Efficient Web Usage Mining Based on K-Medoids Clustering Technique

P. Sengottuvelan, T. Gopalakrishnan

***Abstract*—**Web Usage Mining is the application of data mining techniques to find usage patterns from web log data, so as to grasp required patterns and serve the requirements of Web-based applications. User's expertise on the internet may be improved by minimizing user's web access latency. This may be done by predicting the future search page earlier and the same may be pre-fetched and cached. Therefore, to enhance the standard of web services, it is needed topic to research the user web navigation behavior. Analysis of user's web navigation behavior is achieved through modeling web navigation history. We propose this technique which cluster's the user sessions, based on the K-medoids technique.

***Keywords*—**Clustering, K-medoids, Recommendation, User Session, Web Usage Mining.

I. Introduction

WEB mining is one of the data mining technique to automatically discover and extract information from internet documents and services. Web mining is divided into three differing types, like web usage mining, web content mining and web structure mining.

Web usage mining is that the method of extracting helpful information from server logs. Every user is totally different from each other such as; some users may well be curious about the information that's provided in text format, whereas some others may well be curious about transmission or graphical content like footage, videos, audios, etc. Web Usage Mining discovers the attention by extracting usage patterns from web log so as to know the user behavior and fulfill the web users by applying certain data mining techniques.

Web personalization is the process of modifying user's list of web pages based on individual's preferences or tailoring the contents of list as per the individual's need. This practice is accomplished either by the selecting user from a menu of obtainable alternatives or by pursuit his or her behavior like that pages are accessed or however typically on the geographical location. The goal of personalization systems is to supply users need with what they have or need without specific indication from them. Currently, there are three key classes of personalization systems are available as the mentioned: manual decision rule systems, cooperative filtering system, and content-based filtering system. Personalization is predicated on any of the user attributes such as department or area, or role etc.

The quantity of web-based information offered has improved and exaggerated dramatically, these days. The way to gather helpful information from the online has become a difficult issue for users. Current web operation systems decide to satisfy user needs by capturing their information desires to service them at the fullest needs. For this purpose, user profiles are created for user background description. User profiles represent the conception models possessed by users by gathering web log information's through the user's search pattern. An inspiration model is implicitly possessed by users and is generated from their background. Whereas this idea model cannot be verified in laboratories, several internet ontologies have ascertained it in user behavior. Once users scan through a document, they will simply confirm whether or not it's of their interest or relevancy to them, a judgment that arises from their implicit conception models. If a user's conception model is simulated, then a superior illustration of user profiles can be built for the every user's on the web.

The WWW is taken into account as associate degree information hub and as a result, it contains large quantity of data and that we will access it through totally different web site. However, the storage and show of data and material on web site is not quite enough and specially for e-learner it becomes feverish, hazardous, boring and time intense as a result of one cannot realize or perceive the relevant information from the online. Once if the user spends much of his time for searching the desired information will be a main problem and another problem might be to look for the specific content in a page where the web page designer has dumbed more content on to the single page instead of having more pages. Both these problem would be seriously seen by the web user. That's why the user/learner loses his/her interest and leaves the web site when he is not obtaining its desired information.

One of the main applications of above said methods is at e-Learning platforms. The general advantages of Web-based teaching are training is usually self-paced, highly interactive, which results in increased holding rates, and has reduced costs associated with student to travel to an instructor-led workshop.

II. Related Work

A number of researchers attempted to improve the Web page access prediction precision or coverage by combining different recommendation frameworks [1]-[6]. For instance, many papers combined clustering with association rules [7]-[12]. Lai & Yang [7] have introduced a customized marketing on the Web approach using a combination of clustering and

Dr P. Sengottuvelan is with the Department of Information Technology, Bannari Amman Institute of Technology, Sathy- 638401, India (e-mail: sengottuvelan@rediffmail.com).

