HIGH UTILITY MINING ALGORITHM FOR PREPROCESSED WEB DATA

Swapna Mallipeddi[1], D.N.V.S.L.S.Indira[2]

[1] M.Tech(CSE),Gudlavalleru Engineering College, Gudlavalleru, A.P., India.

[2] Associate professor, Gudlavalleru Engineering College, Gudlavalleru, A.P., India.

Abstract

With the explosive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses.Web usage mining is usually an automated process whereby Web servers collect and report user access patterns in server access logsThrough the world wide web development, the web utility mining, which can be viewed as a mixed term of web mining & utility mining, becomes much more meaningful with the introduction of the emerging concepts of two-phase algorithm & on-shelf utility, already proved to be very effective under their respective fields of data mining. In real applications, however, utility mining may have a bias if items are not always on-shelf. On-shelf utility mining is then proposed, which considers not only individual profit and quantity of each item in a transaction but also common on-shelf time periods of a product combination. With the rapid growth of the Web, the web log data have become an important data source for machine learning and data mining. During preprocessing phase, raw Web logs need to be cleaned, analyzed and converted before further utility mining.

In this paper to improve the efficiency of on shelf utility mining we apply the 2-phase algorithm on preprocessed web transactions.

1. INTRODUCTION

The goal of web usage mining is to find out the useful information from web data or web log files. The other goals are to enhance the usability of the web information and to apply the technology on the web applications, for instance, pre-fetching and caching, personalization etc. For decision management, the result of web usage mining can be used for target advertisement, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise, and marketing analysis etc.

Forecasting the users' browsing behaviors is one of web usage mining issues. In order to achieve the purpose, it is necessary to understand the customers' browsing behaviors through analyzing the web data or web log files. Predicting the most possible user's next requirement is based on the previous similar behavior. There are many advantages to implement the prediction, for example, personalization, building proper web site, improving marketing strategy, promotion, product supply, getting marketing information, forecasting market trends, and increasing the competitive strength of enterprises etc[2].

The terminology of web mining was proposed by Etzioni in 1996. Web mining is applied to web pages and services of Internet in order to find and extract the available knowledge. Web mining can be categorized into three categories (as Fig. 0) which are web content mining, web structure mining and web usage mining

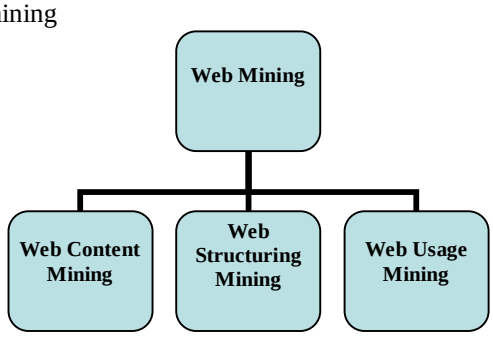


Figure 0. Taxonomy of Web Mining

Web content mining focuses on useful knowledge which is extracted from web pages. Web structure mining is used to analyze the links between web pages through the web structure to infer the knowledge. Web usage mining is extracting the information from web log file which is accessed by users. Lee and Fu proposed a Two Levels of Prediction Model in 2008 (as Fig. 1) . The model decreases the prediction scope using the two levels framework. The Two Levels of Prediction Model are designed by combining Markov model and Bayesian theorem. In level one, Markov model is used to filter the most possible of categories which will be browsed by user. In the level two, Bayesian theorem is used to infer precisely the highest probability of web page.

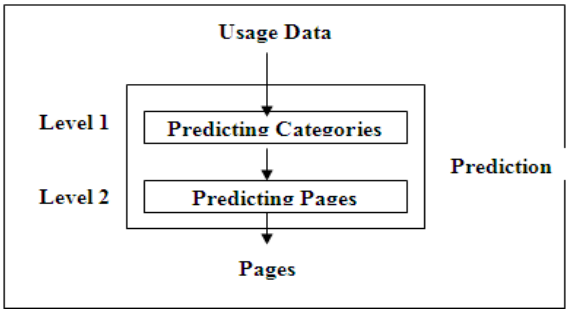


Fig 1

In level one, it is to predict the most possible user's current state (web page) of category at time *t*, which depends on user's category at time *t*-1 and time *t*-2. Bayesian theorem is used to predict the most possible web pages at a time *t* according to user's states at a time *t*-1. In the Two Levels of Prediction Model framework the similarity matrix *S* of category is established. The approach of establishing similarity matrix is to gather statistics and to analyze the users' behavior browsing which can be acquired from web log data.

