Data Preprocessing and Easy Access Retrieval of Data through Data Ware House

Suneetha K.R, Dr. R. Krishnamoorthi

***Abstract*-The World Wide Web (WWW) provides a simple yet effective media for users to search, browse, and retrieve information in the Web. Web Usage Mining is the application of data mining techniques to click stream data in order to extract usage patterns. These patterns are analyzed to determine user's behavior which is an important and challenging research topic in web usage mining. In order to determine which pages of the web site were accessed and how various web pages were reached, requires examining the raw data recorded in the log files created by the web server. An important fundamental task for Web log mining is data preprocessing. This paper presents algorithm for data cleaning, user identification and session identification. The main new approach of this paper is to access the usage pattern of preprocessed data using snow flake schema for easy retrieval.**

***Index terms*-Web Usage Mining, Data Preprocessing, User Identification, Session Identification, Data Warehouse Schema.**

I INTRODUCTION

The World Wide Web has become one of the most important media to store, share and distribute information. At present, Google is indexing more than 8 billion Web pages [1]. The rapid expansion of the Web has provided a great opportunity to study user and system behavior by exploring Web access logs. The WWW is serving as a huge widely distributed global information service center for technical information, news, advertisement, e-commerce and other information service. This makes information retrieval process difficult. Most users may not have good knowledge of the structure of the information network, and may easily get bored by taking many access hops and losing impatience when waiting for the information. These challenges will have been solved efficiently by Web mining, which is the application of data mining technologies. Web mining [2] that discovers and extracts interesting knowledge/patterns from Web is classified into three types as Web Structure Mining that focuses on hyperlink structure, Web Contents Mining that focuses on page contents as well as Web Usage Mining that focuses on Web logs. In this paper, we are concerned about Web Usage Mining (WUM), which also named Web log mining. The process of WUM [3] includes three phases shown in Fig. 1: data preprocessing, pattern discovery, and pattern analysis. Data Preprocessing consists of data cleaning, data integration and transformation, and data reduction. Pattern discovery deals with extracting knowledge from preprocessed data. Some of the techniques used in Pattern discovery are Association rules, Classification, Clustering etc. Pattern Analysis filters out uninteresting rules or patterns from the set found in the pattern discovery phase.

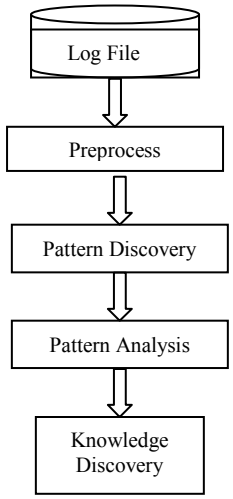

Fig.1: Web log mining structure

A. Web Log Information

A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. Most logs use the format of the common log format. Each entry in the log file consists of a sequence of fields relating to a single HTTP transaction with the various fields separated by a space The following is a fragment from the server logs[4], [5] for loganalyzer.net.

66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html HTTP/1.1" 200 11179 "-" "Mozilla/5.0(compatible;Googlebot/2.1;+http://www.google.com/bot.html)".

This reflects the information as follows,

- Remote IP address or domain name: An IP address is a 32-bit host address defined by the Internet Protocol.

- Authuser: Username and password if the server requires user authentication.
- Entering and exiting date and time.
- Modes of request: GET,POST or HEAD method of CGI(Common Gateway Interface).
- Status: The HTTP status code returned to the client, e.g., 200 is "ok" and 404 is "not found".
- Bytes: The content-length of the document transferred.
- Remote log and agent log.
- Remote URL.
- "Request:" The request line exactly as it came from the client.
- Requested URL.

NASA Server log file is used for the purpose of analyzation. The remainder of this paper is organized as follows. Section 2, presents the related work, data preprocessing algorithms are discussed in detail in the section 3, experimental results and performance analysis are reported in section 4, the storage of data using snow flake schema is shown in the section 5, and paper concludes in section 6.

II RELATED WORK

In this section, we first introduce some related work in data preprocessing and then we focus on need of data warehouse to store the relevant data from the log files created by the web server.

