Novel Web Usage Mining using Modified Expectation Maximization Clustering

Dr. P. Sumathi
Asst. Professor, Govt. Arts College, Coimbatore,
Tamil Nadu, India

B. Uma Maheswari
Research Scholar, Bharathiyar University, Coimbatore
Tamil Nadu, India

***Abstract*— The Web mining field encompasses a wide array of issues, primarily aimed at deriving actionable knowledge from the Web, and includes researchers from information retrieval, database technologies, and artificial intelligence. Most data used for mining is collected from Web servers, clients, proxy servers, or server databases. All of which generate noisy data. Because Web mining is sensitive to noise .So data cleaning methods are necessary. Data preprocessing includes data cleaning, user identification, session identification and path completion. The inexact data in web access log are mainly caused by local caching and proxy servers which are used to improve performance and minimize network traffic. The proposed method uses path completion algorithm to preprocess the data. We collect the data from a college website and it is preprocessed based on the proposed method. The proposed path completion appends the lost information and improves the consistency of access data for further web usage mining calculations. The clustering web data is finding the groups which share common interests and behavior by analyzing the data collected from the web servers and this improves clustering on web data efficiently using Modified Expectation Maximization clustering of incremental and stepwise of EM. This paper experiments about the accomplishment of preprocessing and the precision of the clustering. The experimental result shows the significant performance of the proposed algorithm.**

***IndexTerms* —Data preprocessing, Web usage mining, Path completion algorithm, Data cleaning, User session identification, Modified Expectation Maximization.**

I .INTRODUCTION

Web has newly become a dominant platform for, not only retrieving information, but also discovering knowledge from web data. Web mining can be defined as the application of Data Mining techniques to the web related data. Many web mining algorithms are available to retrieve the WebPages. Web Usage Mining consists of three main steps: data preprocessing, knowledge extraction, and results analysis. Raw data is highly susceptible to noise, missing values. The quality affects the data mining results. In order to improve the data quality, that the data is preprocessed. Data preprocessing deals with the preparation and transformation of the dataset. Cleaning, Integration, Transformation, Reduction are the methods involved in this. Data preprocessing has been studied extensively in the past decade [12] and many commercial products such as Informatics [13] and Data Joiner [2] are applied in many areas.

In order to collect the data for preprocessing, much research has been done so far, e.g. the cookies [13] or the remote agent [2] recognize the user session [12] can do help to user identification, session identification and path completion. Data mining is one of the challenging tasks to discover the large database. Data mining through these data preprocessing is increased and becomes important in industry. In competitive consumer markets, data mining faces the growing challenge of systematic knowledge discovery in large datasets to achieve operational, tactical and strategic competitive advantages. As a consequence, the support of corporate decision making through data mining has received increasing interest and importance in operational research and industry. As an example, direct marketing campaigns aiming to sell products by means of catalogues or mail offers [7] are restricted to contacting a certain number of customers due to budget constraints.

The Web data is stored in Web servers, client machines, proxy servers or organizational databases. The primary data sources used in Web usage mining are the server log files which include Web server access logs, referrer logs and agent logs. Additional data sources that are also essential include the site files and meta-data, operational databases and domain knowledge. In some cases and for some users, additional data may be available in the client-side and proxy-server. Referrer logs contain information about the referring pages for each page reference. There are various types of Web data such as content data, structure data and usage data. Based on the type of data to be mined for analysis, Web mining can be further classified into Web content mining, Web structure mining and Web usage mining.

Data preprocessing is predominantly significant phase in Web usage mining due to the characteristics of Web data and its association to other related data collected from multiple sources. This phase is often the most time-consuming and computationally intensive step in Web usage mining. This process is critical to the success of Pattern discovery and Pattern Analysis. In short, the whole process deals with the conversion of raw Web server logs into a formatted user session file in order to perform Web usage mining [4].

