A NewClustering and Preprocessing for Web Log Mining

1B.Uma Maheswari, 2 Dr. P.Sumathi
1Doctoral student in Bharathiyar University, Coimbatore ,Tamil Nadu, India
2Asst. Professor, Govt. Arts College, Coimbatore, Tamil Nadu, India
1.umasharan7@gmail.com, 2.sumathirajes@hotmail.com

***Abstract*---World Wide Web is a massive repository of web pages and links. It provides information about vast area for the Internetusers. There is tremendous growth and development ininternet. Users' accesses are documented in web logs. Web usage mining is application of mining techniques in logs. Sincedue to tremendous usage, the log files are growing at a faster rate and the size is becoming huge. Preprocessing plays a vital role in efficient mining process as Log data isnormally noisy and indistinct. Reconstruction of sessions and paths are completed by appending missing pages in preprocessing. Additionally, the transactions which illustrate the behavior of users are constructed exactlyin preprocessing by calculating the Reference Lengths of user access by means of byte rate.Using Web clustering several types of objects can be clustered into different groups for various purposes.By using the theory of distribution in Dempster-Shafer's theory, the belief function similarity measure in this algorithm adds to the clustering task the ability to capture the uncertainty among Web user's navigation performance. This paper experiments about the accomplishment of preprocessing and clustering of web log.The experimental result shows theconsiderable performance of the proposed algorithm.**

Keywords---Preprocessing, DataCleaning, Dempster-Shafer, Clustering

I.INTRODUCTION

In this internet era, web sites on the internet are useful source of information in every day life. Therefore there is an enormous development of World Wide Web in its volume of traffic and the size and complexity of web sites. As per August 2010 Web Server survey by Net craft there are 213,458,815 active sites. Web mining is the application of data mining, chart technology, artificial intelligence and so on to the web data and identifies user's visiting behaviors and extracts their interests using patterns. Due to its usual application in Web analytics, e-learning, e-commerce, information retrieval etc., web mining has become one of the important areas in computer and information science. The application of Web Usage Mining techniques in log data to extract the behavior of users which is used in variety of applications like pre fetching, creating attractive web sites, personalized services, adaptive web sites, customer profiling, etc.

Web servers gathers data about user's interactions in log files whenever requests for resources are received. Log files records information such as client IP address, URL requested etc., in different formats such as Common Log format, Extended Common Log format which is issued by Apache and IIS.

Web usage mining includes three main steps: Data Preprocessing, Knowledge Extraction and analysis of extracted results. Preprocessing is an important step because of the complex nature of the Web architecture which takes 80% in mining process. The raw data is pretreated to get reliable sessions for efficient mining. It includes the domain dependent tasks of data cleaning, user identification, session identification, and clustering and construction of transactions. Data cleaning is the task of removing irrelevant records that are not necessary for mining. User identification is the process of associating page references with same IP address with different users. Session identification is breaking of a user's page references into user sessions. Path completion is used to fill missing page references in a session.Classifications of transactions are used to know the users interest ad navigational behavior. The second step in web usage mining is knowledge extraction in which data mining algorithms like association rule mining techniques, clustering, classification etc. are applied in preprocessed data. The third step is pattern analysis in which tools are provided to facilitate the transformation of information into knowledge [4]. Knowledge query mechanism such as SQL is the most common method of pattern analysis.

This paper focuses on data cleaning and session identification process which is used to append lost pages and construction of transactions in preprocessing stage. In this study a referrer-based method is presentedfor efficiently constructing the reliable transactions in data preprocessing. And this result is given to the clustering process which uses Dempster's rule of combination. The refined groups of pages are common user profiles we want. This theory is appropriate for clustering analysis because it provides an aggregation operator, Dempster's rule for merging evidence, which allows the expression of the uncertainty with respect to aggregated components.The paper is organized as follows. Section II discusses the existing work. Section III deals with the preprocessing. Section IV deals with the probability assignment and Section V gives algorithm of clustering .Experimental results are given in Section IV. Finally summary of work is given in Section V.

I. RELATED WORK

Log data differs from other datasets used in data mining, and there are various problems which must be addressed in preparation for data mining. The main problem is to get a reliable dataset for mining. Therefore the data should be pretreated and users' accessing behavior is to be constructed as transactions. These transactions are to be reliable.

