IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Issue-7 E-ISSN: 2347-2693

Enhanced User Interest Level Preprocessing Technique for Efficient Web Page Recommendation

R. Suguna

Theivanai Ammal College for Women, Villupuram, Tamil Nadu, India

Available online at: www.ijcseonline.org

Received:18/Jun/2016 Revised: 26/Jun/2016 Accepted: 16/Jul/2016 Published: 31/Jul/2016

***Abstract*—** Web based applications play a major role in people day to day activities. Monitoring the users actions are really an interesting and necessary job of the website forecaster to familiarize about their performance, classify the likeminded users, understand the website visitor's browsing history, reconstruct the website, web recommendation and web personalization. Web logs are the main source to provide sufficient information about the users and achieve the above requirements. Pattern discovery algorithms are applied to the web logs to extract the desirable information. It is mandatory for website analyst to understand the user behavior and interest for many analytical purposes. Web logs take an important role to know about the user behavior. Several pattern mining techniques were developed to understand the user behavior. But, there are no special preprocessing techniques to identify the user interest level and understand their browsing patterns. A special kind of preprocessing technique is needed to improve the quality and efficiency of the pattern mining algorithms. The proposed preprocessing technique performs the preprocessing activities on web logs and also identifies the similar kind of users. The user similarity helps for efficient web page recommendation technique.

***Keywords*—**Web logs; Preprocessing; Data Cleaning; User Identification; Session Identification; Web page recommendation

I. INTRODUCTION

World Wide Web is a massive collection of information. Information is arranged in proper hierarchy in the form of websites. Website contains collection of web pages that are accessed via hyperlinks (Chhavi 2012). Currently Internet is a major source of information for all kinds of users. It is used by the millions of users every day. Whenever the user interacts with the , the interaction details are automatically recorded in web server in the form of web logs (Sanjay & Sangram 2010). Website analyst use the web log information for variety of purposes such as identifying and understanding the users behavior and expectation, improving the business process, website customization, web personalization and recommendation.

It is mandatory for website analyst to understand the user behavior and interest for many analytical purposes. Web logs take an important role to know about the user behavior. Several pattern mining techniques were developed to understand the user behavior. But, there are no special preprocessing techniques to identify the user interest level and understand their browsing patterns. A special kind of preprocessing technique is needed to improve the quality and efficiency of the pattern mining algorithms.

The existing algorithms have applied the preprocessing activities to reduce the size of the log file and to identify the number of unique users and sessions (Tasawar et al 2010). They are not developed for specific pattern mining algorithms and for particular kind of applications. User Interest Level based Preprocessing (UILP) algorithms were proposed (R Suguna & D Sharmila 2013) to identify the users based on their browsing pattern. Enhanced User Interest Level Preprocessing (EUILP) technique is newly proposed to identify the user interest level and group similar kinds of users. This paper details the literature survey in Section II, Section III describes the basics of Web Logs, existing preprocessing techniques are detailed in Section IV, Section V discusses EUILP technique, the performance of the EUILP technique with UILP algorithms are analyzed in Section VI. Finally conclusion is discussed in Section VII.

II. LITERATURE SURVEY

Web based applications play a major role in people's day to day activities. Monitoring the user's actions are really an interesting and necessary job of the website forecaster to familiarize about their performance, classify the likeminded users, understand the website visitor's browsing history, reconstruct the website, web recommendation and web personalization. Web logs are the main source to provide sufficient information about the users and achieve the above

requirements. Pattern discovery algorithms are applied to the web logs to extract the desirable information.

Cooley et al (1999) clarified that preprocessing is an essential task of web usage mining before applying any pattern discovery algorithms. Preprocessing activities are associated with pattern mining algorithms to improve the quality and accuracy of the discovered patterns. Web usage mining deals with secondary data that are recorded in the form of web logs in various sources. So, careful investigations on web logs improve the quality of the forthcoming algorithms.

