IOSR Journal of Environmental Science, Toxicology and Food Technology (IOSR-JESTFT)
e-ISSN: 2319-2402,p- ISSN: 2319-2399.
PP 59-65
www.iosrjournals.org

Enhanced Pre-Processing Using Definiteuser Identification

Pavithra B[1],Dr.Niranjanamurthy M[2]
[1]*Assistant Professor, Department Of MCA, Jain University, Bangalore, India*
[2]*Assistant Professor, Department Of Computer Applications, MSRIT, Bangalore, India*

***Abstract**: The incredible growth of World Wide Web has made the present users interact with the website on day to day basis. Huge amount of data is been generated which is very much significant for the company to know the behaviour of the user.These issues is been addressed in webmining among which pre-processing is an important phase. In this paper we have discussed an enhanced pre-processing method which involves data cleaning based on definite userip address and agent values. In traditional user identification the heuristic rules upon site structure had flaws in finding the relationship between pages which inturn reduces the efficiency of user identity. So to solve this issue we have proposed a technique for user identification based on IP address and session time which rules out the site structure topology flaws and gives a reliable and efficient process in identifying the user. Our main objective is to simplify the algorithm and give out the reliable unique users informationso as to know who the user is and his access pattern. Which is generally used in fraud detection, unusual access of defended data, terrorismand user behaviour to better the access performance and overall design for farther accessing. Experimental analysis has shown that the proposed enhanced pre-processing technique has intensified the essence of pre-processing and the method is tested on the real data extracted from the server log repositories of a university on the generic content..*
***Keywords:** Pre-processing, Session Identification, Server log, Session Time, Customization*.

I. Introduction

Web mining makes is an introversion application of data mining techniques which fundamentally retrieves and analyse the processed data for information from the web pages. Web usage mining contributes a supportive system for designing the web site, lending personalized server and also helps the business persons to take a reasonable judgement. To impress the users, web mining uses techniques of data mining along with artificial intelligence to follow up the user interest and there patterns. There are three different domains that together club to give out the meaning of web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is series of action involved in extracting knowledge from their contents involved in web documents,text mining using the indexing or agent concept.Web structure mining customises information fetched from web organizations and the references between the client and users on the web.Web usage miningrevolves around tracking the relevant Patterns in web server logs.Web usage mining contributes a supportive system for designing the web site, providing personalized server and also helps the business persons to take a reasonable judgement. Data mining applications is ben vastly speeded all over to extract the relevant information's.

