Advanced Preprocessing Techniques used in Web Mining - A Study

T.Gopalakrishnan
Assistant Professor (Sr.G)
Department of IT
Bannari Amman Institute of Technology
Sathyamangalam, TamilNadu

M.Kavya
PG Scholar
Department of IT
Bannari Amman Institute of Technology
Sathyamangalam, TamilNadu

V.S.Gowthami
PG Scholar
Department of IT
Bannari Amman Institute of Technology
Sathyamangalam, TamilNadu

ABSTRACT
Web based applications are now increasingly becoming more popular among the users across the world. The user interactions with the applications are being tracked by the web log files that are maintained by the web server. For this purpose web usage mining (WUM) is being used. Web usage mining is the process of extracting user patterns from the web usage. In web usage mining, preprocessing plays a key role, since large amount of irrelevant information are present in the web. It is used to improve the quality and efficiency of the data. There are number of techniques available at preprocessing level of WUM. Different techniques are applied at preprocessing level such as data cleaning, data filtering, and data integration. In this paper, we present a survey on the various preprocessing techniques that have been used in order to improve the efficiency.

Keywords
Web usage mining; log cleaning; User identification; sessionization

1. INTRODUCTION
World Wide Web is expanding tremendously day by day by means of increasing websites and the users using them are also relatively increasing. As the websites increase, the amount of data that is available in web is also tremendous. Web mining solves this problem by extracting useful information from web. Generally, three kinds of information have to be handled in a web site: Content, structure, and log data. Content data is nothing but the data that is present in the web page. Structure data is just an organization of the content. And usage data is nothing but the usage patterns of the web sites. These three kinds of information handling is said to be as web content mining, web structure mining and web usage mining.

E-commerce applications also have a rapid growth. So the requirement for the personalized service is more and web usage mining satisfies this need by providing personalized data according to the user's behavior. Web usage mining is the discovery of user access patterns from web server access logs. Web usage mining analyses results of user interactions with a web server including web logs, click streams, and database transactions at the web site. Web usage mining consists of three main steps:

1.1 Preprocessing
According to the client, server and proxy server, the preprocessing is the first approach to retrieve the raw data from web resources and process the data. It automatically transforms the original raw data to the next process.

1.2 Pattern Discovery
According to the data preprocessing, the raw data is used to discover the knowledge and to implement the techniques which will be used for machine learning. This makes use of data mining procedures.

1.3 Pattern Analysis
It is the process after pattern discovery. It checks whether the pattern is correct on the web and guides the process of extraction of the information/ knowledge from the web.

These three steps or phases are sequentially connected to each other to form a complete WUM methodology. Web log file is usually given as input.

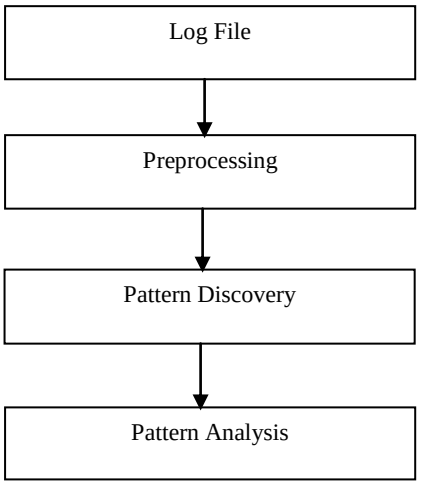

Fig 1: Steps in Web Usage Mining

The goal of the preprocessing step is to transform the raw web log data into a set of user profiles. Each such profile captures a sequence or a set of URLs representing a user session. Web usage data preprocessing exploit a variety of algorithms and heuristic techniques for various preprocessing tasks such as data fusion, data cleaning, user identification, session identification etc.

Data fusion refers to the merging of log files from several web servers. Removing extraneous references to embedded objects, style files, graphics, or sound files, and removing references due to spider navigations are the various tasks involved in data cleaning.

User identification refers to the process of identifying unique users from the user activity logs.

