USER INTEREST LEVEL BASED PREPROCESSING ALGORITHMS USING WEB USAGE MINING

R. Suguna
Assistant Professor
Department of Computer Science and Engineering
Arunai College of Engineering
Thiruvannamalai – 606 603
sugunarajasekar@yahoo.co.in

Dr. D. Sharmila
Professor and Head
Department of Electronics and Instrumentation Engineering
Bannari Amman Institute of Technology
Sathyamangalam – 638 401
sharmiramesh@rediffmail.com

***Abstract* - Web logs take an important role to know about user behavior. Several pattern mining techniques were developed to understand the user behavior. A specific kind of preprocessing technique improves the quality and accuracy of the pattern mining algorithms. The existing algorithms have done the preprocessing activities for reducing the size of the log file and to identify the number of unique users and sessions. In order to identify the user interest level and group similar kind of users, User Interest Level Preprocessing (UILP) algorithm is newly proposed. This paper discusses the basics of web log preprocessing, existing preprocessing techniques, the proposed UILP algorithm and performance of the proposed UILP algorithm with existing algorithms to identify user interest level.**

Keyword – Web logs; Preprocessing; Data Cleaning; User Identification; Session Identification; Path Completion.

1. INTRODUCTION

World Wide Web (WWW) is a massive collection of information. Information is arranged in proper hierarchy in the form of websites. Website contains collection of web pages, which are accessed via hyperlinks. Today, internet is the major source for all kinds of users to obtain their needful. Not even every day, even every minute, millions of users visiting the websites for getting their information. Whenever the user interacts with the website, the interaction details are automatically recorded in web server in the form of web logs [1]. Website analyst use the web log information for variety of purpose such as identifying and understanding the user's behavior and expectation, improving the business process, website customization, web personalization and recommendation.

It is mandatory for website analyst to understand user behavior and interest for variety of reasons. Web logs take an important role to know about user behavior. Several pattern mining techniques were developed to understand the user behavior. But, there is no special preprocessing techniques are developed to identify the user interest level and understand their browsing behavior. A specific kind of preprocessing technique improves the quality and accuracy of the pattern mining algorithms. The existing algorithms have done the preprocessing activities for reducing the size of the log file and to identify the number of unique users and sessions [2]. They were not developed for specific pattern mining algorithms and for particular kind of applications. An Intelligent System Web Usage Preprocessor (ISWUP) is developed by V.V.R. Maheswara Rao and Dr. V. Valli Kumari (2011) which categorize human and search engine accesses before applying the preprocessing techniques. An effective and enhanced preprocessing technique is proposed by K. Sudheer Reddy, et. al., (2012) and D. Nithya and P. Sumathi (2012) which cleans the irrelevant information from web logs, identify users and sessions and perform path completion. In order to identify the user interest level and group similar kind of users, User Interest Level Preprocessing (UILP) algorithm is newly proposed. This paper discusses the basics of web log preprocessing, existing preprocessing techniques, the proposed UILP algorithm and performance of the proposed UILP algorithm with existing algorithms to identify user interest level.

2. BASICS OF PREPROCESSING

Web logs are maintained in the web servers in the form of plain text files which contains the details about user name, Internet Protocol (IP) address, date, time, number of bytes transferred, access request and referrer [1]. Web logs are list of page references by the users or click stream data which contains inconsistent and incomplete data. So, it is difficult to use the web logs directly for pattern mining algorithms to extract the features. Preprocessing techniques are necessary to make them consistent and complete.

