Improved Preprocessing Techniques for Analyzing Patterns in Web Personalization Process

R. Gobinath
Department of Computer Science
Karpagam University
Coimbatore-641021, India

M. Hemalatha
Department of Computer Science
Karpagam University
Coimbatore-641021, India

ABSTRACT
Data preprocessing plays a vital role in Data Mining. In this paper we have adopted the concept of web based mining for cleansing the web server log files. Web mining extracts useful information of hypertext documents. Once a user access the web pages /sites their information are recorded in a file as an entry called log file. The web server log files are used for mining several useful patterns to analyze the access behavior of the user. Before performing the mining process the raw data has to be preprocessed in order to improve the quality of data to be mined.

This paper discusses about the significance of data preprocessing methods and various steps involved in getting the required content successfully. An entire preprocessing technique is being planned to preprocess the web log for extraction of user patterns. Data cleansing algorithm is applied to eliminate the extraneous entries from web log at the same time filtering algorithm is used to discard the impassive attributes from log file. The outlier are detected and removed from the dataset. The User and sessions are identified. The performance of the data cleansing process was evaluated by adapting the wrapper approach in which the resultant cleaned dataset are clustered using five different clustering algorithms namely Farthest First, K-means, COBWEB, make density based algorithm and Expectation maximization algorithm to identify the quality of web log data

Keywords
Web Mining, Field extraction, Data cleansing, User identification, Session identification, Server Log files.

1. INTRODUCTION
In the recent past few years the Internet has urbanized into the prime and most popular means of communication and information dissemination. The WWW provides a platform for exchanging various kinds of information. With the advent of Electronic Commerce and World Wide Web the volume of information available on the internet is rapidly increasing with the explosive growth. Web mining has turned into very imperative for official website management, creating adaptive web sites, business and support personalization, services, network traffic flow analysis and so on [1]. Web content mining is the process to discover useful information from the content of a web page. Fundamentally, the Web content consists of several types of data such as text, audio, metadata, image, video as well as hyperlinks. Web Structure Mining is the process of inferring information from the World Wide Web organization and links between references and referents in the Web. The arrangement of a typical web graph consists of web pages as nodes and hyperlinks as boundaries connecting related pages. Web Structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

Web Usage mining is the application of data mining techniques to discover usage patterns from web data. Data is usually collected from user's interaction with the web, e.g. web/proxy server logs, user queries, registration data [1]. Usage mining tools discover and predict user behavior, in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service. The whole procedure of using Web usage mining for Web recommendation consists of three steps they are data collection, pre-processing and pattern mining well as knowledge application. The preprocessing plays an important role in cleansing unclean data like duplications, graphical files and web crawler. These data can make a challenging task for performing personalization, load balancing in internet, conquers space in log files, etc. The basic preprocessing steps remain constant, and to test the quality of the dataset, various clustering methods can be used [18].

In this paper, we describe a solution to business brainpower to discover the hidden insight of their business and web data. We demonstrate how web mining technology can be effectively applied in business brainpower. The structure we propose takes the outcome of the web mining process as input, and converts these results into actionable knowledge, by elevating them with information that can be willingly interpreted by the business analyst.

2. RELATED WORK
Yan Li [2] presented a detail algorithm method for web usage mining implementation of the data preprocessing system. After identifying the user session, the referrer based method is used to find the user's access path which is attached with an effective solution to the problems with proxy servers and local caching. Hussain .T, Sahail Asghar [3] proposed in the preprocessing level framework for web session cluster of usage mining. It covers the steps to prepare the log data and it converted into numerical data. Doru Tanasa [4] the research describes two main contributions to WUM process (i.e.) for preprocessing the web logs and a divisive with three approaches for the discovery of sequential patterns with a support. The algorithm used for the processing the web log records and obtaining the set of frequent access patterns have been implemented by huiping Peng[5].