The Common log formats or Extended Log Formats only records the visitors browsing activities rather than the details of the visitor's identity. This means that different visitors sharing the same host cannot be differentiated. If there are proxy servers the problem became much severe. Users are identified easily by using Cookies or authentication mechanism. But users are not attracted by these types of sites due to privacy concerns[9].There are two heuristics for the attribution of requests to different visitors.

If two records have varied IP address, then they are distinguished as two different users else if both IP address are same then User agent field is checked. If the browser and

978-1-4799-2877-4/14 $31.00 © 2014 IEEE
DOI 10.1109/WCCCT.2014.67

information of operating system's user agent field is different in two records then they are identified as different users. After users are identified the next step is identification of sessions. A session is a sequence of activities made by one user during one visit to the site.

There are three heuristics available to identify sessions from users. Two heuristics are based on time and other is based on the navigation of users through the web pages.

Time Oriented Heuristics: These are the simplest methods in which one method is based on total session time and the second is based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes [2] to 24 hours [10] while default time is 30 minutes by R.Cooley [9]. The next method is based on the page stay time which is calculated with the difference between two timestamps. If the time exceeds 10 minutes the second entry is taken as a new session.

Navigation Oriented Heuristics: This method uses web topology in graph format. It takes webpage connectivity, but it is not necessary to have hyperlink between consecutive page requests.

Because of proxy servers and cached versions of the pages used by the client using 'Back', the sessions identified have many missed pages. So path completion step is carried out to identify missing pages. Referrer based methods are used to append the missing pages.

After session construction transactions are identified. A transaction is defined as a set of homogenous pages that have been visited in a user session. The transaction identification process depends on a split and merges process in order to look for a appropriate set of transactions that can be used in data mining.

Density-based algorithms: It starts by searching for core objects, and they are growing the clusters based on these cores and by looking for objects which are in a neighborhood within a radius ^{2} of a given object. The advantage of these types of algorithms is that they can identify arbitrary form of clusters and it can filter out the noise. DBSCAN [20] and OPTICS [21] are density-based algorithms.

Grid-based algorithms: This algorithm uses a hierarchical grid structure to decompose the object space into finite number of cells. For every cell statistical information is accumulated about the objects and the clustering is achieved on these cells. The advantage of this approach is the fast processing time that is in commonly independent of the number of data objects. Grid-based algorithms are CLIQUE [16], STING [15], and Wawe Cluster [17].

Model-based algorithms utilize different distribution models for the clusters which should be verified during the clustering algorithm. A model-based clustering method is MCLUST [18].

Fuzzy algorithms deduce that refusal hard clusters exist on the set of objects, but one object can also be assigned to more than one cluster. The best known fuzzy clustering algorithm is FCM (Fuzzy CMEANS) [19].

II. PREPROCESSING

Preprocessing of Web log data is a complex process and takes 80% of total mining process. Log data is pretreated to get reliable data. The aim of data preprocessing is to select necessary features clean data by removing irrelevant records and finally transform raw data into sessions. There are four steps in preprocessing of log data.

A. Data Cleaning

The data cleaning main process is removal of outliers or irrelevant data. Analyzing the huge amounts of records in server logs is a cumbersome activity. Therefore initial cleaning is necessary. If a user requests a specific page from server entries like gif, JPEG, etc., are also downloaded which are not useful for further analysis are eradicated. The records with failed status code are also eradicated from logs. Programs automated such as web robots, spiders and crawlers are also to be removed from log files. Thus removal process in our experiment includes

1. If the status code of each and every record when it is lesser than 200 and greater than 299 then those individual records are removed.

2. The cs-stem-url field is checked for its extension filename. If the filename has jpg, JPEG, CSS,gif andmuch more then they are removed.

3. The records which request robots.txt are removed and if the time taken is very less as like 2 seconds which can be considered as automated programs traversal and they are also removed [5].

