Volume 1, No. 1, August 2010

Journal of Global Research in Computer Science

RESEARCH PAPER

Available Online at www.jgrcs.info

Classification and Clustering of Web Log Data to Analyze User Navigation Patterns

Mrs.Niranjana.Kannan*[1] and Dr (Mrs).Elizabeth Shanthi[2]

*[1]Avinashilingam Deemed University for Women,
Coimbatore, India.
Email: nir_kannan@yahoo.com

[2]Avinashilingam Deemed University for Women,
Coimbatore, India.
Email: shanthianto@yahoo.com

Abstract: The information explosion in World Wide Web has increased the interest in Web usage mining techniques in both commercial and academic areas. Study of interested web users; provide valuable information for web designers to quickly respond to their individual needs and for the efficient organization of the website. Among the several approaches, like, Association rule mining, classification, clustering, to extract knowledge from user's navigation data, this paper uses clustering and classification of log data to discover knowledge from web log files. The proposed algorithm uses Expectation Maximization (EM) clustering along with Maximum Likelihood classification for knowledge discovery from user's navigation patterns. Experiments have been carried out in order to validate the proposed approach and evaluate the proposed algorithm.

Keywords: Web usage mining; Expectation maximization; Maximum Likelihood; Navigation pattern mining.

INTRODUCTION

The success of digital revolution and the growth of the Internet have ensured that huge volumes of high-dimensional multimedia data are available to all users. This information is often mixed with different data types such as text, image, audio, speech, hypertext, graphics and video components interspersed with each other. These enormous amount of data stored in files, databases, and other repositories have abundant of information, which needs to be extracted and analyzed [20]. However, often most of this data are not of much interest to most of the users. The problem is to mine useful information or patterns from the huge datasets. Data mining refers to this process of extracting knowledge that is of interest to the user [17].

Most often, the data mining algorithms are categorized according to the type of data being processed. Examples include text mining [9], image mining [3] and web mining [10]. Out of these, web mining, which is the art of identifying, extracting and analyzing useful information from the browsing and usage patterns from web documents and services has attracted much attention from the webmasters. The web developers use this information for designing and developing better web designs [4]. Web mining is becoming the de-facto technology in computer and information science as they have direct impact on e-commerce applications, information retrieval and filtering applications [11].

Web usage mining, also referred as Web Log Mining, is a category of web wining, is the automatic discovery of user access patterns from data describing the usage of web resources. The access logs of HTTP servers are the best source of such information. The knowledge gained from web usage mining gives guidelines on user behaviour, usage patterns for mass customization and personalization.

The objectives of this paper are to produce a model that helps webmaster and web developers to improve the design of a webpage by analyzing the web log files. The paper is organized as below. Section 1 provided a brief introduction to web usage mining. Section 2 introduces web log files. Section 3 discusses the previous studies related to this paper and Section 4 describes the proposed methodology. The results are presented in Section 5 and Section 6 concludes the work with future directions.

IP Address-Base URL-Date Time-Method of Access-File-Protocol-Status Code-Bytes-Referrer-User Agent

Fig. 2a**.** Common Web Log Format

```
203.30.5.145    www.avinutry.ac.in-[01/May/2010.03.09.21-
0600] "GET /e-learning/index.html HTTP/1.0" 200 3942
http://www.avinuty.ac.in/lessons/cs/c1.jsp "Mozilla/4.5 (en)
"
```

Fig. 2b**.** Sample Entry

WEB LOG FILES

Web log file contains information that are automatically created maintained by a web server. Every "hit" to the Web; site, including each view of a HTML document, image or other object, is logged. The raw web log file format (Figure 2a) is essentially one line of text for each hit to the web site. This contains information about who was visiting the site, where they came from, and exactly what they were doing on the web site. A sample Web log file entry is shown in Figure 2b. Either a client perspective approach or a server perspective approach can be used to extract information from these log files Server perspective approach extracts information about the sites where the service resides and are mainly used to improve the overall design of a website. Client perspective approach, on the other hand, extracts information about client's sequence of clicks, which provides information about a user or group of users. This information can then be used to perform prefetching or caching of pages. The client perspective approaches use, 'click streams', a subpart of web log file for processing. Information regarding, each and every click made by an Internet user in a web page or link is recorded in a clickstream file. Clickstream Data is information that users generate as they move from page to page and click on items within a website, usually stored in log files. This information can be used to evaluate the user trend during browsing, amount of time spent on each page, predict user's next visit, analyze pattern similarity among different users, etc.