An improved preprocessing expertise has been used by ling Zheng [6] for the purpose of solving some existing issues in traditional information preprocessing in web log mining. User identification follows the strategy based on the referred web page for identifying the user. JIANG Chang-bin and Chen Li [7] says that, even if the statistical data are not enough and absence of visiting user history records, the web log data preprocessing algorithm based on collaborative filtering identifies the session flexibly and quickly.

O.R. Zaiane [8] deals with the behavior of the users may attempt to change the frequent pattern from the analysis of a log file which makes the challenge.

The paper [9] focuses on WUM with the interaction behavior of web users and requested web pages to identify navigation patterns. Juan Velasquez Hiroshi Yasuda and Terumasa Aoki [10] in this paper the author says that how to study the visitor behavior on a Web site, based on web content and web usage mining. In web mining to test the data an approximation is supposed to be used that depends on the content, navigation sequence and time spent on each page visited by the user. Carlos G. Marquardt, Karin Becker Duncan D. Ruiz [11] proposed the impact of developing the WUM preprocessing phase according to concepts, problems and goals specific to Web-based learning environments. Xin Jin, Yanzan Zhou, Bamshad Mobasher [12] identified a unified framework for the discovery and analysis of Web navigational patterns based on Probabilistic Latent Semantic Analysis (PLSA).

3. PROPOSED METHODOLOGY

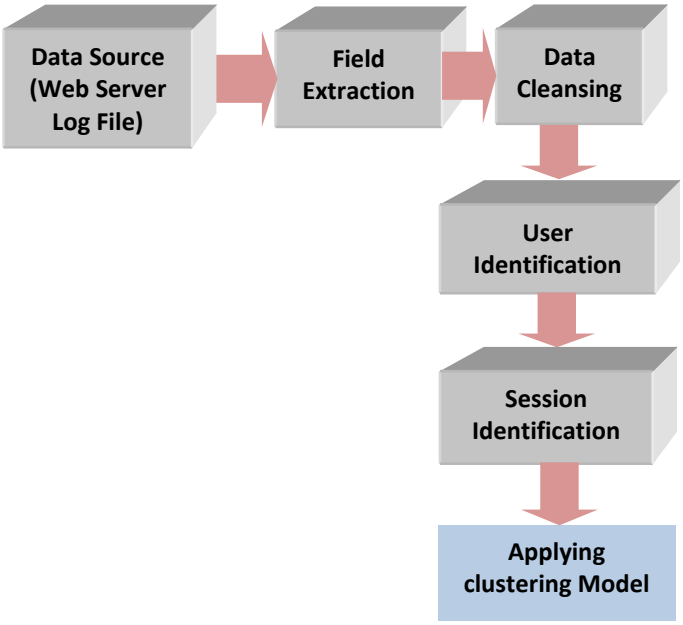

Figure 1: General flow of the preprocessing stage

The Data source consists of a Web server log file used for identifying the pattern of web usage by the user. Using the field extractor the attributes are separated using the delimiters. Then the data cleansing process is applied for filtering the unwanted and irrelevant data entry to increase the quality of data. The cleaned data are then used for user identification, session identification and clustering technique. The traditional web based models where the method followed by various researchers for preprocessing.

This paper shows the new method algorithm for data preprocessing to solve some of the existing problems in traditional web based method. Delimiter based Field Extraction algorithm is used in the field extraction method. The Distinct User Identification strategy of identifying the user from the referred web page is differing from that of the traditional web based method. Time Oriented Heuristic Algorithm for Session Identification which categorizes the users by session also improved from the traditional web based method. The experimental results clearly show the quality of the data can be improvised by following the improved preprocessing techniques for analyzing instead of using the traditional web based method.

3.1 Data Sources

The data collected for usage mining is from diverse sources which represents the navigation patterns of various segments of the overall web traffic. Web server log does not exactly contain adequate information for inferring the behavior on the client side as they relate to the pages provided by the web server.