The enterprise proxy log is access log of the employees visiting the World Wide Web by using the proxy servers. Mining the enterprise proxy log provides a new aspect of analyzing the user behavior of surfing the web, and in addition it helps to optimal the cache strategies of proxy servers and the intranet management. In this paper, we focus
on providing web recommendations. Firstly, a few features of the enterprise proxy log are presented by comparing with the web server log[1]:
(1) Unlimited access to the web sites. Web server log is access log of the users surfing a particular site, while the enterprise proxy log has no limit to the sites which the users
access. It makes that although we have a huge amount of data records, but these records are discrete and make the access patterns more hidden.

(2) Unknown to the information of the web sites. This includes the topology of the site, the classification of the pages, and the mark of the attach pages. This information has an important affect on filtering the access log to make mining algorithms correct.
(3) Diversifying the behavior motivations. The motivation of user browsing one site is usually related to the topic of this site, so we can expect to mine out some access patterns related to this topic. But that's not true while mining the enterprise proxy log for its WWW-oriented feature. We will be confronted with more access patterns and more motivations while mining the enterprise proxy log.
(4) Rapid growth of new pages in log. It's different from the web server log for its WWW-oriented feature.

2 LITERATURE SURVEY

S.Madria et al. [Madria 1999] gave details about how to discover interesting facts describing the connectivity in the Web subset, based on the given collection of connected web documents. The structure information obtained from the Web structure mining has the followings:

- The information about measuring the frequency of the local links in the web tuples in a web table The information about measuring the frequency of web tuples in a web table containing links within the same document
- The information measuring the frequency of web tuples in a web table that contains links that are global and the links that point towards different web sites
- The information measuring the frequency of identical web tuples that appear in the web tables.

Most of the research in Web usage mining is focused on applications using web Server Data. The only remained information after users visits a web site is the data about the path through the pages they have accessed. Most of the Web log analysis tools only use the textual data from log files. They ignore the link information, which is also very important. Web usage mining tries to discover useful information from the data extracted from the interactions of the users while surfing on the Web. It also has focus on the methods predicting user behavior while the user interacts with Web.
Tasawar *et al.*, [3] proposed a hierarchical cluster based preprocessing methodology for Web Usage Mining. In Web Usage Mining (WUM), web session clustering plays a important function to categorize web users according to the user click history and similarity measure. Web session clustering according to Swarm assists in several manner for the purpose of managing the web resources efficiently like web personalization, schema modification, website alteration and web server performance. The author presents a framework for web session clustering in the preprocessing level of web usage mining. The framework will envelop the
data preprocessing phase to practice the web log data and change the categorical web log data into numerical data. A session vector is determined, so that suitable similarity and swarm optimization could be used to cluster the web log data. The hierarchical cluster based technique will improve the conventional web session method for more structured information about the user sessions.
Yaxiu *et al.*, [4 put forth web usage mining based on fuzzy clustering. The World Wide Web has turn out to be the default knowledge resource for several fields of endeavor, organizations require to recognize their customers' behavior, preferences, and future requirements, but when users browsing the Web site, several factors influence their interesting, and various factor has several degree of influence, the more factors consider, the more precisely can mirror the user's interest. This paper provides the effort to cluster similar Web user, by involving two factors that the page-click number and Web browsing time that are stored in the Web log, and the various degree of influence of the two factors. The method suggested in this paper can help Web site organizations to recommend Web pages, enhance Web structure, so that can

draw more customers, and improves customers' loyalty. Web usage mining based on fuzzy clustering in identifying target group is suggested by Jianxi *et al.*, [5].

Data mining deals with the methods of non-trivial extraction of hidden, previously unknown, and potentially helpful data from very huge quantity of data. Web mining can be defined as the use of data mining methods to Web data. Web usage mining (WUM) is an significant kind in Web mining. Web usage mining is an essential and fast developing field of Web mining where many research has been performed previously. The author enhanced the fuzzy clustering technique to identify groups that share common interests and behaviors by examining the data collected in Web servers.