RELATED STUDIES

Web usage mining, the art of analyzing user interactions with a web page, has been dealt by several researchers using different approaches. Many different data mining techniques, like, Association rule mining, graph partitioning, classification, clustering are some techniques that have gained equal attention from academicians and researchers. Association rule mining method ([8], [16]) generates rule describing the relationship between user access and a web page. Though very successful, the number of association rules generated is often very huge and are often difficult to process and analyze. As an alternative clustering and classification techniques are being used to mine web log files.

Some researchers including [15], [14] have used classification algorithms for detecting web usage patterns. In [18], the authors used similarity upper approximation clustering technique on web transactions from web log data to extract the behavior pattern of users page visits and order of occurrence of visits. Graph partitioning clustering is another technique that is frequently used. Jalali *et al*. [6] [7] used graph partitioned clustering algorithm to propose a recommender system for predicting future moves of user. They discovered user navigation patterns from log files to predict future moves. Similarly, [14] used graph partitioning clustering technique to automatically improve their website organization and presentation by mining usage logs. They proposed a new clustering algorithm called 'cluster mining' to cluster log data into groups of behavioral pattern, associating pages in a Web site. All these techniques are reported to be computation expensive [12]. As an alternative to graph partitioning clustering, [2] used EM algorithm based on Markov chains to group user sessions into clusters. The argument behind using EM algorithm was that it is both memory efficient and has simple implementation procedure. Similarly, [1] also used EM algorithm for clustering user navigation paths. The experimental results of both these systems proved that the models produced quality clustering but was found to be very slow at convergence. To solve the convergence problem, [12] proposed a model for mining user's navigation pattern using EM algorithm and used maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. This model is refereed in this paper as EM model. During experimentation, it was found that the performance of EM model could be increased by applying classification algorithm to classify user requests and combine the results with EM clustering to provide a more accurate web usage mining prediction system. This research work is focused on producing such a system.

THE PROPOSED METHODOLOGY

The proposed methodology uses an unsupervised clustering and classification algorithm to learn about the navigation pattern of users. The clustering technique is the same as the one proposed in EM model. The EM model is enhanced by introducing a classification algorithm for classifying the user requests. The main aim of the proposed model is to mine user navigation patterns. User navigation pattern is defined as common browsing characteristics among group of users. Different users have common browsing practices and navigation patterns needs to capture these common interests to identify user needs. In this paper, clustering technique is used to group users with similar browsing characteristics and classification technique is used to associate navigation behaviour with these groups of users. Clustering has the advantage of grouping common interest users together and the user of classification will be able to classify different user requests. This paper uses EM algorithm for clustering and maximum likelihood classification. This enhanced model is referred as MLC EM model. Both the algorithms selected are memory efficient and are easy to implement, with a profound probabilistic background. The architecture used is given in Figure 3.

Data Preprocessing

The main objective of this step is to reformat the raw log data into a format that identifies all web access sessions. Not every access made contain useful information. All those entries that are irrelevant should be removed. Examples of irrelevant entries include button image access, multimedia file access, non-human accesses (accesses made by web crawlers, web robots), etc. Redundant clicks and failed transactions should also be removed. A user session is a delimited set of pages visited by the same user within the duration of one particular visit to a Web site. Session identification is carried out using the assumption that if a certain predefined period of time between two accesses is exceeded, a new session starts at that point. Web usage data is prepared for applying navigation patterns mining algorithms by doing these pretreatment tasks and are done manually in the present work.