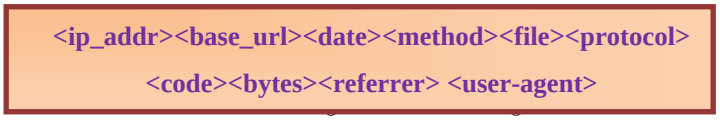
Web logs are maintained in the following places as line of text [6]:

- Web Server – The log files stored in web server provides more complete and accurate information about the user's interaction with the website. The World Wide Web Consortium (W3C) maintains a standard format for web server log files. It records the details about client IP address, request date and time, page requested, Hyper Text Transfer Protocol (HTTP) code, bytes transferred, user agent, and referrer.
- Proxy Server – The proxy servers act as a mediator between the browser and the web server. It takes the HTTP request from the user and sends them to the web server. Web proxy is a caching mechanism which lies between client browsers and web servers.
- Browser – Web Logs for the particular user are stored in the browser machine. The browsers are programmed and scripting languages are used to collect client side data. This implementation of client side data collection requires user support to activate the scripting languages or use the programmed browser.

Commonly, there are three types of web log formats[6] They are:

- W3C Extended Log File (ELF) Format - W3C log format is a default log file format on Internet Information Server (IIS) server. Field are separated by space, time is recorded as Greenwich Mean Time (GMT). This format is personalized by the administrators to add or remove fields depending on the information needed to record. The date format for W3C is YYYY-MM-DD.
- National Center for Supercomputing Application (NCSA) Common Log File (CLF) Format – It records the information pertaining to user name, date, time, request type, HTTP status code and number of bytes. NCSA is fixed format, not customized by administrators.
- Microsoft IIS Log File – Web logs are maintained in American Standard Code for Information Interchange (ASCII) format which is not customized. Fields are separated by comma. Time is recorded in local time. It records more information than NCSA format.

All the log file formats share the common information. The common format for the web log files are CLF and ELF. The ELF Format additionally has two fields at the end which are the referrer Universal Resource Locator (URL) and User Agent. Figure 2.1 depicts the format of ELF.



Where

- ip_addr - user IP address
- user Id - user name (' -', if not specified)
- base_url – requested resource path
- date - access time and date
- method - HTTP request type
- file - html file requested by the user
- protocol - protocol used for transmission
- code - status code
- bytes - number of bytes transferred.
- referrer - previously visited site by the user
- user agent – type and version of the browser

Figure 2.2 shows example web log file. In this example log file, 72.30.252.91 is an ip address of the user, user name is not specified and mentioned as '-', 18/Jun/2006 is date of request, 12:28:33+0000 is time of

request, GET is a request type, robots.txt is URL address, HTTP/1.0 is a protocol, 200 is status code, 52 is number of bytes transferred, Mozilla/5.0 is browser type and version.

Figure 2.2. Example Web Log File

Preprocessing [7] is an important activity in web usage mining and treated as a key to success. Preprocessing techniques eliminate the unwanted information from the web logs and facilitate the effective pattern mining. It consists of Data collection, Data cleaning, User identification, Session identification and Path completion. It is the process of transforming the unformatted interactions details with the website into a suitable form for further processing.

3. EXISTING PREPROCESSING TECHNIQUES

The existing techniques perform the preprocessing operations by using the following techniques:

Data Collection

Data collection [8] is the initial step in web log preprocessing. The user interaction details with the website are recorded in the form of web logs in three different places, (i) Web Server, (ii) Proxy server and (iii) Browser machine. The web logs are collected from multiple data sources and combined into new log file.

Data Cleaning

Data cleaning [9] is the process of removing noisy and irrelevant data that are not helpful for mining the knowledge from the web logs. When the user request the HTML web pages, the embedded images are also be downloaded and stored in the web server. But these are not explicitly requested by the users which are avoided. This is done by checking the suffix of each URL. In addition to that, poor status code and request from auto search engines are also removed.

Data cleaning process consists of removing (i) The records which have the extension *.gif, *.jpeg, *.css, *.cgi, etc., (ii) The records with the failed status code. The status code greater than 299 and lesser than 200 are treated as failure status code, (iii) The request processed by auto search engines such as Crawlers, Robot and Spider are removed.