T Gopalakrihsnan is with the Department of Information Technology, Bannari Amman Institute of Technology, Sathy- 638401, India (phone: 9942832002; e-mail: gopalakrishnan.ct@gmail.com).

association rules. The authors collected information about customers using forms, Web server log Files and cookies. They categorized customers according to the information collected. K-means clustering algorithm works only with numerical data. They then performed association rules techniques on each cluster. They are proved through experimentations that implementing association rules on clusters achieves better results than on non-clustered data for customizing the customers' marketing preferences. Liu et al. [12] have introduced MARC (Mining Association Rules using Clustering) that helps reduce the I/O overhead associated with large databases by making only one pass over the database when learning association rules. The authors group similar transactions together and they mine association rules on the summaries of clusters instead of the whole dataset. Although the authors prove through experimentation that MARC can learn association rules more efficiently, their algorithm does not improve on the accuracy of the association rules learned. Other papers combined clustering with Markov model [13]-[15]. Cadez et al. [13] partitioned site users using a model-based clustering approach where they implemented First order Markov model using the Expectation-Maximization algorithm. After partitioning the users into clusters, they displayed the paths for users within each cluster. They also developed a visualization tool called Web CANVAS based on their model. Zhu et al. [15] constructed Markov models from log Files and use co-citation and coupling similarities for measuring the conceptual relationships between Web pages. Citation Cluster algorithm is then proposed to cluster conceptually related pages. A hierarchy of the Web site is constructed from the clustering results. The authors then combine Markov model based link prediction to the conceptual hierarchy into a prototype called ONE to assist users' navigation. Lu et al. [14] were able to generate Significant Usage Patterns (SUP) from clusters of abstracted Web sessions. Clustering was applied based on a two-phase abstraction technique. First, session similarity is computed using Needleman-Wunsch alignment algorithm and sessions are clustered according to their similarities. Second, a concept-based abstraction approach is used for further abstraction and a First order Markov model is built for each cluster of sessions. SUPs are the paths that are generated from First order Markov model with each cluster of user sessions.

Combining association rules with Markov model is novel to our knowledge and only few of past researches combined all three models together. In [16], Kim et al. improve the performance of Markov model, sequential association rules, association rules, and clustering by combining all these models together. For instance, Markov model is used first. If MM cannot cover an active session or a state, sequential association rules are used. If sequential association rules cannot cover the state, association rules are used. If those association rules cannot cover the state, clustering algorithm is applied. Work of Kim et al. [16] improved recall and it did not improve the Web page prediction accuracy.

III. Proposed Methodolgy

In order to improve the web page prediction accuracy the following system is been proposed. The web server data is given as the dataset here.

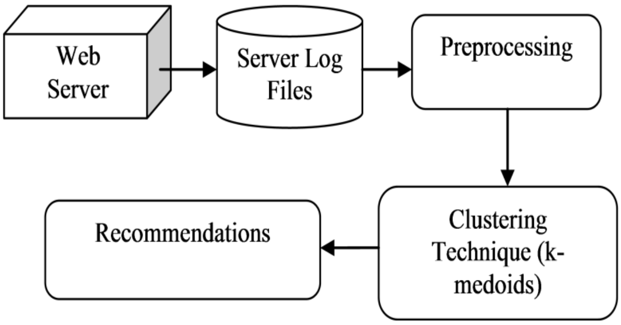


Fig. 1 System Architecture

A. Web Server Data

When any user agent (e.g., IE, Mozilla, Netscape, etc) hits an URL in a domain, the information related to that operation is recorded in an access log file. In the data processing task, the web log data can be preprocessed in order to obtain session information for all users. Access log file on the server side contains log information of user that opened a session [17]-[20]. These records have seven common fields, which are:

- User's IP address
- Access date and time
- Request method (GET or POST)
- URL of the page accessed
- Transfer protocol (HTTP 1.0, HTTP 1.1,)
- Success of return code
- Number of bytes transmitted

B. Preprocessing

The quality of the patterns discovered in web usage mining process highly depends on the quality of the data used in the mining processes [17]. When the web browser traces the web pages and stores the Server log file. Web usage data contains information about the Internet addresses of web users with their navigational behavior the basic information source for web usage [18].

The cleaning process of web log data is conducted in order to remove the unwanted data items (irrelevant data).This data cleaning process is done based on four criteria.

1) File extension
2) Respond code from web server
3) Access methods
4) User access frequency

C. Clustering

Based on the access of the user, the documents are clustered. They are clustered by checking those documents with their threshold values already assumed. This kind of clustering makes it easy for the users further search and makes the search more easy and fastest. They are clustered by checking whether those documents match the same format which the user has accessed previously. By that way of grouping, only user interested document formats are been recommended to the users. This kind of clustering makes it easy for the users further search and makes the search easy and fast.

