and noisy data in the form of multimedia files, error files as stated above. The data needs to be preprocessed in order to make it suitable for analysis. Therefore, preprocessing was done for the astrological web logs. After the completion of preprocessing, it was found that file size reduced considerably. Once the data was cleansed, it was fed as an input to Microsoft Excel and SQL Server 2008 Data Mining

Add-ins for Office 2007 were used for analyzing the data. The Add-ins provides wizards and tools that make it easier to extract meaningful information from data. They help in deriving patterns and trends hidden in complex data, visualize those patterns in charts and then generate rich, colorful summaries for presentation and for analysis.

DATA UNDERSTANDING

The web logs that were taken for analysis had 21 attributes in the dataset namely <Date>,<Time>,<c-ip>, <cs-username>, <s-sitename>,<s-computername>,<s-ip>,<s-port>,<cs-method>,<cs-uri stem>,<cs-uri-query>,<sc-status>,<sc-win32-status>,<sc-bytes>,<cs-bytes>,<time-taken>,<cs-version>,<cs host>,<cs(User-Agent)>,<cs(Cookie)>,<cs(Referer)>.The data recorded was in Microsoft IIS W3C extended log format.

DATA PREPROCESSING

The data that was extracted in the form of web logs was in the .log format. This data was imported to Excel sheets. It was cleansed by applications of filter (feature of MS Excel). The cleansing of data reduced the instances from 1203 to 238. After the data has been cleansed, the table analysis tools are applied to perform data mining.

PATTERN MINING AND ANALYSIS USING KEY INFLUENCERS

The classification model analyze key influencers have been used to perform analysis. Analyze Key Influencers tool analyzes the correlation between all the columns in the table and a specified target column. The result is a report that classifies the columns having significant influence on the target and explains in detail how this influence matters itself. The major key influencers report consists of a table with four columns: Column, Value, Favors and Relative Impact. The table contains multiple sections, identified by different colors in the Relative Impact column and different values in the Value column. Each of the sections represents the key influencers for one distinct value of the target column. The relative impact indicates the strength of the association of this attribute with the outcome. The length of the bar indicates the probability that the factor contributes to the outcome; therefore, the longer the shaded bar, the stronger the association

RESULT AND DISCUSSION

After analyzing the value of cs_uri_stem it has been found that the websites' most frequently accessed web pages are the yantra page and lalkitab page. All the yantra's provided are having more or less the same relative impact value. That shows these two are the most visited web pages amongst all. It has also been observed that maximum number of users is not accessing the website through the index page as its relative impact

value is low. This indicates that the users are not aware of the website or they are not the regular users. They are searching the page using some search engine. Moreover, the index page is not user friendly or poorly structured. The client with i.p. address 66.249.72.11 has the highest relative impact value w.r.t time taken. Moreover, the client has access maximum pages of the website including the shipping page as shown in Figure 1. This means that the user has accessed the website the maximum without wasting much of time. As the client has used <537 ms on the website and still has a higher relative impact value it shows that if the time taken by a user on the website is more it is not necessary that he is interested in accessing the information on the website. As can be seen the other c_ip have a relatively higher value of time taken but a lesser relative impact.

The start of the week or mid of the week does not have any impact on the access behavior of a client for the website. As can be seen by the relative impact values the date and day on which a client is accessing an astrological website, the values are almost same for the same client on different dates. That means the client behavior does not changes due to the start of the week or mid of the week. It has also been found that Mozilla has the highest relative impact as a . In other words, it was the most used browser for accessing this website. So, the compatibility of the website with other browsers need to be checked.

CONCLUSION

With the advent of web, astrology has also been modernized. Now, the users can get much information while surfing the stars. But unlike other websites, astrological websites are facing the challenge of acquiring and retaining the customers. The analysis of the customer behavior helps in understanding the requirements of the visitors in a better way. The websites are an important source of data for the analysis. Businesses are using web logs for studying the behavior of the visitors accessing their website and discovering interesting patterns. According to this study, the majority of the customers are accessing the yantra and lalakitab web page. So, branding of website can be done on these web pages. Day of the week does not have any significant impact on the access of the website. Therefore, the promotional offers need to be launched for all seven days. Strong emphasis should be taken while designing the index page as it can be considered as mirror image of the business or website owner. The study has revealed that maximum users are not surfing the website using index page. Mozilla is the only browser used for accessing the website. It shows that the website is less compatible with all the other browsers.