The role of fuzzy sets in Web mining holds promise mainly in document and user clustering, deduction and summarization, handling of fuzzy queries involving natural language and/or linguistic quanti8ers like almost, about, and information fusion in multimedia data. Fuzzy logic may serve as the backbone of the Semantic Web, an extension of the current Web in which information is given well-defined meaning, thereby better enabling computers and people to work in cooperation [6].

The input for the Web Usage Mining process is a user session file, which is basically a pre-processed file and consists of information such as who accessed the website and what pages were accessed and for how long with their respective order. This user session file is first processed by removing outliers and irrelevant items from the raw server logs, identifying genuine and unique users from the server log and finally keeping the meaningful transactions within a user session file [11].

The data cleaning process removes the data tracked in web logs that are useless or irrelevant for mining purposes. The request processed by auto search engines, such as Crawler, Spider, and Robot, and requests for graphical page content (e.g., jpg and gif images) are deleted because these image files are auto-downloaded with the requested pages. The user identification process analyzes the log file and clusters the users so that every user in the same group has the same access characteristics. Sessions identification Once the log files have been cleaned, the next step in the data preprocessing is the identification of the session. Session identification is the process of segmenting the user activity log of each user into groups of page references during one logical period called session [14].

The paper can be organized as follows. Section II describes the related works, Section III describes the methodology used and section IV describes conclusion of the proposed work.

II. RELATED WORKS

In [16] presented a data preprocessing system for constructing the transactions in web usage mining. To implement transaction identification, the user sessions and the user access paths are extracted from the web access log and missing information is appended. These tasks are accomplished with the application of the referrer-based method, which is an effective solution to the problems introduced by using proxy servers, local caching and firewall. Meanwhile, the reference length of accessed pages is calculated with the consideration of the time spent on data transfer over internet. Then two kinds of transactions are defined, i.e. travel-path transactions and content-only transactions. These two kinds of transactions are constructed by the maximal forward references (MFR) algorithm and the reference length (RL) algorithm, respectively.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services [19]. In the literature, three main axes of Web mining have been identified, according to the Web data used as input in the data mining process, namely Web structure, Web content and Web usage mining.

In [17] presented an implementation of data preprocessing system for web usage mining and the details of algorithm for path completion. After user session identification, the missing pages in user access paths are appended by using the referrer-based method which is an effective solution to the problems introduced by using proxy servers and local caching. The reference length of pages in complete path is modified by considering the average reference length of auxiliary pages which is estimated in advance through the maximal forward references and the reference length algorithms.

In [10] introduced a new session reconstruction heuristic which is based on user web page requests logs. Smart-SRA has been experimentally shown to be better than previously developed reactive, time and navigation oriented heuristics. They did not allow page sequences with any unrelated (without any hyperlinks from the preceding page(s) to the next page) consecutive requests to be in the same session. Navigation oriented heuristics will insert artificial browser (back) requests into a session in order to guarantee that consecutive requests will have connectivity between each other. They also extend navigation oriented heuristics by using two time oriented heuristics. Another advantage of Smart-SRA is that it guarantees that all sessions generated will be maximal sequences and do not subsume any other session. They also implemented a novel agent simulator for generating simulated user sessions. They have compared the sessions reconstructed by Smart-SRA and previous heuristics against the simulated sessions generated by the agent simulator. They also defined a method to calculate the accuracy of the reconstructed sessions as a sequence – subsequence relationship.

In [14] presented a brief introduction to WUM, apart from the data mining technologies and also the implementation of the preprocessing of web log files in NASA's web server. This study focuses on methods that can be used for the task of session identification from web log files. The work in this study also produces statistical information of user session. After preprocessing is completed, the result will be used for mining user access pattern, the future work involves various data transformation tasks that are likely to influence the quality of the discovered patterns resulting from the mining techniques like Association, Clustering, and classifications that may be applied only on to a group of sessions according to assumptions of users' intentions.

In [15] performed a survey on a selection of web usage methodologies in preprocessing proposed by research community. More concentration is done on preprocessing stages like session identification and path completion and has presented various works done by different researchers. In [4] focused data preprocessing as a significant and prerequisite phase in Web mining.

