Efficient Web Usage Mining with Clustering

K.Poongothai[1] M.Parimala[2] and Dr. S.Sathiyabama[3]

[1] Asst Prof, Department Of Information Technology, Selvam College of Technology, Namakkal, Tamilnadu, India.

[2]Lecturer, Department Of MCA, M.Kumarasamy College of Engg, Karur, Tamilnadu, India.

[3]Assistant Professor of Computer Science, Thiruvalluvar Govt Arts and science college, Rasipuram, Tamilnadu, India.

Abstract

Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis etc., Web site under study is part of a nonprofit organization that does not sell any products. It was crucial to understand who the users were, what they looked at, and how their interests changed with time. To achieve this, one of the promising approaches is web usage mining, which mines web logs for user models and recommendations. Web usage mining algorithms have been widely utilized for modeling user web navigation behavior. In this study we advance a model for mining of user's navigation pattern.

The proposal of our work proceeds in the direction of building a robust web usage knowledge discovery system, which extracts the web user profiles at the web server, application server and core application level. The proposal optimizes the usage mining framework with fuzzy C means clustering algorithm (to discover web data clusters) and compare with Expected Maximization cluster system to analyze the Web site visitor trends. The evolutionary clustering algorithm is proposed to optimally segregate similar user interests. The clustered data is then used to analyze the trends using inference system. By linking the Web logs with cookies and forms, it is further possible to analyze the visitor behavior and profiles which could help an e-commerce site to address several business questions. Experimentation conducted with CFuzzy means and Expected Maximization clusters in Syskill Webert data set from UCI, shows that EM shows 5% to 8% better performance than CFuzzy means in terms of cluster number.

1. Introduction

Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc. Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages.

Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited. The structure of a typical web graph consists of web pages as nodes, and hyper-links as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and a hyperlink that connects

two different pages is called an inter-document hyperlink. In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered i.e web server data, application server data and application level data.

In Web Server Data, user logs are collected by the web server and typically include IP address, page reference and access time. In Application Server Data, commercial application servers such as Web logic, Story Server, have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs. In Application Level Data, new kinds of events can be defined in an application, and logging can be turned on for them, generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

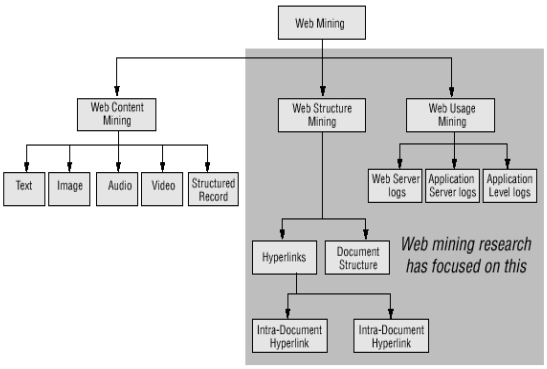

Fig 1: Web Mining Taxonomy

2) Literature Review

The WWW continues to grow at an amazing rate as an information gateway and as a medium for conducting business. Web mining is the extraction of interesting and useful knowledge and implicit information from artifacts or activity related to the WWW [4], [6]. Based on several research studies we can broadly classify Web mining into three domains content , structure and usage mining [8], [9]. This work is concerned with Web usage mining. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the Web access logs can help understand the user behavior and the web structure. From the business and applications point of view, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e business, e-services, e-education and so on [10]. Accurate Web usage information 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 [3]. User profiles could be built by combining users' navigation paths with other data features, such as page viewing time, hyper- link structure, and page content [12]. What makes the discovered knowledge interesting had been addressed by several works [11] and [12]. 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.

Whenever a visitor accesses the server, it leaves the IP, authenticated user ID, time/date, request mode, status, bytes, referrer, agent and so on. The available data fields are specified by the HTTP protocol. There are several commercial software that could provide Web usage statistics[1]. These stats could be useful for Web administrators to get a sense of the actual load on the server. However, the statistical data available from the normal Web log data files or even the information provided by Web trackers could only provide the information explicitly because of the nature and limitations of the methodology itself. Generally, one could say that the analysis relies on three general sets of information given a current focus of attention past usage patterns, degree of shared content and inter-memory associative link structures. After browsing through some of the features of the best trackers available it is easy to conclude that rather than generating statistical data and texts they really do not help to find much meaningful information.

