Data Preprocessing in Web Usage Mining

Vijayashri Losarwar, Dr. Madhuri Joshi

***Abstract*--** It is well known that over 80% of the time required to carry out any real world data mining project is usually spent on data preprocessing. Data preprocessing lays the groundwork for data mining. Web mining is to discover and extract useful information from the world wide web. It involves the automatic discovery of patterns from one or more Web servers. This helps the organizations to determine the value of specific customers, cross marketing strategies across products and the effectiveness of promotional campaigns, etc. This paper discusses the importance of data preprocessing methods and various steps involved in getting the required content effectively. A complete preprocessing technique is being proposed to preprocess the web log for extraction of user patterns. Data cleaning algorithm removes the irrelevant entries from web log and filtering algorithm discards the uninterested attributes from log file. User and sessions are identified.

***Keywords*--** Preprocessing, Web usage, Web log.

I. INTRODUCTION

DURING the past few years the World Wide Web has become the biggest and most popular way of communication and information dissemination. It serves as a platform for exchanging various kinds of information. The volume of information available on the internet is increasing rapidly with the explosive growth of the World Wide Web and the advent of e-Commerce. While users are provided with more information and service options, it has become more difficult for them to find the "right" or "interesting" information, the problem commonly known as information overload.

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. A common taxonomy of web mining defines three main research lines: content mining, structure mining and usage mining. The distinction between those categories is not a clear cut, and very often approaches use combination of techniques from different categories.

Vijayashri Losarwar, Associate Professor, Department of Computer Science & Engineering, P.E.S. College of Engineering, Aurangabad, Maharashtra, India, v_a_losarwar@yahoo.com

Dr. Madhuri Joshi, Head of Computer Engineering Department, Govt. College of Engineering and Research, Awasari (Khurd), Dist - Pune, Maharashtra, India, madhuris.joshi@gmail.com

Web content mining is the process to discover useful information from the content of a web page. Basically, the Web content consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks.

Web Structure Mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Web Structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

Web Usage mining is the application of data mining techniques to discover usage patterns from web data. Data is usually collected from user's interaction with the web, e.g. web/proxy server logs, user queries, registration data. Usage mining tools discover and predict user behavior, in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service[1].

The whole procedure of using Web usage mining for Web recommendation consists of three steps, i.e. data collection and pre-processing, pattern mining (or knowledge discovery) as well as knowledge application. Fig 1.1 depicts the architecture of the web usage mining.

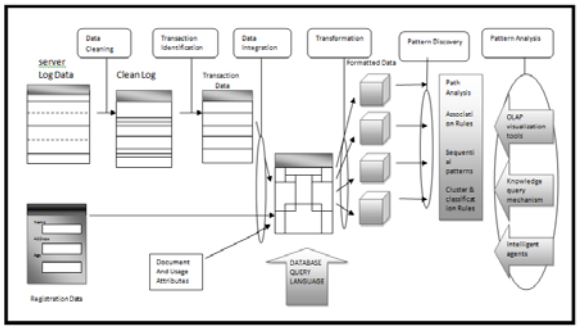

Fig.1 The Architecture of Web usage mining

For the data on the web, it has its own distinctive features compared to the data in conventional database management systems. Web data usually exhibits the following characteristics:

– The data on the Web is huge in amount. Currently, it is hard to estimate the exact data volume available on the Internet due to the exponential growth of Web data every day.

− The data on the Web is distributed and heterogeneous. Due to the essential property of the Web being an interconnection of various nodes over the Internet, Web data is usually distributed across a wide range of computers or servers, which are located at different places around the world.

− The data on the Web is unstructured. There are, so far, no rigid and uniform data structures or schemas which Web pages should strictly follow that are common requirements in conventional database management.

− The data on the Web is dynamic. The implicit and explicit structure of Web data is updated frequently.

The rest of this paper is organized as follows: In section 2, we present a review on existing preprocessing techniques. Section 3, explains the proposed methodology of preprocessing of WUM. Section 4 concludes the paper and ponders the future directions.