3.1.1 Web log file

Web server log file records the information about each user in a simple plain text file. This file contains information about user name, IP address, date, time, bytes transferred, access request. Each time when a user requests a resource from a particular web site the concern web server writes the information of the request in a web log file.

When a user submits a request to a web server that activity are recorded in the web log file. Log file ranges 1KB to 100MB

Example of a web log file

```
ip2283.unr - - [16/Nov/2005:00:01:03 -0500] "GET /dmcourse/dm.css HTTP/1.1" 200
155 "http://www.kdnuggets.com/ dmcourse/data_mining_course/ assignments/
assignment-3.html " "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
```

a) ip1664.com: host name that can be converted to the IP Address.

b) - : The name of the remote user

c) - : Login of the remote user. Both name of remote users and login of remote user usually omitted and replaced by a dash "-"

d)[16/Nov/2005:00:01:03 -0500]: Date: DD/Mon/YYYY, Time: HH:MM:SS Time Zone: (+|-) HH00 relative to GMT - 0500 is US EST

e)" GET /dmcourse/dm.css HTTP/1.1" : Method: GET and POST are methods . URL: relative to domain,HTTP protocol: protocol version e.g. HTTP/1.0 or HTTP/1.1

f)200: Status code for successful.

Classes of Status code

Success – 200

Redirect – 300

Failure – 400

Server error – 500

g)155: size of the object returned to the client, in bytes Can also be "-" if status code is 304 (not modified)

h)http://www.kdnuggets.com/ dmcourse/data_mining_course/ assignments/ assignment-3.html URL the visitor came from the website kdnuggets.com in that they navigated to assignment 3.html by start navigating from dmcourse, data

_mining _course, assignments .The referrer can also be a static page, internal (same domain) or external (different domain).

i) Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)" Almost all browsers start with Mozilla. Browser type, version: MSIE 6.0 - Internet Explorer 6.0, OS: window NT 5.1 with .NET framework 1.1 installed.

3.1.2 Log file format

Web log file is a simple plain text file which record information about each user. Display of log files data in three different formats

- W3C Extended log file format
- NCSA common log file format
- IIS log file format

NCSA and IIS log file format the data logged for each request is fixed. W3C format allows user to choose properties, the user wants to log for each request.

3.2 Data Pre-Processing

In Web page personalization to perform web usage mining preprocessing is necessary, because Log file encloses noisy and unclear data which may affect results of the mining process. Some of the web log file data are unnecessary for analytical process and could affect the performance of personalization.

Data preprocessing is an important step to filter and organize the appropriate information before applying any web mining algorithm. Preprocessing increases the quality of available data by reducing the log file size. The primary use of data preprocessing is to improve data quality and increase mining accuracy. The Preprocessing consist of following steps

- Field extraction
- Data cleansing
- User identification
- Session identification

In this paper main task is to ”clean” the raw web log files and apply the clustering technique to find the quality of log file data and pattern analysis. So the main steps of this phase are:

1) Extract the web logs that collect the data on the web server.
2) Clean the web logs and remove the redundant information.
3) Applying the five different clustering Techniques for finding the quality of data of log files

- The raw web log data after pre-processing and cleansing could be used for pattern discovery, pattern analysis, web usage statistics, and generating association/ sequential rules.

3.2.1 Field Extraction

Each user entry is represented as a single line of the log file. The log entry contains many fields as discussed in the earlier section which has to be separated out for further processing. The filed extraction is the process of separating the field from the single line of the server log file. The server used different characters which work as separators. The most used separator character is ',' or 'space' character.

Delimiter based Field Extraction algorithm is given below.

Input: Log File

Output: DB

Open Log File

Read all fields contain in Server Log File

Separate out the Attribute using the delimiter

Extract all fields and Add into the Server Log Table (SLT)

Close

An example of Server Log Table (SLT) is shown in Table 1. Each and every section of the log entries has been separated to attributes for easy cleansing of unclean data form huge data source. The detailed description of the letters mentioned in the table is given below the table.