User session identification is the process of segmenting the user activity log of each user into sessions, each representing a single visit to the site. The goal of a sessionization heuristic is to reconstruct, from the click stream data, the actual sequence of actions performed by one user during one visit to the site.

In web mining, web log mining is high-flying due to effective use in numerous web related application. These applications may include modification of web site design, schema modifications, web site and web server performance; improve web personalization, fraud detection and future prediction.

While surfing the web sites, users' interactions with web sites are recorded in web log file. There are three main sources to get the raw web log file[4] such as 1) Client Log File, 2) Proxy Log File and 3) Server Log File. Usage of these sources has its own pros and cons but their importance to collect the data for WUM is invaluable. The true user behavior can be portrayed from client log file [1]. Client log files are most authentic and accurate [6] to depict the user behavior but it is difficult task to modify the browser for each client and requires users' essence and collaboration as well. It is in the form of one-to-many relationships of client and web sites visited by that particular user. Proxy log file is also used to capture user access data. Proxy server log files are most complex and more vulnerable to user access data in log file. To unleash the true picture of user behavior is difficult. Same IP address is used by many users but on the other hand we can have unique user login. Proxy server is in many-to-many relationships. One user can access many sites and many users can visit one site. To capture the real user and users' browsing behavior is difficult. Server log files are in relationship of many to one. Many users may visit one web site and user behavior about that particular web site can be captured accordingly. In this survey, we observed that most of the researcher considered web server log file as most reliable and accurate for WUM process. Server log file do not record the cashed pages [4] requests and that's why it is considered as incomplete and some time it is completed through the topological structure [8] of the web site.

The common server log file types are Access Log; Agent Log; Error Log; and Referrer Log [5, 6].

Referrer log file contains the information about the referrer. As someone jumps from any side to www.google.com by clicking the link, referrer log file of google server will record a referrer entry that a user came from that particular web site. In this regard google has implemented the PageRank algorithm for assigning the weights to referrer sites [1].

Error log file records the errors of web site especially when user clicks on particular link and link does not locate the promised page or web site and user receives ''Error 404 File Not Found''. Error Log file is more helpful for the web page designer to optimize the web site links.

Agent log file records the information about the web site users' browser, browser's version and operating system [6]. This information is again utilized by the web site designer and administrator for the analysis that users are using which specific browser to access the web site. There are number of browser available to users and each browser has its own properties and advantages to their users. Different version of same browser can different added utilities and benefits to its users, so web site can be modified accordingly. Information about the users' operating system is also help for designer and web site changes are made accordingly.

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2010-09-16 16:17:18
#Fields: date time s-sitename s-ip cs-uri-query s-port
 cs-usename c-ip cs(User-Agent) sc-status sc-substatus
 sc-win32-status
2010-09-16 16:17:18 W3SVC1 127.0.0.1 GET/Board
EzLog/Service1.asmx WSDL 80 – 127.0.0.1
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT
+5.2;+SV1;+.NET+CLR+1.1.4322)2000
```

Fig 2: Example of common log file

Access Log File is major log of web server which records all the clicks, hits and accesses made by any web site user. There are number of attributes ([5], [8]) in which information is captured about users. Information about the user is then processed for WUM and user behaviour and interest can be mined. Table 1 elaborates the different attributes of access log file along with their description.

There are three main types of web server log file formats available to capture the activities of user on web site [3]. All the three log files are in ASCII text format. Log files act as health monitor for the web sites and are main source of user access data and user feedbacks. These are Common Log File Format (NCSA); Extended Log Format (W3C); and IIS Log Format (Microsoft). NCSA Common log file format is most widely used to capture user data. It is standardized format but not customizable [5]. Only fixed numbers of attributes are available for raw data of users. Figure 2 elaborates the example of common log file with basic necessary information of log entries.