Various heuristics are employed in each step so as to remove irrelevant items and identify users and sessions along with the browsing information. The output of this phase results in the creation of a user session file. Nevertheless, the user session file may not exist in a suitable format as input data for mining tasks to be performed. They also focused on a design that can be adopted for preliminary formatting of a user session file so as to be suited for various mining tasks in the subsequent pattern discovery phase.

Accurate Web usage information [20] could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, track leaving customers and find the most effective logical structure for their Web space [21]. User profiles could be built by combining users' web paths with other data features, such as page viewing time, hyper-link structure and page content [22]. What makes the discovered knowledge interesting had been addressed by several works [15, 23]. Results previously known are very often considered as not interesting. So the key concept to make the discovered knowledge interesting will be its novelty or unexpected appearance.

There are numerous commercial software packages usable to obtain statistical patterns from web logs, such as [18]. They focus mostly on highlighting log data statistics and frequent navigation patterns but in most cases do not explore relationships among relevant features.

III.METHODOLOGY

1.Preprocessing

Data Preprocessing plays a major role in Web usage mining process. Data preprocessing mainly depends on server log file.

The goal of preprocessing is to transform the raw click stream data into a set of user profiles [5]. Data preprocessing performs a series of processing of web log file which includes data cleaning, user identification, session identification and path completion. Here we collect the data's from our college website for preprocessing. The process involved in the data preprocessing is shown in the figure 1.

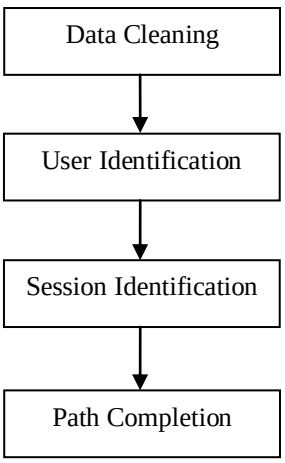

Figure-1. Steps involved in the data preprocessing

Data cleaning

The primary data sources used in Web usage mining are the server log files, which include Web server access logs and application server logs. An additional data may be available from client-side or proxy-server. The content data in a site is the collection of objects and relationships that is conveyed to the user. The structure data represents the designer's view of the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. The operational database for the site may include additional user profile information. 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, including references to style files, graphics, or sound files. The cleaning process also may involve the removal of at least some of the data fields. The status code return by the server is three digit number. There are four class of status code: Success (200 Series), Redirect (300 Series), Failure (400 Series), Server Error (SOD Series). The most common failure codes are 401 (failed authentication), 403 (Forbidden request to a restrict subdirectory, and the dreaded 404 (file not found) messages. Such entries are useless for analysis process and therefore they are cleaned form the log files [1].

a. User Identification

User identification deals with associating page references with different users. To reduce network traffic and improve performance, the pages that are requested are cached by most Web browsers. Hence, when the user navigates backwards by using the "back" button, the repeat page access is not recorded in Web server log. Proxy servers provide an intermediary solution but the difficulty of user identification still persists. All requests coming from a proxy server have the same identifier even though the requests are put forth by multiple users. Two solutions for this problem are user registration data and use of cookies. One method to identify users is by means of the user id field in the server log files. The user registration data helps in capturing additional demographic information in addition to the data which is automatically collected in the server log. However, due to privacy reasons, many users prefer not to browse sites that require registration and logins. Sometimes user registration data is not compulsory and users may often provide incorrect information [4]. Hence, user identification becomes a complex task unless an exact user id is provided. In the absence of authentication mechanisms, the most well-known approach is the use of cookies.

A user is defined as the key using a client to interactively retrieve and render resources or resource manifestations. But in Proxy servers many users are sharing the same address and same user uses many browsers. An Extended Log Format overcomes this problem by referrer information, and a user agent. The Web Usage Mining methods that rely on user assistance are the easiest ways to deal with this problem.

