Web Usage Mining for Predicting Users' Browsing Behaviors by using FPCM Clustering

R.Khanchana and M. Punithavalli.

Abstract--**World Wide Web is a huge storehouse of web pages and links. It offers large quantity of data for the Internet users. The growth of web is incredible as around one million pages are added per day. Users' accesses are recorded in web logs. Web usage mining is a kind of mining techniques in logs. Because of the remarkable usage, the log files are growing at a faster rate and the size is becoming very large. This leads to the difficulty for mining the usage log according to the needs. This provides a vast field for the researchers to provide their suggestion to develop a better mining technique. Then the researchers propose the hierarchical agglomerative clustering to cluster users' browsing patterns. The provided prediction by two levels of prediction model framework work healthy in general cases. On the other hand, two levels of prediction model experience from the heterogeneity user's behavior. In this paper, the author enhances the two levels of Prediction Model to achieve higher hit ratio. This paper uses Fuzzy Possibilistic algorithm for clustering. The experimental result shows that the proposed techniques results in better hit ratio than the existing techniques.**

***Index Terms*---Web Usage Mining, Hierarchical Agglomerative Clustering, Fuzzy Possibilistic Clustering**

I. Introduction

In this internet era web sites on the internet are useful source of information in day to day activities. So there is a rapid development of World Wide Web in its volume of traffic and the size and complexity of web sites. As per August 2010 Web Server survey by Netcraft there are 213,458,815 active sites. Web mining is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces user's visiting behaviors and extracts their interests using patterns. Because of its direct application in e-commerce, Web analytics, e-learning, information retrieval etc., web mining has become one of the important areas in computer and information science. Web Usage Mining uses mining methods in log data to extract the behavior of users which is used in various applications like personalized services, adaptive web sites, customer profiling, prefetching, creating attractive web sites etc.

Web servers accumulate data about user's interactions in log files whenever requests for resources are received. Log files records information such as client IP address, URL requested etc., in different formats such as Common Log format, Extended Common Log format which is issued by Apache and IIS.

Manuscript received March 30, 2011, revised August 15, 2011.

R khanchana is with a research scholar, Department of Computer science, Karpagam University, Coimbatore, TamilNadu, India.(E-mail: kanchusri@gmail.com).

M Punithavalli is with Dean and Director, Department of Computer science, Sns College, Bharathiar University, Coimbatore, Tamil, India. (E-mail: mpunitha_srcw@yahoo.co.in).

The intention of web usage mining is to identify the useful data from web data or web log files. The additional purposes are to improve the usability of the web data and to apply the techniques on the web applications like, pre-fetching and caching, personalization etc. For assessment management, the outcome of web usage mining can be utilized for target advertisement, enhancing web design, enhancing satisfaction of customer, guiding the strategy decision of the enterprise, and marketing analysis etc.

Predicting the users' browsing pattern is one of web usage mining technique. For this purpose, it is required to recognize the customers' browsing behaviors by means of analyzing the web data or web log files. Predicting the exact user's next needs is according to the earlier related activities. There are several merits to employ the prediction, for example, personalization, building proper web site, enhancing marketing strategy, promotion, product supply, getting marketing data, forecasting market trends, and increasing the competitive strength of enterprises etc.

This paper focuses on Web Usage Mining with the help of clustering technique. The clustering will perform classification in the browsing features. This paper uses Fuzzy Possibilistic algorithm for the purpose of clustering.

II. Related Works

Tasawar *et al.,* [1] proposed a hierarchical cluster based preprocessing methodology for Web Usage Mining. In Web Usage Mining (WUM), web session clustering plays a important function to categorize web users according to the user click history and similarity measure. Web session clustering according to Swarm assists in several manner for the purpose of managing the web resources efficiently like web personalization, schema modification, website alteration and web server performance. The author presents a framework for web session clustering in the preprocessing level of web usage mining. The framework will envelop the data preprocessing phase to practice the web log data and change the categorical web log data into numerical data. A session vector is determined, so that suitable similarity and swarm optimization could be used to cluster the web log data. The hierarchical cluster based technique will improve the conventional web session method for more structured information about the user sessions.