User Identification

User Identification [10] is the complex job of web log preprocessing. But it is essential to distinguish the users. Because, grouping the users based on their visiting behavior is one of the important application of web usage mining. Different techniques such as using IP address, referrer log and user agent are used to identify the users. The following methods are used to identify the user:

(i) Unique IP address represents one user.

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

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

Session Identification

Whenever user interacts with the website, they spend some time in each web page. Session is time duration spending on each web page by the single user. Session identification [10] is the process of dividing the

individual user access logs into sessions. The login and logout time is considered for identifying the starting and ending time of each session. The following are the common rules to identify user session:

(i) If the user is identified as new user, there is a new session;

(ii) For the same session, if the refer page is null, there is a new session;

(iii) If the time between page requests exceeds 30 or 25.5 minutes, it is considered as new session.

Path Completion

There are chances of missing pages after constructing transactions due to proxy servers and caching problems in web server logs. In such condition it becomes mandatory for identifying the user's access path, and adding the missing paths. Because of local buffers existence, some requested pages are not recorded in access log [3]. The goal of path completion is to fill all the missing references that are not recorded. The solution for path completion is, if a requested page is reachable by a hyperlink from any of the visited pages by the user, it is assumed that it is added in the session.

Data Formatting

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

4. PROPOSED PREPROCESSING METHODOLOGY

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

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

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

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

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

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

User Interest Level Preprocessing (UILP) Algorithm

The preprocessing steps are considered as the initial process and web logs are formatted according to the pattern mining algorithm to group the users based on their website visiting behavior. According to the pattern mining algorithm, the web logs are preprocessed with the following five attributes:

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

Where, "b" is boid, "ip" is the ip address, "user" is the user name, "url" is web address, "session" is session duration of the user, and "frequency" is the number of visits by the user. There are many techniques by which to reduce the density of log content in a log file. This work considers five entities namely IP address, user name, website name, session and frequency. The UILP algorithm effectively performs the preprocessing techniques which support the next level clustering process. Cookies based web logs are taken as the input which effectively classify the unique users, the explicit user request for image and multimedia fields are taken into consideration for data cleaning which helps to identify the user interest level. Site topology is used to identify the user and for completing the missing path for the user.

Data Cleaning Algorithm

Algorithm for Data Cleaning

Algorithm : Data Cleaning

Input: Web Logs (Web Server Logs and Cookies)

Output: Filter Log Table

Step 1: Read the web logs record

Step 2: If (suffix.url represent image and multimedia file extensions) then
If (the request is explicit request) then
Add the records to FilterLog Table
If (status code not equal to failure) and
(user agent not equal to crawler, spider, robot) and
(method equals 'GET') then
Add the records to Filter Log Table
Step 3: Repeat Step 1 and 2 till end of Log File

User Identification Algorithm

Algorithm for User Identification

Algorithm: User Identification
Input: FilterLog Table
Output: Log file with distinguished users
Step 1: If IP address is unique then new user;
Step 2: If IP address is same and
user name is not unique,
agent log, operating system and browser are different
then distinguish users.
Step 3: Construct site topology to verify access path and identify users
Step 4: Repeat step 1,2 and 3 till end of FilterLog Table

Session Identification and Frequency Calculation

The extraction process of the session timing and the frequency is calculated by taking the time difference and the total number of clicks on a particular web site given in a log file.

To label the Session, the time duration is calculated between two nearby website visited by the particular user. It is calculated each and every time when a user switches from one website to another and the amount of time spends in each website. Equation 4.1 shows the formula for calculating the session time.

$$session = \sum time(site_i \rightarrow site_j) \quad (4.1)$$

$$frequency = \sum w_v \quad (4.2)$$

The session is calculated as the time taken to traverse from one site to another site by the user and the proposed approach take the whole sum of the duration of particular web site. The sum is taken as the total session duration collected for a website. Frequency is calculated using the equation 4.2 ; w_v represents the visit of the user to a website *w*.