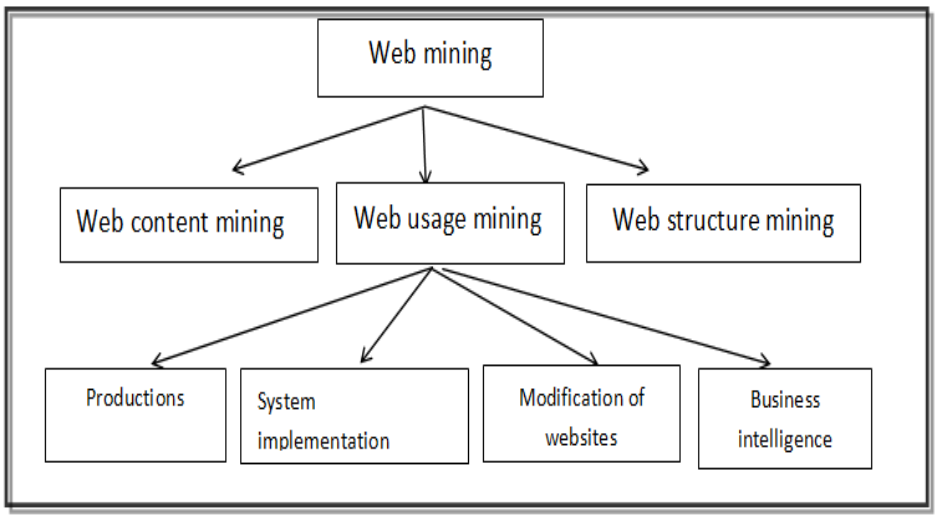

Figure 1: Web Mining Categories

Web Personalizationis defined as any process that personalises the interest of the web users based on his browsing data.Based on the behaviour of the user it will integrate the contents and based on their interest reference will be showcased to them to interact with the web pages[1].are the elements of Web Personalization.The latest survey done on netcraft states that the usage of the internet is growing rapidly, as on July 2015 survey states that there are around 989,000,456 websites that is been multiplied into larger number when compared to some previous years.The size of the web users have raised up to 678.76% in 2015 from 2013 as per the survey done statistically and stated by the internet world status.As and when the web users opens the web pages and uses it , the required information will be recorded based on some format into a file which is said to be as log file and in conserves as the usage of web builds up the log files also gets piled up. The given below states the different stages involved in mining the web log or cleaning the web log files

- **Data Collection:** Web Log files which will distributed in different servers will be segregated[4].
- **Pre-Processing:** The segregated log file contains so many raw data with all information mixed up with relevant and irrelevant data which is referred as noisy data. Removing noisy data is an important task which controls the efficiency.Pre-processing is a carried out with sequence of process like data filtering, identifying the user, identifying the session, path completion and identifying the transaction.
- **Pattern Discovery :** Many data mining techniques is been applied on the filtered data to obtain the pattern
- **Pattern Analysis:** Based on the pattern interested values is fetched and uninterested data will be removed out using queries.

Session identification means to find out all the web pages visited by a particular user in a particular time period. Generally to identify the session many methods are available to find out the exact time spent by the user on the particular web page.This paper presents an enhanced techniquethat can be used in pre-processing to identify the user to perform Web Mining.

Generally visualization methods are used for result analysis, the proposed algorithm and methods is implemented for bettering the performance of a server which recommends the modifications of the websites and draw the interest of the user. Usage mining tools are used to predict the user behaviour in preference to improve the website, attract more and more users or to give personalized attempts. The applications will retrieve users oriented dataand discover the user behavioural pattern. Identification of the distinct user which will be helpful for online products, enhancing the performance and efficiency of network service towards the other user, which betters the network server quality and performance.

II. Web Mining

Web data contains compilation of data fromweb server, proxy, browser and also user track records.The knowledge used on the server side can be split up into three categories depending on source collection. The client side data it involves the complete description of active services and which client is holding that particular service[5].The proxy side will be imparted somewhere in between the client and user.

Web server logs are generally lapped up with plain text which results again depends on the log data, thereexists difference among the various servers software's used, so far three formats of server logs are used to trace the log data.

- W3C Extended log format
- Microsoft IIS log format
- NCSA Common log format

The above mentioned log patterns are in ASCII code format.The W3C Extended and NCSA f patterns maintains the login data and records it in a format of four digit, the Microsoft IIS format is stored in two digit format in older version and later extended it to four digit format ,it is also used to provide reversecongeniality when compared to older versions[4].A networks log file holds on the request made to the network in parametric order which is a standard one. The most commonly and popularly used format is NCSA Common Log Format in short it is termed as CLF format.This log files are originated from the web servers to serve the impact of the requests made by the web servers from a particular web site. A common log file format is as follows :

<ip_addr><base_url><date><method><file><Protoc-ol><code><bytes><referrer><user_agent>