II. LITERATURE SURVEY

An implementation of data preprocessing system for web usage mining and the details of algorithm for path completion are presented in Yan Li's paper[2]. After user session identification, the missing pages in user access paths are appended by using the referrer-based method which is an effective solution to the problems introduced by using proxy servers and local caching. The reference length of pages in complete path is modified by considering the average reference length of auxiliary pages which is estimated in advance through the maximal forward references and the reference length algorithms. As verified by practical web access log, the algorithm path completion, proposed by Yan LI, efficiently appends the lost information and improves the reliability of access data for further web usage mining calculations.

In Web Usage Mining (WUM), web session clustering plays a key role to classify web visitors on the basis of user click history and similarity measure. Swarm based web session clustering helps in many ways to manage the web resources effectively such as web personalization, schema modification, website modification and web server performance. Tasawar Hussain, Dr. Sohail Asghar[3] proposed a framework for web session clustering at preprocessing level of web usage mining. The framework covers the data preprocessing steps to prepare the web log data and converts the categorical web log data into numerical data. A session vector was obtained, so that appropriate similarity and swarm optimization could be applied to cluster the web log data. Author says that the hierarchical cluster based approach enhances the existing web session techniques for more structured information about the user sessions.

Doru Tanasa[4], in his research brought two significant contributions for a WUM process. They proposed a complete methodology for preprocessing the Web logs and a divisive general methodology with three approaches (as well as associated concrete methods) for the discovery of sequential patterns with a low support.

Huiping Peng[5] used FP-growth algorithm for processing the web log records and obtained a set of frequent access patterns. Then using the combination of browse interestingness and site topology interestingness of association rules for web mining they discovered a new pattern to provide valuable data for the site construction.

In order to solve some existing problems in traditional data preprocessing technology for web log mining, an improved data preprocessing technology is used by the author ling Zheng[6]. The identification strategy based on the referred web page is adopted at the stage of user identification, which is more effective than the traditional one based on web site topology. At stage of Session Identification, the strategy based on fixed priori threshold combined with session reconstruction is introduced. First, the initial session set is developed by the method of fixed priori threshold, and then the initial session set is optimized by using session reconstruction. Experiments have proved that advanced data preprocessing technology can enhance the quality of data preprocessing results.

JIANG Chang-bin and Chen Li[7] brought about a Web log data preprocessing algorithm based on collaborative filtering. It can perform user session identification fast and flexibly even though statistic data are not enough and user history-visiting records are absence.

TABLE I ANALYSIS MATRIX FOR PREPROCESSING

Author	Source of Log File	Preprocessing Technique	Algorithm Applied
Yan LI, Boqin FENG and Qinjiao MAO[2]	English Study Web site Log File	Data Cleaning User Identification Session Identification Path Completion Transaction Identification	Maximal Forward References(MFR), Reference Length
Tasawar Hussain, Sohail Asghar, Nayyer Masood[3]	Server Log File	Data Cleaning Log File Filtering User Identification Session Identification	NA
Doru Tanasa and Brigitte Trousse[4]	Log Files from INRIA web sites	Data Fusion Data Cleaning Data Structuration Data Summarization	NA
Ling Zheng, Hui Gui and Feng Li [6]	IIS Server Log File	Data Cleaning User Identification Session Identification Path Completion	Based on referred web page and fixed priori threshold
JING Chang-bin and Chen Li[7]	Web server Log file	Data Preprocessing	Based on Collaborative Filtering
Fang Yuankang and Huang Zhiqiu [8]	Chizhou College Website	Data Filtering Session Identification	Frame page and Page Threshold
J. Vellingiri and S. Chenthur Pandian[9]	College Web Site	Data Cleaning User Identification Session Identification Path Completion Transaction Identification	MFR RL & Time Window

III PROPOSED PREPROCESSING METHODOLOGY

Web Personalization is the process of customizing the content and structure of a web site to the specific and individual needs of each user taking advantage of the user's navigational behavior.

The steps of the web personalization process include:

a) The collection of web data.
b) The modeling and categorization of these data (preprocessing phase)
c) The analysis of the collected data.
d) The determination of the actions that should be performed.[10]

Web pages belonging to a particular category have some similarity in their structure. This general structure of web pages can be deduced from the placement of links, text and images (including images and graphs). This information can be easily extracted from a HTML document. [11]

The main data source in the web usage mining and personalization process is the information residing on the web sites logs. Web logs record every visit to a page of the web server hosting it. The entries of a web log file consists of several fields which represent the date and the time of the request, the IP number of the visitor's computer(client), the URI request , the HTTP status code returned to the client, and so on. The log data collected at Web access or application servers reflects navigational behaviour knowledge of users in terms of access patterns. .