A request using GET should only retrieve data and should have no other effect. POST submits data to be processed in a form of an HTML to the identified resource. The data is included in the body of the request. This may result in the creation of a new resource or the updates of existing resources or both. The session time is calculated from the beginning of the connection till it is switched to another website. In the same way, the frequency is also calculated in the similar manner. It depends upon the number of clicks clicked for a particular web site in the web browser. The number of clicks retrieves the frequency of that particular web site.

Path Completion

Site topology is constructed to fill the missing page references. Referrer page is taken from the site topology for path completion. Figure 4.1 shows the site topology for path completion.

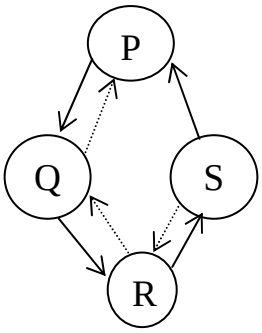


Figure 4.1 Site Topology

The actual page reference sequence is P – Q – R – S – R – Q - P. The user is accessing the page P using back button. But it is recorded in the web server as P – Q – R – S – P. So, this type of missing paths is completed by using site topology and referrer log.

5. EXPERIMENTAL SETUP AND PERFORMANCE ANALYSIS

The web log files are collected from college web server and browser machine during 01st August 2012 to 10th September 2012. Totally 27962 records were extracted from web log files. Java (jdk 1.6) is used for implementing the preprocessing algorithms. The system has Intel core i3 processor with 4GB RAM.

Performance Metrics

The performance of proposed preprocessing algorithms is compared with the existing techniques for (i) Data cleaning, (ii) User identification and (iii) Session identification.

Performance evaluation of UILP Algorithm with existing techniques

The performance of proposed UILP algorithm is analyzed with the existing techniques. Figure 5.1 shows the comparison result of data cleaning process. The UILP algorithm takes consideration of explicit user request for image and multimedia files. So, ULIP only removes 5084 image files and 3813 multimedia files, whereas in the case of data cleaning existing algorithms removes entire image and multimedia request. The ISWUP is an intelligent learning algorithm, which removes 5950 image files and 4784 video files; other two existing algorithms remove 6330 image files and 5140 video files. Removal of auto search engine request and failure status code is common for both proposed UILP algorithm and existing algorithms. By the result, the proposed UILP algorithm outperforms to classify explicit and implicit user requests.

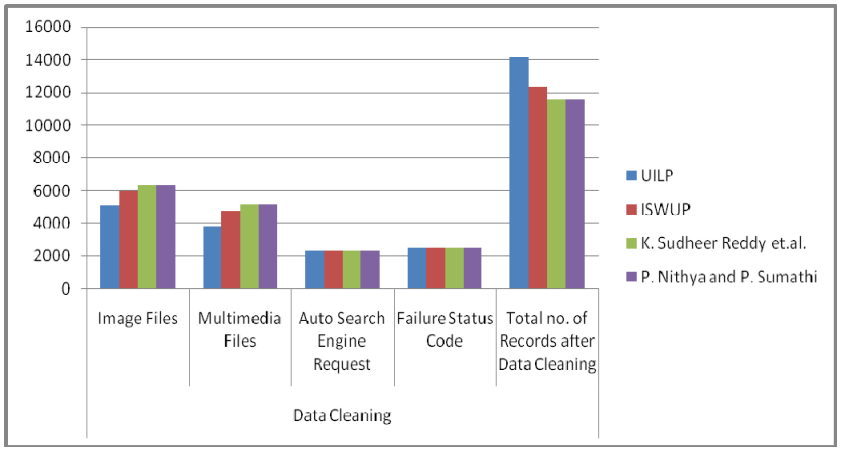


Figure 5.1 Performance comparisons of data cleaning

The UILP algorithm identifies the users based on the data collected from cookies and constructs a site topology to analyze user navigation behavior. Cookies and site topology helps to identify exact number of users and unique users. Figure 5.2 show that, the proposed algorithm proves its efficiency for identifying users and unique users. The proposed UILP algorithm identifies 1271 users and 550 unique users. The ISWUP identifies 1190 users and 457 unique users. The other two existing algorithms identify 1190 users and 428 unique users. Cookies based user identification gives better performance for user identification.