```
127.0.0.1—[06/Dec/2010:14:09:32 +0500] "GET / HTTP
/1.1" 200 1494
127.0.0.1—[06/Dec/2010:14:09:32 +0500] "GET
/apache_pb.gif HTTP/1.1"200 2326
127.0.0.1—[06/Dec/2010:14:09:32 +0500] "GET /
favicon.ico HTTP/1.1" 404 283
127.0.0.1—[06/Dec/2010:14:13:00 +0500] "GET /
phpinfo.php HTTP /1.1" 200 41959
127.0.0.1—[06/Dec/2010:14:13:01 +0500] "GET /
phpinfo.php?=PHPE9568F34-D428-11d2-A769-
0DAA001ACF42 HTTP /1.1" 200 2524
```

Fig 3: W3C Extended Log File

W3C extended log file format is more flexible and can be customized according to requirements. Different attributes can be added to collect the user access data. Figure 3 shows a segment of W3C extended log file.

Microsoft IIS Log file format Figure4 is non-customizable but as compare to common log file format have more attributes and record more data of users' accesses.

```
192.168.114.201, -, 03/20/01, 7:55:20, W3SVC2, SERVER,
172.21.13.45, 4502, 163, 3223, 200, 0, GET, /DeptLogo.gif, -,
```

Fig 4: Microsoft IIS Log File Format

In the upcoming sections literature review, conclusion and the future directions will be presented.

2. LITERATURE REVIEW

The focus of literature review is to study, compare and contrast the available preprocessing techniques. Due to large

amount of irrelevant entries in the web log file, the original log file cannot be directly used in WUM process. Therefore, the preprocessing of web log file becomes more essential.

Theint Theint Aye (2011) [11] states that preprocessing of web log file plays an important role in WUM. In this research, log file structure is presented stating the fields available in them. Then the field extraction technique was presented to separate field from the single line of the log file. The server used different characters which work as separators. The most used character is ',' or 'space' character.

After extracting the fields data cleaning is done in order to eliminate the irrelevant or unnecessary items in the analysed data. The data including references to style files, graphics or sound files, the records with failed HTTP status code also may include in log data. These entries are useless for analysis process and hence it is cleaned from the log files. A filtered web log is obtained by applying field extraction and data cleaning technique to the raw data.

Only for specific and limited analysis within a particular domain, this technique can be implemented. The authors performed data cleaning; but no other preprocessing technique was applied such as path completion, filtering, data integration and data grouping. Users and user sessions were not identified which play vital role in upcoming steps of WUM.

Web log data preprocessing is a complex process and takes 80% of total mining process. This view has been supported in the research work of Nithya (2012) [12]. In this research author takes a raw data and applies some cleaning techniques to remove the irrelevant records and finally tranforms them into sessions. For their experiment, they took the UCI machine learning repository datasets and a real dataset collected from reputed college. In data cleaning step, first the global and the local noises is removed. Global noises include the mirror sites, duplicated web pages, previous versioned web pages, noise words such as "ad-serving", "contact", "company profiles", "copyright", "all rights reserved". Local noises include the irrelevant items in the web page such as banner ads, navigational guide, decoration pictures, etc. Then the graphics and video formats such as gif, JPEG, CSS, etc are removed. Then by checking the status code and method field, some records are resulted. Then the web robots are identified and two different techniques are followed to remove them. First, all records containing the name "robots.txt" must be removed. Next if the browsing speed is more than the threshold then those requests are removed. Thus finally some amount of records is obtained after applying robot cleaning process.

Authors performed preprocessing based on data cleaning. Just performing one step cannot guarantee the reliable results for other phases of WUM. Session identification is another very important technique at preprocessing level, which authors did not apply.

According to Wahab, et al., (2008) [5]**,** proper analysis of log file can be used for proper management of bandwidth and server capacity. Preprocessing step is complex and laborious job. Authors also discussed the various types of log file in detail. In addition to that authors also discussed 19 attributes of log file in detail.