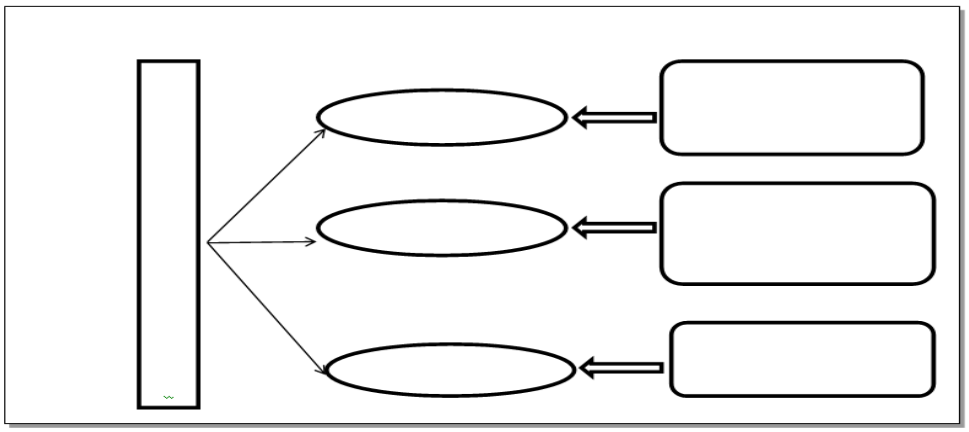

Figure 2: Sources of the data log

2.1 Server Log

Whenever a user requests or in need of a particular page on web, it will be recorded into a file called as server log file.These special files are not general fetched by the users instead administrative persons or the server head can fetch these files[8].Server log files are considered to be most reliable and standard source of information to know the users interest but it has many flaws with certain concepts like privacy and redundancy server pages are cached in order to attained fast response time forexample if the requesting page is available in cache then no record will be imparted in server log file for that page.so increasing the cache becomes a issue here and it can be solved using path completion methods[10].

2.2 Client/Browser Log

By merging java applet or java scripts with respect to the website, web log data is been collected from client machine which can be used to solve the issues that depends on server logs and session reconstruction[12][13].

2.3 Proxy Server Log

Servers which reacts as an intermediary between users request and web servers are referred as proxy servers. In order to improve the navigation speed and privacy proxy server caching is been done, it is lagging behind with some problems in user identification to predict the log source contents access[14].

III. Data Pre-Processing

Log files contains lot of information which will be irrelevant for mining so ruling out such irrelevant data is an important process. Data pre-processing is done involving three steps like cleaning, user identification, session identification and finally path completion The first process is to eradicate the irrelevant http request which will be identified by the code status in the log files of the cleaning phase[16]. The second process is to filter the graphicscontents like audio, video and images which is easily identified with their file extensions, as they just occupy more space in the log file and nothing to do with navigational behaviour. The final process involves removal of indexed pages generally referred as web spider using any of the robot detection techniques[17][18].

Then after cleaning the robots file heuristic techniques is applied based on the following situations :

- Same URL request is been repeated again and again by the same host
- Interval of the time is too short or limited between two requests
- Request from the single host referrer is by chance empty

3.1 User Identification

User identificationis the most interlaced task because of emerging local and proxy servers, buffering systems, security and privacy[18, 19]. User identification involves in tracking out the various users by considering the log files as the source file. User identification is a process of identifying who access the websites what pages are they accessing.The main aim of user session is to help session identification so as to split up the accessed pages of every user at a particular period into separate private periods. The issues in identifying the distinct user is the proxy servers which records the same users as different users depending on the session and also one more issues is that different users will be using the system of same IP address but it will be registered in the log file as a single user even though different users are using it. So a referrer based technique is used to solve this issues which is been addressed in this paper.There are plentiful methods to find out the unique users based on IP address, topologies, cookies and authenticity which is been discussed below.

3.1.1 User identification using IP Address:

IP Address are the unique address of the devices running on the network, which will be registered into the log files when the user of a particular system hits the page, but if proxy servers exists then it registers the same IP address when multiple users are requesting for same web pages. Apart from that caching also creates a problem to recognize the unique user suppose if the user request for the page that was previously accessed then those pages will be fetched from cache without even making an entry into the log file. So these issues can be solved by rejecting the proxy server entries by considering their domain name[20].