Many authors have done the research on web log preprocessing and provide the necessary outline for better preprocessing activities. It has the following phases like (i) Data collection (ii) Data cleaning (iii) User identification (iv) Session identification and (v) Path completion. The authors mentioned that the aim of preprocessing activities is to identify the user's interest level with websites. So, during data cleaning unwanted information are removed. Users are identified by using IP address, agent log, browser machine and operating system. Session is constructed by using time in and time out mechanism. It is considered as new session when the page request time exceeds 25 or 30 minutes. Path completion is performed by using site topology and referrer log entry. Finally, data are formatted in a suitable manner for applying pattern mining algorithm.

A complete preprocessing technique which includes data collection, data cleaning, user identification, session identification and path completion is developed by Srivastava et al (2000). They stated that preprocessing is the most important and critical job of web usage mining due to the incomplete and inconsistent data. During data collection phase, client and proxy level data collections are suggested for better preprocessing. Server side web logs are not reliable due to the caching mechanism that is available many places in the web environment.

Ramya & Shreedhara (2012) mentioned that data collection from multiple web servers provide better result for preprocessing the web logs. During processing activities web logs are collected from web server, proxy server and client machine. User identification, session identification and path completion are performed by the authors with the multiple sources of web logs. Sheetal & Shailendra (2012) and Malarvizhi & Sahaaya (2012) recommend that extraction of web logs from multiple data sources such as web server, proxy server and client side machine improves the efficiency of preprocessing technique.

Distinct user identification is proposed by Sheetal & Shailendra (2012). They suggested that user identification is an important and necessary step of web log preprocessing and it helps in applying various pattern mining algorithms. They used distinct user identification algorithm to identify the users and unique users. Users are identified based on IP address, user name, operating system and browser version.

Suguna R & Sharmila D (2013) proposed UILP preprocessing algorithms which identify the user interest level based on their browsing history.

III. BASICS OF WEB LOGS

Web logs are maintained in the web servers in the form of plain text files which contains the details about user name, IP address, date, time, number of bytes transferred, access request and referrer log (Sanjay & Sangram 2010). Web logs are list of page references by the users or click stream data which contains inconsistent and incomplete information (Mohd et al 2008). So, it is difficult to use the web logs directly for pattern mining algorithms to extract the features. Preprocessing techniques are necessary to make them consistent and complete.

Web logs are maintained in the following places as line of text (Srivastava et al 2000): (i) Web server (ii) Proxy server and (iii) Browser machine.

The log files stored in web server provides more complete and accurate information about the user's interaction with the website. The World Wide Web Consortium (W3C) maintains a standard format for web server log files. The logs are added at the end of the log file. It records the details about client ip address, request date and time, page requested, Hyper Text Transfer Protocol (HTTP) code, bytes transferred, user agent and referrer (Raju & Sathyanarayana 2008).

Web logs for the particular user are stored in the browser machine. The browsers are programmed and scripting languages are used to collect client side data. This implementation of client side data collection requires user support to activate the scripting languages or use the personally programmed browser.

There are three types of web log formats (Srivastava et al 2000). They are

- W3C Extended Log File (ELF)

- National Center for Supercomputing Application (NCSA) Common Log File (CLF)
- Microsoft Internet Information Server (IIS) Log File.

W3C ELF log format is a default log file format for IIS server. Field are separated by space, time is recorded as Greenwich Mean Time (GMT). This format is personalized by the administrators to add or remove fields depending on the information needed to record. The date format for W3C is YYYY-MM-DD.

NCSA CLF records the information pertaining to user name, date, time, request type, HTTP status code and number of bytes. NCSA is fixed format and not customized by the administrators. The date format is DD/MMM/YYYY. Fields are separated by space and follow the local time.

In Microsoft IIS log format, web logs are maintained in American Standard Code for Information Interchange (ASCII) format which is not customized by the administrators. Fields are separated by comma. Time is recorded in local time. It records more information than NCSA format. Fields in IIS log file are client ip address, user name, data, time, server name, server IP address, time, client bytes sent, server bytes sent, service status code, request type, target of operation and parameters.