The authors proposed an algorithm to read the log file from any of the three given log file formats and convert the log file data into a database. Through another algorithm, authors filtered out the all the un-interested attributes of web log file.

Authors performed an experiment and applied proposed algorithm on web log file. There were 18 attributes in web log file, out of which 17 attributes were dropped. Only *"URL"* attributes was declared interested. *Date, Time, IP Address,* and *User Agent* are some other useful attributes were also dropped. By dropping out such important attributes, the reliability of later phases of WUM cannot be secured. For later phases of WUM, only URL attribute is not sufficient. Authors also mentioned the future trends for grouping the similar data on semantic information. Weighing the pros and cons, we come to the conclusion that proposed algorithms for data cleaning and data filtering techniques are very weak and needs to be modified.

Chang-bin and Chen Li (2010) [9] state that the main function of web data preprocessing is to sort out the mutual records between users and websites into user sessions and transaction file. And also adds that the study of session identification is of practical significance. Authors used a collaborative filtering method to categorize the users based on their interests and preferences. Collaborative filtering algorithm contains the measure of user's similarity, inquiry of nearest-neighbors and predicting scores by means of K-nearest neighbor classifier. Main methods of measurement for user's similarity are cosine similarity, adjusted cosine similarity and person correlation coefficient.

Cosine Similarity:

$$sim(u,v) = \frac{\vec{u},\vec{v}}{|\vec{u}||\vec{v}|} = \frac{\sum_{c\in I_{u,v}} r_{u,v} r_{v,c}}{\sqrt{\sum_{c\in I_u}(r_{u,c})^2} \times \sqrt{\sum_{c\in I_v}(r_{v,c})^2}} \quad (1)$$

Adjusted Cosine Similarity:

$$sim(u,v) = \frac{\sum_{c\in I_{u,v}}(r_{u,c} - \vec{r_u}) \times (r_{v,c} - \vec{r_v})}{\sqrt{\sum_{c\in I_u}(r_{u,c} - \vec{r_u})^2} \times \sqrt{\sum_{c\in I_v}(r_{v,c} - \vec{r_v})^2}} \quad (2)$$

Person Correlation Coefficient:

$$sim(u,v) = \frac{\sum_{c\in I_{u,v}}(r_{u,c} - \vec{r_u}) \times (r_{v,c} - \vec{r_v})}{\sqrt{\sum_{c\in I_{u,v}}(r_{u,c} - \vec{r_u})^2} \times \sqrt{\sum_{c\in I_{u,v}}(r_{v,c} - \vec{r_v})^2}} \quad (3)$$

Some log data are randomly took from the web server log file as the data source and collaborative filtering algorithm is applied and the results are derived. Though it does session identification, clustering of the users with similarity could still improve the efficiency of the algorithm. Therefore it would be better if clustering is done.

Alam, et al., (2008) [7] states that web session clustering is an important WUM technique to predict the user access behavior. From the web log file, main attributes such as IP Address, Data, Time, Object requested, Page size, Response, Referrer were taken, and data cleaning was performed. Authors performed the web sessions clustering by applying Particle Swarm Optimization (PSO) algorithm based on time and browsing sequence dimensions. For PSO, sessions were taken as particles and ParticleID; DistanceFromEachSession; WonSessionVectors; SessionAttributeValuses; and PBest

attributes were used. To measure the distance between sessions, Euclidean Distance measure was used.

Euclidean Distance (ED) measure is used for numerical data. For categorical data, there are other similarity measures, which can produce batter results. Alam, et al., (2008) [7] performed the experiment and compare the results with K-mean. While PSO and K-Mean are different in nature and produce different results. Authors did not compare the results with any other PSO based session clustering. Furthermore, authors were also unable to give any other preprocessing techniques such as data cleaning, data filtering.

Clusters group the similar data in a natural way and reduce huge amount of data and hierarchical clustering can provide more information about the sessions. Authors did not perform hierarchical clustering as well.

Lu and Nguyen, (2009) [8] state that for web personalization, web session clustering play an important role in WUM. In this paper authors proposed the PSO based sequence clustering technique. Similarity measure for sequences clustering was defined as ratio of common items and unique items in two sequences then set the similarity in the order of occurrence of items in two sequences.