One of the problems of the k-means algorithm is that it gives a hard partitioning of the data, i.e. to say each point is attributed to one and only one cluster. But points on the edge of the cluster, or near another cluster, may not be as much in the cluster as points in the center of cluster.

1) K-medoids

The *k*-medoids algorithm is clustering algorithm related to the *k*-means algorithm and the medoids shift algorithm. Both the *k*-means and *k*-medoids algorithms are partitioned (breaking the dataset up into groups) and both attempt to minimize squared error, the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the *k*-means algorithm, *k*-medoids chooses data points as centers (medoids or exemplars).

K-medoids is also a partitioning technique of clustering that clusters the data set of *n* objects into *k* clusters with *k* known *a priori*. A useful tool for determining *k* is the silhouette.

It could be more robust to noise and outliers as compared to *k*-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances. The possible choice of the dissimilarity function is very rich but in our applet, we used the squared Euclidean distance.

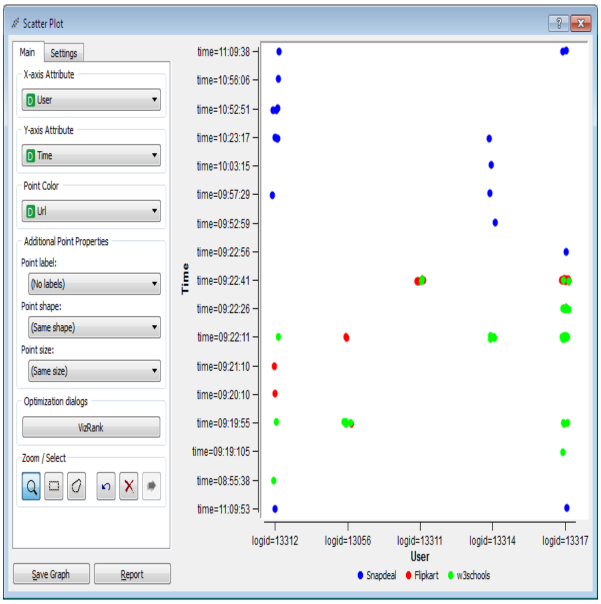
A medoid of a finite dataset is a data point from this set, whose average dissimilarity to all the data points is minimal i.e. it is the most centrally located point in the set.

The common realization of *k*-medoid clustering is the Partitioning around Medoids (PAM) algorithm and is as:

1) ***Initialize:*** randomly select *k* of the *n* data points as the medoids
2) ***Assignment:*** Associate each data point to the closest medoid.
3) ***Update step:*** For each medoid *m* and each data point *o* associated to *m* swap *m* and *o* and compute the total cost of the configuration (that is, the average dissimilarity of *o* to all the data points associated to *m*).
4) Select the medoid *o* with the lowest cost of the configuration.
5) Repeat alternating steps 2 and 3 until there is no change in the assignments.

IV. Experimental Results

The proposed system is developed by using the specific data mining tool called Orange. This is a very capable open source visualization and analysis tool with an easy to use interface. Most analysis can be achieved through its visual programming interface (drag and drop of widgets) and most visual tools are supported including scatter plots, bar charts, trees, dendrogram and heat maps.



X-axis Attribute: User

Y-axis Attribute: Time

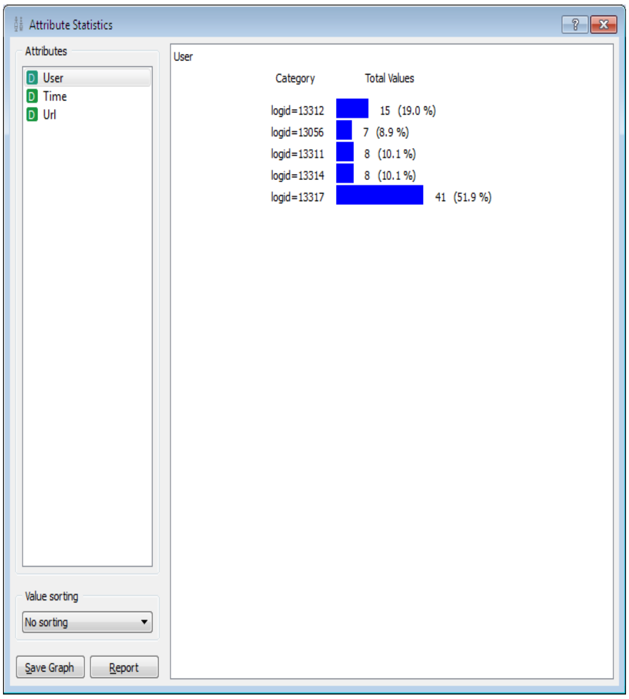
Point Color: Url

Point label: (No labels)