3.1.2 User identification by data authentication:

This method works by making use of log name and referrer name by asking password each and every time whenever the user requests for the pages[16]. This method is not much popular because users generally avoid these method due to many steps

3.1.3User Identification by cookies:

This methods uses cookies to identify the users .cookies are small information generated when user requests for a web page which is sent by server to client system, these information will be stored in form of a text file on the client system with browser details[14].Again there are some issues in this method where cookies doesn't work, or cookies will not be supported by some browsers, cookies will be disabled by browsers or cookies will be deleted by the web servers or the users.

3.1.4User Identification by client information:

To look at the agent field in the log files which has the details about the OS and versioning, if request is made for a web page then it stores the information about the OS and browsers version so that might be helpful in sorting out the proxy servers, which is been proceeded by heuristic methods.The main aim of user session is to help session identification so as to split up the accessed pages of every user.

31.5 User Identification by site topology:

This methoduses topology of the site and each and every time it considers as a new users again the solution is not feasible

IV. Related work

User identification is a major research topic to identify the unique users apart from all those hurdles of cookies, proxy servers and browsers topology.Most partially the users are identified with their IP address which can hold good for few time period may be for an hour or minutes or in certain scenario when data mining doesn't require precise information about the unique user. In other ways users are identified using heuristic techniques[9]. The identification method are categorised as two classes called as proactive and reactive. Reactive class aims at identifying the users individually from the raw file when once the entry is registered in case of proactive classes aims at identifying the user during the page request[11] [12]. Proactive involves simple user authentication along the form, cookies which is clubbed with the client browser which is requesting In Reactive class the users will be differentiated based on the navigational behaviour, time stamping and involves heuristic methods based on some assumptions[5].

4.1 Problem in User Identification

User's identification role is to find out who is using the web page and sites.There are lot of users who will be accessing the websites, generally tracking out the users will be based on the information that they register when they login, at the same time there are many users who will not register and also there are users who access the websites through agent were the information is not correct and many other issues like firewall existence, browser different mechanism and cookies. All these problems makes the user identification process very complicated. So cookies can be used to solve major problem but then again considering users privacy, many users delete the cookies and some don't use it. So in [9] they have come up with heuristic method to address the issue in which testing is done in a scenario that when a user is in need of a page, that will not be easily fetched by a hyperlink the heuristics algorithm whichis been referred in [9] imagines some other user is using with the same IP address.In Ref [10] they have proposed a method called navigational patterns to track the user automatically. Cookies are small information generated when user requests for a web page which is sent by server to client system, these information will be stored in form of a text file on the client system with browser details

All the above mentioned algorithm are not accurate because they only consider few factors which imparts user identification. The success of web sites not only depends on the hit and page visited but on various other factors. All these factors leads to the issues in finding out discrete user session, user identification construction, collecting important web pages for analysis. So many web log mining tool is been developed to solve these issues

4.2 Proposed Method for User Identification

Considering the above mentioned issues, we have presented an enhanced algorithm which leads to an efficient way of identifying the user based on IP address, site topology, OS, Referrer Page and session timing, which will be useful in fraud detection, unusual access of defended data, terrorism and user behaviour to better the access performance and overall design of websites. Experiments have proved that the proposed enhanced pre-processing technique has intensified the quality of pre-processing results. Cookies are small information generated when user requests for a web page which is sent by server to client system, these information will be stored in form of a text file on the client system with browser details. These details are collected from the log repositories of the servers files including the proxy files, by applying the clara's algorithm with setting k maximum and k minimum values so that the filtered values lies between these ranges. After applying the algorithm data is filtered out by eliminating unwanted and irrelevant data by ignoring the tuple values method, when once the data is filtered out that is noise is removed from the files and it is set according to the format required by our proposed algorithm, in case of proxy servers if data is found missing then it is considered as noisy data and if no data is missing then it is given some order and considered to be a new user.

