International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:2, 2009

Navigation Patterns Mining Approach based on Expectation Maximization Algorithm

Norwati Mustapha, Manijeh Jalali, Abolghasem Bozorgniya, Mehrdad Jalali

***Abstract*—**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 model makes user model based on expectation-maximization (EM) algorithm.An 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 experimental results 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 treatment.

***Keywords*—**Web Usage Mining, Expectation maximization, navigation pattern mining.

I. Introduction

INTERNET in the analysis of user behavior on the web has been increasing rapidly. There are many research and algorithm for modeling and analyzing the user behavior that it can be used in many applications.

The user modeling based on user navigation data is challenging task that is continuing to gain importance as the size of the web and its user-base increase. A web navigation behavior is helpful in understanding what information of online users demand. Following that, the analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps, web based personalization system and improving searching accuracy when seeking information. Nevertheless, an online user navigation behavior grows each passing day, and thus extracting information intelligently from it is a difficult issue. Web usage mining (WUM) is the process of extracting knowledge from Web user's access data by exploiting data dining technologies[1]. It can be used for different purposes such as personalization, system improvement and site modification.

In this research, user navigation patterns are described as the common browsing behaviors among a group of users. Since many users may have common interests up to a point during their navigation, navigation patterns should capture the overlapping interests or the information needs of these users. In addition, navigation patterns should also be capable to distinguish among web pages based on their different significance to each pattern.

In this study we advance a model for mining of user's navigation pattern. The model makes user model based on expectation-maximization(EM)algorithm.An 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 experimental results 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 treatment. These results can be use for predicting user's next request in the huge web sites.

The rest of this paper is organized as follows: In section 2, we review some researches that advance in navigation pattern mining. Section 3 describes the proposed method for the clustering of navigation pattern mining by EM algorithm. Results of an experimental evaluation are reported in section 4. Finally, section 5 summarizes the paper and introduces future work.

II. Related Works

Much research has focused on the web usage mining algorithms for mining user navigation behavior. In the following we review some of the most significant navigation pattern mining systems and algorithm in web usage mining area that can be compared with our system.

A partitioning method was one of the earliest clustering methods to be used in Web usage mining by Yan et al.[2]. They used an incremental algorithm that produces high quality clusters. Each user session is represented by an n-dimensional feature vector, where n is the number of Web pages in the session. The value of each feature is a weight, measuring the degree of interest of the user in the particular Web page. The calculation of this figure is based on a number of parameters, such as the number of times the page has been accessed and the amount of time the user spent on the page. Based on these

Norwati Mustapha is Assist. Prof. in faculty of computer science and information technology, Department of computer science, Putra University of Malaysia (UPM), Malaysia, email address: norwati@fsktm.upm.edu.my

Manijeh Jalali is Master student in faculty of science, Department of statistic, Islamic Azad University of Mashhad, Iran ,email address: manijehjalali@yahoo.com

Abolghasem Bozorgniya is Prof. in faculty of Science, Department of statistic, Islamic Azad University of Mashhad, Iran, email address: bozorgniya@math.um.ac.ir

Mehrdad Jalali is PhD candidate in faculty of computer science and information technology , Department+ of computer science, Putra University of Malaysia (UPM), email address: mehrdadjalali@ieee.org

vectors, clusters of similar sessions are produced and characterized by the Web pages with the highest associated weights. The characterized sessions are the patterns discovered by the algorithm. One problem with this approach is the calculation of the feature weights. The choice of the right parameter mix for the calculation of these weights is not straightforward and depends on the modeling abilities of a human expert

Cadez et al. [3] in the Web CANVAS tool proposed a partitioning clustering method, which visualizes user navigation paths in each cluster. In this system, user sessions are represented using categories of general topics for Web pages. A number of predefined categories are used as a bias, and URLs from the Web server log files are assigned to them, constructing the user sessions. The Expectation-Maximization (EM) algorithm[4], based on mixtures of Markov chains is used for clustering user sessions. Each Markov chain represents the behavior of a particular subgroup. EM is a memory efficient and easy to implement algorithm, with a profound probabilistic background. However, there are cases where it has a very slow linear convergence and may therefore become computationally expensive, although in the results in Cadez et al.[3], it is shown empirically that the algorithm scales linearly in all aspects of the problem.