Point shape: (Same shape)

Point size: (Same size)

Fig. 2 Scatter plot based on the access of the user



Value sorting: No sorting

User

Category	Total Values
logid=13312	15 (19.0 %)
logid=13056	7 (8.9 %)
logid=13311	8 (10.1 %)
logid=13314	8 (10.1 %)
logid=13317	41 (51.9 %)

Fig. 3 Attribute statistics of the users based on the user id

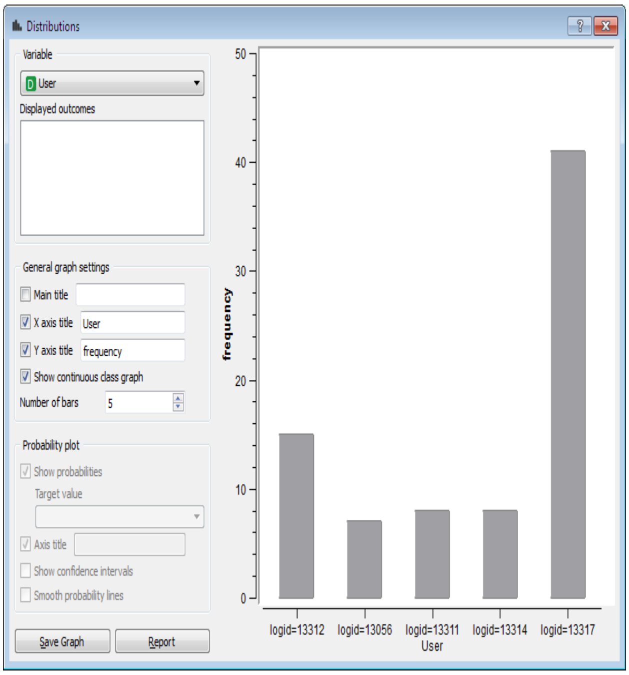
A large number (over 100) of widgets are supported. These includes most advanced features like data transformation using classification, regression, visualization, unsupervised learning methods etc.. . There are various specialized add-ons covering different applications namely bioinformatics, text mining and other specialist requirements.

The web log file is given as an input to the system. Log files are generated by web servers automatically based on the searches made by the user at the end of the every search request. Each time a visitor requests any file (page, image, etc.) from the site, information on his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text. Fig. 2 shows the access of different users at various time intervals and they are clustered based on the access to the websites.

Fig. 3 shows the attribute statistics of the users and it shows which user has accessed for more number of times.

Fig. 4 shows the distributions of different users based on the frequency of the user's access and which user has accessed the particular website frequently.

Table I shows sample records in given web log file is 200.The data is collected from the different user id's and the analysis is done to form the clusters.



Variable: D User

Displayed outcomes

General graph settings

☐ Main title

☑ X axis title: User

☑ Y axis title: frequency

☑ Show continuous class graph

Number of bars: 5

Probability plot

☑ Show probabilities

Target value

☑ Axis title

☐ Show confidence intervals

☐ Smooth probability lines

Fig. 4 Distributions of the users based on the frequency of the user's access

TABLE I
THE NUMBER OF RECORDS IN EACH CLUSTERED LOG ID

Log id	NO. OF RECORDS
13312	38
13056	17
13311	21
13314	21
13317	103

V. CONCLUSION AND FUTURE ENHANCEMENTS

In this paper, we presented a clustering of web usage data, which is most useful in finding the user access patterns and the order of visits of the hyperlinks of the each user. The suggested approach was used for efficiency contained a hard clustering of the data set and as the analysis indicate each of the individual clusters seems to contain observations with specific common features and improve the algorithm efficiency with help of k-medoids clustering algorithm. Experiments prove that this system has high prediction accuracy by using the appropriate data mining tool orange. As a further improvement, we can still enhance the quality of data by applying two level clustering techniques.