FIGURES AND TABLES

FIGURES

```
2010-12-27 14:36:09 W3SVC1 H-ASTROSCIENCES 69.16.245.169 GET /index.html - 80 - 69.16.234.108 HTTP/1.0 check_http/1.81+(nagios-plugins+1.4.2) - - 69.16.245.169 200 0 64 091 93
2010-12-27 15:02:49 W3SVC1 H-ASTROSCIENCES 69.16.245.169 GET /index.html - 80 - 69.16.234.108 HTTP/1.0 check_http/1.81+(nagios-plugins+1.4.2) - - 69.16.245.169 200 0 64 091 0
```

Fig. 1 Example of Weblog generated

2/13/2012	4:54:02	/privacy-policy.asp
2/13/2012	4:54:31	/lal-kitab-amrit.asp
2/13/2012	4:54:59	/genral-remedies.asp
2/13/2012	4:55:24	/Index.asp
2/13/2012	4:55:24	/style-index.css
2/13/2012	4:55:26	/images/img1.png
2/13/2012	4:55:27	/images/devider.png

Fig. 2 Before Filtering .png files present

2/13/2012	4:54:02	/privacy-policy.asp
2/13/2012	4:54:31	/lal-kitab-amrit.asp
2/13/2012	4:54:59	/genral-remedies.asp
2/13/2012	4:55:24	/Index.asp
2/13/2012	4:55:24	/style-index.css
2/13/2012	4:55:27	/lalkitab-booking.asp
2/13/2012	4:55:29	/favicon.ico

Fig. 3 after filtering .png files removed

Key influencers and their impact over the values of 'cs_uri_stem'

Filter by 'Column' or 'Favors' to see how various columns influence 'cs_uri_stem'

Column	Value	Favors	Relative Impact
c_ip	66.249.72.11	/shipping.asp	
c_ip	66.249.72.11	/saket-yantra.html	
c_ip	66.249.72.11	/santan-yantra.html	
c_ip	66.249.72.11	/sanket-yantra.html	
c_ip	66.249.72.11	/sarbsidhi-yantra.html	
c_ip	66.249.72.11	/order-yantra.asp	

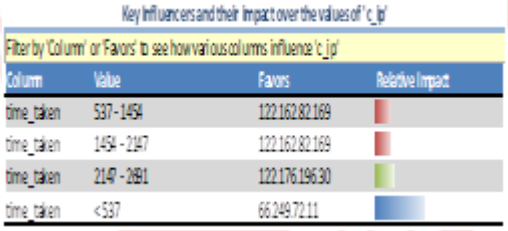
Fig. 4 Key influencers for cs_uri_stem


Key influencers and their impact over the values of 'cs_uri_stem'

Filter by 'Column' or 'Favors' to see how various columns influence 'cs_uri_stem'

Column	Value	Favors	Relative Impact
c_ip	122.162.82.169	/index.asp	
c_ip	122.161.44.33	/index.asp	

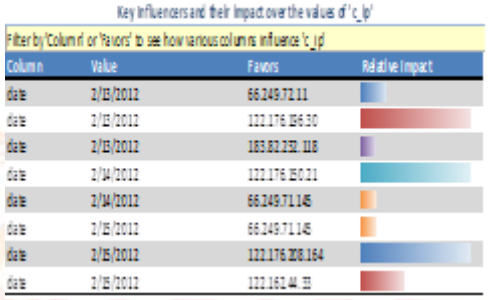
Fig. 5 Key influencers showing relative impact of index.asp


Key Influencers and their impact over the values of 'c_ip'

Filter by 'Column' or 'Favors' to see how various columns influence 'c_ip'

Column	Value	Favors	Relative Impact
time_taken	537 - 1454	122.162.82.169	
time_taken	1454 - 2147	122.162.82.169	
time_taken	2147 - 2891	122.176.196.30	
time_taken	<537	66.249.72.11	

Fig. 6 Impact of attribute time_taken on c_ip along with relative impact.