The EM algorithm is also employed by Anderson et al. [5] in two clustering scenarios, for the construction of predictive Web usage models. In the first scenario, user navigation paths are considered members of one or more clusters, and the EM algorithm is used to calculate the model parameters for each cluster. The probability of visiting a certain page is estimated by calculating its conditional probability for each cluster. The resulting mixture model is named Naive Bayes mixture model since it is based on the assumption that pages in a navigation path are independent given the cluster. The second scenario uses a similar approach to[3]. Markov chains that represent the navigation paths of users are clustered using the EM algorithm, in order to predict subsequent pages.

Improving quality of clustering is main objective in all previous works. These works attempt to find architecture and algorithm for this purpose, but the quality still does not meet satisfaction. In this work we advance a model based on EM algorithm for improving accuracy of user navigation clustering.

III. System Design

The main goal of the proposed system is to find valuable information from the user access data (clickstream data) collected in web server log. In the next step these information is exploited to mining user navigation pattern based on EM algorithm.

According to different function, the system is partitioned into two main modules; Data pretreatment and navigation pattern mining. The model of the system is shown in Figure1.

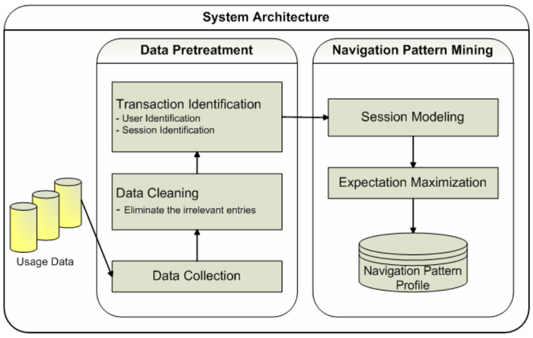

Fig. 1 Model of the system for unsupervised clustering

A. Data Pretreatment

There are several tasks in this module of the system. Data pretreatment in a web usage mining model, aims to reformat the original web logs to identify all web access sessions. The Web server usually registers all users' access activities of the website as Web server logs. Due to different server setting parameters, there are many types of web logs, but typically the log files share the same basic information, such as: client IP address, request time, requested URL, HTTP status code, referrer, etc.

Applying data cleaning algorithm is the next step in this module .Not every access to the content should be taken into consideration. We need to remove accesses to irrelevant items (such as button images, multimedia files), redundant items, accesses by Web crawlers (i.e. non-human accesses), and failed requests. Data cleaning also identifies Web robots and removes their request. Web robots (also called spiders) are software tools that scan a web site to extract its content.[6]

For web usage mining, to get knowledge about each user's identity is not necessary. However, a mechanism to distinguish different users is still required for analyzing user access behavior [7].

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. Sessions can have some missing parts. This is due to the browser's own caching mechanism and also because of the intermediate proxy-caches. The missing parts can be inferred from the site's structure [8]. Web usage data is prepared for applying navigation patterns mining algorithms by doing these pretreatment tasks.

B. Navigation Pattern Mining

The main objective of this module is mining of user's navigation pattern. A user navigation pattern is common browsing characteristics among a group of users. Since many users may have common interests up to a point during their navigation, navigation patterns should capture the overlapping interests or the information needs of these users. In addition, navigation patterns should also be capable to distinguish among web pages based on their different significance to each pattern.

The large majority of methods that have been used for navigation pattern mining from Web data are clustering methods. Clustering aims to divide a data set into groups that are very different from each other and whose members are very similar to each other. In this paper, a partitioning method is used for clustering of user's navigation patterns. Expectation maximization (EM) is a clustering algorithm that works based on partitioning methods. The EM is a memory efficient and easy to implement algorithm, with a profound probabilistic background. The detail of the algorithm is described in the next section.