In the recent years, there has been much research on Web usage mining [9], [10], [11], [12]. However, data preprocessing has received far less attention than it deserves. Methods for user identification, session zing, and path completion, are presented in [13]. In another work [14] the authors compared time-based and referrer-based heuristics for visit reconstruction. They found out that a heuristic's appropriateness depends on the design of the Web site (i.e. whether the site is frame-based or frame-free) and on the length of the visits (the referrer-based heuristic performs better for shorter visits). In [15] Marquardt et al. addressed the application of WUM in the e-learning area with a focus on the preprocessing phase. In this context, they redefined the notion of visit from the e-learning point of view. Moreover, in the same paper, the authors have presented several data preparation techniques to identify Web users, i.e., the path completion and the use of site topology. To identify the user sessions, a fixed time period, say thirty minutes [6], is used to be the threshold between two sessions. Zaïane et.al[16] have applied various traditional data mining techniques to Internet log files in order to find different types of patterns, which can be harnessed as electronic commerce decision support knowledge. The pre-processed data is then loaded into a data warehouse which has an *n*-dimensional web log cube as basis. From this cube, various standard OLAP techniques are applied, such as drilldown, roll-up, slicing, and dicing. The authors [17] employ the data warehousing technology as a preprocessing step to apply piecewise regression as a predictive data mining technique that fits a data model which will be used for prediction

III DATA PREPROCESSING

The first phase of this paper discusses on data preprocessing algorithms used to clean raw log data. The purpose of data preprocessing is to extract useful data from raw web log and then transform these data in to the form necessary for pattern discovery. Due to large amount of irrelevant information in the web log, the original log cannot be directly used in the web log mining procedure, hence in data preprocessing phase, raw Web logs need to be cleaned, analyzed and converted for further step. The data recorded in server logs, such as the user IP address, browser, viewing time, etc, are available to identify users and sessions. However, because some page views may be cached by the user browser or by a proxy server, we should know that the data collected by server logs are not entirely reliable. This problem can be partly solved by using some other kinds of usage information such as cookies.

By cleaning the data, we can create the database according to our application which includes the information about user identification, session identification, path completion etc. Because of the proxy servers and Web browser cache existing, to get an accurate picture of the web-site access is difficult. Web browsers store pages that have been visited and if the same page is requested, the Web browser will directly displays the page rather than sending another request to the Web server, which makes the user access stream incomplete. By using the same proxy server, different users leave the same IP address in the server log, which makes the user identification rather difficult. [6] presented the solution to these problems by using Cookies or Remote Agents.

In our work the various processes involved in data preprocessing is shown in Fig. 2. This module includes the identification of users and user's sessions, which are used as basic building blocks for pattern discovery.

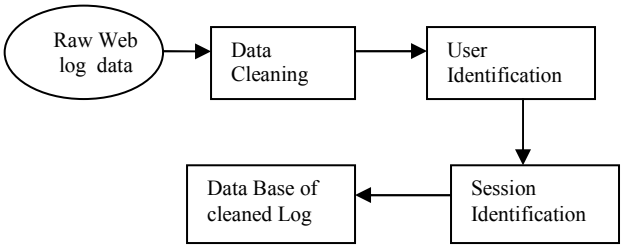

Fig.2: Preprocessing process

Algorithms are proposed for each block in the following sections. To validate the efficiency of preprocessing methodology, several experiments have been conducted on the log files of NASA web site.

A. Data cleaning

Data cleaning means eliminate the irrelevant information from the original Web log file. Usually, this process removes requests concerning non-analyzed resources such as images, multimedia files, and page style files. For example, requests for graphical page content (*.jpg & *.gif images) and requests for any other file which might be included into a web page or even navigation sessions performed by robots and web spiders. By filtering out useless data, we can reduce the log file size to use less storage space and to facilitate upcoming tasks. For example, by filtering out image requests, the size of Web server log files reduced to less than 50% of their original size. Thus, data cleaning includes the elimination of irrelevant entries like:

- Requests for image files associated with requests for particular pages; an user's request to view a particular page often results in several log entries because that page includes other graphics, while we are only interested in what the users explicitly request, which are usually text files.
- Entries with unsuccessful HTTP status codes; HTTP status codes are used to indicate the success or failure of a requested event, and we only consider successful entries with codes between 200 and 299.
- Entries with request methods except GET and POST.

The proposed algorithm for data cleaning is given below.

```
Data Cleaning Algorithm
Input: Web Server Log File
Output: Log Database
Step1: Read LogRecord from Web Server Log File
Step2: If((LogRecord.url-stem(gif,jpegjpg,cssjs)) AND
            (LogRecord.mehod='GET') AND
              (LogRecord.Sc-status<>(301,404,500)AND
          (LogRecord.Useragent<>Crawler,Spider,Robot))
          then Insert LogRecord in to LogDatabase.
          End of If condition.
  Step 3: Repeat the above two steps until eof (Web Server
        Log File)
  Step 4: Sop the process.
```

The outcome of this algorithm is the LogDatabase consisting relevant set of records with the entries as user name, userIP address, Date, Time, and URLdetails etc. By filtering out the irrelevant entries the size of web log files reduces to more than 50% of its original size.

