ISSN:0975-9646

Identifying the User Access Pattern in Web Log Data

NeetuAnand
(Research Scholar, Lingayas University)
Asst. Professor(Deptt. of Computer Sc.)
Maharaja Surajmal Institute
Delhi

Prof(Dr.)SabaHilal
Professor &Head
Deptt.of Computer Application
Lingayas University
Faridabad

***Abstract*- Web Usage Mining (WUM) is the research area combining both data mining and WWW.The users' accesses to Web sites are stored in Web server logs. Theobjective behind WUM is to analyse web log files for extracting usage patterns. The web log files contain the raw data that need to be pre-processed first for discovering knowledge. Mining techniques are then applied for clustering users, to identify frequent item set, for classification of users and association rule mining.This paper present an approach to identify user access pattern from web log data .In the first phase ,the server raw log data is pre-processed. In second phase analysis is performed to identify access pattern of users.**

I. INTRODUCTION

The Web is a massive, varied, dynamic and mostly unstructured data repository, which provides incredible amount of data information, and also increases the complexity of how to deal with the information from the different perceptions of users, Web service providers, business analysts etc. [1].Web mining is divided into three areas: Web content mining (WCM), Web structure mining (WSM), and Webusage mining (WUM). Web content mining is a process of picking up information from texts, images, audio, video, or structured records such as lists and tables and scripts. Web structure mining is a process of discovering structure information from linkages of web pages (inter page structure/hyper link structure).The web usage or log mining is defined as the process of extracting interesting patterns from the log data.The log data is consists of textual data and is represented in standard format (common log format or extended log format).The main goal of web usage mining is to capture, model and examine the web logdata in such a way that it inevitablydetermines the usage behaviour of web user [10].

II. APPLICATIONS AREAS OF WUM

The main application areas of WUM are shown in figure1.

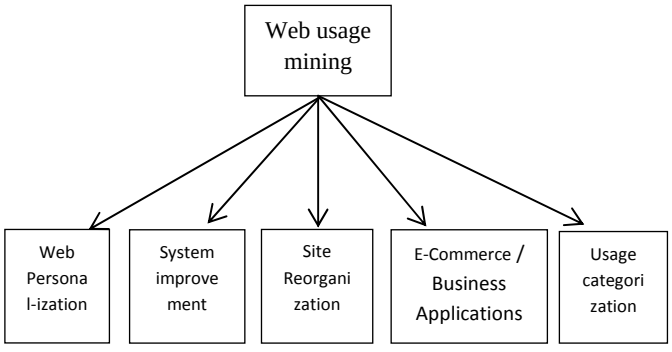


Fig 1: Major Application area of Web usage mining[11]

Web personalization:Web server logsare used to cluster web users having similar interests.It is alsodefined as the task of adapting services and information available on a website to the needs and the expectations of a target user, the *active user*; the personalization task might benefit from the knowledge gained from an analysis of the user's navigational behaviour combined with other features which are peculiar to a Web context, namely its structure and content.

System Improvement:Web caching, load balancing, network transmission or data distribution are the common application areas of web mining for improving the system performance.

Site reorganization: The link structure and content structure of any website are two significant factors for any web site. The recent development in web mining technologies go towards shorter navigation sequences, for that purpose the ease to access target page in any web domain needs to be increased The reorganization task can be performed with respect to the frequent patterns extracted. Web usage data also gives information about the design of any web site with respect to users behaviours. Web site owner can redesign these pages and observe the behaviour of users on these pages.

E-Commerce/Business intelligence: The use of WUM allow different organizations to understand its customers and built customer profiles on the basis of customer's habits, their needs and interests so that companies can increase their profit by "cross selling" or selling items correlated to their demands. Hence, knowledge about the customers' preferences and needs make the CRMmore effective .The main goal of companies are retaining their old customer and attract new customers to beat their competitor's.

Usage Categorization: In this process the information storedin Web server logs is processed by applying various data miningtechniques so at to (a) extracting statistical information and discovering interesting usage patterns, (b) clustering the users intogroups according to their navigational behaviour, and (c) determine possible links between Web pages and user groups. Others data mining techniques are also used for finding useful patterns.

III. MINING TECHNIQUES

Web usage mining is the "Applying data mining techniques to web data repositories to extract patterns ".Data mining techniques that are commonly used includes association rules, sequential pattern , clustering, and classification.