For small web servers, the usage statistics provided by conventional Web site trackers may be adequate to analyze the usage pattern and trends. However as the size and complexity of the data increases, the statistics provided by existing Web log file analysis tools may prove inadequate and more intelligent knowledge mining techniques will be necessary[2], [3]. In the case of Web mining, data could be collected at the server level, client level, proxy level or some consolidated data. These

data could differ in terms of content and the way it is collected etc. The usage data collected at different sources represent the navigation patterns of different segments of the overall Web traffic, ranging from single user, single site browsing behavior to multi-user, multi-site access patterns. Web server log does not accurately contain sufficient information for inferring the behavior at the client side as they relate to the pages served by the Web server.

To demonstrate the efficiency of the proposed frameworks, Web access log data at the Monash University's Web site were used for experimentations. The University's central web server receives over 7 million hits in a week and therefore it is a real challenge to find and extract hidden usage pattern information. To illustrate the University's Web usage patterns, average daily and hourly access patterns for 5 weeks are shown. The average daily and hourly patterns nevertheless tend to follow a similar trend the differences tend to increase during high traffic days (Monday - Friday) and during the peak hours (11:00 - 17:00 Hrs). Due to the enormous traffic volume and chaotic access behavior, the prediction of the user access patterns becomes more difficult and complex.

Previous work presented approaches for discovering and tracking evolving user profiles. It also describes how the discovered user profiles can be enriched with explicit information need that is inferred from search queries extracted from Web log data. Profiles are also enriched with other domain-specific information facets that give a panoramic view of the discovered mass usage modes. An objective validation strategy is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behavior. However the previous work concentrated only on user profiling at the application level data but not associating it to the web server. The user profile maintained by the web server enriches the user's session of authenticity at different spatial entities. The previous work used conventional web log profile analyzers weakened at the linkage of web user profiling to its server.

3) Cluster based Web Usage Knowledge Discovery Framework

The rapid e-commerce growth has made both business community and customers face a new situation. Due to intense competition on the one hand and the customer's option to choose from several alternatives, the business community has realized the necessity of intelligent marketing strategies and relationship management. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis and so on.

The proposed cluster based framework presents the important concepts of Web usage mining and its various practical applications. Further a novel approach called Web usage miner is presented. Web Usage Miner could optimize the concurrent architecture of a fuzzy clustering algorithm (to discover web data clusters) and a fuzzy inference system to analyze the Web site visitor trends. A hybrid evolutionary fuzzy clustering algorithm is proposed to optimally segregate similar user interests. Proposed approach is compared with hierarchical patterns (to discover patterns) and several function approximation techniques.

The Miner hybrid framework optimizes a fuzzy clustering algorithm using an evolutionary algorithm. The raw data from the log files are cleaned and pre-processed and a fuzzy C means algorithm is used to identify the number of clusters. The developed clusters of data are fed to a fuzzy inference system to analyze the trend patterns. The if-then rule structures are learned using an iterative learning procedure by an evolutionary algorithm and the rule parameters are fine tuned using a back propagation algorithm. The optimization of clustering algorithm progresses at a faster time scale in an environment decided by the inference method and the problem environment.