3.2.2 Data cleansing

This stage consists of removing the entire data track in Web logs that are ineffective for mining purposes. Graphic file requests, agent/spider crawling etc. could be simply removed by only seeming for an HTML file requests. Normalization of URL's is regularly required to make the requests consistent.

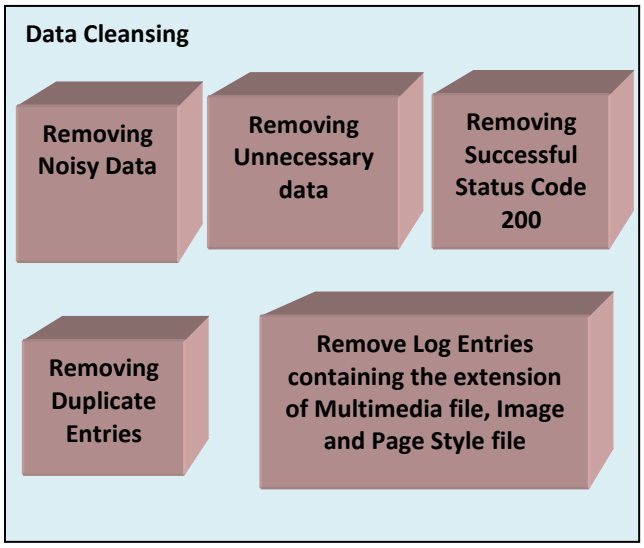

Figure 2: Data cleansing process constrains

IP	RU	RUL	S	URL	SC	SOR	R	B &OS
ip1664.com	-	-	[16/Nov/2005:00:00:43 -0500]	"GET /gpspubs/sigkdd-kdd99-panel.html HTTP/1.0"	200	14199	-	"msnbot/1.0 (+http://search.msn.com/msnbot.htm)"
ip1115.unr	-	-	[16/Nov/2005:00:01:00 -0500]	"GET /news/99/n23/i12.html HTTP/1.1"	200	3171	"http://discount-blah1.professional-doctor.com/"	"Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; SAFEXPLORER TL)"

Table 1: Server Log Table (SLT)

[IP-Internet Protocol, RU-Remote User, RUL-Remote User Login, S-Session, URL- Uniform Resource Locator, SC- Status Code, SOR- Size of the object returned, R- Referrer, B & OS- Browser and Operating System.]

In Data Cleansing the log entry involves the irrelevant references to embedded objects like multimedia files which may not be necessary for analyzing purposes. Therefore such kind of useless entries has to be removed from log files before performing any analysis process. By performing Data cleansing process, errors and discrepancy will be discovered and removed to improve the quality of data In the Data cleansing process following steps are performed

Remove Noisy and Unnecessary data

- Remove log entry nodes contain extension like jpg, gif means remove request such as multimedia files, image, page style file
- Remove successful status code 200.
- Remove Duplicate Entries

Algorithm for Data Cleansing

Read Entries in SLT

For each Entry in SLT

Read fields (Status code, method)

If Status code=‘200’and method= ‘GET’

Then

Get IP_address and URL_link

If suffix.URL_Link= {*.gif,*.jpg,*.css} Then

Remove suffix.URL_link

Save IP_sddress and URL_Link

End if

Else

Next Entry

End if

3.2.3 User Identification

In this stage the individual user is identified using their IP address. While reading the entry in the sever log table if the IP address is new then consider it as new user. If the IP address already exists but either the browser or operating system differs then it is also considered as different users. The algorithmic representation for the User Identification is given bellow.

Distinct User Identification algorithm using Server Log Table (DUI)

Read each entry in SLT

If an IP address not exist then

Consider the user as new user

End if

If IP address exists and the ((browser version or Operating System) is not exist) then

Consider the user entry as new user

Elseif

Next entry

3.3 Session Identification

The time duration spent on web pages are called Session. To identify the new session a referrer-based method is used. When the IP address, browser version and operating system are same the referrer information should be taken. A new user session is identified if the URL in the Refer URI – field is a

larger interval usually more than 30 minutes between the accessing times on this record.