All the log file formats share the common information. The common format for the web log files are CLF and ELF. The ELF format additionally has two fields at the end which are the referrer Universal Resource Locator (URL) and user agent.

IV. EXISTING PREPROCESSING TECHNIQUE

Preprocessing (Cooley et al 1999 and Chitraa & Antony 2011) is an important activity in web usage mining and treated as a key to success. Preprocessing techniques eliminate the unwanted information from the web logs and facilitate the effective pattern mining process.

It consists of data collection, data cleaning, user identification, session identification and path completion.

Data Collection

Data collection (Malarvizhi & Sahaaya 2012) is an initial step in web log preprocessing. The user interaction details with the website are recorded in the form of web logs in three different places. They are (i) Web server (ii) Proxy server and (iii) Browser machine. The web logs are collected from multiple data sources and combined into new log file.

Data Cleaning

Data cleaning (Vijayashri & Madhuri 2012) is the process of removing noisy and irrelevant data that are not helpful for mining the knowledge from the web logs. When the users request the HTML web pages, the embedded images are also to be downloaded and stored in the web server. But these are not explicitly requested by the users and therefore avoided. This is done by checking the suffix of each URL. In addition to this, poor status code and request from auto search engines are also removed.

Data cleaning process consists of removing (Surbhi & Rinkle 2012) the following records from log files: (i) The records which have the extension *.gif, *.jpeg, *.css, *.cgi (ii) The records with the failed status code. The status code greater than 299 and lesser than 200 are treated as failure status code and (iii) The requests processed by auto search engines such as crawlers, robot and spider that are removed.

User Identification

User identification is a complex job of web log preprocessing. But it is essential to distinguish the users (Vijayashri & Madhuri 2012 and Sheetal & Shailendra 2012).

Grouping the users based on their visiting behavior is one of the important applications of web usage mining. Different techniques such as IP address, referrer log and user agent are used to identify the users. The following methods are used to identify the user:

i) Unique IP address represents one user.

ii) If IP address is same and agent log is different, then it is considered as distinguish users.

iii) To construct the browsing path using the access log and referrer logs. If there is a mismatch in the browsing path, then consider another user in same ip address.

Session Identification

Whenever user interacts with the website, they spend time to visit each web page. Session is the time duration spent by the users to visit the set of web pages during their login period. Session identification is the process of dividing the

individual user access logs into sessions (Sheetal & Shailendra 2012 and Thanakorn et al 2012).

The login and logout time are considered to identify the starting and ending time of each session. The following are the common rules to identify user session: If the user is identified as new user then there is a new session. For the same session, if the referrer page is null then there is a new session. If the time between page requests exceeds 30 or 25.5 minutes then it is considered as new session.

Path Completion

There are chances of missing pages after constructing transactions due to proxy servers and caching problems in web server logs (Marathe 2012). In such condition it becomes mandatory for identifying the user's access path, and adding the missing paths. Because of local buffer existence, some requested pages are not recorded in access log (Maheswara & Valli 2011). The goal of path completion is to fill all the missing references that are not recorded. The solution for path completion is, if a requested page is reachable by a hyperlink from any of the visited pages by the user then it is assumed that it is added in the session.

Data Formatting

The data formatting is the final step in preprocessing. The preprocessed web log information is properly formatted suitable for applying the pattern discovery algorithms.

V. EUILP PREPROCESSING METHODOLOGY

In the recent days voluminous amount of users use the internet services for their necessity. It is essential and important to realize their website surfing practice in order to make the websites user friendly. It motivates the research activities and stokes the user interest to draw information from web log files. The web logs are one of the most utilized features to extract the user's interest measure. The web log mining is used more frequently in order to identify the user behavior based on the extent to which a user is visiting a particular web site. The web logs are updated every time whenever the user visits a particular web site.