Yaxiu *et al.,* [2] put forth web usage mining based on fuzzy clustering. The World Wide Web has turn out to be the default knowledge resource for several fields of endeavor, organizations require to recognize their customers' behavior, preferences, and future requirements, but when users browsing the Web site, several factors influence their interesting, and various factor has several degree of influence, the more factors consider, the more precisely can mirror the user's interest. This paper provides the effort to cluster similar Web user, by involving two factors that the page-click number and Web browsing time that are stored in the Web log, and the various degree of influence of the two factors. The method suggested in this paper can help Web site organizations to recommend Web pages, enhance Web structure, so that can draw more customers, and improves customers' loyalty.

Web usage mining based on fuzzy clustering in identifying target group is suggested by Jianxi *et al.,* [3]. Data mining deals with the methods of non-trivial extraction of hidden, previously unknown, and potentially helpful data from very huge quantity of data. Web mining can be defined as the use of data mining methods to Web data. Web usage mining (WUM) is an significant kind in Web mining. Web usage mining is an essential and fast developing field of Web mining where many research has been performed previously. The author enhanced the fuzzy clustering technique to identify groups that share common interests and behaviors by examining the data collected in Web servers.

Houqun *et al.,* [4] proposed an approach of multi-path segmentation clustering based on web usage mining. According to the web log of a university, this paper deals with examining and researching methods of web log mining; bringing forward a multi-path segmentation cluster technique, that segments and clusters based on the user access path to enhance efficiency.

III. Methodology

A. Two Levels of Prediction Model

A Two Levels of Prediction Model is suggested by Lee and Fu which is shown in figure 1[17]. This technique reduces the prediction scope with the help of the two levels framework. The Two Levels of Prediction Model are created by merging the Markov model and Bayesian theorem. In first level, Markov model is utilized for the purpose of filtering the highly probable of categories that will be surfed by user. In the second level, Bayesian theorem is utilized to assume precisely the maximum probability of web page.

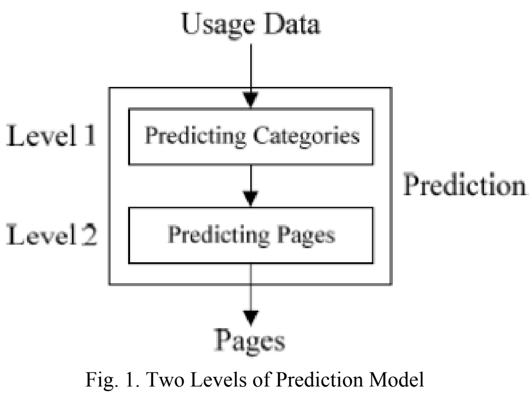

Fig. 1. Two Levels of Prediction Model

In first level, it is to forecast the highly probable user's present state (web page) of group at time t, that depends on user's category at time t-1 and time t-2. Bayesian theorem is utilized to forecast the highly probable web pages at a time t based on user's states at a time t-1. At last, the prediction result of two levels of prediction model is provided.

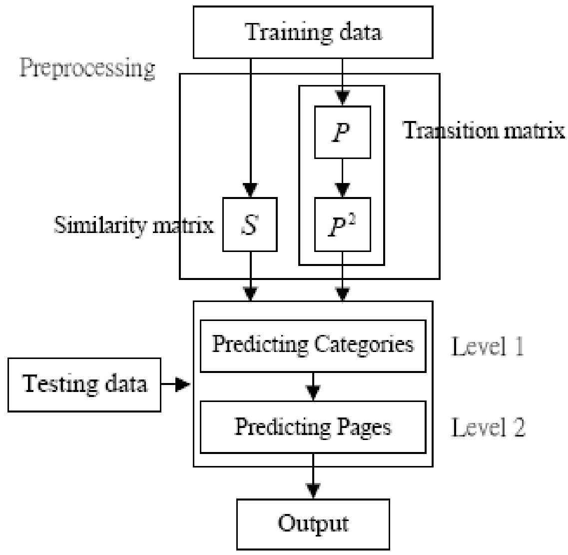

Fig. 2. Two Levels of Prediction Model Framework

The Two Levels of Prediction Model framework is provided in figure 2. In the first step, the similarity matrix S of category is created. The technique of creating similarity matrix is to collect statistics and to examine the users' behavior browsing that can be obtained from web log data. In the second step, it is to create the first-order transition matrix P and second-order transition matrix P2 of Markov model. The transition matrix of Markov is created by the similar technique, statistical method, from web log file.