B. Computing the Reference Length

The time taken by the user to view a particular page is called as Reference Length [13]. This plays an important role in the following procedures. Usually it is calculated by the dissimilarity between access time of a record and the next record. But this is not correct since the time includes data transfer rate over internet, launching time to play audio or video files on the web page and much more. The user's actual browsing time is complex to analyze. The data transfer rate and size of page is also considered and the reference length is calculated as

$RL_{time} = RLT' - bytes\_sent / c$

Where RLT' denotes the difference of access time between a record and the next one and bytes_sent is taken from log entry of a record and c is the data transfer rate.

C. User Identification

The log file after cleaning is considered as Web Usage Log Set

WULS = {UIP, Date, Method, URI, Version, Status, Bytes, ReferrerURL, BrowserOS}. The next important and complex step is unique user identification. The difficulty is due to the local cache and proxy servers. To overcome this cookies are used. But users may disable cookies. [8] Another solution is to collect registration data from users. But users neglect to give their information due to privacy concerns. So majority of records does not contain any information in the user-id and authentication fields. The fields which are useful to find unique users and sessions are

- IP address
- User agent
- Referrer URL

Users and sessions are identified by using these fields as follows. Iftwo records has same IPaddress check for browser information. If user agent value is same for both records then they are identified as from same user.

D. Session Identification

The goal of session identification is to partition the page accesses of each user into individual sessions. These sessions are used as data vectors in various prediction, classification, clustering into groups and other tasks. If the URL in referrer URL field in present record is not accessed previously or if referrer URL field is empty then it is considered as a new user session. Reconstruction of accurate user sessions from server access logs is a confronting task and time oriented heuristics with a time limit of 30 minutes is followed.

From WULS, the set of user sessions are extracted as referrer based method and time oriented heuristics.

$$USS = \{USID, (\mathrm{URI}_1, Referrer\mathrm{URI}_1, \mathrm{Date}_1) \ldots\ldots (\mathrm{URI}_k, Referrer\mathrm{URI}_k, \mathrm{Date}_k))\},$$

Where1 $\leq k \leq n$ and n denotes the amount of records in WULS. Every record in WULS must belong to a session and every record in WULS can belong to one user session only.

III.Basic Probability Assignment for Each User

Basic probability assignment (bpa) is appointed to every user. Once data preprocessing, we discover that some sessions from identical user will overlap as a result of a user might perform identical task in several sessions. A probability is appointed to every unique session once data preprocessing; it is the fraction of this unique session to the entire range of user sessions. This probability measures however likely the user can perform the tasks known in the unique session. The entire probability has measure one. It offers an enormous image of what the user typically will, likewise as however usually she will it within the web site. This assignment is cheap since it captures the uncertainty among visits to distinct pages. In our observation, session by itself could be a semantically meaningful unit. It represents one or many tasks users tend to perform in one visit. Users typically have to be compelled to browse a group of pages, instead of a single page, to accomplish one task. Therefore, assigning a probability to a group of pages looks to suit perfectly the semantic which means of session.

IV. User Profile Clustering Algorithm

In order to further Clustering analysis to mine the Web it is quite different from traditional clustering due to the inherent difference between Web usage data clustering and classic clustering. In this process the users are clustered based on their profile.

A. Belief Function as Similarity Measure

Suppose m (A_i), m (B_j), are two bpas for two users, A and B. We also use A, B to represent the set of unique content pages in the user's profile, respectively. Then definition is made as *bel* (A) as the total belief that user A's profile can represent user B's profile:

$$bel(A) = \sum_{B_i \subseteq A} m(B_i)$$

In some cases, if B is contained in A, bel(A) = 1. However the reverse is not true. So the similarity between A and B is defined as:

$$sim(A, B) = \min\left(\sum_{A_i \subseteq A} m(A_i), \sum_{B_i \subseteq A} m(B_i)\right)$$

this measures the similarity between two user profiles.