Key Influencers and their impact over the values of 'c_ip'

Filter by 'Column' or 'Favors' to see how various columns influence 'c_ip'

Column	Value	Favors	Relative Impact
date	2/13/2012	66.249.72.11	
date	2/13/2012	122.176.196.30	
date	2/13/2012	183.82.232.118	
date	2/14/2012	122.176.150.21	
date	2/14/2012	66.249.71.145	
date	2/15/2012	66.249.71.145	
date	2/15/2012	122.176.208.164	
date	2/15/2012	122.161.44.33	

Fig. 7 Impact of date attribute on c_ip along with relative impact.

REFERENCES

[1]. Chhavi Rana, "A Study of Web Usage Mining Research Tools", International Journal of Advanced Networking and Applications, ISSN: 0975-0290, Vol: 03, Issue: 06, pp.1422-1429, 2012.

[2]. Ford Lumban Gaol, "Exploring the Pattern of Habits of Users Using Web Log Sequential Pattern", in 2010 Second International Conference on advances in Computing, Control and Telecommunication Technologies IEEE, 2010, p. 161.

[3]. Ida Mele, "Web Usage Mining for Enhancing Search Result Delivery and helping Users to find Interesting Web content", ACM, WSDM '13, pp.765-769, 2013.

[4]. Jaideep Srivastava , Robert Cooley, Mukund Deshpande and Pang- Ning Tan, "Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD, Vol1, Issue 2, pp.1-12,Jan2000.

[5]. Jianli Duan and Shuxia Liu, "Research on Web Log Mining Analysis", in 2012 International Symposium on Instrumentation & Measurement , Sensor Network and Automation, 2012 IEEE , p. 515.
[6]. L.K. Joshila Grace, V.Maheswari, DhinaharanNagamalai, "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011, pp 99-110
[7]. Liu Kewen , "Analysis of Preprocessing Methods for Web Usage Data", in 2012 International Conference on Measurement, Information and Control, 2012 IEEE, p. 383.
[8]. Mahendra Pratap Yadav , Mhd Feeroz and Vinod Kumar Yadav, "Mining the Customer Behavior using Web Usage Mining in e-commerce", in ICCCNT 2012, 26-28 July 2012, IEEE.
[9]. Naga Lakshmi, Raja Sekhara Rao and Sai Satyanarayana Reddy, "An Overview of Preprocessing on Web Log data for Web Usage Analysis", IJITEE, ISSN: 2278-3075, Vol.2, Issue-4, pp.274-279, March 2013.
[10]. Neha Goel and Dr. C.K. Jha, "Analyzing Users Behavior from Web Access Logs Using Automated Log Analyzer Tool", International Journal of Computer Applications (0975 – 8887) ,Vol. 62– No.2,pp.29-33, January 2013.
[11]. P. Nithya and Dr. P. Sumathi , "Novel Preprocessing Technique for Web Log Mining by removing Global Noise and Web Robots", in 2012 National Conference on Computing and Communication Systems, 2012 IEEE.
[12]. Ramakrishnan Srikant and Yinghui Yang, "Mining Web logs to Improve Website Organization", ACM 2001.
[13]. Ramya C, Dr. Shreedhara K S and Kavitha G," Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process ", in (ICCEI 2011) IEEE, 2011, p. 317.
[14]. Ravindra Gupta and Prateek Gupta, "Application specific web log pre-processing", IJCTA, ISSN: 2229-6093, Vol 3(1), pp.160-162, Jan- Feb 2012.
[15]. S. K. Pani, L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal and S.K.Padhi, "Web Usage Mining: A survey on Pattern Extraction from Web Logs",IJICA,Vol1,Issue1, ,pp.15-23,2011. [Online]. Available: http://www.loganalyzer.net/log-analyzer/w3c-extended.html

[16]. Swapna Mallipeddi, and D.N.V.S.L.S. Indira , "High Utility Mining Algorithm for Preprocessed Data", International Journal of Computer Trends and Technology, Vol 3, Issue 3, pp. 379-386,2012.
[17]. Yoon Ho Choa, Jae Kyeong Kimb and Soung Hie Kima , "A personalized recommender system based on web usage mining and decision tree induction" , 2002 Elsevier Science Ltd.
[18]. Zhenglu Yang, Yitong Wang, Masaru Kitsuregawa, "An Effective System for Mining Web Log", APWeb, Volume 3841 of Lecture Notes in Computer Science, Springer,2006, pp 40-52

– END –

International Research Journal
p-ISSN 2202-2821 e-ISSN 1839-6518 (Australian ISSN Agency)