In the final step, the relevance matrix R is created from first-order and second-order transition matrix of Markov model and similarity matrix. When the value of relevance is higher then the value of transition probability and similarity among categories is also higher. The relevance matrix is an significant feature of forecasting. Relevance matrix can be utilized to assume the users' browsing pattern among web categories.

B. Modified Prediction Model

The author spotlights on the preprocessing process and alters the Two Levels of Prediction Model framework additional. As the users' browsing patterns are varied, the hierarchical agglomerative clustering is utilized to cluster users' browsing patters and obtain several various user clusters. The data of clusters can be projected as cluster view for replacing of the global. As a result, the author presents a altered Prediction Model. In the new model, the view selection will be utilized by which user's browsing patterns is matched and utilized for forecasting and enhancing the accuracy confidently.

The steps of proposed technique are provided in figure 3 that are explained below. The user database will be separated into training data and testing data. In the first step of preprocessing, training data is passed to the hierarchical agglomerative clustering. The k number of cluster view (as in figure 4) will be resulted that consist of k similarity matrices S, k first-order transition matrices P and k second-order transition matrices P^2 among clusters. Hence, k relevant matrices R are obtained to indicate k cluster views. In the second step, the centric vector of clusters will be free for generating an index table. The index table is utilized for view selection according to the user's browsing behavior in time.

In the third step, after view selection, testing data will be passed to the prediction model. The prediction result will be obtained from the model as a result.

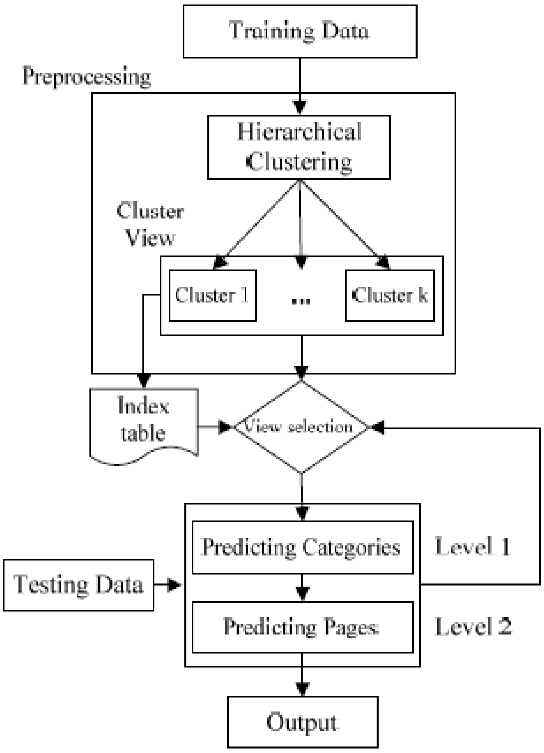

Fig. 3. Research Framework

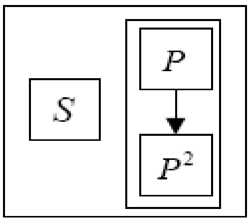

Fig. 4. Cluster View of Users

C. Preprocessing: Hierarchical Agglomerative Clustering

The hierarchical agglomerative clustering is used for classifying the users' browsing behaviors. The similarity is determined by Euclidean distance function. In the initialization, the entire users are predicted to be a cluster. The comparably same users' browsing characteristic will be identified and combined into a cluster till terminal condition is contented. At last, the user clusters and index table will be obtained as a result.