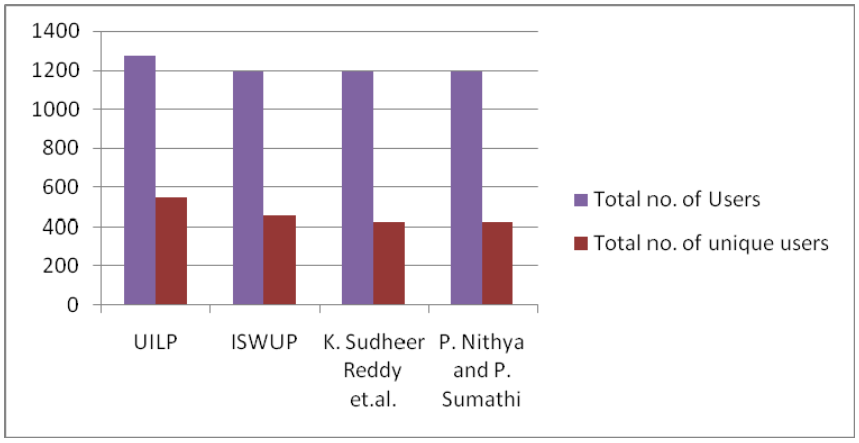


Figure 5.2 Performance comparison of User Identification

Session identification is considered as an important work of UILP algorithm and it helps to group the users based on their website surfing behavior. Figure 5.3 shows the number of sessions identified by proposed and existing algorithms. Session identification and frequency calculation is considered as an important of proposed UILP algorithm. Here, Session is calculated based on the movement from one website to another website and frequency value is calculated with respect to the number of pages visited by the user in the particular website. Session time and frequency value is calculated using the equation 3.1 and 3.2. The proposed system takes the whole sum duration for session and frequency is based on the number of clicks on the particular website. The session and frequency is considered as an important factor for clustering the users based on their interesting measure. From the figure 3.8, the UILP identifies 61783 sessions and existing algorithms consider the session time duration is 30 minutes and there is no calculation for identifying the frequency. 4760 sessions are identified by ISWUP algorithm and 3570 sessions are identified by other two existing algorithms.