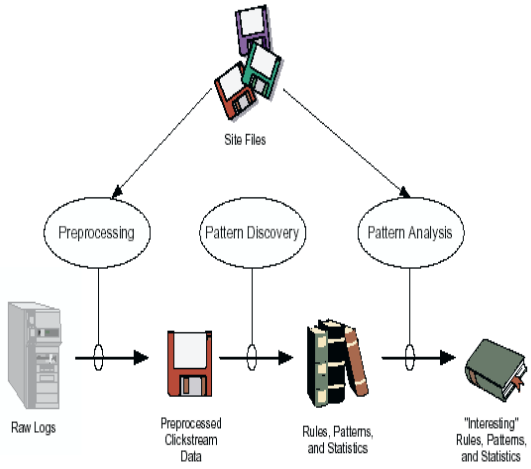

Figure 2: High Level Web Usage Mining Process

3.1 Optimization of Fuzzy Clustering Algorithm

Usually a number of cluster centers are randomly initialized and the FCM algorithm provides an iterative approach to approximate the minimum of the objective function starting from a given position and leads to any of its local minima. No guarantee ensures that FCM converges to an optimum solution (can be trapped by local extrema in the process of optimizing the clustering criterion). The performance is very sensitive to initialization of the cluster centers. An evolutionary algorithm is used to decide the optimal number of clusters and their cluster centers. The algorithm is initialized by constraining the initial values to be within the space defined by the vectors to be clustered. In the Miner approach, the fuzzy clustering algorithm is optimized jointly with the trend analysis algorithm (fuzzy inference system) in a single global search.

3.2 Expectation Maximization

Expectation maximization (EM) is used for clustering in the context of mixture models. This method estimates missing parameters of probabilistic models. Generally, this is an optimization approach, which had given some initial approximation of the cluster parameters, iteratively performs two steps, i.e., the expectation step computes the values expected for the cluster probabilities, and second, the maximization step computes the distribution parameters and their likelihood given the data. It iterates until the parameters being optimized reach a fix point or until the log-likelihood function, which measures the quality of clustering, reaches its maximum. To simplify the discussion we first briefly describe the EM algorithm.

The algorithm is similar to the Fuzzy C-means procedure in that a set of parameters are re-computed until a desired convergence value is achieved. The parameters are re-computed until a desired convergence value is achieved. The finite mixtures model assumes all attributes to be independent random variables. A mixture is a set of *N* probability distributions where each distribution represents a cluster. An individual instance is assigned a probability that it would have a certain set of attribute values given it was a member of a specific cluster. In the simplest case *N=2*, the probability distributes are assumed to be normal and data instances consist of a single real-valued attribute. Using the scenario, the job of the algorithm is to determine the value of five parameters are the mean and standard deviation for cluster 1, the mean and standard deviation for cluster 2 and the sampling probability *P* for cluster 1 (the probability for cluster 2 is *1-P*)

Algorithm Procedure

a. Guess initial values for the five parameters.

b. Use the probability density function for a normal distribution to compute the cluster probability for each instance. In the case of a single independent variable with mean μ and standard deviation $\acute{o}$, the formula is:

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)e^{\frac{-(x-\mu)^2}{2\sigma^2}}}$$

In the two-cluster case, we will have the two probability distribution formulas each having differing mean and standard deviation values.

c. Use the probability scores to re-estimate the five parameters.

d. Return to Step b.

The algorithm terminates when a formula that measures cluster quality no longer shows significant increases. One measure of cluster quality is the likelihood that the data came from the dataset determined by the clustering. The likelihood computation is simply the multiplication of the sum of the probabilities for each of the instances.

4. Experimental Evaluation of EM and CFuzzy Cluster for Web usage Mining

Measuring the quality of the EM and CFuzzy clustering in navigation patterns mining systems needs to characterize the quality of the results obtained. The experimental evaluation was conduced using UCI repository data sets of Zoo Data Set. The data is in the original arff format used by Weka tool. The characteristics of the dataset used are given in the Table 1. Expectation Maximization Clustering Algorithm and Cfuzzy means algorithm are used for User Modeling in Web Usage Mining System.

Table 1: Dataset used in the experiments

Data set	Size (Mb)	Record- Instances
Zoo Data Set	12	1010

All evaluation tests were run on a dual processor Intel CPU 2.5 GHz Pentium Core 2 Duo

with 4GBytes of RAM, operating system Windows XP. Our implementations run on Weka tool, a data mining software for evaluation part of the system. In this study, there are two steps of data converting before applying EM clustering algorithm. There are around 800 URLs in DePaul dataset. Assigning each URL address in the session to sequential numeric values is the first step. It is impossible to assign 800 attributes to Weka so for reducing the number of attribute, each eight sequence of attributes is assigned to one attribute based on EM algorithm. Table 2 shows some basic statistics on user and sessions after cleaning, filtering and session the Zoo dataset.