D. Pattern representation

In the process of pattern depiction, user sessions are generated from web log files. User sessions can be restructured as a m x k matrix as in table 1, each row can be represented by $session^u = (p_{u,1}, p_{u,2}, p_{u,k})$where , $p_{u,i} = 1$ indicates that user u browsed the web page else , $p_{u,i}=0$. The k represents the number of web pages. The session shows the user's browsing position that is also user's browsing characteristics.

TABLE 1. USER SESSIONS

	P_1	P_2	…	P_k
session^1	1	0	…	1
session^2	0	1	…	0
. . .	. . .	. . .	…	. . .
session^m	1	1	…	1

E. Definition Similarity Measure

The similarity among any two users can be determined with the help of distance measure. Euclidean distance function (1) is utilized for computing the similarity amond user i and user j, the similarity can be indicated by $Sim(user_i, user_j) = (session^i\ session^j)$. Euclidean distance is additional normalized by equation (2). Additional, the m x m matrix of user similarity will be resulted.

Euclidean distance：

$$D(user_i, user_j) = \sqrt{\sum_{l=1}^{k}(p_{i,l} - p_{j,l})^2} \qquad (1)$$

Normalization：

$$ND(user_i, user_j) = 1 - \sqrt{\frac{\sum_{l=1}^{k}(p_{i,l} - p_{j,l})^2}{k}} \qquad (2)$$

F. Clustering

In the hierarchical agglomerative clustering technique, the distances are calculated among centroids of clusters. The two clusters are combined by the shortest distance among two centroids. Finally, the new centroid vector of newly obtained cluster will be determined by equation (3).

It is implicited there are n objects in a cluster, the characteristics of every object can be indicated by $(p_{i,1}, p_{i,2}, \ldots, p_{i,k})$, where $1 \leq i \leq n$. The ceotroid vector of cluster can be determined by:

$$centroid^{cluster} = \left(\frac{\sum_{i=1}^{n} p_{l,1}}{n}, \frac{\sum_{i=1}^{n} p_{l,2}}{n}, \frac{\sum_{i=1}^{n} p_{l,k}}{n}\right) \qquad (3)$$

Hierarchical agglomerative clustering used are as follows:

(1) Initialization cluster:
  (1.1) Each object be a cluster.
  (1.2) Creating similarity matrix of users.
(2) Clustering:
  (2.1) Finding a pair of the most similar clusters and merging.
  (2.2) Computing the new centroid vector of new cluster.
  (2.3) Computing the distances between new cluster and others.
  (2.4) Pruning and updating the similarity matrix.
  (2.5) If the terminal condition is satisfied then output, else repeating 2.1 to 2.4.
(3) Cluster output:
  (3.1) Output index table.
  (3.2) Output all clusters.

G. Fuzzy Possibilistic Clustering Algorithm

The fuzzified version of the k-means algorithm is Fuzzy C-Means (FCM). It is a clustering approach which allows one piece of data to correspond to two or more clusters. Dunn in 1973 developed this technique and it was modified by Bezdek in 1981. The algorithm is an iterative clustering approach that brings out an optimal c partition by minimizing the weighted within group sum of squared error objective function J_{FCM}:

$$J_{FCM}(V,U,X)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\,d^{2}(X_j,v_i),1<m<+\infty \quad (4)$$

In the equation $X = \{x_1, x_2,...,x_n\} \subseteq R^p$ is the data set in the p-dimensional vector space, the number of data items is represented as p, c represents the number of clusters with $2 \leq c \leq n-1$. $V = \{v_1, v_2, . . . ,v_c\}$ is the c centers or prototypes of the clusters, v_i represents the p-dimension center of the cluster i, and $d^2(x_j, v_i)$ represents a distance measure between object x_j and cluster centre v_i. $U = \{\mu ij\}$ represents a fuzzy partition matrix with $u_{ij} = u_i (x_j)$ is the degree of membership of x_j in the ith cluster; x_j is the jth of p-dimensional measured data. The fuzzy partition matrix satisfies:

$$0\prec\sum_{j=1}^{n}\mu_{ij}\prec n,\forall i\in\{1,\dots,c\} \quad (5)$$

$$\sum_{i=1}^{c}\mu_{ij}=1,\forall j\in\{1,\dots,n\} \quad (6)$$

m is a weighting exponent parameter on each fuzzy membership and establishes the amount of fuzziness of the resulting classification; it is a fixed number greater than one. Under the constraint of U the objective function J_{FCM} can be minimized. Specifically, taking of J_{FCM} with respect to u_{ij} and v_i and zeroing them respectively, is necessary but not sufficient conditions for J_{FCM} to be at its local extrema will be as the following:

$$\mu_{ij}=\left[\sum_{k=1}^{c}\left(\frac{d(X_jv_i)}{d(X_jv_k)}\right)^{2/(m-1)}\right]^{-1},1\leq i\leq c,1\leq j\leq n. \quad (7)$$

$$v_i=\frac{\sum_{k=1}^{n}\mu_{ik}^{m}x_k}{\sum_{k=1}^{n}\mu_{ik}^{m}},1\leq i\leq c. \quad (8)$$

In noisy situations, the memberships of FCM do not constantly correspond better to the degree of belonging of the data, and may be mistaken. This is usually results from the real data unavoidably includes few noises. To overcome this difficulty of FCM, the constrained condition (6) of the fuzzy c-partition is not taken into account to obtain a possibilistic type of membership function and PCM for unsupervised clustering is proposed. The factor produced by the PCM belongs to a dense region in the data set; every cluster is not dependent of the other clusters in the PCM technique. The following formulation is the objective function of the PCM.

$$J_{PCM}(V,U,X)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ik}^{m}\,d^{2}(X_j,v_i)+\sum_{i=1}^{c}\eta_i\sum_{j=1}^{n}(1-u_{ij})^{m} \quad (9)$$

where

$$\eta_i=\frac{\sum_{j=1}^{n}\mu_{ik}^{m}||x_j-v_i||^2}{\sum_{j=1}^{n}\mu_{ij}^{m}} \quad (10)$$

η_i is the scale parameter at the ith cluster,

$$u_{ij}=\frac{1}{1+\left[\frac{d^2(x_j,v_i)}{\eta_i}\right]^{\frac{1}{m-1}}} \quad (11)$$

u_{ij} represents the possibilistic typicality value of training sample x_j belong to the cluster i. $m \in [1,\infty]$ is a weighting factor said to be the possibilistic parameter. PCM is also according to the initialization typical of other cluster techniques. The clusters do not have much mobility in PCM method, as each data point is categorized as only one cluster at a time rather than all the clusters simultaneously. Consequently, a appropriate initialization is necessary for the algorithms to converge to nearly global minimum.

The features of both fuzzy and possibilistic c-means techniques is combined for better result. Memberships and typicalities are very significant characteristics for the proper characteristics of data substructure in clustering problem. Consequently, an objective function in the FPCM depending on both memberships and typicalities can be represented as below:

$$J_{FPCM}(U,T,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij}^{m}+t^{n})\,d^{2}(X_j,v_i) \quad (12)$$

with the following constraints :

$$\sum_{i=1}^{c}\mu_{ij}=1,\forall j\in\{1,\dots,n\} \quad (6)$$

$$\sum_{j=1}^{n}t_{ij}=1,\forall i\in\{1,\dots,c\} \quad (13)$$

A result of the objective function can be resulted by means of an iterative procedure where the degrees of membership, typicality and the cluster centers are update with the equations as below.

$$\mu_{ij}=\left[\sum_{k=1}^{c}\left(\frac{d(X_j,v_i)}{d(X_j,v_k)}\right)^{2/(m-1)}\right]^{-1},1\leq i\leq c,1\leq j\leq n. \quad (7)$$

$$t_{ij}=\left[\sum_{k=1}^{n}\left(\frac{d(X_j,v_i)}{d(X_j,v_k)}\right)^{2/(\eta-1)}\right]^{-1},1\leq i\leq c,1\leq j\leq n. \quad (14)$$

$$v_i=\frac{\sum_{k=1}^{n}(\mu_{ik}^{m}+t_{ik}^{\eta})X_K}{\sum_{k=1}^{n}(\mu_{ik}^{m}+t_{ik}^{\eta})},1\leq i\leq c. \quad (15)$$

FPCM constructs memberships and possibilities simultaneously, together with the normal point prototypes or cluster centers for every cluster. Hybridization of

Possibilistic C-Means (PCM) AND Fuzzy C-Means (FCM) is the FPCM that frequently rejects several drawbacks of PCM, FCM . The noise sensitivity fault of FCM is solved by FPCM, which conquers the concurrent clusters drawbacks of PCM.