Physically, a page is a collection of Web items, generated statically or dynamically, contributing to the display of the results in response to a user action. A page set is a collection of whole pages within a site. User session is a sequence of Web pages clicked by a single user during a specific period. A user session is usually dominated by one specific navigational task, which is exhibited through a set of visited relevant pages that contribute greatly to the task conceptually. The navigational interest/preference on one particular page is represented by its significant weight value, which is dependent on user visiting duration or click number. The user sessions (or called usage data), which are mainly collected in the server logs, can be transformed into a processed data format for the purpose of analysis via a data preparing and cleaning process. In one word, usage data is a collection of user sessions, which is in the form of weight distribution over the page space.

3.1 Data Collection

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. When exploiting log information from Web servers, the major issue is the identification of users' sessions. Because the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side, this task is usually quite difficult. Probably, the best approach for tracking Web access consists of directly accessing the server application layer.

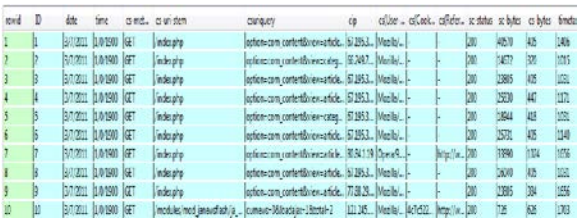

Fig. 2 Format of Web Log File from Server Side.

B. The Proxy Side

Many Internet Service Providers (lSPs) give to their customer proxy server services to improve navigation speed through caching. The main difference with the server side is that proxy servers collect data of groups of users accessing huge groups of web servers. Even in this case, session Identification is difficult and not all users' navigation paths can be identified. However, when there is no other caching between the proxy server and the clients, the identification of users' sessions is easier.

C. The Client Side

Access data can be tracked also on the client side by using JavaScript, Java applets, or even modified browsers. These techniques avoid the problems of 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 raise many issues concerning the privacy laws, which are quite strict [12]

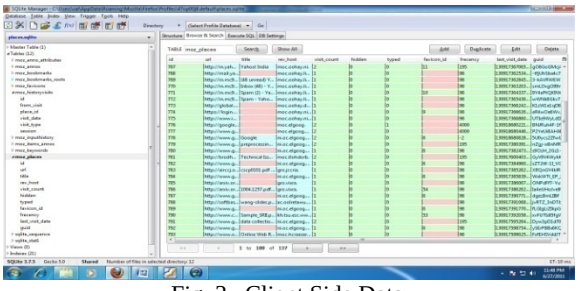

Fig. 3 Client Side Data

3.3 Data Preprocessing

Data preprocessing transforms data into a format that will be more easily, and efficiently processed for the purpose of the user. The main task of data preprocessing is to select standardized data from the original log files, prepared for user navigation pattern discovery algorithm[13]. The stage of data preprocessing includes data cleaning, user identification and session identification.