Time Oriented Heuristic Algorithm for Session Identification (TOH)

For each entry in SLT
Sort the log data by IP address , agent and time
Next entry
Read each entry in SLT
If the IP address and agent not exist then

If (requested_time$_i$ - requested_time$_{i-1}$) > Session Time Out or Session time-out does not belong to Session history then
Increment the value of session by 1
Consider the user session as new
Endif

Endif

Next entry

After performing Data cleansing of the raw data set the quality of the cleaned data was validated with the existing clustering models namely Partitioned clustering, Hierarchical methods, Density based clustering and Sub Space Clustering. The results show on applying these models the quality of the datasets was highly improvised. So the obtained cleaned datasets will be used for further personalization process.

4. CLUSTERING MODELS

The similar data items were grouped and categorized these phenomena is called as clustering. The absence of class label which makes clustering as unsupervised learning technique. The successive clustering methods can be found by Hierarchical algorithms using earlier established clusters. These algorithms are commonly followed the bottom-up or top-down approach. Bottom-up approach algorithms initial element is clustered separately and merging as larger clustering. Top-down algorithm approach is to start with the complete set and division of smaller clustering have been made [18].

4.1 Farthest First Algorithm(FFA)

Farthest First algorithm [13] is the variant picks up the center of cluster randomly next to the point furthest from the first center of the cluster. The third point is furthest most from the previous cluster centered points. The next center is argmaxx Minc d (x, c). The points should fall within the data area. This process can speed up the clustering since adjustment and reassignment are needed. Approximation algorithm where created by Hochbaum and Shmoys [14] from the farthest-first traversal of a data set. This construction is known to be k-center problem that can perform the finding of k-clustering optimal solution using the cost function, with the cluster radius high. The outcomes of clusters between two points are optimal, if the distance function is a metric. The proper understanding of hierarchical clustering can be made only after understanding of the farthest-first traversal in detail.

4.2 K-mean

In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm, particularly in the computer science community. Initially, the number of clusters must be known, or chosen, to be K say. The initial step is the chosen a set of K instances as centers of the clusters. Often chosen such that the points are mutually "farthest apart", in some way. Next, the algorithm considers each instance and assigns it to the cluster which is closest. The cluster centroids are recalculated either after each instance assignment, or after the whole cycle of re-assignments. This process is iterated [15].

4.3 COBWEB

COBWEB [16] is an incremental system for hierarchical conceptual clustering. COBWEB incrementally organizes observations into a classification tree. Each node in a classification tree represents a class (concept) and is labeled by a probabilistic concept that summarizes the attribute-value distributions of objects classified under the node. This classification tree can be used to predict missing attributes or the class of a new object. There are four basic operations COBWEB employs in building the classification tree. Which operation is selected depends on the category utility of the classification achieved by applying it. The operations are:

i. Merging Two Nodes IT merging two nodes means replacing them by a node whose children is the union of the original nodes' sets of children and which summarizes the attribute-value distributions of all objects classified under them.

ii. Splitting a node A node is split by replacing it with its children.

iii. Inserting a new node a node is created corresponding to the object being inserted into the tree.

iv. Passing an object down the hierarchy IT effectively calling the COBWEB algorithm on the object and the sub tree rooted at the node.

4.4 Make Density Based(MDB) algorithm

The cluster will be constructed based on the density properties of the database are derived from a human natural clustering approach. The clusters and consequently the classes are easily and readily identifiable because they have an increased density with respect to the points they possess. The elements of the database can be classified in two different types: the border points, the points located on the extremities of the cluster, and the core points, which are located in its inner region.