Houqun *et al.*, [6] proposed an approach of multi-path segmentation clustering based on web usage mining. According to the web log of a university, this paper deals with examining and researching methods of web log mining;

bringing forward a multi-path segmentation cluster technique, that segments and clusters based on the user access path to enhance efficiency.

Data Preprocessing

The normal procedure of data preprocessing includes 5 steps :data cleaning, user identification, user session identification, path completion and user transaction identification. While applying these to the enterprise proxy log, we encounter some new challenges. Web pages are becoming more and more colorful with attachments such as advertisements. However, as mentioned in section 1, we have no information about the web sites. It makes the normal data cleaning methods still have a lot of noisy pages which affects the data mining. Besides, this also disables the step of path completion for lacking information. For these reasons, the data preprocessing we used in this paper makes some modifications, which includes data cleaning, user identification, incremental filtering, user session identification and user transaction identification. These steps are shown as Fig. 2. We use the method of data cleaning according to [7] and user identification is easier because of the authentication information. Through the observation of the content pages and the attached pages, based on the feature that the attached pages are requested automatically when the related content pages are requested, we present some hypotheses. These hypotheses are as follows:

(1) Because with the feature above, the requested time of an attached page is immediately after the requested time of the related content page. We set this interval to be 1 second.

(2) An attached page usually can refer from many different pages. Although a content page can also refer from more than one other pages, but such records are much fewer in logs. So we assume that a page referred from more than 10 different pages is an attached page.

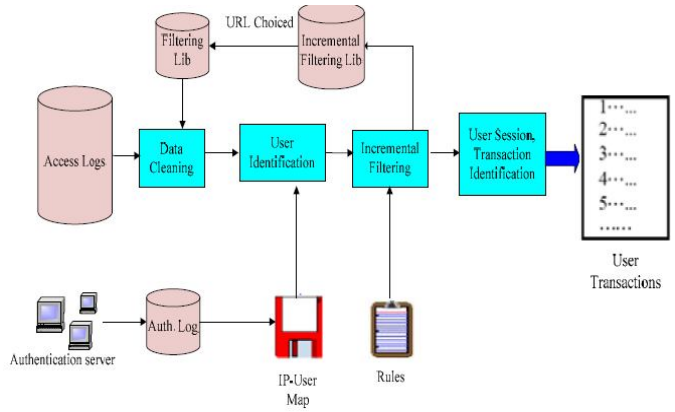


Fig.2. Data Preprocessing

(3) Through our observation, the size of a content page surely is larger than 4500 bytes. Even if a few non-attached pages are smaller than 4500 bytes, they usually have too little content to attract user and can be ignored without affecting the mining results. Based on the above hypotheses, we propose a method of incremental data filtering and put those filtered pages into an incremental filtering lib. In this lib, we choice the pages referred from more than 10 different pages into the filtering lib to assist the data cleaning. The result shows that after a period of time, the

filtering lib and the incremental filtering lib can gain the feature of attached pages. It helps a lot to find the advertisements or to find the naming rules of the attached pages. After incremental filtering, we apply the process of user session identification using 30 minutes as interval . A transaction is a subset of related pages that occur within a user session, which targets at creating meaningful clusters of references for each user. Although lacking the topology of the web sites, we can construct a visit tree by using the records expressed by (URL, Referrer URL). As the Fig.3 shows: The transactions generated from Fig.3 are as follows. We express a transaction as: (user, transaction-id, {(navigational pages), (index page), (content pages)} to keep the relation of pages. *(user, transaction-id, transaction={(A,B,D), ,(H)}) (user, transaction-id, transaction={(A), C, (E, G)})*

(user, transaction-id, transaction={(A, C), F, (I, J, K)})

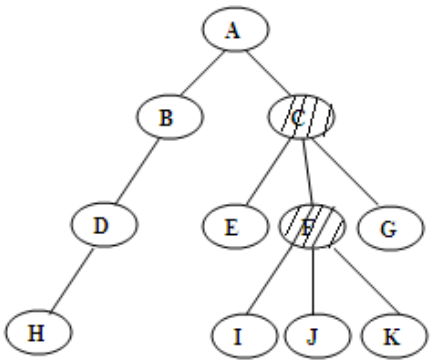


Figure 3 Example of A visit to Tree

3 BASIC WEB MINING STRUCTURE

In personal Web usage mining, two kinds of user Web activities are recorded for analysis: remote activities and local activities. The remote activities include requests sent by a user to a Web server. Such kind of click stream data includes the URLs of pages as well as any keywords, queries, forms, and cookies sent with the URL. The local activities include actions the user can take at his or her desktop without the knowledge of Web servers. They include, but are not limited to, the following.