However, it's difficult because of privacy and security. In our experiment, we use the following heuristics to identify the user: For more logs, if the IP address is the same, but the agent log shows a change in browser software or operating system, an IP address represents a different user; Each IP address represents one user; Using the access log in combination with the referrer logs and site topology to construct browsing paths for each user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, there is another user with the same IP address.

b. Session Identification

To group the activities of a single user from the web log files is called a session. As long as user is connected to the website and it is called the session of that particular user. Most of the time, the 30 minutes time-out was taken as a default session time-out. This type of session is a set of page references from one source site during one logical period. Historically a session would be identified by a user logging into a computer, performing work and then logging off. That the login and logoff represent the logical start and end of the session.

Let L be a log. A session S is an ordered list of pages accessed by a user that is

$S =< ((p1, t1), (p2, t2), \ldots\ldots, (pn, tn))$ Where there is a user $ui \epsilon U$ such that

$\{< ui, p1, t1 >, < ui, p2, t2 >, \ldots. < ui, pn, tn >\}$ is part of L. than the session S as $< p1, p2, \ldots., pn >$.

Following rules to identify user's sessions

- If there is a new user there is new session.
- In one user session, if the referrer page is null, there is a new session.

If the time between page requests exceeds a certain limit (30 minutes) It is assumed that user is starting a new session. The reference page is estimated by access time of this page and the next one i.e. the reference length of an accessed page equals the difference between the access time of the next and the present page. If this time is few seconds than that page can be considered as an auxiliary page and otherwise that page can be considered as a content page.

c. Path Completion

Path completion is necessary to be carried out due to the existence of local caching and proxy servers. The task of path completion is to fill missing page references. Methods similar to those used for session identification can be used for path completion. If a page request is made that is not directly linked to the last page a user requested, the referrer log can be checked to see what page the request came from. The user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in user access paths should be appended. Path completion is depends on mostly URL and REFF fields in server log file. If the page is in the user's recent request history, the best guess is that the user backtracked with the "back" button; calling up cached versions of the pages awaiting a new page was requested. If the referrer log is not clear, the site topology can be used to the same effect. If more than one page in the user's history contains a link to the requested page, it implicit that the page closest to the previously requested page is the source of the new request. The procedure of the path completion algorithm can be described as follows [17]:

```
Last USID = 0; // The USID value of the previous record
Now ReferURI = ""; //The ReferURI of the current record

        L1: Getting the next record (i) in PS;
        if (Record(i).USID != LastUSID)
                {
                GOTO L3;
                }
        NowReferURI = Record(i).ReferURI;
        if (NowReferURI == Record(i-1).URI)
                {
                GOTO L3;
                }
        Getting Record(j); // j = i - 2
        L2: Record(j).RLength = ARLAP;
        Record(i-1).RLength=Record(i-1).RLength -
ARLAP;
        Inserting Record(j) into PS' according to the
        USID value;
        if (NowReferURI != Record(j).URI)
                {
                j - -;
                GOTO L2;
                }
        L3: Inserting Record(i) into PS' according to the
        USID value;
        LastUSID = Record(i).USID;
        if (Record(i) is the last record in PS)
                {
                Outputting PS';
                }
        else GOTO L1; //The End
```

This process makes certain, where the request came from and what all pages are involved in the path from the start till the end. The referrer plays an important role in determining the path for a particular request. The problem faced in this process is of the missing entries that mislead in tracking the request. But with the help of the referrer, the site topology and proper tracking of the web page requests, one can easily get the details of the path followed.

All of these processes of user identification, session identification and path completion together form the data-structuring phase of the classical data preprocessing scheme.

2.Efficient EM algorithm

The first is incremental EM (iEM) [3], in which we not only keep track of μ but also the sufficient statistics $S_1, \ldots, S_n$ for each example ($\mu = \sum_{i=1}^{n} S_i$). When we process example i, we subtract out the old s_i and add the new S_i'.

In [9] developed another variant, later generalized by [8], which it call stepwise EM (sEM). In sEM, it interpolate between μ and S_i'based on a stepsize η_k (k is the number of updates made to μ so far).