B. Greedy Clustering Using Belief Function (GCB)

A GCB algorithm for this clustering task using the similarity measure is defined in belief function. The greedy technique has been widely used in many algorithms as an efficient and effective way to approach a goal. In this process representatives of the clusters are done iteratively, so thispresent representative is totally separated from those that have been done in literature. An outline of the algorithm follows:

```
Input: K- number of clusters; S- a simple set of users,
Output: M- set of cluster representatives
 begin
M = { ·}
//select random users  m1 into the common profile set
M = { m1}
For each user profile x̣ S – M, calculate the distance
between x and m1
Dist(x) = -ln (sim (x, m1))
For i=2 to K
begin
    //choose representative m1 to be far from previous
    representatives
     Let m1 ∈ S − M, such that dist(m1) = max(dist(x)|x∈ S-
    M)
        M = M*{mt}
     // Update the similarity of each point to the closest
representatives
            for each x ∈ S-M
                dist(x) = min (dist(x), -ln(sim(x, m1)))
            end
    return M                    // M will contain a set of distinct
    cluster representatives
    end
```

C. Common User Profile Creation

A user is allocated to the cluster whose representative is most similar to this user based on the similarity measure. After we assign each user to different groups, we apply Dempster's rule [22] of combination to get the common user profiles

$$m_1 \oplus m_2 \ldots\ldots \oplus m_n$$

In the definition, if $1-\sum_{A_i \cap B_j=\emptyset} m_1\,(A_i) m_2(B_j)\;=\;0, m_1$ and m_2 are said to be incompatible, and $m_1 \oplus m_2$ is undefined. Here in this application, to restrict the condition to get higher quality clusters two *bpa* is defined as incompatible if one subset of one *bpa* has an empty intersection with any subset of the other *bpa.* In this case, these two *bpas* should belong to two different profiles. The probability value for an empty set may get larger after iterations. Normalization is used to eliminate the empty set. The probability portion for the empty set is subtracted and the probability distribution is recalculated, so that the total measure is one. After iterations, the sets in the common profile become separated and stable. Thus we get groups of pages in each regular user profile. It is expected that most of these groups will represent a single task and some of them may contain numerous tasks. Association rules can be found in the co-occurrences of multiple tasks.

V. Experimental Results

Experiment was carried out using a log retrieved from a reputed college web site for a period of 20 days in April 2010.[14] Initially there are 750 records in the log file. Cleaning of data is done by removing noisy data from a log file.

After cleaning records with gif, JPEG etc and status code less than 200 and greater than 299 there are 351 records in the log file. In the proposed method the records accessed by robots, agents are also cleaned by considering the access time limit of 2 seconds. Finally a log of 332 records is obtained. Samples of 4 log files are taken and performance of the enhanced cleaning proposed is compared and depicted as follows:

Users and sessions are identified after finding reference length of all records. There are 20 users identified by comparing IP address and Browser and Operating Systems and 28 sessions are identified by using referrer URL method and divided with a time limit of 30 minutes. Then GCB algorithm to the access log data for clustering. Here a clustering factor of 3 is choosing is made because the amount of data is small.

TABLE I
USERS AND SESSIONS

ClientIp	No of User	No of Sessions
117.254.157.152	1	1
117.195.161.21	4	8

After sessions are identified incomplete paths are identified from sessions and by using Referrer URL method missing pages are appended and complete path set is framed. Reference Lengths of appended pages are calculated and lengths of neighbor pages are adjusted.

TABLE II
CLUSTERS RESULTS FOR LOG (K = 3)

Cluster	Members	Common User Ip
1	1,3,5,13,14,39,45,47	GET /images/ISOLOGO.gif HTTP/1.1,GET / HTTP/1.0 GET /contactushostels.htm HTTP/1.1
2	2, 6, 7,10, 11,27, 28,29, 32,49	GET /hometxt.swf HTTP/1.1, GET / HTTP/1.0, GET /cmsCollege.swf HTTP/1.1
3	11, 8, 26	GET /favicon.ico HTTP/1.1,GET / HTTP/1.0,GET /placement.htm HTTP/1.1

The results shown from table II that:

Users of different groups can be identified by the common courses they selected, such as cluster 1, 2, 3, 4.Some groups of users are discerned by their general interest, such as cluster 3./GET / HTTP/1.0/ appear in every profile because as the entry page to the site, it has both high hit rate and long viewing time.

