Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 3, March -2017

PREDICTING USERS WITH SIMILAR BEHAVIOUR THROUGH SESSIONS

Reeny Zackarias[1], Nicy K S[2]

[1]*M. Tech student, Computer Science and Engineering, IES College of Engineering, Thrissur, Kerala, India*
[2]*Assistant Professor, Computer Science and Engineering, IES College of Engineering, Thrissur, Kerala, India.*

Abstract — *Weblog mining is the method to extract the user sessions from the given log files. Each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Server-side logs and client-side logs are commonly used for web usage and usability analysis. Server-side logs are generated automatically by every user when users click to the corresponding web servers. Weblog mining process includes three process, namely data preprocessing, pattern analysis and pattern discovery. Data preprocessing includes 3 stages, namely Data cleaning, User identification and Session identification. In this paper, we are implementing these three processes and finding the users with same behaviour.*

Keywords *- Weblog mining, User Identification, Session Identification, Buyer Pattern*.

I. INTRODUCTION

Web Usage mining applies data mining technique to extract knowledge from web log files automatically. Web mining can be classified into web content mining, web structure mining and web usage mining. Web content mining is the process of extracting information from large amount of collected data. This technique extracts the information from the contents of web pages. Web content mining is again classified into multimedia mining and web textual mining. Web structure mining is the process of analyzing the links between webpages through the web structure. Web structure mining is again categorized into hyper link mining and inside structure mining. Web usage mining analyses weblog files for finding browser patterns of user. Web usage mining is also called weblog mining.

Web log mining has important value and meaning in the field of science and technology management. This technique uses information resources from the internet and help technological workers and decision makers. Web log mining helps to discover useful patterns of major customers which reduces competition and simultaneously increases business profit in E-commerce

II. RELATED WORKS

G. Neelima and Dr. Sireesha Rodda members of IEEE have proposed an idea for predicting user behaviour through sessions using the web log mining. They deal with web server logs of NCSA common log file format for mining. The proposed methodology consists of data processing, user identification and session identification processes. They are using three different algorithms for performing each of these processes [1].

Anshul Bhargav and Munish Bhargav also the members of IEEE proposed frame work for web usage mining. The frame work is to perform users classification based on discovered patterns and to find the characteristics of users. The frame work consists of three main steps: preprocessing, pattern discovery and user classification. The preprocessing stage includes log file cleaning, user identification and session identification. The pattern discovery process discovers new pattern from the data set. In user classification process, each user is classified according to their characteristics. The classifications are based on country based classification, site entry based classification and access time based classification [2].

S. S. Patil and H.P Khandagale proposed cognitive user model which specify the anticipated usage behaviour based on the patterns discovered in previous preprocessing, phase. The frame work includes data cleaning, user identification, session identification, path completion, transaction identification and pattern discovery and extraction processes. Path completion refers to crating rules for missing references based on site structure, referrer and other heuristic information. By analysing the discovered pattern the behaviour of the user is identified and navigation updates are provided in the web page [3].

Virendra R. Rathod and Govind V. Patel proposed a frame work based on FCM clustering and Markov model for predicting user behaviour using web log. After preprocessing fuzzy c means (FCM) algorithm is applied for pattern discovery and analysis. Markov model is used to next page prediction and better web page prediction accuracy. Fuzzy c means clustering algorithm provides better result than k- means clustering. Fuzzy c means clustering is one of the most widely used fuzzy clustering algorithms [4].

III. WEB LOG FILES

Weblog files are the files which contains complete information about the user's browser activities. There are three types of log files:

3.1 Web server logs

Web server logs contain information about user request history including client IP address, request date/time, HTTP code, bytes reserved etc. These files are not accessible by general internet users. Only the webmaster or other administrative person can access these files. A statistical analysis of the web server log may be used to examine navigation patterns by time of day, day of week, referrer, or user agent. Efficient web site administration, adequate hosting resources and fine tuning of sales efforts can be aided by the analysis of web server logs.