The two algorithms are motivated in different ways. Recall that the log-likelihood can be lower bounded as follows [3]

Incremental EM (iEM)

$S_i \leftarrow$ Initialization for 1=1,…, n

$\mu \leftarrow \sum_{i=1}^{n} S_i$

for each iteration t=1,..,T

for each example i=1,…,n in random order

$S_i' \leftarrow \sum_Z p\left(Z\middle|X^{(1)};\theta(\mu)\emptyset\left(X^{(i)},Z\right)\right)$

$\mu \leftarrow \mu + S_i' - S_i; S_i \leftarrow S_i'$

Stepwise EM (sEM)

$\mu \leftarrow$ initialization; k=0

for each iteration t=1,..,T

for each example i=1,…,n in random order

$S_i' \leftarrow \sum_Z p\left(Z\middle|X^{(1)};\theta(\mu)\emptyset\left(X^{(i)},Z\right)\right)$

$\mu \leftarrow (1-\eta_k)\mu + \eta_k S_i'; k \leftarrow k+1$

$$l(\theta) \geq L(q1, \dots, qn, \theta)$$

$$\underline{def} \sum_{i=1}^{n}\left[\sum_Z qi\left(Z\middle|X^{(i)}\right) logp\left(X^{(i)},Z;\theta\right) + H(qi)\right]$$

where H(qi) is the entropy of the distribution $qi\left(Z\middle|X^{(i)}\right)$. Batch EM alternates between optimizing L with respect to $q_1, \dots., q_n$ in the E-step and with respect to θ in the M-step. Incremental EM alternates between optimizing with respect to a single $q_i and\ \theta$.

Stepwise EM is motivated from the stochastic approximation literature, in this think of approximating the update μ' in batch EM with a single sample S_i'. Since one sample is a bad approximation, then interpolate between S_i' and the current μ. Thus, sEM can be seen as stochastic gradient in the space of sufficient statistics.

For both iEM and sEM, we also need to efficiently compute $\theta(\mu)$. We can do this by maintaining the normalizer for each multinomial block (sum of the components in the block). This extra maintenance only doubles the number of updates we have to make but allows us to fetch any component of $\theta(\mu)$ in constant time by dividing out the normalizer.

IV. Experiment Results

Experiment was carried out using a log dataset. That dataset are collected from the college website for a period of 20 days in April 2010 [22]. Initially there are 1750 records in the log file. Data Cleaning is done by removing noisy data from a log file. Log data differs from other datasets used in data mining, and there are several problems that 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.

EM Modified algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models then EM algorithm, where the model depends on unobserved latent variables. The process of the algorithm repeats until likelihood is stable.

No of User	Size(MB)	No of Sessions	No of Repeat User
125	10	1028	234

Table I: Data Set For Clustering

Table I gives the number of user and size of the data and the number of sessions for the web log dataset. Then the numbers of repeated user in the sessions are fined

No of Record Instances	EM precision	Modified EM precision
1000	0.690	0.750
820	0.550	0.650
300	0.360	0.450

Table II
Performance of Cluster Precision

Table II gives the precision of the cluster formation for Log data set for Modified EM algorithm and EM algorithm are depicted in it. Then performance of precision is measured with respect to number of instances in the data set to form the cluster with their attributes.

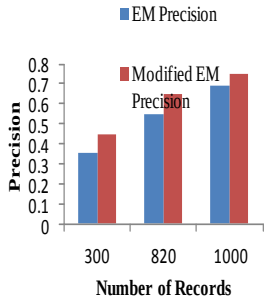

Figure 2 Performance of EM Precision

Figure 2 shows the Performance of Modified EM Precision and EM precision with Number of Records. That the higher precision is obtained for 1000 records its precision is 0.750 by Modified EM algorithm

V. Conclusion

A data preprocessing system for web usage mining has been proposed in this paper. The process used in data preprocessing such as data cleaning, user identification, session identification, path completion. The algorithm used to preprocess the data has been analyzed. . It not only reduces the log file size but also increases the quality of the available data. The proposed algorithms avoid the complicated procedure of mining site topology and don't produce the user privacy issues. The proposed method of data preprocessing system can prepare reliable transactions for the further web usage mining tasks. Thus our proposed method of effective incremental and stepwise algorithm of EM for Web Usage Mining indicate the EM approach can improve accuracy of clustering