The proposed method gives out the result considering the path traversed by the user and User_IP on specific network.

Prerequisites: collectserver log file which has to be noise free data and also all the set of records from the server logs

RecordSetrs= {rs1,rs2,rs3……rsn},when x is greater then 0

- **First:** Consider Database ofAUsers with x Records as input
- **Second:** Definite Our User identification
- **Third:** AUser = a<url, ip_addr, agent, method, OSstatus,sessionid,time_stamp>
- **Fourth:** AUser=<rs1,rs2,rs3…rsn> where x!=0,a=0
- **Fifth:** where i is greater then x
- **Sixth:** clean database from log AUser
- **Seventh:** check if rs(a).User Id does not belongs to user then our User identification base then it is considered as a new user and copy the User Id in our User identification .
- **Eight:** stop the if loop
- **Ninth:** (a= a + 1);
- **Tenth:** looping stops

End of algorithm

Table: Result analysis

No of Records in raw Log	57256
No of records after cleaning	17125
No of sessions	8792
No of Users	8238
No of Unique users	5147

Data, Result And Analysis Of Experiment

To authenticate the reliability and performance of the proposed algorithm, we have considered server log file of the library of some university. The main data is been collected form the log files repositories were in all the information about the users who has used the library system in a particular website is collected first and numbered in ordered were in it contains the information about the ip address of the system, users session time, length of the session ,pages visited and duration of the time spent on each page and length of the users session is been measured using Clara's algorithm in which each records will be considered as one single unit and all the values will be filtered out separately by using the existing simulators, then filtered data is taken into consideration. IF proxy servers exists then proxy data will be eliminated if repetition of values is seen ,otherwise it is taken into consideration as a fresh entry, the same procure is repeated for the data of about six months, every month the data is collected and filtered out to get the users first and then apply the algorithm thus segregated the unique users and the page hits visited by them. The proposed algorithm filters the data as the unique users and generall users by considering the proxy servers log file also. The size of the file was approximately 175MB, which is performed on 2.8GH_Z ,510 MB of main memory with SQL server and JDK every month and came up with the following conclusion which is illustrated in the table given below.

The number of page request is decreased from 843354 to 144367 after data cleaning

Table: Result of the proposed method

Month	Unique Users	Number of users	Pages	Hits	Bandwidth
July	648	930	1280	5339	52.58MB
August	653	907	1234	5340	53.59MB
September	672	927	1412	5283	47.85MB
October	623	870	1226	5038	1.79GB
November	515	709	1148	4121	5.10GB

The above mentioned algorithm can be used in finding out the users, which also gives out the information about the users browsing interest and the pages visited frequently by them so in turn that helps out e-commerce oriented retailers to track the users interested behaviour to provide advertisement and related information about their products and also helps in improving their websites designs according the users taste. It also has its application in fields of fraud detection, unusual access of defended data, terrorism and user behaviour to better the access performance and overall design for farther accessing. It is also useful to give out an enhanced way of pre-processing the data for easier access. The limitations of the proposed algorithm is that it slows down in accessing huge amount of data and also the data has to treated by setting the limitation on the number of files and algorithm lags behind if the data is not filtered out well , that is the data should be completely noise free data. To elaborate this research the algorithm has to be improvised so as to take huge data and there should not be any limitations set on the file size .

V. Conclusion

In this research paper we have proposed an enhanced algorithm of pre-processing technique for user identification in data mining. We have implemented two of the pre-processing techniques combined into a single step to find out the user with reference to session time. The proposed algorithm is very efficient when compared to other available techniques and when it comes to accuracy the result is comparatively more accurate. Based on the above proposed algorithm we can personalize the website and also helps in improving the performance of the website. It has to be completed with certain methodologies such as pattern discovery and pattern analysis which results out the faster identification of the user.