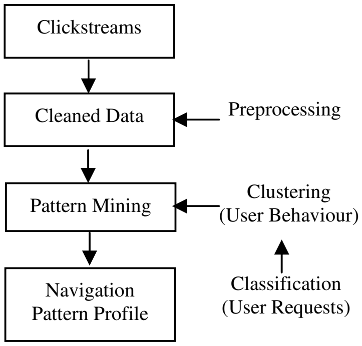

Fig. 3. Proposed Methodology

Pattern Mining

In the proposed method, the pattern mining is performed in two steps. The first step classifies the user requests and then in the second step, the classified results are clustered into groups. The classification is performed using maximum likelihood classification algorithm and the clustering is performed using Expectation Maximization algorithm. The main reason behind combining classification and clustering is to increase the speed of convergence.

Maximum likelihood classification

Maximum likelihood classification (MLC) algorithm is a statistical decision rule that examines the probability function of a data for each of the classes and assigns the pixel to the class with the highest probability. This method is most often used in image classification and is used to classify user requests from web log data in the present study. The maximum likelihood equation used is called Mahalanobis minimum distance (MD) and is defined equation (1).

$$MD = (x\text{-}m_I)^T C_I^{-1} (x\text{-}m_I)$$

Where CI is the covariance matrix for the particular imagined movement considered, left or right and T stands for the transposition operator. The Mahalanobis distance is used in a minimum-distance classifier as follows: Let **m**R, **m**L be the means for the right and left imagined movement classes, and let CR, CL be the corresponding covariance matrices. A feature vector x is classified by measuring the Mahalanobis distance d from **x** to each of the means, and assigning **x** to the class for which the Mahalanobis distance is minimum. In this paper, the full covariance matrix is used to calculate the MD.

The MLC algorithm is used to classify user requests into NI and I, based on the amount of time spent by them in a website. If the amount of time spent is more than 30 seconds, then they are considered as genuine users. According to [19], interested users exhibit certain access patterns; they access certain web pages for a rather long time because they need time to spend on its contents. The ratio between the time spent on content reading and the amount of time they navigate is large. The interested users often use the HTTP POST mode, because they are interested in registering with websites and are willing to fill out forms with their own information. The user who does not have interest simply accesses many pages quickly to browse contents. These users do not often use POST method because they are not interested in registering at websites. Thus, the web log files are classified as Interested Users (IU) and Not Interested Users (NIU) based on the Time Stamp parameter (30 min),method used(GET/POST)and number of pages referred(min 5) .

In order to use the MD to classify a test request as belonging to one of two classes, the covariance matrix of each class is estimated, based on requests known to belong to each class. Then, given a request, the MD to each class is calculated and classifies the request as belonging to that class for which the MD is minimal. Using the probabilistic interpretation given above, this is equivalent to selecting the class with the maximum likelihood. More information on MD can be found at [5].

EM Clustering Algorithm

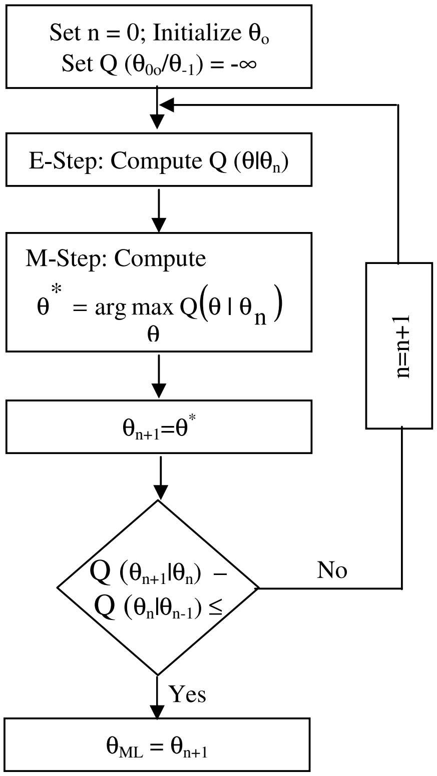
Set n = 0; Initialize θ_o
Set Q $(\theta_{0o}/\theta_{-1})$ = -∞
E-Step: Compute Q $(\theta|\theta_n)$
M-Step: Compute
$\theta^* = \arg\max_\theta Q(\theta \mid \theta_n)$
$\theta_{n+1}=\theta^*$
Q $(\theta_{n+1}|\theta_n)$ –
Q $(\theta_n|\theta_{n-1}) \leq$
No
n=n+1
Yes
$\theta_{ML} = \theta_{n+1}$

Fig. 4. EM Algorithm

After classifying the user into two classes, the NI request class is ignored and the I class user requests, which are genuine requests made by the user are taken into consideration. The EM clustering algorithm is then applied to this class. The procedure of EM algorithm is given in Figure 4, where θ_{ML} is the maximum likelihood estimate and ξ is the termination threshold.