3.2 Proxy server logs

Proxy server is a caching mechanism which lies between client browsers and web servers. Proxy caching is used to decrease the loading time of webpage and to reduce the network traffic at the server and client side. Proxy server may be a computer system or an application. In computer networks, a proxy server acts as an intermediary for requests from clients seeking resources from other servers. A client that connects to the proxy server requests some service, such as a file, connection, web page, or other resource available from a different server. Then the proxy server evaluates the request as a way to simplify and control its complexity.

3.3 Browser Logs

Browser logs are client side data. JavaScript and Java applets are used to collect browser logs. User cooperation is needed to implement client side data collection. To improve the E-commerce usability web server logs are used. This log is stored by the browser on user's device's local hard drive and can be utilized for a number of purposes. It provides on-the-fly suggestions as you type a URL or website name into the address bar.

IV. LOG FILE FORMAT

Most popular log file formats are W3C Extended log file format, NCSA common log file format and IIS log file format. In this paper NCSA log file format is used for mining process. The data logged for each request is fixed in NCSA log file format. The following is the NCSA log format:

216.67.1.91 - - [01/Jul/2002:12:11:52 +0000] "GET /index.html HTTP/ 1.1" 200 431

It contains IP address of the user, date and time of the request, zone, method, requested page, protocol version, status of the request and bytes served.

V. PROPOSED METHODOLOGY

The main aim is to do classification of users based on discovered patterns and finding users with same behaviour. The proposed methodology consists of the following steps:

5.1 Log files cleaning

In data preprocessing it takes web log data as input and removes unnecessary data. The web log data contains records of graphics, video and the format information. That is every record of URL field contains GIF, JPEG, PNG and CSS file extensions. The URL containing these extensions has to be eliminated from the log file. We will keep only the URL which contains extensions like .asp, .php, .html etc. We can also delete log entries with empty URL and with HTTP status code other than 200.

5.2 User Identification

In user identification process each user accessing the website is identified by different IP addresses. By identifying each user we can retrieve each user's access characteristics. With this we can make user clustering and provide recommendation service to the users.

5.3 Session Identification

Session refers to a sequence of web pages viewed by a user during one visit and this session is recorded in the log file. The time between the entry to the site and exit from the site is known as session. If the time between the entry and exit on the site is less than say 1 hour, then there is no new session. But if the time exceeds 1 hour, the new session of the same user has been started.

5.4 Classification of users

The cleaned log files are input for the classification process. The users are classified according to different criteria. The process identifies the number of users in each session and also the number of sessions for each user.