Furthermore, standard PSO was used with the similarity *S3M and* Total Benefit *TB* to cluster the web session data. In this research, only similarity measure was changed from Euclidean Distance to S3M and authors calculated the similarity based on the longest common sequence.

According to Nichele and Becker, (2006) [19] session similarity is major issue while clustering the web usage session data. The page similarity measure on the concept hierarchy is given in Eq 4.

$$sim(c_1, c_2) = \frac{2 \times depth(LCA(c_1,c_2))}{depth(c_1)+depth(c_2)} \quad (4)$$

Where c1and c2 are concept hierarchy and depth is number pages. After calculating similarity, algorithm is applied to find the session clusters.

According to Tasawar Hussain (2010) [10] web session clustering is an emerging and common technique at preprocessing level of WUM, which not only extract the hidden behaviour from web usage but also groups the sessions based on some common properties such as similarity. From the web log file, main attributes were taken, and data cleaning was performed. Then, user identification is performed where the distinct and unique users of website are identified. After that session identification is performed where a session i.e., a collection of activities of a single user is identified.

Further, authors performed a hierarchical sessionization based on PSO and used a new parameter user agent. To measure the similarity between the clusters similarity measures such as Angular Separation (AS) Eq 5, Canberra Distance (CD) Eq 6 and Spearman Distance (SD) Eq 7 is used instead of Euclidean Distance (ED) which was used by Alam, et al. [7].

$$s_{ij} = \frac{\sum_{k=1}^{n} X_{ik}.X_{jk}}{\left(\sum_{k=1}^{n} X_{ik}^2.X_{jk}^2\right)^{1/2}} \quad (5)$$

$$d_{ij} = \sum_{k=1}^{n} \frac{|X_{ik}-X_{jk}|}{|X_{ik}|+|X_{jk}|} \quad (6)$$

$$D_{ij} = \sum_{k=1}^{n} \left(X_{ik} - X_{jk}\right)^2 \quad (7)$$

Here, the velocity and position of particles are updated in each iteration and at last adapted agglomerative algorithm is applied to the winning sessions. Winning sessions are single input clusters to agglomerative and City Block (Manhattan) distance between two sessions is calculated. Next, the pair wise distance based on average linkage to link the pair of clusters to hierarchical clusters of sessions is calculated.

Table 1: **Comparison of Various Techniques**

Author Name	Technique	Algorithm
Nithya Sumathi	Data cleaning	NA
Wahab	Extraction Cleaning	A
Lu & Nguyen	Hierarchical sessionization	Particle Swarm Optimization
Tassawar Hussain	Data Cleaning Log File Filtering User Identification Session Identification	Agglomerative
Nichele & Becker	Session Identification	NA
Chang – bin Chen Li	Session identification	Collaborative Filtering
Theint Theint Aye	Extraction Cleaning	A

3. CRITICAL EVALUATION

In the previous section, we described the survey of literature on web log preprocessing of WUM. We found that data cleaning, data filtering, path completion, user identification, session identification, and web session clustering are the commonly used techniques at preprocessing level of WUM. We also observed that different parameters (attributes) from log are used in different preprocessing techniques. Commonly used attributes from web log file are IP Address, Date, Time, URL, and User agent.

In most of surveyed preprocessing techniques, web server log file is used as compared to client and proxy web log files. The research on client access log file is limited. Only Murata and Saito (2006) [1] performed their experiment on client log.

Different log files have different log attributes and different preprocessing techniques are based on these attributes. The selection of web log is independent of preprocessing technique being applied. Web log format is based on Web Server being used on which web site has been launched.

In data cleaning technique, irrelevant entries from web log are removed. This is widely used preprocessing technique and it has been performed as mandatory preprocessing technique.

In most of the session identification techniques, 30 minutes timeout was taken and transactions made by user with web site in 30 minutes are grouped as session. Stermsek., et al.,