· Save a page
· Print a page
· Click **Back** on browser
· Click **Forward** on browser
· Click **Reload** on browser
· Click **Stop** on browser
· Email a link/page
· Add a bookmark
· Minimize/maximize/close window
· Change visual settings such as font size

The remote activities can be captured by almost all Web browsers. Besides, the browsers also cache the Web pages in most cases. The local activities can be recorded by an activity recorder, which is a client side program running on top of the browser. These two kinds of activities are put together into an activity log. Each entry in the activity log will contain a timestamp and an activity. Some will contain extra information such as URL, cache address, keyword, cookie, email address, and font size. The schema of the log looks like this: (timestamp, activity, [URL], [cache address], [keyword], [cookie], [email address], [other optional fields])

The framework for personal Web usage mining is given in Figure 4. There are four major modules in the framework: logging, data warehousing, data mining, and tool/application. In the logging module, user Web activities are stored into the activity, as well as the cached pages. In the data warehousing module, the logs and cached pages are cleansed, extracted, transformed, aggregated, and stored in a data warehouse.

A user's actions are recorded by the Web browser and the activity recorder. The browser will also cache the Web pages requested by the user. As shown in Figure 4, the user's activities will be the source for data warehousing and data mining, whose results will be employed by the tools and applications, which, in turn, aim to help the user with his or her Web activities. In such a user-centric way, personal Web usage mining has great potentials in bringing Web to people.

Several kinds of patterns can be discovered from the data using various data mining techniques. Onet among them is is Summarization.

Summarization

The data can be summarized and abstracted to find general patterns. For example, if the user prints ten pages about music from the Yahoo site on March 12, 2012, the individual activities can be summarized as "10 pages on Yahoo related to music printed on March 12, 2012". This can be done using the OLAP operations and data summarization techniques such as attribute-oriented induction

4. PROPOSED APPROACH

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process. This system only describe it shortly and say that its main task is to "clean" the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process. So the main steps of this phase are:

1) Extract the web logs that collect the data in the web server.
2) Clean the web logs and remove the redundant information.
3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in frequency analysis to create summary reports.

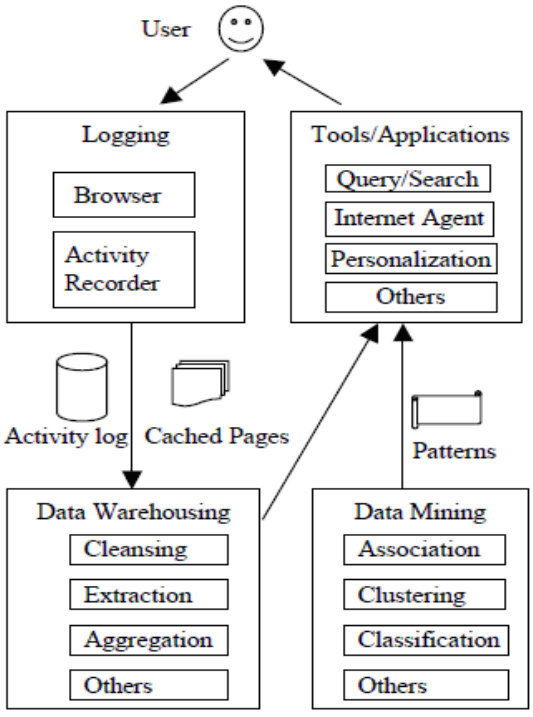


Figure 4. A framework for personal Web usage mining.

Data Preprocessing

The purpose of Data Preprocessing is to change a web data mining into reliable data, including four phases: data cleaning,

user identification, session identification as well as fragments identification [2]

1) Data Cleaning: Data cleaning is usually site-specific, and involves tasks such as, removing extraneous references to embedded objects that may not be important for the purpose of analysis.