C. Expectation Maximization Algorithm

Expectation maximization (EM) is a well-known algorithm used for clustering in the context of mixture models. EM was proposed by Demster et al. [4].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: first, 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 fixpoint 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 K-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, specifically:

1. The mean and standard deviation for cluster 1
2. The mean and standard deviation for cluster 2
3. The sampling probability *P* for cluster 1 (the probability for cluster 2 is *1-P*)

And the general procedure states as follow:

1. Guess initial values for the five parameters.
2. 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 σ, the formula is:

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

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

3. Use the probability scores to re-estimate the five parameters.
4. Return to Step 2 .

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. With two clusters *A* and *B* containing instances $x_1, x_2\ x_3, \ldots, x_n$ where $P_A=P_B=0.5$ the computation is:

$$[.5P\ (x_1|A)\ +.5P\ (x_1|B)]\ [.5P\ (x_2|A)\ +.5P\ (x_2|B)] \ldots [.5P\ (x_n|A)\ +.5P\ (x_n|B)]$$
(2)

C.1. navigation Pattern Mining by EM

After data preprocessing and exploratory data analysis have been completed, we can finally begin the modeling phase. This is the favorite part for many web usage miners, since it allows them to apply the range of their data mining skills and attack the problem at hand using an array of data mining methods, algorithms, and models.

For modeling of user navigation patterns, we need to apply training dataset to specific algorithm. In this study after preparing usage data in pretreatment phase we use EM clustering algorithm for mining of user's navigation patterns. In the next section there are some experimental evaluations that illustrate the efficiency of the system.

IV. Experimental Evaluation

Measuring the quality of the EM clustering in navigation patterns mining systems needs to characterize the quality of the results obtained.

The experimental evaluation was conduced using DePaul University CTI Log file. The only cleaning step performed on this data was the removal of references to auxiliary files (e.g., image files). No other cleaning or preprocessing has been performed in the first phase.

The data is in the original log format used by Microsoft IIS. The characteristics of the dataset we used are given in table1.

TABLE I DATASET USED IN THE EXPERIMENTS

Dataset	Size (Mb)	Records (Thousand)	Period (Days)
CTI	260	1051	14

Table 2 shows some basic statistics on user and sessions after cleaning, filtering and sessioning the CTI dataset.

TABLE II DATASET AFTER PREPROCESSING

Number of users	Size (Mb)	Number of sessions	Number of repeat users
5446	10	20950	2734

All evaluation tests were run on a dual processor Intel CPU 1.8 GHz Pentium 4 with 2GBytes of RAM, operating system Windows XP. Our implementations run on .net framework 2 and also we used Weka 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 attributes, each eight sequence of attributes is assigned to one attribute based on bitmap algorithm. In this case there are only 100 attribute for applying to EM algorithm. Tables 3 to Table 6 illustrate the processes of the data converting.

TABLE III URLS ADDRESS ASSIGN TO NUMERIC VALUE

Index	URLs Address
0	/admissions/
1	/admissions/career.asp
2	/admissions/checklist.asp
3	/admissions/costs.aspsalam
4	/admissions/default.asp
5	/admissions/general.asp
6	/admissions/helloworld/arabic.asp
7	/admissions/helloworld/chinese.asp
8	/admissions/helloworld/italian.asp
9	/admissions/helloworld/portugese.asp
…	…

TABLE IV *URLS THAT APPEAR IN EACH SESSION*

Session	URLs Number appear in the sessions									
1		1		3				8		..
2	0	1			4	5			9	..
3	0		2	3			7	8		..
4						5	6			..
5				3					9	..
…	.. .	..	..	..	..	..	..	..	..	..	...

TABLE V *URLS ASSIGN TO BINARY CODE*

Session	URLs Number appear in the sessions									
1	0	1	0	1	0	0	0	0	1	0	..
2	0	1	0	0	1	1	0	0	0	1	..
3	0	0	1	1	0	0	0	1	1	0	..
4	0	0	0	0	0	1	1	0	0	0	..
5	0	0	0	1	0	0	0	0	0	1	..
…	..	..	..	..	..	..	..	..	..	..	..