VI. Conclusion

A data preprocessing treatment system for web usage mining has been analyzed and implemented for log data. It has undergone various steps such as data cleaning, user identification, session identification and clustering. Dissimilar from usual implementations records are cleaned effectively by removing robot entries. By considering the byte transfer rate the reference length is computed. Apart from using Maximal Forward Reference (MFR) and Reference Length (RL) algorithm Time Window concept is also combined to find content pages.This preprocessing step is used to give a reliable input for data mining tasks. Web personalization method and introduce asuccessful clustering technique using belief function based on Dempster-Shafer's theory. Perfect input can be createdwhen the byte rate of each and every record is found.

This algorithm lacks in scalability problem. Usage data collection on the Web is incremental. Therefore, there is a need for mining algorithms to be scalable.This can be focused in future.

References

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

[2] Catlegde L. and Pitkow J., "Characterising browsing behaviours in the world wide Web," Computer Networks and ISDN systems, 1995.

[3] Chungsheng Zhang and Liyan Zhuang , “New Path Filling Method onData Preprocessing in Web Mining ,“ Computer and InformationScience Journal , August 2008.

[4] Cyrus Shahabi, Amir M.Zarkessh, Jafar Abidi and Vishal Shah“Knowledge discovery from users Web page navigation, “ In.Workshop on Research Issues in Data Engineering, Birmingham,England,1997.

[5] Istvan K. Nagy and Csaba Gaspar-Papanek “User Behaviour Analysis Based on Time Spenton Web Pages,”Web Mining Applications in E-commercce and E-Services, Studies in Computational Intelligence, 2009, Volume 172/2009, 117-136, DOI: 10.1007/978-3-540-88081-3_7 -Springer

[6] Jaideep Srivastave, Robert Cooley, Mukund Deshpande, Pang-Ning Tan“Web Usage Mining:Discovery and Applications of Usage Patterns fromWeb Data,” SIGKDD Explorations. ACM SIGKDD,2000.

[7] Peter I. Hofgesang , “Methodology for Preprocessing and Evaluating the Time Spent on Web Pages,” Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence ,2006.

[8] Robert.Cooley,Bamshed Mobasher and Jaideep Srinivastava,“DataPreparation for Mining World Wide Web Browsing Patterns,“ journalof knowledge and Information Systems,1999.

[9] Robert.Cooley,Bamshed Mobasher, and Jaideep Srinivastava, “Webmining:Information and Pattern Discovery on the World Wide Web,”,InInternational conference on Tools with Artificial Intelligence, Newport Beach, IEEE,1997, pages 558-567.

[10] Spilipoulou M.and Mobasher B, Berendt B.,“A framework for theEvaluation of Session Reconstruction Heuristics in Web UsageAnalysis,” INFORMS Journal on Computing Spring ,2003.

[11] Suresh R.M. and Padmajavalli .R. ,“An Overview of Data Preprocessing in Data and Web usage Mining ,“ IEEE, 2006.

[12] Yan Li, Boqin FENG and Qinjiao MAO, “Research on Path CompletionTechnique in Web Usage Mining,,” International Symposium onComputer Science and Computational Technology, IEEE,2008.

[13] Yan Li and Boqin FENG “The Construction of Transactions for Web Usage Mining,” International Conference on Computational Intelligence and Natural Computing, IEEE, 2009.

[14] http://news.netcraft.com/

[15] W. Wang, J. Yang, and R. Muntz, “Sting: A statistical information grid approach to spatial data mining,” 1997.

[16] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications,” pp. 94– 105, 1998.

[17] G. Sheikholeslami, S. Chatterjee, and A. Zhang, “WaveCluster: A multi-resolution clustering approach for very large spatial databases,” in Proc. 24th Int. Conf. Very Large Data Bases, VLDB, pp. 428– 439, 24–27 1998.

[18] C. Fraley and A. Raftery, “Mclust: Software for model-based cluster and discriminant analysis,” 1999.

[19] J. C. Bezdeck, R. Ehrlich, and W. Full, “Fcm: Fuzzy c-means algorithm,” Computers and Geoscience, vol. 10, no. 2-3, pp. 191–203, 1984.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise.” in KDD, pp. 226–231, 1996.

[21] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “Optics: ordering points to identify the clustering structure,” in SIGMOD ’99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data, (New York, NY, USA), pp. 49–60, ACM Press, 1999.

[22] Dempster, A. P. A Generalization of Bayesian Inference J. Roy. Stat. Soc. B, 30(1968), 205-247.