B. User identification

User identification means identifying each user accessing Web site, whose goal is to mine every user's access characteristic. This paper works with the assumptions, that each user has unique IP address and each IP address represents one user. But user identification is greatly complicated by the existence of local caches, corporate firewalls and proxy servers. In order to over come these problems some rules are proposed for user identification i) If there is a new IP address, then there is a new user ii) If the IP address is same, but the operating system or browsing software are different, a reasonable assumption is that each different agent type for an IP address represents a different user. These rules are used in our proposed algorithm to identify users mentioned below.

```
 User Identification Algorithm
Input: Log Database
Output: Unique Users Database

Step1: Initialize
       IPList=0; UsersList=0; BrowserList=0;
       OSList=0; No-of-users=0;
Step2: Read Record from LogDatabase
Step3: If Record.IP address in not in IPList
        then add new Record.IPaddress in to IPList
             add Record.Browser in to BrowserList
           add Record.OS in to OSList
           increment count of No-of-users
           insert new user in to UserList.
     Else
          If Record.IP address is present in IPList OR
             Record.Browser not in BrowserList OR
             Record.OS not in OSList
          then
              increment count of No-of-users
             insert as new user in to UserList.
          End of If
        End of If
Step 5: Repeat the above steps 2 to 3
         until eof (Log Database )
Step 6: Stop the process.
```

The outcome of this algorithm is Unique Users Database gives information about total number of individual users, users IPaddress, user agent and browser used.

C. Session Identification

After user identification, the pages accessed by each user must be divided into individual session, which is known as session identification. The goal of session identification is to find each user's access pattern and frequently accessed path. The simplest method is using a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout. Once a site log has been analyzed and usage statistics obtained, a timeout that is appropriate for the specific Web site can be fed back in to the session identification algorithm.

The rule about session identification is as follows (1) If there is a new user, then there is a new session. (2) In one user session, if the refer page is null, we can draw a conclusion that there is a new session. (3) If the time between page requests exceeds a certain limit, it is assumed

that the user is starting a new session. The following is the algorithm designed by using the above rules.

Session Identification Algorithm
Input: Log Database
Output: Session Database
Step1: Initialize
SessionList=0
UserList=0
No-of-Sessions=0
Step2: Read LogRecord from Log Database
Step3: If (LogRecord.Refer='-') OR
LogRecord.time-taken>30min OR
LogRecord.UserID not in UserList)

then
Increment No-of-Sessions
Get Url address of corresponding Session and
Insert in to SessionList
End of If
Step4: Repeat the above steps 2 and 3 till eof (Log Database)
Step5: End of process.

This algorithm gives user IPaddress along with page accesses performed by individual users during a visit in a Web site. The Session Database gives details about total number of sessions, Session key with start time and end time details.

IV EXPERIMENTAL RESULTS

In this work the NASA server log file of size 195MB is analyzed. Several experiments have been conducted on log files. Through these experiments it is shown that the proposed preprocessing methodology reduces the size of the initial log files significantly by eliminating unnecessary requests and also increases quality through better structuring of the Web data. This is shown in the Table I. It is observed that the size of the log file is reduced to 73% of the original size. The Table II shows data analysation results of seven days which includes overall entries, total number of IPaddresses, total number of unique users and total number of sessions. Fig. 3 indicates total number of unique users identified, and number of sessions created by the users is shown in Fig. 4.

Table I: Results of Preprocessed data

Web Server Log File	NASA Jul-95
Duration	1 - 7days
Original Size	69.84MB
Reduced Size After Preprocessing	19.23MB
Percentage in Reduction	72.47

Table II: Day wise data after preprocessing step.