Table 2: Dataset for Clustering (EM and CFUZZY)

No of Users	Size (Mb)	No of Sessions	No of Repeat Users
125	10	1028	234

5. Result and Discussion on Clusters effect on web usage patterns

EM algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. The process of the algorithm repeats until likelihood is stable. Table 2 shows the cluster detection rate of the EM algorithm on the Zoo data set. The cluster detection rate is measured in terms of True positive, which is indicated in the Table 3, as the number of data record-instances increases, true positive rate of cluster object gets higher. The experiment was accomplished by maximum 114 iterations. It is complex to define the number of clusters in the initial cluster formation. In our experiment, we tried several times to tune the cluster size with other parameters to get higher true positive rate of clustering. Finally cluster the user's navigation patterns into 20 groups. Meanwhile, that the EM algorithm will get a local optimization after 26 iterations.

Table 3: Performance of EM with True Positive rate against number of record instances

No of Record Instance	Cluster True Positive Rate
1010	0.9
840	0.713
286	0.217

Table 4 depicts the performance of EM for true negative rate of cluster formation of the Zoo data set. For instance the percentage of the largest cluster is 34, while the experiment creates 20 clusters. With the increased record size, False Positive of cluster object formation also raises. However the false positive rate is comparatively very small compared to that of total number of cluster object. This in turn reduces the impurity in cluster formation. The performance graph of EM clustering on zoo data set is shown in Figure 3. Percentage of maximum cluster in clusters set will be decreased if the number of the cluster object purity is increased for every iteration.

Table 4: True Negative rate of cluster objects in web usage mining

No of Record Instances	Cluster False Positive Rate
1011	**841**

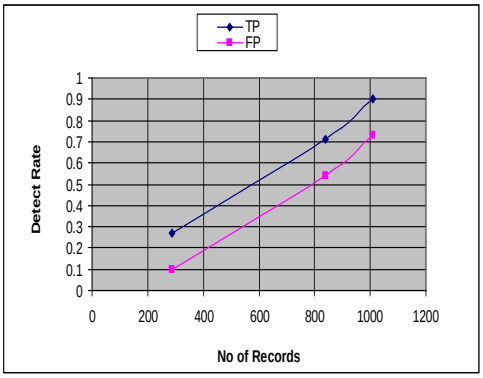

Figure 3: Performance of EM on Zoo Data set with TP and FP

The precision of the cluster formation for Zoo data set with CFuzzy cluster and EM algorithm are depicted in the Table 5. The performance of precision is measured with respect to number of instances in the data set to form the cluster with all the 10 attributes. The precision of EM is higher than the CFuzzy cluster model which is shown in Table 2 and Figure 4 indicates that the precision rate even increases for higher number of record instances of the data set.

Table 5: Performance of Cluster Precision on EM and CFuzzy algorithm

No of Record Instances	EM Precision	CFuzzy Precision
1012	0.745	0.713
842	0.682	0.645
286	0.535	0.526

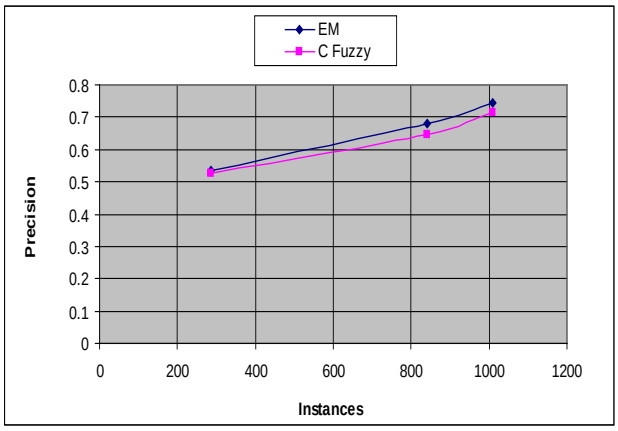

Figure 4: Comparative performance of Cluster Precision on EM against CFuzzy

Visit-coherence is utilized to evaluate the quality of the clusters (navigation pattern) produced by the EM clustering algorithm. In addition visit-coherence quantifies a session intrinsic coherence. As in the page gather system, the basic assumption here is that the coherence hypotheses holds for every session. To evaluate the visit-coherence, split dataset using 10 cross validation. The percentage of user usage visit coherence precision in the EM is approximately 11% higher than C Fuzzy means clustering algorithm for difference values of attributes.