3.3.1 Data Cleaning

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2011-03-08 00:31:51
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem
cs-uri-query s-port cs-username c-ip cs-version cs(User-Agent) cs(Cookie)
cs(Referer) cs-host sc-status sc-substatus sc-win32-status sc-bytes cs-
bytes time-taken
2011-03-08 00:31:50 W3SVC23804 C36280-2214 74.52.2.99 GET /robots.txt -
80 - 77.88.29.246 HTTP/1.1
Mozilla/5.0+(compatible;+YandexBot/3.0;++http://yandex.com/bots) - -
www.abc.org 200 0 0 615 182 203
2011-03-08 00:38:24 W3SVC23804 C36280-2214 74.52.2.99 GET /Index.php - 80
- 114.80.93.57 HTTP/1.0 Sosospider+(+http://help.soso.com/webspider.htm)
- - www.abc.org 200 0 0 28541 161 1968
2011-03-08 00:45:18 W3SVC23804 C36280-2214 74.52.2.99 GET /md10-11.pdf -
80 - 66.249.71.201 HTTP/1.1
Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)
- - www.abc.org 304 0 0 274 287 46
```

Fig. 4 Web Log File in Text format.

The Web Log file is in text format then it is required to convert the file in database format and then clean the file. First, all the fields which are not required are removed and finally we will have the fields like date, time, client ip, URL access, Referrer and Browser used/ Access log files consist of large amounts of HTTP server information. Analyzing, this information is very slow and inefficient without an initial cleaning task. Every time a web browser downloads a HTML document on the internet the images are also downloaded and stored in the log file. This is because though a user does not explicitly request graphics that are on a web page, they are automatically downloaded due to HTML tags. The process of data cleaning is to remove irrelevant data. All log entries with file name suffixes such as gif, JPEG, jpeg, GIF, jpg, JPG can be eliminated since they are irrelevant [14].Web robot (WR) (also called spider or bot) is a software tool that periodically a web site to extract its content. Web robot automatically follows all the hyper links from web pages. Search engines such as Google periodically use WRs to gather all the pages from a web site in order to update their search indexes. Eliminating WR generated log entries simplifies the mining task. To identify web robot requests the data cleaning module removes the records containing “Robots.txt” in the requested resource name (URL).

The HTTP status code is then considered in the next process of cleaning by examining the status field of every record in the web access log, the records with status code over 299 or under 200 are removed because the records with status code between 200 and 299, gives successful response.

As we are interested in users request for information from server, the records with POST or HEAD method should be removed. Log files should have the records with GATE methods. By implementing above mentioned techniques original log files is cleaned. Near about 50-60% irrelevant records are removed.

3.3.2 User Identification

Once HTTP log files have been cleaned, next step in the data preprocessing is the identification of users. Different methods for this are 1)by converting IP address to domain name. 2) The web server randomly assigns an ID to web browser while it connects first time to the site. This is called cookies. The web browser sends same ID back to web server effectively telling the web site that a specific user has returned. Cookies help the website developer to easily identifying individual visitors which results in a greater understanding of how the site is used.

We can use special internet services such as finger services which provide username about the client accessing web server. Identd can be used for user identification. It is a protocol defined in RFC 1413. It allows us to identify connected users with unit TCP connection. The problem with Identd users should configure with this protocol. Another way for user detection is through usernames added in log file in field authuser. But this field can be empty (default value -) according to server/user command.

In this work, we assumed that each combination of IP address/Agent/Operating system as a single user. Also we added the login information to the log file to get username.

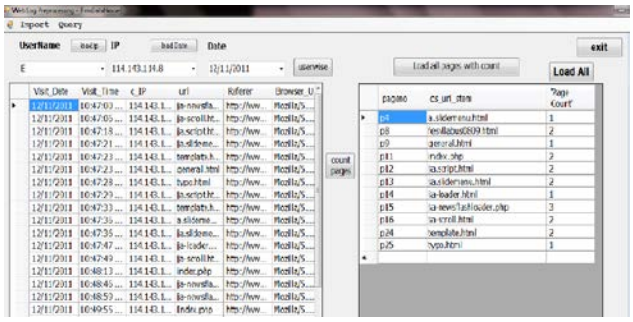

Fig. 5 Web Log file after cleaning and user identification.

For getting the username, we used the user login detail table which can be maintained by the server. Attributes in this table are Date, User ID, Client IP, In-time, Out-time.

If ((logtable.date = userlogindetail.date) and (logtable.time between userlogindetail.intime and userlogindetail.outtime)
Then logtable.userid <= userlogindetail.userid

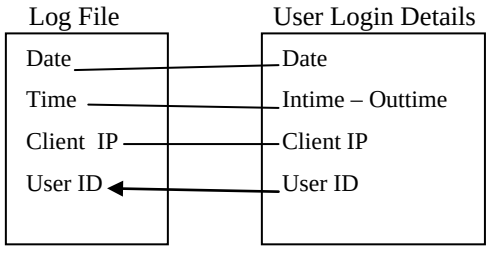

Fig. 6 User Identification

A unique number is given to all the pages visited by users (p1,p2,..). The count, how many times the particular page is visited by a user is calculated. Also the count, how many times the page is visited by various users is calculated.