TABLE VI FINAL ASSIGNING BEFORE CLUSTERING

Session	Attribute 1	Attribute 2	…	Attribute 100
1	144	200	…	..
2	12	134	…	..
3	49	25	…	..
4	10	0	…	..
5	16	184	…	..
…	…	..	…	..

As we know An 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.

Figure 2 plots the Log likelihoods based on number of cluster in each training data. This experiment was accomplished by maximum 10 iterations. As shown in figure, by decreasing the number of clusters, the likelihood converges toward lower values. As we know, the number of cluster is difficult to decide. In our experiment, we tried several times to tune the parameters in order to get higher performance of clustering. We finally cluster the user's navigation patterns into 20 groups. Meanwhile, that the EM algorithm will get a local optimization after 10 iterations.

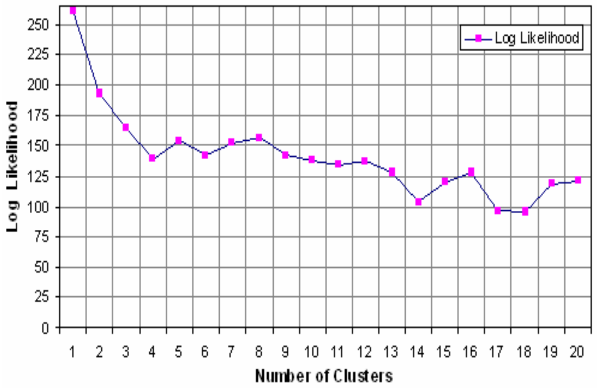

Fig. 2 Log Likelihood

Figure 3 plots probability of largest cluster in each treatment. For instance the percentage of the largest cluster is 34, while the experiment creates 20 clusters. Figure shows percentage of maximum cluster in clusters set will be decreased if the number of the cluster increases in each treatment.

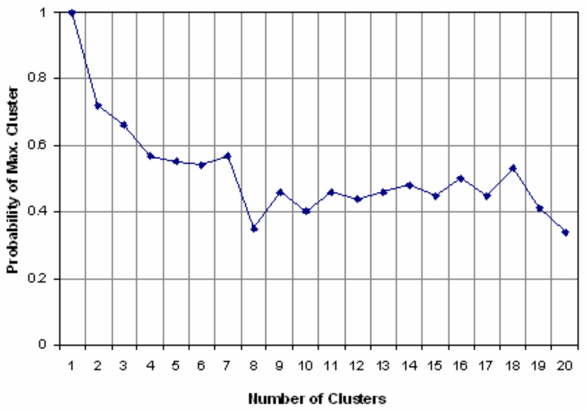

Fig. 3 Maximum cluster in each treatment

V. CONCLUSION

In this study we advance a model for mining of user's navigation pattern. The model makes user model based on expectation-maximization (EM) algorithm. The experimental results 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 treatment.

For the future, we would perfect the algorithm and apply some classification methods for classifying user request. This can be used in web usage mining-based prediction systems.

REFERENCES

[1] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on Web usage mining," *Communications of the ACM,* vol. 43, pp. 142-151, 2000.

[2] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, "From user access patterns to dynamic hypertext linking," *Computer Networks and ISDN Systems,* vol. 28, pp. 1007-1014, 1996.

[3] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "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, 2000.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society,* vol. 39, pp. 1-38, 1977.

[5] C. R. Anderson, P. Domingos, and D. S. Weld, "Adaptive Web Navigation for Wireless Devices," 2001, pp. 879-884.

[6] D. Tanasa and B. Trousse, "Advanced data preprocessing for intersites Web usage mining," *Intelligent Systems, IEEE,* vol. 19, pp. 59-65, 2004.

[7] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis," *INFORMS Journal on Computing,* vol. 15, pp. 171-190, 2003.

[8] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," *Knowledge and Information Systems,* vol. 1, pp. 5-32, 1999.