References

[1]. S BamshadMobasher Robert Cooley, JaideepSrivastava, "Automatic Personalization Based on web Usage Mining", Communications of the ACM, New York, Volume 43,Issue 8, 20014

[2]. Zahid Ansari, M.F. Azeem, Waseem Ahmed, A.VinayaBabu, "Quantitative Evaluation of Performance and Validity indices for Clustering the Web Navigational Sessions", proceedings of the IEEE /ACM International Conference on Data Mining & Knowledge Management Process, Vol. 3, No. 2, pp. 1-21, 2015.

[3]. Dr.AntonySelvadossThanamani and V.Chitraa "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2013

[4]. Alberto Cano, Amelia Zafra, Sebastián Ventura "Weighted data gravitation classification for standard and imbalanced data", Cybernetics, IEEE Transactions on Cybernatics.2014

[5]. SG Mathews,MAGongora and AA Hopgood "Web useage mining with evolutionary extraction of temporal fuzzy association rules", IEEE Transactions on Knowledge and Data Engineering, 17(6):734–749, 2013

[6]. Abdul-Aziz and Rashid Al-Azmi, "Data, Text and Web Mining for Business Intelligence: A Survey", proceedings of the IEEE /ACM International Conference on Data Mining & Knowledge Management Process, Vol. 3, No. 2, pp. 1-21, 2014.

[7]. Baoyao Zhou, Siu Cheung Hui and AlvisC.M.Fong, "An Effective Approach for Periodic Web Personalization", Proceedings of the IEEE/ACM International Conference on Web Intelligence. IEEE,2014.

[8]. J Guo, V Keselj& Q Gao, "Integrating web content clustering into web log association rule mining", In Proc. Springer, CCIS, Volume 3501, pp. 182-193, 2013.

[9]. Robert F.Dell ,Pablo E.Roman, and Juan D.Velasquez, "Web User Session Reconstruction Using Integer Programming" , IEEE/ACM International Conference on Web Intelligence and Intelligent Agent,2008

[10]. G. Arumugam and S. Suguna, 20015, "Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs", International Conference on Network and Service Security, IEEE, 1- 6.

[11]. JaideepSrivastava, Robert Cooleyz, MukundDeshpande& Pang-Ning Tan, "Web Usage Mining Discovery and Applications of Usage Patterns from Web Data", ACM-SIGKDD, Jan-2000.

[12]. K. Sudheer Reddy, M. Kantha Reddy & V. Sitaramulu, "An Effective Data preprocessing Method for Web Usage Mining", Feb-2013, IEEE

[13]. BrijeshBakariya, Krishna K. Mohbey and G.S. Thakur, "An Inclusive Survey on Data Preprocessing Methods Used in Web Usage Mining", Springer-2011.

[14]. Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya&Jayesh N. Rathod, "Web Usgae Mining : A Review on Process, Methods and Techniques", Feb-2013, IEEE

[15]. Liu Kewen, "Analysis of Preprocessing Methods for Web Usage Data", IEEE 2012
[16]. Sheetal A. Raiyani, Shailendra Jain and Ashwin G. Raiyani, "Advanced Preprocessing using Distinct User Identification in web log usage data", ISSN : 2278 – 1021, IJARCCE, Vol. 1, Issue 6, August 2012
[17]. V. Sujatha and Punithavalli, "Improved User Navigation Pattern Prediction Technique From Web Log Data", ELSEVIER-2012
[18]. HongzhouSha, TingwenLiub, Peng Qin, Yong Sun and Qingyun Liu, "EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining", ELSEVIER-2013
[19]. MofrehHogo, MiroslavSnorek&PawanLingras, "Temporal Web Usage Mining", IEEE 2003.
[20]. Sourabh Jain, Susheel Jain &Anurag Jain, "An Assessment of Fuzzy Temporal Association Rule Mining", IJAIEM-2013
[21]. Jiawei Han and MichelineKamber, "Data Mining: Concepts and Techniques", Second Edition, ELSEVIER Inc