Association rules are used to find the relationship between attributes from the item set. In web usage mining item set is set of pages .Rules are applied to discern pages which are often looked together In order to reveal associations between groups of users with specific interests. The resulted Knowledge can be used in marketing and business or as

guidelines to web designers for reorganizing Websites. [3] Used association rules to decide the next likely web page requests based on significant statistical correlations.

Sequential pattern is used to discover sequential navigational pattern for user session . Using this approach, useful users' trends can be discovered, and forecast concerning visit patterns can be made. [6] Used sequential patterns in web usage data for predicting the possible next move in browsing sessions for web personalization.

Clustering is a technique to group together items that have similar features. In Web usage domain, there are two clustering groups, user clusters and page clusters. Page clustering generates the group of pages that are considered to be related according to user view. In user clustering the goal is to group users which have same browsing patterns. Such understanding can be used in business to perform market segmentation and Web site personalization.[7] created a model by applying clustering algorithm, and then the model is adjusted by statistical approach based on the change of behaviour of users or data domain of website periodically.[12] proposed to integrate Markov model based sequential pattern mining with clustering. .[8] experimented for many of the tuneable parameters, such as the time delta involved in sessionizing logs, confidence and support for associations, initializing of the medoids in clustering.

Classification is a method that maps a data item into one of several predefined classes. In Webusages mining the users are in different classes according to their profiles.

IV. WEB LOG FILE

The main sources of data in WUM are server - side, proxy - side and client- side.Server log files include access logs, referrer logs and agent logs. Different servers have different log formats .The typical format of web log have the following fields for the portion of log data[15]:

---

ppp931.on.bellglobal.com--[26/Apr/2000:00:16:12 -0400]"GET /download/windows/asctab31.zip HTTP/1.0"200 1540096 http://www.htmlgoodies.com/ downloads/ freeware/webdevelopment/15.html" Mozilla/4.7 [en] C-SYMPA (Win95; U)"

- *Remotehost*: Domain name or IP address of host
- *Username etc*: "- -" **:** Username etc. Only relevant when accessing password-protected content.
- *Timestamp*: "[26/Apr/2000:00:16:12 -0400]" - Time stamp of the visit as seen by the web server.
- *Mode of request*: "GET /download/windows/asctab31.zip HTTP/1.0". The requestis made. In this case "GET" request (i.e. "show me the page") for the file "/download/windows/asctab31.zip" byusing "HTTP/1.0" protocol. Other possible requests are POST or HEAD.
- *Statuscode*: "200".The status code. "200" means success. If not found then "404" is the status code.
- *Bytes transferred*: "1540096" numbers of bytes are transferred.
- *Referrer URL*:http://www.htmlgoodies.com/downloads/freeware/webdevelopment/15.htm". The referring page. Not all user agents supply this information.
- *User Agent*: "Mozilla/4.7 [en]C-SYMPA (Win95; U)". The User Agent isthe software the visitor used to access this site. It's usually a browser, but it could equally be a web robot, a link checker, an FTP client or an offline browser.. In this case "Mozilla/4.7" probably means Netscape 4.7, "[en]" probably implies it's an English version, "Win 95" indicates Windows 95 etc.

V. PROPOSED FRAMEWORK OF WUM MODEL

To obtain the user profile in web usage mining the following steps are to be undertaken