(2007) [15] and Raju, and Satyanarayana, (2008) [16] followed the same strategy to identify the sessions as proposed by Yuan, et al., (2008) [3], 30 minutes timeout to identify the sessions. Alam, et al., (2008) [7] further enhanced the session identification added the data downloaded attribute from log file. For path completion technique, structure of web site is used (Castellano, 2007 [14], Yuan, 2003 [13]).

4. CONCLUSION AND FUTURE WORK

Preprocessing of web log file is most essential and must step for web usage mining. For pattern mining and pattern analysis cleaned data after preprocessing is concrete base. Quality of pattern mining is fully dependent on preprocessing step. In this survey, we summarized the existing web log preprocessing techniques. Server log file is considered most authentic source for web usage mining. So it must be standardized and needs to be updated to capture user access data. For future work we should explore preprocessing techniques and use them with the combination of existing techniques to make the whole process more robust. And more concentration should be made on session identification part of preprocessing where some new algorithms can be used to improve the quality of data.

In order to gain better understanding of log file we need hierarchical clustering by using proposed clustering technique. The user or data mining expert can have more knowledge of log file since impartial grouping of data exists. With this enhanced information, the web log user can be more focused.

5. REFERENCES

[1] Murata, T. and K. Saito (2006). Extracting User's interests from Web Log Data. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

[2] Pabarskaite, Z. (2002). Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining. 24th Int. Conf. information Technology Interfaces /TI 2002, June 24-27, 2002, Cavtat, Croatia.

[3] Yun, L., W. Xun, et al. (2008). A Hybrid Information Filtering Algorithm Based on Distributed Web log Mining. Third International Conference on Convergence and Hybrid Information Technology 978-0-7695-3407-7/08 © 2008 IEEE DOI 10.1109/ICCIT.2008.39.

[4] Suneetha, K. R. and D. R. Krishnamoorthi (2009). "Identifying User Behavior by Analyzing Web Server Access Log File." IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[5] Wahab, M. H. A., M. N. H. Mohd, et al. (2008). Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm. World Academy of Science, Engineering and Technology 48 2008.

[6] Stermsek, G., M. Strembeck, et al. (2007). A User Profile Derivation Approach based on Log-File Analysis. IKE 2007: 258-264.

[7] Alam, S., G. Dobbie, et al. (2008). Particle Swarm Optimization Based Clustering Of Web Usage Data. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 978-0-7695-3496-1/08 DOI 10.1109/WIIAT.2008.292 IEEE/WIC/ACM International Conference on Web.

[8] Lu. H. and Nguyen. T.T.S., 2009, "Experimental Investigation of PSO Based Web User Session Clustering", 2009 International Conference of Soft Computing and Pattern Recognition 978-0-7695-3879-2/09. IEEE DOI 10.1109/SoCPaR.2009.127

[9] JIANG Chang-bin and Chen Li., 2010, "Web Log Data Preprocessing Based On Collaborative Filtering", 2010 International Conference On Education Technology and Computer Science 978-0-7695-3967-4/10. © IEEE DOI 10.1109/ETCS.2010.588.

[10] Tasawar Hussain, Sohail Asgar and Nayyer Masood., 2010, " Hierarchical Sessionization At Preprocessing Level of WUM Based On Swam Intelligence", 6th International Conference on Emerging Technology (ICET) 978-1-4244-8058-6/10 .© 2010 IEEE

[11] Theint Theint Aye., 2011," Web Log Cleaning For Mining Of Web Usage Patterns", 978-1-61284-2/11 © 2011 IEEE.

[12] Nithya.P and Dr.P.Sumathi., 2012, " Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots" , 2012 National Conference on Computing and Communication Systems 978-1-4673-1953-9/12 © 2012 IEEE.

[13] Yuan, F., L.-J. Wang, et al. (2003). Study on Data Preprocessing Algorithm in Web Log Mining. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.

[14] Castellano, G., A. M. Fanelli, et al. (2007). LODAP: A LOg DAta Preprocessor for mining Web browsing patterns. Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Corfu Island, Greece, February 16-19, 2007.

[15] Stermsek, G., M. Strembeck, et al. (2007). A User Profile Derivation Approach based on Log-File Analysis. IKE 2007: 258-264.

[16] Raju. G. T. and Satyanarayana. P. S., 2008, "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", IJCSNS International Journal of Computer Science and Network Security, V|OL. 8 No. 1, January 2008.