Day	No. of Entries	No. of IP Address	No of Unique Users	No. of Sessions
1	64567	7597	576	652
2	60264	6630	613	693
3	89565	19193	1766	1989
4	65536	9340	868	982
5	65535	17638	1039	1865
6	68342	24706	2033	2301
7	87233	22657	980	1109

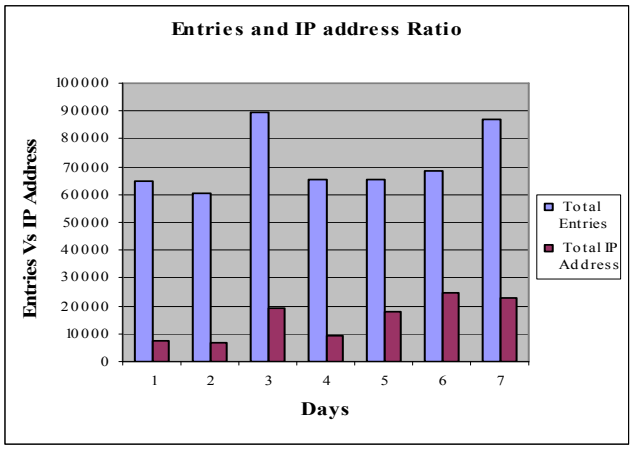

Fig.3: Total number unique users identified from overall entries

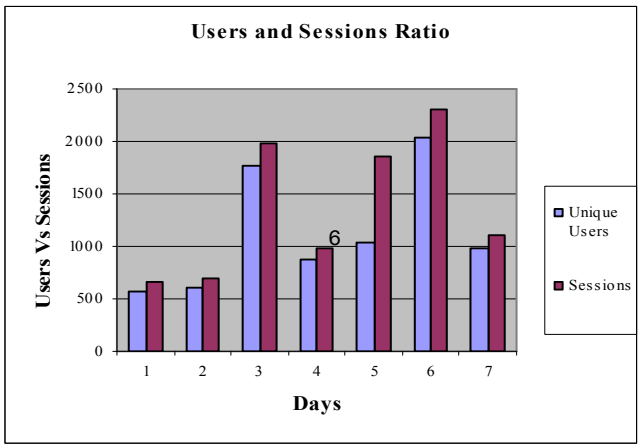

Fig.4: Total number of Sessions created by unique users

The main focus or main idea of this paper is construction of snow flake schema of data ware house using user and session details for easy access retrieval is discussed in next section.

V DATA WAREHOUSE CONSTRUCTION

The results of preprocessing the web server logs will be stored using snow flake schema of data warehouse to facilitate easy retrieval and analysis. Figure 3 illustrates the structure of the data warehouse that will be used to facilitate the mining of web usage data.

The design of the data warehouse [7] plays an important role in finding the useful patterns during the mining process. Data warehouse will contain the main table and dimension tables. The main table normally stores the dynamic data, whereas the dimension table stores the relatively static data, which facilitates querying, calculating

and selecting useful information. Fact tables act as connecting element in a data model representing keys and summarization information**.** The snowflake schema is used to store preprocessed data as it supports multiple granularities and also dimensional hierarchy is explicitly represented by normalizing dimension tables.

Fig. 5 depicts a web log snowflake schema that is centered on a fact table which is typical for analyses of Web log to identify user behavior. It contains key fields (user Key, session Key, Date Key, Access Key) as well as some statistical summarization information. The user information is easily accessed through these key fields and by use of dimension table.

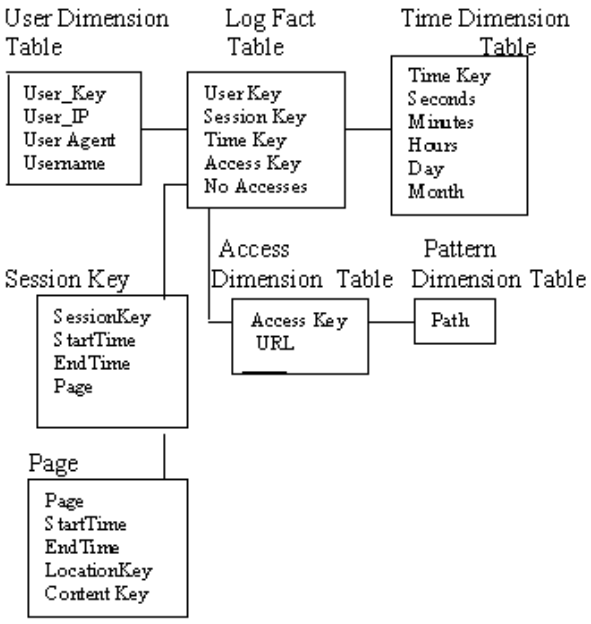

Fig.5: Snow Flake Schema for Preprocessed Data

Once the data warehouse has been created and populated, various statistical and data mining techniques will be used in order to identify any web usage patterns that exist. An existing application that may be able to assist with this pattern discovery phase is 123LogAnalyzer [8]. These patterns will then be analyzed, interpreted and used to determine how well the web site is being used.