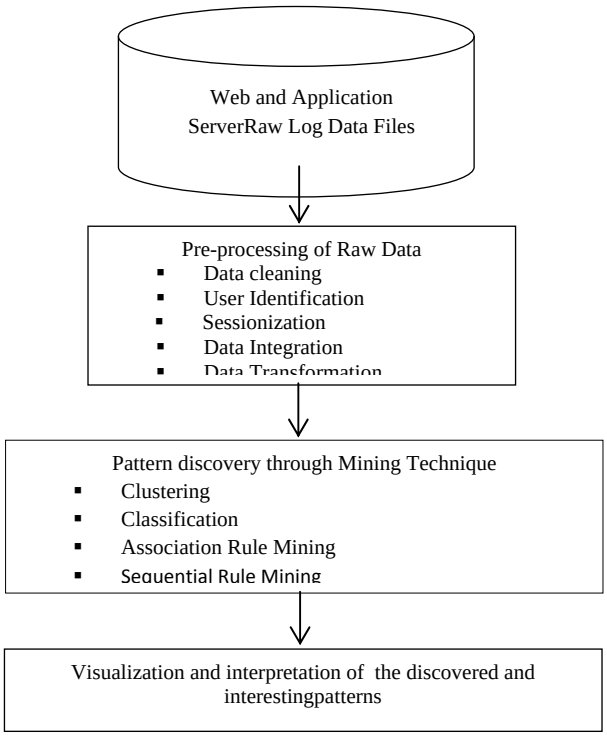


Fig2:Framework of WUM

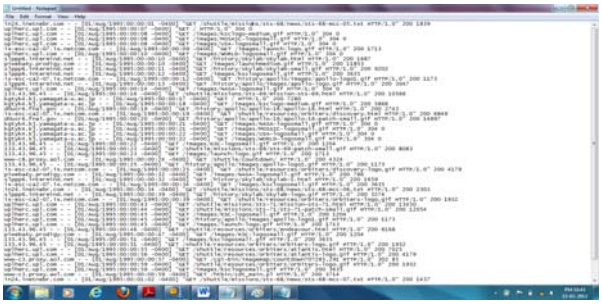
VI. EXPERIMENTS

A. *Pre-processing Log data*

In the pre-processing phase, sample server log file, was processed to transform the raw data into structured information. The purpose of data cleaning is to eliminate irrelevantitems.

B. *Log Data File*

The raw data for mining purpose is collected from NASA website .The records of ten days are considered for further analysis. It contains approximately 4,00000 records in Common log file format. The sample of raw web log data is as shown:



C.Data cleaning

Log data is stored in database for further processing of data by means of queries and program .Data file obtained was very huge and it takes almost 80% of total time to mine the data. In data cleaning process,the unwanted information is removed from the log database. The data cleaning takes the following steps:

Step1: Removal of the entries having image files, graphic or multimedia files. The records which are accessing file with extension gif,jpg, jpeg etc. are to be removed .After performing this step around 1,23785 records left.

Step 2: The removal of entries with failed status code.The various status codes for HTTP 1.1 are as shown [14]:

In this step the entries having status code of 200 will be retained, rest are removed.

Step3: Removal of records with bytes transferred field zero. The records having entries zero in the byte transferred field indicates that the requested page is not opened, and is to be removed. After performing the above two steps the number of records left are 80307.

Step4: Removal of records with less than three times visits the websites. The IP address which access the website less than three times are not of any use for further analysis and hence to be removed from log file.After performing this step the numbers of active records left are 4000.

C. User Identification

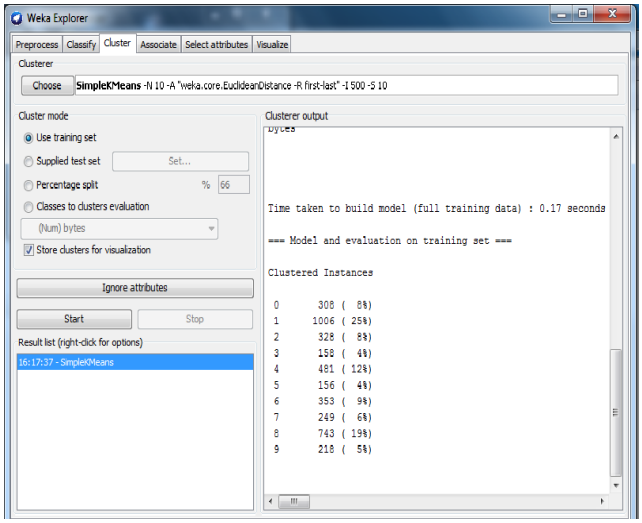
It is very important step to further refine the data for mining purpose. The unique user is identified by using IP address,User agent and Referred URL field.The users with the same IP address field are considered to be same.If IP address is same but user agent is different thaneach different agent represent the different user. These rules are used for making the program to identify users.After grouping users,486 different usersare identified.

D. Sessionization

The goal of session identification is to divide the page accesses of each user into individual sessions. Sessions are constructed using *time based heuristics*. 30 minutes is usually considered as the default timeout of page viewing time .The set of pages visited by a specific user at a specific time is called page viewing time.After30 minutes a new session begins. The other method depends on the difference between two timestamps. If it exceeds 10minutes then the second entry is assumed as a new session. Time heuristic based methods are not accurate because users may involvein some additional activities after opening the website andissues liketraffic problem in line, loading time of web page, content size of web pages are not considered. The second method is*Navigation based Heuristics:* If a web page is not linked with prior visited pages in a session, then it is considered as a new session.