Based on the thickness properties of the database the cluster can be built by undergoing normal human clustering approach. The identification of the classes in the cluster can be easily processed from the points of the cluster from the increasing bulk of the cluster. The elements in the database can be separated into two types, they are:

- Border points
- Core points

Border points are the points which are present furthest point of the cluster and core points are the points which are present inside the inner region.

4.5 Expectation Maximization (EM)

Clustering performs division also by expectation maximization method. It is one of the important Data mining algorithms. This EM algorithm follows a well formalized statistical method including some ideas of class membership in partial [17]. Usually the EM algorithm in prescribing to fuzzy c-means. This EM algorithm is used only after satisfying k-mean method. The EM algorithm can also accommodate categorical variables. The technique will at first randomly assign different probabilities to each class or section, for each cluster. In successive iterations, these probabilities are refined to maximize the likelihood of the information given the specified number of clusters. The result gathered from EM algorithm is different from the k-means clustering method. After this observation are assigned to maximize the cluster distance. In other words, each observation belongs to each cluster with a sure probability. Of work, as a final result you can usually review an actual task of observations to clusters, based on the classification probability. The EM algorithm extends this basic approach to clustering in important ways:

i. The alteration for observing the maximum difference in means for continuous variables of the cluster, the EM clustering algorithm computes probabilities of cluster memberships based on or more probability distributions. The aim of the clustering algorithm then is to maximize the general probability or the likelihood of the information, given the clusters.

ii. The EM algorithm can be applied to categorical variables and continuous variables, where k-means implementation differs from EM algorithm in modification of accommodate categorical variables.

5. EXPERIMENTAL RESULTS

In this study, we have analyzed the log files of Web server of Apache Combined Log entry format with the help of rapid miner. The cleaned dataset is used for performing the clustering model and the result shows that the performance of Make density based algorithm works fine than the remaining algorithms in the proposed DUI and TOH model.

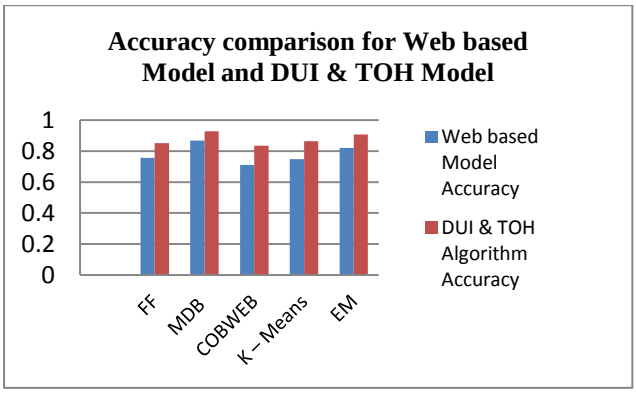

Figure 3: Accuracy Comparison for Web Based and DUI&TOH Model using Clustering models.

The data set is performed with the traditional preprocessing based on website topology with the proposed DUI and TOH model. The original data size is 39974 Instances after applying the preprocessing process of proposed model such as removing duplicates, robotic files and image files it was reduced to the percentage of 50.2. The number of users identified was 2307, the number of sessions identified was 1787 and the Accuracy 92.7. When we consider the same data set in traditional based web topology models, the record cleaned log files are 14678, the percentage of reduction is 36.7, number of users identified is 1508, number of sessions identified is 678 and the Accuracy is 86.8 which is much less compared to the proposed model.

Web log file information	Web based Model	DUI & TOH Algorithm
Record in original web log file	39974	39974
Records in cleaned log file	14678	20065
No. of. Users Identified	1508	2307
No. of. Session Identified	678	1787
Reduction (%)	36.7	50.2
Accuracy (%)	86.8	92.7

Table 3: Web log file result for Web Based and proposed DUI&TOHalgorithm.

Clustering Models	Web based Model Accuracy	DUI & TOH Algorithm Accuracy
Farthest First	75.67%	85.09%
Make Density Based	86.89%	92.73%
COBWEB	71.06%	83.56%
K – Means	74.87%	86.32%
EM algorithm	82.05%	90.78%

Table 2: Clustering Model results for Web Based and DUI&TOH algorithm accuracy.