User's interest level is identified mainly based on their website and webpage navigation behavior. The UILP algorithms considered the following four features to identify the user interest level.

i) During data cleaning process explicit image and multimedia requests from users are considered and those requests are not removed from web logs.

ii) Users are identified based on site topology and cookies.

iii) Session time is calculated based on the time spent on each website by a particular user.

iv) Frequency value is calculated based on the number of web pages visited by the user on particular website.

v) Quality rate of the website is considered to identify the likeliness of the website by the users.

EUILP Algorithms

The preprocessing steps are considered as the initial process and web logs are formatted according to the proposed clustering algorithm to group the users based on their website visiting behavior which is to be explained in next chapter. According to the clustering algorithm, the web logs are preprocessed with the following five attributes:

b = *<ip, user, url, session, frequency, quality_rate>*

Where b is boid, ip is the ip address, user is the user name, url is website address, session is time duration spent on each website by the user and frequency is the number of visits by the user, quality_rate is quality rate of each website given by the user.

The preprocessing techniques are applied to the web log files in various perspectives to make them reliable.

VI. PERFORMANCE ANALYSIS

The web log files are randomly generated by the algorithm.

Performance Metrics

The performance of EUILP preprocessing algorithms is compared with the existing technique namely UILP (2013) with respect to (i) Data Cleaning (ii) Similar User Identification.

Performance of EUILP Algorithms with Existing Algorithms

The performances of EUILP algorithms and existing algorithm is summarized in table 6.1 and 6.2. Totally 5,40,350 records are randomly generated for performance evaluation. Initially data cleaning process is applied to the

web logs to eliminate the unwanted records. The size of the log files are considerably reduced after performing the data cleaning process.

Algorithm Name	Total Number of Records	Total Number of Records after Data Cleaning
UILP	540350	522424
EUILP	540350	522424

Table 6.1 Performance Comparison of EUILP Algorithms with UILP

UILP algorithm could not identify the similar users. The algorithm only considered user and unique user identification. But the proposed technique is specific for considering similar users based on frequency, session and quality rate of the website. So, EUILP identified 5678 similar users.

Algorithm Name	User Identification	
	Total Number of Users	Total no of Similar Users
UILP	10256	10256
EUILP	10256	5678

Table 6.2 Performance Comparison of EUILP Algorithms with UILP

Figure 6.1 shown the performance of data cleaning process. Both the systems followed the same method for data cleaning. After data cleaning process 522424 records were considered for further process.

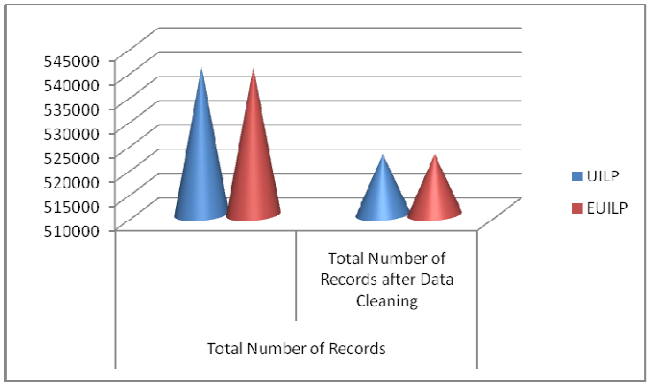

Figure 6.1. Performance comparison of Data Cleaning

Figure 6.2 shown the performance of similar user identification. User similarity identification is base for all the web page recommendation technique. User similarity is identified with respect to their browsing behavior. The EULIP has quality rate measure for classifying similar users. The figure has shown the result of EUILP.