3.3.2 Session Identification

To group the activities of a single user from the web log files is called a session. As long as user is connected to the

website, it is called the session of that particular user. Most of the time, 30 minutes time-out was taken as a default session time-out. A session is a set of page references from one source site during one logical period. Historically a session would be identified by a user logging into a computer, performing work and then logging off. The login and logoff represent the logical start and end of the session.

Let L be a log. A session S is an ordered list of pages accessed by a user that is S=<((p1,t1),(p2,t2),…..,(pn,tn))[15], where there is a user ui € U such that {<ui,p1,t1>, <ui,p2,t2>, …. <ui,pn,tn>} is part of L. Thus we write a session S as <p1,p2,…,pn>.

We used following rules to identify users sessions.

1.If there is a new user there is new session.

2.In one user session, if the referrer page is null, there is a new session.

3.If the time between page requests exceeds a certain limit (30 minutes) It is assumed that user is starting a new session.

The reference page is estimated by access time of this page and the next one i.e. the reference length of an accessed page equals the difference between the access time of the next and the present page. If this time is few seconds than that page can be considered as an auxiliary page and otherwise that page can be considered as a content page.

IV. CONCLUSIONS

An important task in any data mining application is the creation of a suitable target data set to which data mining and statistical algorithms can be applied. This is particularly important in Web usage mining due to the characteristics of click stream data and its relationship to other related data collected from multiple sources and across multiple channels. The data preparation process is often the most time consuming and computationally intensive step in the Web usage mining process, and often requires the use of special algorithms and heuristics, not commonly employed in other domains.

TABLE II : SUMMARY OF PREPROCESSED DATA

Records in original Web Log File	1,000
Records in cleaned log file	420
Attributes in original Log file	22
Attributes in Filtered Log file	6
No. of Users Identified	26
No. of different pages	27

REFERENCES

[1] Magdalini Eirinaki and Michalis Vazirgiannis , "Web Mining for Web Personalization" *Communications of the ACM, vol. 3, No. 1*, pp.2-21, Feb. 2003.

[2] Yan LI, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique In Web Usage Mining", *IEEE International Symposium On Computer Science and Computational Technology*, pp. 554-559, 2008.

[3] Tasawar Hussain, Dr. Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence ", *6th International Conference on Emerging Technologies (ICET) IEEE*, pp. 21-26, 2010.

[4] Doru Tanasa and Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining ", *Published by the IEEE Computer Society*, pp. 59-65, March/April 2004.

[5] Huiping Peng, "Discovery of Interesting Association Rules Based On Web Usage Mining", *IEEE Coference*, pp.272-275, 2010.

[6] Ling Zheng, Hui Gui and Feng Li, " Optimized Data Preprocessing Technology For Web Log Mining", *IEEE International Conference On Computer Design and Applications(ICCDA)*, pp. VI-19-VI-21,2010.

[7] JING Chang-bin and Chen Li, " Web Log Data Preprocessing Based On Collaborative Filtering ", *IEEE 2nd International Workshop On Education Technology and Computer Science*, pp.118-121, 2010.

[8] Fang Yuankang and Huang Zhiqiu, " A Session Identification Algorithm Based on Frame Page and Pagethreshould", *IEEE Conference*, pp.645-647, 2010.

[9] J. Vellingiri and S. Chenthur Pandian, "A Novel Technique for Web Log Mining with Better Data Cleaning and Transaction Identification", *Journal of Computer Science*, pp. 683-689, 2011.

[10] Rana Forsati, Mohammad Reza Meybodi and Afsaneh Rahbar, "An Efficient Algorithm for Web Recommendation Systems", *IEEE Conference*, pp. 579-586, 2009.

[11] D.Vasumathi, D.Vasumathi and K.Suresh, "Effective Web Personalization Using Clustering", *IEEE IAMA*, 2009.

[12] Zhiguo Zhu and Liping Kong, " A Design For Architecture Model Of Web Access Patterns Mining System ",*IEEE International Conference on Computer and Communication Technologies In Agriculture Engineering*, pp.288-292, 2010.

[13] Chaoyang Xiang, Shenghui He and Lei Chen, "A Studying System Based On Web Mining", *IEEE International Symposium On Intelligent Ubiquitous Computing and Education*, pp.433-435, 2009.

[14] R.M.Suresh and Padmajavalli, " An Overview Of Data Preprocessing In Data and Web Usage Mining", *IEEE Conference*, pp.193-198, 2006.

[15]Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education.