REFERENCES

[1] Mojtaba Salehi, Isa Nakhai Kamalabadi, and Mohammed B. Ghaznavi Ghoushchi, "An effective Recommendation Framework for Personal Learning Environments using a Learner Preference Tree and a GA," IEEE Transactions on learning technologies, vol. 6, No. 4,2013

[2] M. Salehi, M. Pourzaferani, and S.A. Razavi, "Hybrid Attribute-Based Recommender System for Learning Material Using Genetic Algorithm and a Multidimensional Information Model," Egyptian Informatics J., vol. 14, no. 1, pp. 67-78, 2013.

[3] Yi Li,Jian Wang,Lin Mei, "A Personalized Recommendation System in E-Learning Environment based on Semantic Analysis", Information Science and Service Science and Data Mining (ISSDM), 6th International Conference on New Trends 2012.

[4] J. kay, "Lifelong Learner Modeling for Lifelong Personalized Pervasive Learning," IEEE Trans. Learning Technology, vol. 1, no. 4, pp. 215-228, Oct. 2008.

[5] Kumar, J. Nesbit, and K. Han, "Rating Learning Object Quality with Distributed Bayesian Belief Networks: The Why and the How," Proc. Fifth IEEE Int'l Conf. Advanced Learning Technologies (ICALT '05), pp. 685-687, 2005.

[6] N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel, and R. Koper, "Recommender Systems in Technology Enhanced Learning," Recommender Systems Handbook, P.B. Kantor, F. Ricci, L. Rokach, and B. Shapira, eds., pp. 387-415, Springer, 2011.

[7] Lai, H. & Yang, T. C. (2000), "A group-based inference approach to customized marketing on the web integrating clustering and association rules techniques" Hawaii International Conferenceon system sciences pp. 37 – 46.

[8] W. Chen, Z. Niu, X. Zhao, and Y. Li, "A Hybrid Recommendation Algorithm Adapted in E-Learning Environments," World Wide Web, Sept. 2012, doi:10.1007/s11280-012-0187-z.

[9] F. Masseglia, P. Poncelet, and M. Teisseire, "Using data mining techniques on web access logs to dynamically improve hypertext structure". In ACM SigWeb Letters, 8(3): 13-19, 1999.

[10] V elasquez, Bassi J D, YasudaA. "Mining Web data to create online navigation recommendations". Data Mining, 2004:166-172. Proceedings of the Fourth IEEE International Conference on data mining (ICDM'04) 0-7695-2142-8/04 IEEE.

[11] R. Vaarandi, "A Data Clustering Algorithm for Mining Patterns from Event logs" in Proceedings of the 3rd IEEE Workshop on IP Operations and Management.

[12] Liu, F., Lu, Z. & Lu, S. (2001), `Mining association rules using clustering', Intelligent Data Analysis (5), 309 - 326.

[13] Cadez, I., Heckerman, D., Meek, C., Smyth, P. & White, S. (2003), "Model-based clustering and visualization of navigation patterns on a web site", Data Mining and Knowledge Discovery .

[14] Lu, L., Dunham, M. & Meng, Y. (2005), "Discovery of significant usage patterns from clusters of clickstream Data", WebKDD '05 .

[15] Zhu, J., Hong, J. & Hughes, J. G. (2002), "Using markov models for web site link prediction", HT'02, USA pp. 69 - 170.

[16] Kim, D., Adam, N., Alturi, V., Bieber, M. & Yesha, Y. (2004), "A clickstream-based collaborative Filtering personalization model: Towards a better performance", WIDM '04 pp. 88 - 95.

[17] Faten Khalil, Jiuyong Li, Hua Wang, "Integrating Recommendation Models for Improved Web Page Prediction Accuracy", in Proceedings of the thirty-first Australasian conference on Computer science, Vol. 74, 2008.

[18] P Sengouttuvelan, T Gopalakrishnan, V. S. Gowthami"A Pattern Recognition Technique for Learning Style Prediction System", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 9 (2015)

[19] S. Rafaeli, Y. Dan-Gur, and M. Barak, “Social Recommender Systems: Recommendations in Support of E-Learning,” Int’l J. Distance Education Technologies, vol. 3, no. 2, pp. 29-45, 2005.

[20] T Gopalakrishnan,Dr P Sengouttuvelan ,”Discovering user profiles for web personalization using EM with Bayesian Classification”Australian Journal of Basic and Applied Sciences, l.8(3) March 2014, Pages: 53-60