H. View Selection and Prediction

The appropriate view will be chosen when forecasting a user's browsing pattern. That is to select the appropriate relevant matrix for user. The view selection is carry out by taking into consideration of the similarity among user session vector and the vectors of index table. Predicting the current user's browsing behavior through the Two Levels of Prediction Model after the suitable view is chosen.

IV. Experimental Results

For evaluating the proposed technique, the database is selected from UCI Dataset Repository [18] which is a popular dataset repository for various research fields. The dated of the database is September, 28, 1999 on msnbc.com website and a part of news data is from msn.com website. Every sequence data is consequent to users' page views. The time interval is 24 hours. The categories are presented in sequential data. The 17 types of categories are frontpage, news, tech, local, opinion, on-air, music, weather, health, living, business, sports, summary, bbs, travel, msn-news, and msn-sports. The number of category is represented from 1 to 17 in the sequential data.

The resulted hit ratio for the level one is indicated in figure 5, there are five cases in level one which are Top-1 to Top-5. The Hit Ratio of global view is from 70.1% on Top-1 relevance to 85.2% on Top-5 relevance. The Hit Ratio of hierarchical cluster view is from 75.3% on Top-1 relevance to 89.2% on Top-5 relevance. The Hit Ratio of FPCM cluster view is from 75.3% on Top-1 relevance to 89.2% on Top-5 relevance. From this result, it can be observed that the FPCM cluster view is better than hierarchical cluster view and global view, which can be selected for predicting.

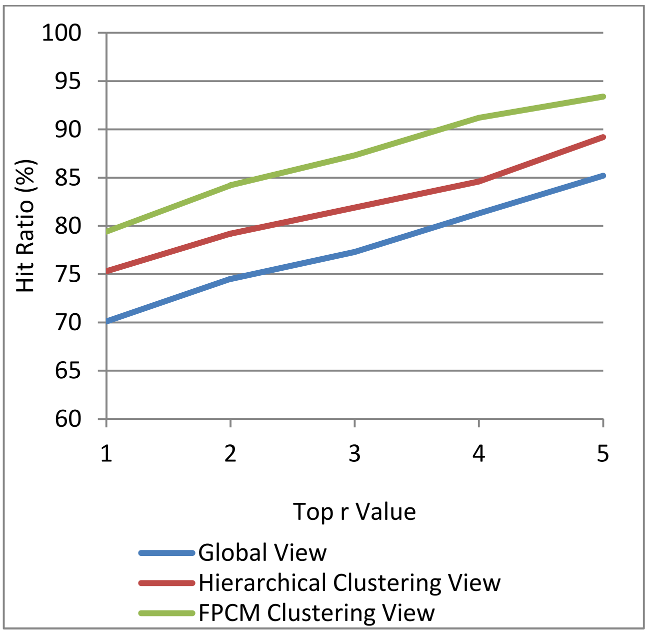

Fig. 5. Hit Ratio of Level One

The resulted hit ratio for the level two is indicated in figure 6 in which every category consists of 40 web pages. The Hit Ratio of global view is from 57.8% on Top-5 probability to 84.5% on Top-20 probability. The Hit Ratio of hierarchical cluster view is from 59.9% on Top-5 probability to 87.7% on Top-20 probability. The Hit Ratio of FPCM cluster view is from 63.6% on Top-5 probability to 91.8% on Top-20 probability. As the results, the FPCM cluster view is also better than hierarchical cluster view and global view. From this result, it can be observed that the proposed technique has better predicting ability of prediction than Two Levels of Prediction Model.