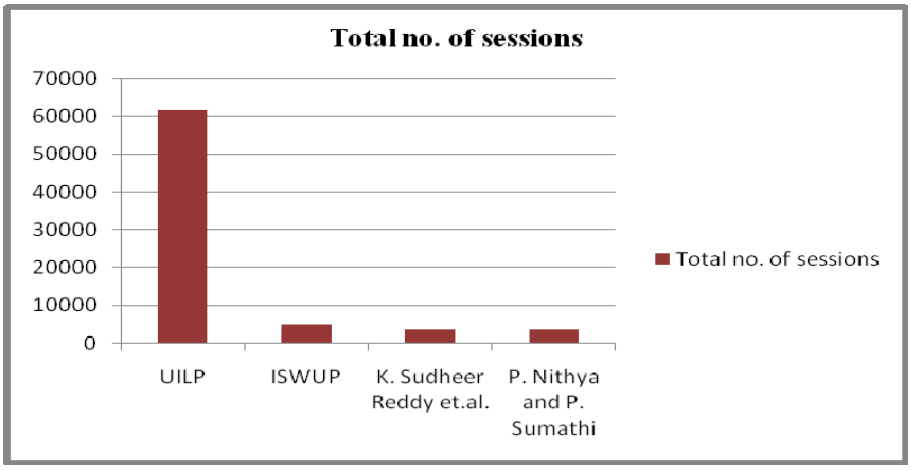
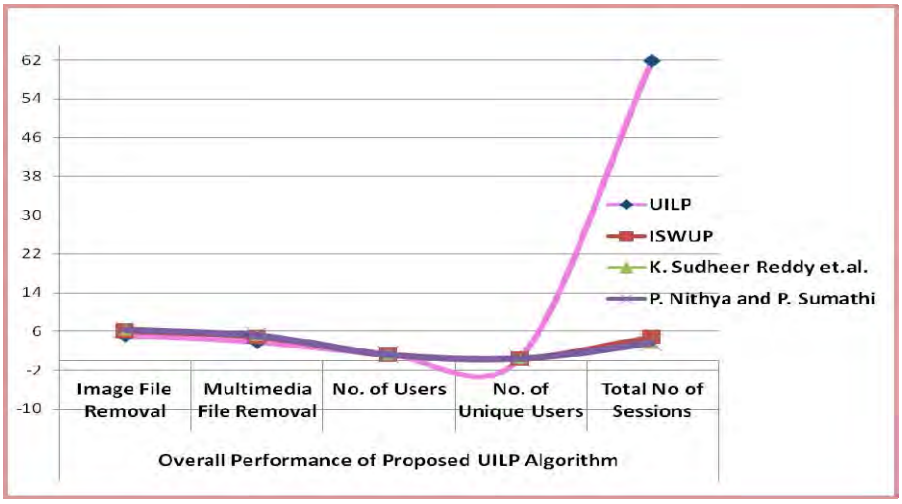


Figure 5.3 Performance comparison of Session Identification

Figure 5.4 shows the overall performance enhancement of proposed UILP algorithm when compared with the three existing algorithms. UILP algorithm significantly improves its performance for identifying unique users and number sessions. It also classifies the explicit and implicit requests during data cleaning.



5.4 Overall Performance of the proposed UILP Algorithm

6. CONCLUSION

User Interest Level Preprocessing algorithm is proposed to perform data cleaning, user identification, session identification and path completion. The UILP algorithm extract fields such as ip address, user name, website address, session and frequency. There are several methodologies and techniques are applied by the researchers to preprocess the web log files to make them consistent. The proposed system effectively performs preprocessing which support user interest level grouping. Session and frequency values are considered as the key for identifying user interest level.

REFERENCES

[1] Mr. Sanjay Bapu Thakare, and Prof. Sangram. Z. Gawali, "A Effective and Complete Preprocessing for Web Usage Mining", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 848-851.

[2] Tasawar Hussain, Dr. Sohail Asghar, and Simon Fong, "A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining", IEEE Conference, 2012.

[3] 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.

[4] K. Sudheer Reddy, M. Kantha Reddy and V. Sitaramalu., "An effective Data Preprocessing method for Web Usage Mining", IEEE Conference Proceedings, 2012.

[5] P. Nithya and P. Sumathi, "An Enhanced Pre-Processing Technique for Web Log Mining by Removing Web Robots", IEEE International Conference on Computational Intelligence and Computing Research, 2012.

[6] J.Srivatsava, R.Cooley, M.Deshpande, and P.N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." ACM SIGKDD Explorat. Newsletter, 2000.

[7] V.Chitraa, Dr.Antony Selvadoss Devamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications, Volume 34– No.9, 2012

[8] M. Malarvizhi S. A. Sahaaya Arul Mary, "Preprocessing of Educational Institution Web Log Data for Finding Frequent Patterns using Weighted Association Rule Mining Technique", European Journal of Scientific Research ISSN 1450-216X Vol.74 No.4 ,617-633, 2012.

[9] Vijayashri Losarwar and Dr. Madhuri Joshi, Data Preprocessing in Web Usage Mining, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, Singapore, 2012.

[10] Sheetal A. Raiyani and, Shailendra Jain, "Efficient Preprocessing technique using Web log mining, International Journal of Advancements in Research & Technology", 1(6) ISSN 2278-7763, 2012.