The Expectation-Maximization (EM) algorithm is a method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. EM is an iterative method which alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood, evaluated using the current estimate for the latent variables, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. The EM algorithm can be applied to problems where the observable data provide only partial information or where some data is missing (Kung *et al*., 2005). As web log data often displays this characteristic, EM algorithm was used for clustering.

EXPERIMENTAL RESULTS

To analyze the quality of the proposed method, several experiments were conducted. Data preparation done by [12] used Microsoft IIS log format and reformatted the log files in the following method.

Index	*URLs Address*
0	/e-learning/
1	/e-learning/FAQ
2	/e-learning/index.jsp
3	/e-learning/lessons/cs.jsp
4	/e-learning/lessons/stat.jsp

Fig. 5. URLs and Numeric value

As a cleaning process, all the irrelevant entries were removed. After cleaning, each URL address of a session is assigned a numeric code (Fig. 5) and the length of the attribute was restricted to a depth of 8, to reduce the number of attributes.

Session	URLS in a Session				
1	0	3	5	0	0
2	0	0	1	0	0
3	1	6	2	0	0

Fig. **6.** URLs in Each Session

Session	URLS in a Session				
1	0	1	1	0	0
2	0	0	1	0	0
3	1	1	1	0	0

Fig. **7.** Binary Coded

Figure 6 shows the log file after assigning the numeric value. This information is further codified by assigning a binary value of one if a visit has been made, zero otherwise (Figure 7). Each page visited is considered as a user request and in the present study, a slight deviation is made to classify this user request to either interested or not interested category is performed. Only those requests that belong to the interesting category are taken to the next step. The number of times each page is visited is calculated only for cell regions having a value of one and is shown in Figure 8.

Both the classification and clustering algorithm are used to find the maximum and minimum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. The process of the algorithm is iteratively repeated until likelihood reaches convergence, that is still it reaches a stable condition.

Session	Attribute 1	Attribute 2	…
1	123	200	…
2	150	75	…
3	2	0	…

Fig. **8.** Final Preprocessing Log File

The log likelihoods obtained based on the number of cluster during training is shown in Figure 9. Further from the figure, it can be seen by decreasing in number of clusters, reduces the likelihood convergence to lower values. The experiments were conducted for 20 clusters, that is, the user navigation patterns were clustered into 20 groups. However, the convergence point without classification was 12 and was 8 while using classification, which proves that the inclusion of classification arrives to a stable state at a faster pace.

Visit coherence is another parameter that was utilized to evaluate the quality of clusters. The visit coherence was applied by the procedure given by [12] and the result is shown in Figure 10.

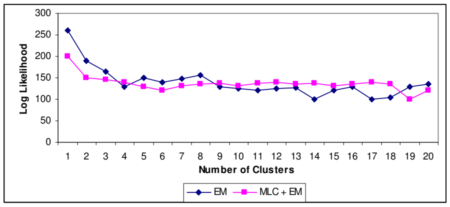

Fig. **9.** Log Likelihood

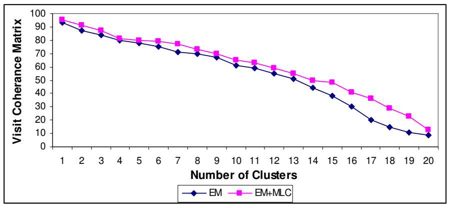

Fig. **10.** Visit Coherence

From the figure, it can be seen that the visit coherence of the proposed MLC + EM model is higher than the EM model. The results obtained indicate that the proposed method is superior to EM clustering model.

CONCLUSION

In this study, the usage of EM clustering for studying user's navigation pattern was enhanced to include user request classification. The experiments conducted results showed a positive response to the inclusion of classification. The accuracy of clustering had increased after classification was included. The system was developed using MATLAB 7.3 and the log files used were simulated. Efforts are being made to obtain web log files from Avinashilingam Deemed University Website and in future the system will be further tested with these log files.