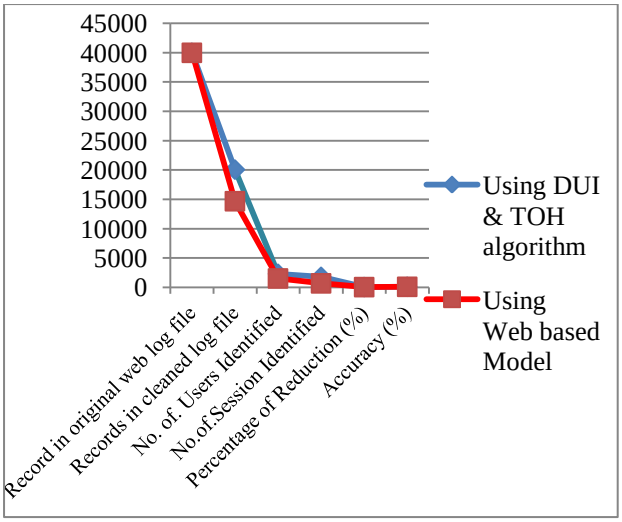

Figure 4: Data Graph for Web Based and DUI&TOH algorithm.

Web log file information	Web based Model	DUI & TOH Algorithm
No. of. Users Identified	1508	2307
No. of. Session Identified	678	1787

Table 4: Comparison for Web based and DUI&TOH algorithm in User and Session Identification.

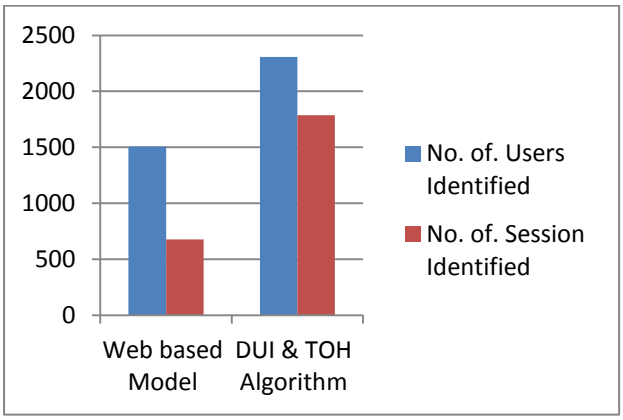

Figure 5: User and Session Identification improvement graph

6. CONCLUSION

We focus on web log file format, its type and location. This web log file records information of each user request. Log files usually contain noisy and ambiguous data. Preprocessing involves removal of unnecessary data from log files. Log file used for debugging purpose. Data preprocessing is an important step to filter and organize appropriate information before using web mining algorithm. This paper presents algorithms for user Identification and session Identification. Preprocessing web log file is used in Web usage Mining and for personalization techniques. The cleaned log file entries are processed by the clustering techniques to perform pattern analysis and to identify the quality of data.

7. REFERENCES

[1] Pirolli, P., Pitkow, J., and Rao, R. 1996. Silk from a Sow's Ear: Extracting Usable Structures from the Web. In Proceedings on Human Factors in Computing Systems, ACM Press, pp. 118-125

[2] Yan LI, Boqin FENG and Qinjiao MAO.2008. Research on Path Completion Technique in Web Usage Mining. IEEE International Symposium On Computer Science and Computational Technology, pp. 554-559.

[3] Hussain, T., S. Asghar, et al. 2010. Web Usage Mining: A Survey on Preprocessing of Web Log File. IEEE, International Conference on (ICIET),pp. 1 - 6

[4] Doru Tanasa and Brigitte Trousse. 2004. Advanced Data Preprocessing for Intersites Web Usage Mining. Published by the IEEE Computer Society, pp. 59-65.

[5] Huiping Peng. 2010. Discovery of Interesting Association Rules Based On Web Usage Mining. IEEE conference, pp. 272-275.