REFERENCES

[1] Anand Sharma, "Determining Usage Patterns on RIT Web Data", March 2008.
[2] C. Shahabi, A. Zarkesh, J. Adibi et al, "Knowledge discovery from users Web-page navigation". In Workshop on Research Issues in Data Engineering, Birmingham, England, 1997.
[3] R. Neal and G. Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Learning in Graphical Models..
[4] C.P. Sumathi, R. Padmaja Valli , T. Santhanam, "An Overview Of Preprocessing Of Web Log Files For Web Usage Mining", Journal of Theoretical and Applied Information Technology, vol. 34 ,no.2, December 2011. ISSN: 1992-8645.
[5] Demin Dong, "Exploration on Web Usage Mining and its Application", IEEE, 2009.
[6] Dragos Arotariteia, Sushmita Mitra, "Web mining: a survey in the fuzzy framework", Department of Computer Science, Aalborg University Esbjerg, Niels Bohrs Vej 8, 6700 Esbjerg, Denmark, 2004.
[7] E.L. Nash, The Direct Marketing Handbook, second ed., McGraw-Hill, New York, 1992.
[8] O. Capp´e and E. Moulines. 2009. Online expectation maximization algorithm for latent data models. Journal of the Royal Statistics Society: Series B (Statistical Methodology), 71.
[9] M. Sato and S. Ishii. 2000. On-line EM algorithm for the normalized Gaussian network. Neural Computation, 12:407–432.
[10] Murat Ali Bayir, Ismail H. Toroslu, Ahmet Cosar, "A New Approach for Reactive Web Usage Data Processing", Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006.
[11] P.Tan, and V. Kumar, "Discovery of Web Robot Sessions Based on Their Navigational Patterns", Data Mining and Knowledge Discovery, 6:9-35, 2002.
[12] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. Knowl. Inf. Syst., 1(1):5–32, 1999.
[13] S. Elo-Dean and M. Viveros. Data mining the IBM official 1996 Olympics web site. Technical report, IBM T.J. Watson Research Center, 1997.
[14] Thanakorn Pamutha, Siriporn Chimphlee, Chom Kimpan, and Parinya Sanguansat, "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns", International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol. 2, No. 2, June 2012, ISSN: 2046-6447.
[15] V.Chitraa and Antony Selvdoss Davamani, "A Survey on Preprocessing Methods for Web Usage Data", (IJCSIS) International Journal of Computer Science and Information Security, vol. 7, no. 3, 2010.
[16] Yan Li Bo-Qin Feng and Yan Li, "The Construction of Transactions for Web Usage Mining", International Conference on Computational Intelligence and Natural Computing, 2009.
[17] Yan Li, Boqin Feng, Qinjiao Mao, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, 2008.
[18] Web trends, Retrieved February 12, 2004 from http://www.netiq.com/products/log.
[19] O. Etzioni, The world wide web: Quagmire or gold mine?, Communications of the ACM, 39(11):65-68, 1996.
[20] M. AbuJarour and A. Awad, "Discovering linkage patterns among web services using business process knowledge", Proceedings of the IEEE International Conference on Services Computing, July 4-9, 2011.
[21] L.K.J. Grace, V. Maheswari and D. Nagamalai,, "Analysis of web logs and web user in web mining", Int. J. Netw. Security Appli., 3: 99-110, 2011.
[22] Susanne, G., I. Terrizzano, A. Lelescu and J. Sanz, "Systematic web data mining with business architecture to enhance business assessment services", Proceedings of the Annual SRII Global Conference, March 29,2011.
[23] L. Tantan, and G. Agrawal, "Active learning based frequent item set mining over the deep web", Proceedings of the IEEE ICDE Conference, Apr. 11-16, 2011.