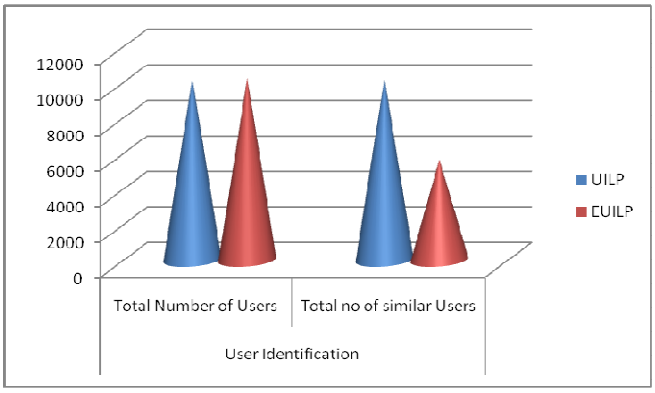

Figure 6.2. Performance comparison of Similar User Identification

VII. CONCLUSION

EUILP technique is proposed to perform data cleaning, user identification, session identification and path completion. The UILP algorithms extract fields such as IP address, user name, website address, session, frequency and quality rate. There are several methodologies and techniques are applied by the researchers to preprocess the web log files and make them consistent. From the performance analysis carried out between UILP and EUILP, it is concluded that EUILP technique outperforms to preprocess the weblogs and identify the similar users.

REFERENCES

[1] Chhavi, R, 2012, "A Study of Web Usage Mining Research Tool", International Journal of Advanced Networking and Applications, vol. 3, no. 6, pp. (1422-1429), 2012.

[2] Raju, GT & Sathyanarayana, PS, "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", International Journal of Computer Science and Network Security, vol. 8, no.1, pp.(179-186), 2008.

[3] Sanjay, BT & Sangram, ZG, "An Effective and Complete Preprocessing for Web Usage Mining", International Journal on Computer Science and Engineering, vol. 2, no. 3, pp. (848-851), 2010.

[4] Srivastava, J, Desikan, P & Kumar, V, "Web Mining - Concepts, Applications and Research Directions", AHPCRC Technical Report, pp. (51-70), 2003.

[5] Ramya, C & Shreedhara, KS, "Clustering of Web Users using ART1 NN based Clustering Approach with a Complete Preprocessing Methodology", International

Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 1, pp. (71-77), 2012.

[6] Sheetal, AR & Shailendra, J, "Efficient Preprocessing Technique using Web Log Mining", International Journal of Advancements in Research & Technology, vol. 1, no. 6, pp.(418-422), 2012.

[7] Malarvizhi, M & Sahaaya, AM, "Preprocessing of Educational Institution Web Log Data for Finding Frequent Patterns using Weighted Association Rule Mining Technique", European Journal of Scientific Research, vol. 74, no. 4, pp. (617-633), 2012.

[8] Cooley, R, "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", Ph.D. thesis, University of Minnesota, 2000.

[9] Chitraa, V & Antony, SD, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications, vol. 34, no. 9, pp. (78-83), 2011.

[10] Tasawar, H, Sohail, A & Nayyer, M, "Web Usage Mining: A Survey on Preprocessing of Web Log File", IEEE Conference on Information and Emerging Technologies, pp. (1-6), 2010.

[11] Mohd, HW, Mohd, NM, Hafizul, FH, Mohamad, F & Mohamad, M , "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", Proceedings of World Academy of Science, Engineering and Technology, vol. 36, pp. (970-977), 2010.

[12] Vijayashri, L & Madhuri, J, "Data Preprocessing in Web Usage Mining", Proceeding in International Conference on Artificial Intelligence and Embedded Systems, pp. (1-5), 2012.

[13] Suguna, R & Sharmila, D, "User Interest Level based Preprocessing Algorithms using Web Usage Mining", International Journal of Computer Science and Engineering (IJCSE), vol. 5, no. 9, pp. (815-822), 2013.

AUTHORS PROFILE

Dr. R. Suguna has completed her doctorate degree under Anna University, Chennai, Tamil Nadu, India. She has published many National and International Journals and presented papers in various conferences. She has 9 years of teaching experience. Her area of interest includes Data Mining, Software Engineering and Networking.