2) User Identification: The analysis of Web usage does not require knowledge about a user's identity. However, it is necessary to distinguish among different users. Using a similar modification of the paper as web log records, and as is shown in table 1 after data preprocessing [3]. Identifying the user's browsing path A1-B1-C1-C2-D3-C3-D4, A1-B2 - C4-B3-D4-B2, A1-B3

3) Session identification: Session identification is the process of segmenting the user activity record of each user into sessions, each representing a single visit to the site. After session identification, the ideal heuristic can re-construct the exact sequence of user navigation during a session.Server log consists of 19 attributes. The attributes are:-

a) Date
The date from Greenwich Mean Time (GMT x 100) is recorded for each hit. The date format is YYYY-MM-DD.

b) Time
Time of transactions. The time format is HH:MM:SS.

c) Client IP Address
Client IP is the number of computer who access or request the site.

d) User Authentication
Some web sites are set up with a security feature that requires a user to enter username and password. Once a user logs on to a Website, that user's "username" is logged in the fourth field of the log file.

e) Server Name
Name of the server. Ex **CSLNTSVR20.**

f)Server IP Address
Server IP is a static IP provided by Internet Service Provider. This IP will be a reference for access the information from the server[8].

g) Server Port
Server Port is a port used for data transmission. Usually, the port used is port 80.

h) Server Method (HTTP Request)
The word request refers to an image, movie, sound, pdf, txt, HTML file and more. The GET in front of the path name specifies the way in which the server sends the requested information. Currently, there are three formats that Web servers send information [8] in GET, POST, and Head. Most HTML files are served via GET Method while most CGI functionality is served via POST.

i) URI-Stem
URI-Stem is path from the host. It represents the structure of the websites. For examples:- /tutor/images/icons/fold.gif

j) Server URI-Query
URI-Query usually appears after sign "?". This represents the type of user request and the value usually appears in the Address Bar. For example:- ?**q=tawaran+biasiswa&hl=en&lr=&ie=UTF-8&oe=UTF-8&start=20&sa=N**

k) Status
This is the status code returned by the server; by definition this will be the three digit number [2]. There are four classes of codes:

i. Success (200 Series)
ii. Redirect (300 Series)
iii. Failure (400 Series)
iv. Server Error (500 Series)

A status code of 200 means the transaction was successful. Common 300-series codes are 302, for redirect from http://www.mydomain.com to http://www.mydomain.com, and 304 for a conditional GET. This occurs when server checks if the version of the file or graphics already in cache is still the current version and directs the browser to use the cached version. The most common failure codes are 401 (failed authentication), 403 (Forbidden request to a restrict subdirectory, and the dreaded 404 (file not found) messages. In the above transmission a status is 200 means that there was a successful transmission.

a) Bytes Sent The amount of data revisited by the server, not together the header line.

b) Bytes Received
Amount of data sent by client to the server.

c) Time Stamp
This attribute is used to determine how long a visitor spent on a given page.

d) Protocol Version
HTTP protocol being used (e.g. HTTP/1.1).

e) Host

- **T**his is either the IP address or the corresponding host name
- (www.tutor.com.my) of the remote user requesting the page.

Data Preprocessing
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

Algorithm: Data cleaning
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.method='GET') AND
(LogRecord.Sc-status<>(301,404,500)AND
(LogRecord.Useragent<>Crawler.Spider.Robot))
Then insert LogRecord in to LogDatabase.
End of If condition.
Step3: Repeat the above two steps until eof(Web Server Log File)
Step4: Stop the process.

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 overcome these.

Algorithm: User Identification Algorithm.
Input: Log Database
Output: Unique Users Database
Step 1: Initialize
IPLIst=0;UsersList=0;BrowserList=0;
OSList=0;No-Of-users=0;
Step 2: Read Record From LogDatabase
Step 3: If Record.IP address is 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.IPaddress is present in IPList OR
Record.Browser not in BrowserList OR
Record.OS not in ORList
Then
Increment count of No-Of-users
Insert as new user in to UserList.
End of If
End of If
Step 4: Repeat the above step 2to 3
Until eof(Log Database)
Step 5: Stop the process.