[6] Ling Zheng, Hui GUI and Feng Li. 2010. Optimized Data Preprocessing Technology For Web Log Mining. IEEE International Conference On Computer Design and Applications (ICCDA), pp. 19-21.

[7] JING Chang-bin and Chen Li. 2010. Web Log Data Preprocessing Based On Collaborative Filtering. IEEE 2nd International Workshop on Education Technology and Computer Science, pp. 118-121.

[8] Zaiane, Web, O., R. 2001. Usage mining for a better web-based learning environment, proceeding of Conference on Advanced Technology for Education, pp. 450-455.

[9] Leticia dos Santos Machado, Karin Becker. 2003. Distance Education: a Web Usage Mining Case Study for the Evaluation of Learning Sites,In Proceedings of ICALT,pp. 360-361

[10] Juan Velasquez., Hiroshi Yasuda and Terumasa Aoki. 2003. Combining the web content and usage mining to understand the visitor behavior in a web site. In proceeding of: Data Mining, Third IEEE International Conference.

[11] Carlos G. Marquardt, Karin Becker Duncan D. Ruiz, 2004. A Preprocessing Tool for Web Usage Mining in the Distance Education Domain, Proceedings of the International Database Engineering and Applications Symposium, pp. 78 - 87

[12] Xin Jin, Yanzan Zhou, Bamshad Mobasher. 2004. Web Usage Mining Based on Probabilistic LATENT Semantic Analysis, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 197-205

[13] Sanjoy Dasgupta , 2005. Performance guarantees for hierarchical clustering, Journal of Computer and System Sciences - Special issue on COLT 2002: 70 (4) PP. 555-569.

[14] Hochbaum and Shmoys, 1985.A Best Possible Heuristic for the k-center Problem, Mathematics of Operations Research: 10 (2) PP.180-184.

[15] Hewijin Christine Jiau., Yi-Jen Su.,Yeou-Min Lin and Shang-Rong Tsai,,2006. "MPM: a hierarchical clustering algorithm using matrix partitioning method for non-numeric data", J Intell Inf Syst (2006) 26: pp.185–207.

[16] Cheng, Y., &, Fu, K. (1985). Conceptual clustering in knowledge organization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 7, pp. 592-598.

[17] Shantakumar B.Patil., Y.S.Kumaraswamy.,2009 "Warehouses for Heart Attack Prediction",International Journal of Computer Science and Network Security, 9(2-),pp. 228-235

[18] Cooley, R., Mobasher, B., and Srivastava, j. 1999.Data preparation for mining World Wide Web browsing patterns, journal of knowledge and Information Systems, 1 (1).

[19] Buchner, A. And Mulvenna, M.D.1999. Discovering Internet marketing intelligence through online analytical Web usage mining, SIGMOD Record. 4(27).pp.27-35.

[20] B. Mobasher, R. Cooley, J. Srivastava.2000,Automatic Personalization Based on Web Usage Mining, Communications of the ACM, 43(8).PP, 142-151.

[21] Ling Zheng, Hui Gui and Feng Li, ,2010 " Optimized Data Preprocessing Technology For Web Log Mining", IEEE International Conference On Computer Design and Applications(ICCDA), pp. 19-21.

AUTHORS PROFILE

R. Gobinath, Pursuing Ph.D Research in Computer Science, under the guidance of Dr. M. Hemalatha, Professor and Head, Dept. Software System, Karpagam University, Coimbatore, Tamilnadu. He has completed MCA degree in Anna University and has completed a Bachelor of Computer Science in Bharathiar University. Area of Research is Data mining.

Dr. M. Hemalatha completed MCA., M. Phil., PhD in Computer Science and currently working as a Professor and Head, Dept. Software Systems Karpagam University. Twelve years of Experience in teaching and published more than ninety papers in International Journals and also presented more than eighty papers in various national conferences and international conference. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several National and International Journals