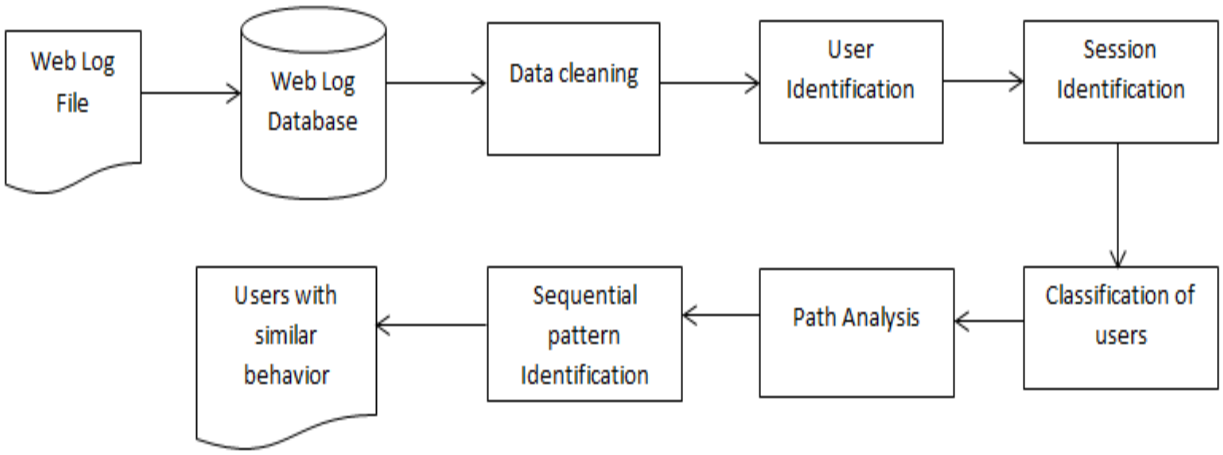

FIG 1: PROPOSED ARCHITECTURE DIAGRAM OF PREDICTING USERS WITH SIMILAR BEHAVIOUR THROUGH SESSIONS

5.5 Path analysis

The cleaned log file from data preprocessing stage is the input for pattern discovery process. The users and sessions are identified in the cleaned data. In the next stage we have to discover the web usage patterns. There are different methods for discovering patterns. The methods include classification, clustering, statistical analysis, association rules, sequential pattern analysis etc. In this paper sequential pattern analysis technique will be used. By using sequential pattern analysis technique useful patterns will be discovered in this process. The pattern discovered during pattern discovery process is analysed for finding users with similar behaviour and similar web usage pattern.

VI. CONCLUSION

The web log mining is based upon the discovery and analysis of web usage patterns. It helps website administrators to provide the needs of their website users in a better way. Lots of frame works and methodologies have been already proposed for efficient mining of web log files. Here we propose the methodology for efficient analysis of web log file. The proposed methodology includes three steps: data preprocessing, pattern discovery and pattern analysis. The preprocessing phase includes data cleaning, user identification and session identification. It is important to carry out the data preprocessing stage efficiently, because the data used for mining must be useful and efficient. So in data cleaning phase all unwanted and noisy data needs to be removed. After cleaning any pattern discovery technique such as classification, clustering, sequential analysis can be applied to discover useful pattern. The web log mining has various applications. Some of them are: (i) Web personalization – according to the user behavior a website can be designed and re-structured to make it more user friendly. (ii) System improvement – web mining can be applied to areas of load balancing, web caching, network transmission and data distribution. (iii) E-commerce /Business intelligent – the web log mining helps many organizations to understand its customers behavior and built customer profiles on the basis of customer's habits. Based on the interests and needs companies can increase their profit by selling items correlated to customer's demand.

REFERENCES

[1] G. Neelima , Dr. Sireesha Rodda, "*Predicting user behavior through sessions using the web log mining*", Conference on Advances in Human Machine Interaction (HMI), R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India, March 2016.

[2] Anshul Bhargav ,Munish Bhargav, "*Pattern discovery and users classification through web usage mining",* International Conference on Control Instrumentation, Communication and Computational Technologies, IEEE 2014.

[3] S. S. Patil , H.P.Khandagale,"*Survey paper on enhancing web navigation usability using web usage mining techniques*", International journel of modern trends inenginnering and research.2016.

[4] Virendra R. Rathod and Govind V Patel, "*Prediction of user behaviour using web log mining in web usage mining*", International journel of computer application vol. 139- No. 8, April 2016.

[5] Wei, Shirly Kho, Wahidah Husain, Zurinahni Zainol ,"*A study of customer behaviour through web mining* ", Journal of Information Sciences and Computing Technologies ISSN 2394-9066 , Volume 2, Issue 1 February, 2015.

[6] Mirghani. A. Eltahir ,Anour F.A. Dafa-Alla*," Extracting Knowledge from Web Server Logs Using Web Usage Mining",* International conference on computing, electrical and electronic engineering (ICCEEE), 2013.

[7] Guandong Xu,"*Webmining techniques for recommendation and personalization*",Victoria University, Australia, March 2008.

[8] Dr. R. Krishnamoorthi , K. R. Suneetha, "*Identifying User Behavior by Analyzing Web Server Access Log* File", International Journal of Computer Science and Network Security, April 2009 .

[9] Dilip Singh Sisodia, Shrish Verma, "*Web Usage Pattern Analysis Through Web Logs: A Review",* International Joint Conference on Computer Science and Software Engineering, pp. 49 - 53, 2012