Session Identification
A session ID is a unique number that a Web site's server assigns a specific user for the duration of that user's visit (session). The session ID can be stored as a cookie, form field, or URL (Uniform Resource Locator). Some Web servers generate session IDs by simply incrementing static numbers. However, most servers use algorithms that involve more complex methods, such as factoring in the date and time of the visit along with other variables defined by the server administrator.
Every time an Internet user visits a specific Web site, a new session ID is assigned. Closing a browser and then reopening and visiting the site again generates a new session ID. However, the same session ID is sometimes maintained as long as the browser is open, even if the user leaves the site in question and returns. In some cases, Web servers terminate a session and assign a new session ID after a few minutes of inactivity.

Algorithm: Session Identification algorithm
Input: Log Database
Output: Session Database
Step 1: Initialize
SessionList=0
UserList=0
No-Of-Sessions=0
Step 2:Read LogRecord from Log Database
Step 3: 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
Step 4: Repeat the above steps 2 and 3 till eof(Log Database)
Step 5: End of process.

Algorithm: Two Phase Algorithm
Input:
1. A set of m items I={i1,i2…im},each Ij with a profit value Pj,j=1 to m;
2. A transaction database D={T1,T2,…Tn} in which each transaction includes a subset of items with quantities;
3. The minimum threshold thres.

Output: A set of utility itemsets[9].
Step1: Calculate the utility value Ujk of each item Ij in each transaction Tk as
Ujk=Qjk*Pj,
Where Qjk is the quantity of Ij in Tk for j=1 to m and k=1 to n.
Step2: Find the maximal utility value MUk in each transaction Tk as MUK=max{U1k,U2k,….Umk} for k=1 to n.
Step3:Calculate the Utility upper bound UBj of each item Ij as the summation of the maximal utilities of the transactions which include Ij. That is :

$$ub_j = \sum_{i_j \in T_k} mu_k$$

Step4: Check whether the utility upper bound of an item Ij is larger than or equal to thres. If Ij satisfies the above condition,put it in the set of candidate utility 1- itemsets,C1. That is:

$$C_1 = \{i_j \mid ub_j \geq \lambda, 1 \leq j \leq m\}$$

Step5: Set r=1,where r is used to represent the number of items in the current candidate utility itemsets to be processed.
Step6: Generate the candidate set Cr+1 from Cr with all the r-subitemsets in each candidate in Cr+1 must be contained in Cr.
Step7: Calculate the Utility upper bound UBs of each candidate utility (r+1) itemset as the summation of the maximal utilities of the transactions which include s. That is :

$$ub_s = \sum_{s \subset T_k} mu_k$$

Step8: Check whether the average-utility upper bound of each candidate (*r*+1)-itemsets *s* is larger than or equal to thres. If *s* does not satisfy the above condition, remove it from *Cr+1*. That is:

$$New\ C_{r+1} = \{s \mid ub_s \geq \lambda, s \in original\ C_{r+1}\}.$$

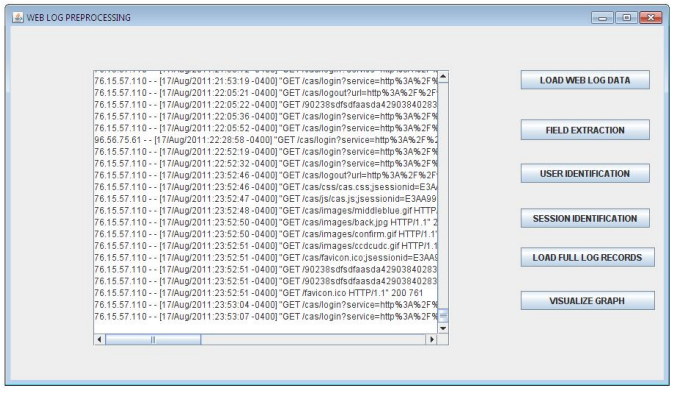
Step9: IF *Cr+1* is null, do the next step; otherwise, set *r* = *r* + 1 and repeat STEPs 6 to 9.
Step10:For each candidate average-utility itemset *s*, calculate its actual average-utility value *aus* as follows:

$$au_s = \sum_{s \subset T_k} \sum_{i_j \in s} u_{jk} \Big/ |s|,$$

where *ujk* is the utility value of each item *ij* in transaction *Tk* and |*s*| is the number of items in *s*.
Step11:Check whether the actual average-utility value *aus* of each candidate average-utility itemset *s* is larger than or equal to thres. If *s* satisfies the above condition, put it in the set of high average-utility itemsets, *H*.