REFERENCES

1. Anderson, C. R., Domingos, P. and Weld, D.S. (2001) Adaptive Web Navigation for Wireless Devices, Pp. 879-884.

2. Cadez, I., Heckerman, D., Meek, C., Smyth, P. and White, S. (2000) Visualization of navigation patterns on a Web site using model-based clustering, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Pp. 280-284.

3. Cao, L. (2010) Domain-Driven Data Mining: Challenges and Prospects, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 6, Pp. 755-769.

4. Heydari, M., Helal, R.A. and Ghauth, K.I. (2009) A graph-based web usage mining method considering client side data, International Conference on Electrical Engineering and Informatics, 2009. ICEEI '09, Vol. 01, Pp.147-153.

5. http://en.wikipedia.org/wiki/ Mahalanobis_distance.

6. Jalali, M., Mustapha, N., Sulaiman, N. B. and Mamat, A. (2008b) A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems," 12th International on Information Visualisation,IV'08, London, UK, Pp. 302-307.

7. Jalali, M., Mustapha, N., Sulaiman, M. N. B. and Mamat, A. (2008a) OPWUMP: An Architecture for Online Predicting in WUM-Based Personalization System, Communications in Computer and Information Science, Advances in Computer Science and Engineering, Springer Berlin Heidelberg, Vol. 6, Pp. 838–841.

8. Kazienko, P. (2009) Mining Indirect Association Rules for Web Recommendation, International Journal of Applied Mathematics and Computer Science, Vol. 19, Issue 1, Pp. 165-186.

9. Khandelwal, M., Shakarmani, R. and Kedar. N. (2010) Article: Performance Assessment using Text Mining, International Journal of Computer Applications, Published By Foundation of Computer Science Vol. 1, No.12, Pp.1–6.

10. Kolari, P. and Joshi, A. (2004) Web Mining: Research and Practice, Computing in Science and Engineering, Vol. 6, No. 4, Pp. 49-53

11. Kosala, R. and Blockeel, H. (2000) Web mining research: A survey, SIGKDD Explorations, Vol. 2, No. 1, Pp. 1-15.

.

12. Mustapha, N., Jalali, M. and Jalali, M. (2009) Expectation Maximization Clustering Algorithm for User Modeling in Web Usage Mining Systems, European Journal of Scientific Research, Vol. 32, No. 4, Pp.467-476.

13. Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R. (2008) A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites, IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 2, Pp. 202-215.

14. Perkowitz, M. and Etzioni, O. (2000) Adaptive Web sites, Communications of the ACM, Vol. 43, Pp. 152-158.

15. Picariello, A. and Sansone, C. (2008) A web usage mining algorithm for web personalization, Intelligent Decision Technologies, Vol. 2, Issue 4, Pp. 219-230.

16. Raju, V.V.R., Rao V.M. and Kumari, V. (2010) Article: Understanding User Behavior using Web Usage Mining, International Journal of Computer Applications, Published By Foundation of Computer Science, Vol.1, No. 7, Pp. 55–64.

17. Roughan, M. and Zhang, Y.(2006) Secure distributed data-mining and its application to large-scale network measurements, ACM SIGCOMM Computer Communication Review, Vol.36 , Issue 1, Pp.7- 14.

18. Santhisree, K. and Damodaran, A. (2010) Clustering on Web usage data using Approximations and Set Similarities, International Journal of Computer Applications (ICJA), Published By Foundation of Computer Science, No. 4, Article 5, Pp. 27-31.

19. Suneetha and Krishnamoorthi (2010) International Conference on Computing, Communications and Information Technology Applications, International Conference on Computing, Communications and Information Technology Applications, (CCITA 2010), Coimbatore, India.

20. Washio, T., Suzuki, E., Ting, K.M. and Inokuchi, A. (Eds.) (2008) Advances in Knowledge Discovery and Data Mining, Proceedings of12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, Lecture Notes in Computer Science, Pp. 1-1102. SBN: 978-3-540-68124.