VII. PATTERN DISCOVERY PHASE

The goal of this stage is to find unknown relationships in the data.The technique applied to the data is Cluster analysis.A cluster is a collection of records which are adjacent to eachother and relatively distant from other clusters. Clustering algorithms *partition* the database into a set of mutually exclusive clusters. With this technique, the most frequently requested pages are extracted from the database and forms a cluster.The k-means clustering inWEkA software is used for performing the clustering .The resultant cluster of URL are:



VIII. CONCLUSION

Pre-processing the web log data is a significant and prerequisitephase in Web mining. It removesirrelevant items and identifies users and sessionsalong with the browsing information. The output ofthis phase results in the creation of a user sessionfile. The different patterns can be then discoveredpatterns by applying the mining techniques. Thediscovered patterns can then be used for variousWeb

Status Code	Reason Phrase	Status Code	Reason Phrase
100	Continue	404	Not Found
101	Switching Protocols	405	Method Not Allowed
200	OK	406	Not Acceptable
201	Created	407	Proxy Authentication Required
202	Accepted	408	Request Timeout
203	Non-Authoritative Information	409	Conflict
204	No Content	410	Gone
205	Reset Content	411	Length Required
206	Partial Content	412	Precondition Failed
300	Multiple Choices	413	Request Entity Too Large
301	Moved Permanently	414	Request-URI Too Long
302	Found	415	Unsupported Media Type
303	See Other	416	Requested Range Not Satisfiable
304	Not Modified	417	Expectation Failed
305	Use Proxy	500	Internal Server Error
306	(unused)	501	Not Implemented
307	Temporary Redirect	502	Bad Gateway
400	Bad Request	503	Service Unavailable
401	Unauthorized	504	Gateway Timeout
402	Payment Required	505	HTTP Version Not Supported
403	Forbidden		

usage applications such as user profiling,usage categorization, site improvement,business intelligence and recommendations.

REFERENCES

[1] Yan Wang, Web Mining and Knowledge Discovery of Usage Patterns, CS 748T Project (Part I), February 2000.

[2]Sumathi, Padmaja valli, Santhanam,An Overview Of Preprocessing Of Web Log Files For Web Usage Mining,Journal of Theoretical and Applied Information Technology,31st December 2011. Vol. 34 No.2.

[3]Qiang Yang,Building Association-Rule Based Sequential Classifiers for Web-document Prediction, Data Mining and Knowledge Discovery, 8, 253–273, 2004.

[4] María J. Martín-Bautista, María-Amparo Vila, Víctor H. Escobar-Jeria, obtaining user profiles via web usage mining, Iadis european conference data mining,2008.

[5] K. R. Suneetha, Dr. R. KrishnamoorthiIdentifying User Behavior by Analyzing Web Server Access Log File , IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[6] Sang T.T. Nguyen, Efficient Web Usage Mining Process for SequentialPatterns, Proceedings of iiwas,2009.

[7] Saeed R. Aghabozorgi, Teh Ying Wah, Recommender Systems: Incremental Clustering on WebLog Data,ICIS, November 24-26**,** 2009 Seoul, Korea.

[8] Karuna P Joshi, Anupam Joshi, Yelena Yesha, RaghuKrishnapuram, Warehousing and Mining Web Logs,Workshop on Web Information and Data Management 1999.

[9] J. Vellingiri and S. Chenthur Pandian, A Novel Technique for Web Log mining with Better DataCleaning and Transaction Identification, Journal of Computer Science7 (5): 683-689, 2011.

[10] V.V.R. Maheswara Rao and Dr. V. Valli Kumari, An Enhanced Pre-Processing Research Framework For Web Log Data Using A Learning Algorithm, netcom 2010,CSCP 01, pp. 01-15, 2011.

[11]Robert Walker Cooley, Web usage mining: Discovery and application of interesting patterns from web data, 2000.

[12] A. Anitha, A New Web Usage Mining Approach for Next Page AccessPrediction, International Journal of Computer Applications (0975 – 8887) Volume 8– No.11, October 2010.

[13]Liping Sun,Xiuzhen Zhang, Efficient Frequent Pattern Mining on Web Log Data,2011.

[14]http://en.wikipedia.org/wiki/List_of_HTTP_status_codes.

[15]*www*.jafsoft.com/searchengines/log_sample.html.