VI CONCLUSIONS

Data preprocessing is an important basic work in web mining. Unnecessary and junk requests were cleaned from the log file. Only valid and important requests were regrouped in to user sessions and finally the results were saved using snowflake schema for easy retrieval and analysis. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Furthermore, many other data mining functions, such as classification, prediction, association, and clustering, can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction. This paper concludes that the proposed algorithms for data preprocessing have been proved efficiency and validity. The new approach of storing preprocessed data using snow flake schema is an increasingly important platform for data analysis for on-line analytical processing which will provide an effective platform for data mining.

REFERENCES

[1] Google Website. http://www.google.com.

[2] R. Kosala, H. Blockeel. "Web Mining Research: A Survey," In SIGKDD Explorations, ACM press, 2(1): 2000, pp.1-15.

[3] R. Srikant, R. Agrawal. "Mining sequential patterns: Generalizations and performance improvements," In 5th International Conference Extending Database Technology, Avignon, France, March1996, pp. 13-17.

.

[4] W.W.W Consortium the CommonLogFileformat http://www.w3.org/Daemon/User/Config/ Logging.html#common-Log file-format, (1995)

[5] W3C Extended Log File Format, Available at http://www.w3.org/TR/WD-logfile.html (1996).

[6] R. Cooley, B. Mobasher and J. Srivastava. "Data Preparation for Mining World Wide Web Browsing Patterns," In Journal of Knowledge and Information Systems, vol. 1, no. 1, 1999. pp. 5-32.

[7] Jiawei Han and M. Kamber. "Data Mining: Concepts and Techniques," In Morgan Kaufmann publishers, 2001

[8] ZY COMPUTING-2003 ,123 Log Analyzer. San Jose USA. Available at http://www.123loganalyzer.com

[9] B. Mobasher, H. Dai, T. Luo and M. Nakagawa. " Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," In proceedings of Data Mining and Knowledge Discovery,2002, pp 61-82.

[10] R. Kosala, H. and Blockeel. "Web mining research: a survey," In proceedings of special interest group on knowledge discovery & data mining, SIGKDD:2000 , ACM 2 (1), pp.1–15.

[11] R. Kohavi and R. Parekh. "Ten supplementary analyses to improve e-commerce web sites," In Proceedings of the Fifth WEBKDD workshop, (2003).

[12] B. Mobasher, R. Cooley and J. Srivastava. "Creating Adaptive Web Sites through usage based clustering of URLs," In proceedings of Knowledge and Data Engineering Exchange, 1999, Volume 1, Issue1, 1999, pp.19-25.

[13] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan. "Web usage mining: discovery and applications of usage patterns from web data," In SIGKDD Explorations, 2002, pp. 12–23.

[14] B. Berendt, B. Mobasher, M. Nakagawa and M. Spiliopoulou. "The Impact of Site Structure and User Environment on Session reconstruction in Web Usage Analysis," In Proceedings of the Forth WebKDD 2002 Workshop, At the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2002), Edmonton, Alberta, Canada,pp.1-13.

[15] C. Marquardt, K. Becker and D. Ruiz. "A Pre-Processing Tool for Web Usage Mining in the Distance Education Domain," In Proceedings of the International Database Engineering and Applications, IDEAS 2004, pp. 78-87.

[16] Zaïane, O.R. Xin and M. Han. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," In Proceedings of Advances in Digital Libraries Conference(1998), pp. 19-29.

[17] Kamal A. ElDahshan, Hany Maher Said Lala Kamal "Data Warehouse based Statistical Mining," In ICGST-AIML Journal, ISSN: 1687-4846, Volume 9, Issue I, February (2009), pp.41-48

Dr. R Krishnamoorthi is currently working as Professor and Head of Information Technology Department, Bharathidasan Institute of Technology, Anna University, Tiruchirapplli. He has received his Ph.D. in Image Processing from Indian Institute of Technology, Karagpur in the year 1995. His research interest includes Software Engineering, Image Compression, Image Encryption and Authentication, Pattern Recognition and Knowledge Discovery & Management. rkrish_25@hotmai..com

Suneetha K. R is a research student at Anna University, Tiruchirappalli, Tamil Nadu. She has received her M.Tech. Degree in Computer Science & Engg. from B.M.S.C.E under Visvesvaraya Technological University, Karnataka. She is currently working in Computer Science & Engineering Department, Bangalore Institute of Technology, V.V.Puram,Bangalore-04. Her research interest includes Knowledge Discovery from Web Usage Data, Classification, and Intelligent Agents. krs_mangalore@hotmail.com