6. Conclusion

The web usage mining framework presented in this work evaluates the performance of expectation-maximization (EM) and CFuzzy means cluster algorithms. The proposed Miner framework is an initial effort to patch up some of the weaknesses of the conventional web log file analyzers. The experimental results of EM represent that by decreasing the number of clusters, the log likelihood converges toward lower values and probability of the largest cluster will be decreased while the number of the clusters increases in each web usage pattern. The experimentation on the K-means clustering is also conducted. The results indicate the EM approach can improve accuracy of clustering to 11 more. By linking the Web logs with cookies and forms, it is further possible to analyze the visitor behavior and profiles which could help an e-commerce site to address several business questions. The further scope can be made in the direction of applying some classification methods for request. This can be used in web usage mining-based prediction systems.

7. References

[1] A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites**,** " Nasraoui, O. Soliman, M. Saka, E. Badia, A. Germain, R.", Knowledge and Data Engineering, IEEE Transactions on Feb. 2008

[2] Abraham, A (2001). Neuro-fuzzy systems: State-of-the- art modeling techniques, connectionist models of neurons, learning processes, and artificial intelligence. In Lecture Notes in Computer Science 2084, J Mira and A Prieto (eds.), Germany, Spain: Springer-Verlag, pp. 269-276.

[3] Bocca, M Jarke and C Zaniolo Analog Website Tracker (2003). (http://www.analog.cx/) [3 October 2003].

[4] Chakrabarti, S (2003). Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers.

[5] Chang, G, MJ Healey, JAM McHugh and JTL Wang (2001) Web Mining, Mining the World Wide Web Chapter 7, pp. 93-104. Kluwer Academic Publishers

[6] Chen, PM and FC Kuo (2000). An information retrieval system based on an user profile. The Journal of Systems and Software, 54, 38.

[7] Chi, EH, A Rosien and J Heer (2002). LumberJack: Intelligent discovery and analysis of web user traffic composition. In Proc. of ACM-SIGKDD Workshop on Web Mining for Usage Patterns and User Pro_les. Canada: ACM Press.

[8] Cho, YH, JK Kim and SH Kim (2003). A personalized recommender system based on web usage mining and decision tree induction, Expert Systems with Applications, 23(3), 329 - 342.

[9] Coenen, F, G Swinnen, K Vanhoof and G Wets (2000). A framework for self adaptive websites: Tactical versus strategic changes. In Proc. of the Workshop on Web mining for E-commerce: Challenges and Opportunities (KDD'00), pp. 75-80.

[10] Hay, B, G Wets and K Vanhoof (2003) Segmentation of visiting patterns on web sites using a sequence alignment method, Journal of Retailing and Consumer Services, 10(3), pp. 145-153

[11] Heer, J and EH Chi (2001) Identification of web user traffic composition using multi-modal clustering and information scent, Workshop on Web Mining, SIAM Conference on Data Mining, pp. 51 - 58.

[12] Heinrichs, JH and JS Lim (2003) Integrating web-based data mining tools with business models for knowledge management, Decision Support Systems, 35(1), pp. 103- 112.

K.Poongothai received the M.Sc (IT). Degree in Information Technology from M.Kumarasamy College of Engineering,Karur in 2006 respectively. Presently she is working in Selvam College of Technology, Namakkal, and Tamilnadu, India as Assistant Professor in Department of Information Technology

M.Parimala received the MCA. Degree in Computer Application from Mother Theresa Women's University, Kodaikanal in 2005 respectively. Presently she is working in M.Kumarasamy College of Engineering,Karur Tamilnadu, India as Lecturer in Department of Computer Application

Dr.S.Sathyabama received the M.Sc.in Avinashilingam Deemed University, Coimbatore in 1997, M.Phil in Bharathiar University, Coimbatore in 2001 and Ph.D. degree in Periyar University in 2007, Salem. She worked as Lecturer from 1997 to 2001 in karuppannan Mariappan College, Muthur. , she worked as a Professor in the Department of Master of Computer Application from 2001 to 2011 at K.S.Rangasamy College of Technology and presently she is working as Assistant Professor of Computer Science, Thiruvalluvar Govt Arts and Science College, Rasipuram.