5. EXPERIMENTAL RESULTS

All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 2 GB RAM, and the operation system platform is Microsoft Windows XP Professional (SP2). The dataset is taken from real time php based real time web analyzer. Some of the results in the dynamic web statistics are given below in fig 5 and 6



Monthly Statistics for March 2012		
Total Hits		12543
Total Files		9266
Total Pages		2244
Total Visits		1064
Total KBytes		1656278
Total Unique Sites		769
Total Unique URLs		258
Total Unique Referrers		193
Total Unique User Agents		218
	Avg	Max
Hits per Hour	20	154
Hits per Day	482	807
Files per Day	356	586
Pages per Day	86	192
Visits per Day	40	55
KBytes per Day	63703	315205
Hits by Response Code		
Code 200 - OK		9266
Code 206 - Partial Content		121
Code 301 - Moved Permanently		6
Code 304 - Not Modified		414
Code 404 - Not Found		2736

Fig 5 : number of pages viewed by the users

- Countries Interests

Countries Summary	Number	Pages per visit	Pages per significant visit	1 page visit rate	Average visit duration
France	878	2.5	4.6	59 %	00:02:17
United States	132	2.4	3.8	49 %	00:02:20
Germany	56	3.6	6.0	48 %	00:03:59
Unknown	55	2.1	4.6	69 %	00:01:40
Belgium	49	2.2	4.6	65 %	00:02:37
Czech Republic	38	3.8	5.8	42 %	00:03:36
Sierra Leone	33	4.2	5.7	33 %	00:06:12
Italy	28	2.0	3.7	64 %	00:02:02
Switzerland	26	3.0	5.6	58 %	00:01:29
Canada	24	3.6	6.2	50 %	00:02:49

Include all the population in statistics

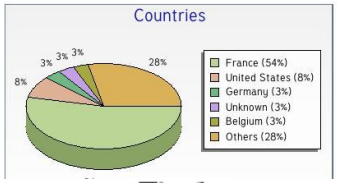


Fig 6

6 CONCLUSION

Data preprocessing is an important task of WUM application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. This paper presents two algorithms for

preprocessing and utility mining. In this paper web log data file is preprocessing and results are Not every access to the content should be taken into consideration. So this system removes accesses to irrelevant items and failed requests in data cleaning. After that necessary items remain for purpose of analysis. Speed up extraction time when users' interested information is retrieved and users' accessed pages is discovered from log data. The information in these records is sufficient to obtain session information. In future Two phase algorithm is implemented on preprocessed data in order to get the high utility in the web preprocessed data.

REFERENCES

[1] A New Perspective Of Web Usage Mining: Using Enterprise Proxy Log Yu Zhang.

[2] Two Levels of Prediction Model for User's Browsing Behavior1 Chu-Hui Lee, Yu-Hsiang Fu.

[3] Jianxi Zhang, Peiying Zhao, Lin Shang and Lunsheng Wang, "Web usage mining based on fuzzy clustering in identifying target group", ISECS International Colloquium on Computing, Communication, Control, and Management, Vol. 4, Pp. 209-212, 2009.

[4] Houqun Yang, Jingsheng Lei and Fa Fu, "An Approach of Multi-path Segmentation Clustering Based on Web Usage Mining", Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 4, Pp. 644-648, 2007.

[5] Bamshad Mobasher, "Data Mining for Web Personalization," LCNS, Springer-Verleg Berlin Heidelberg, 2007.

[6] Jaideep Srivastava , Robert Cooley , Mukund Deshpande, Pang-Ning Tan- "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", 2008.

[7] Honghua Dai and Bamshad Mobasher-"Integrating Semantic Knowledge with Web Usage Mining for Personalization", 2007.

[8] Mark E. Snyder, Ravi Sundaram, Mayur Thakur-"Preprocessing DNS Log Data for Effective Data Mining", 2008.

[9] Two Phase Utility Mining Algorithm G. Sunil Kumar, C.V.K Sirisha, Kanaka Durga.R, A.Devi.