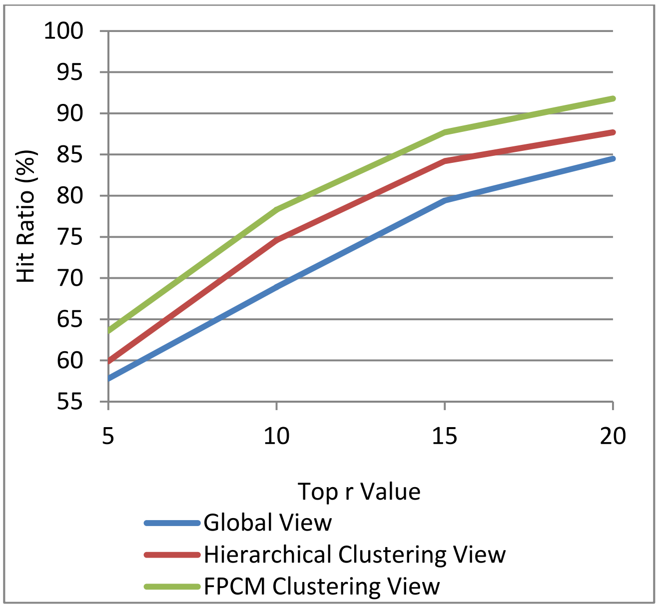

Fig. 6. Hit Ratio of Level Two

V. Conclusion

Forecasting the user's browsing pattern is a significant technique for many applications. The Forecasting results can be utilized for personalization, building proper web site, enhancing marketing strategy, promotion, product supply, getting marketing data, forecasting market trends, and enhancing the competitive strength of enterprises etc. This paper uses web usage mining technique for predicting the user's browsing behavior. One of the effective existing techniques for web usage mining is the usage of hierarchical agglomerative clustering to cluster users' browsing behaviors. The usage of Two Levels of Prediction Model framework is explained in this paper which works better for general cases. However, Two Levels of Prediction Model suffer from the heterogeneity user's behavior. To overcome this difficulty, this paper uses Fuzzy Possibilistic algorithm for clustering. The experimental result shows that the proposed technique results in higher hit rate.

References

[1] Hussain Tasawar, Asghar Sohail and Fong Simon, "A hierarchical cluster based preprocessing methodology for Web Usage Mining", 6th International Conference on Advanced Information Management and Service (IMS), Pp. 472-477, 2010.

[2] Yaxiu Yu and Xin-Wei Wang, "Web Usage Mining Based on Fuzzy Clustering", International Forum on Information Technology and Applications, Pp. 268-271, 2009.

[3] Jianxi Zhang, Peiying Zhao, Lin Shang and Lunsheng Wang, "Web usage mining based on fuzzy clustering in identifying target group", ISECS International Colloquium on Computing, Communication, Control, and Management, Vol. 4, Pp. 209-212, 2009.

[4] Houqun Yang, Jingsheng Lei and Fa Fu, "An Approach of Multi-path Segmentation Clustering Based on Web Usage Mining", Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 4, Pp. 644-648, 2007.

[5] Bamshad Mobasher, "Data Mining for Web Personalization," LCNS, Springer-Verleg Berlin Heidelberg, 2007.

[6] Catlegde. L and Pitkow. J, "Characterising Browsing Behaviours in the World Wide Web," Computer Networks and ISDN systems, 1995.

[7] Cyrus Shahabi, Amir M.Zarkessh, Jafar Abidi and Vishal Shah, "Knowledge discovery from users Web page navigation, " Workshop on Research Issues in Data Engineering, Birmingham, England,1997.

[8] Istvan K. Nagy and Csaba Gaspar-Papanek "User Behaviour Analysis Based on Time Spent on Web Pages," Web Mining Applications in E-commercce and E-Services, Studies in Computational Intelligence, 2009, Volume 172/2009, 117-136, DOI: 10.1007/978-3-540-88081-3_7 -Springer

[9] Jaideep Srivastave, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations. ACM SIGKDD, 2000.

[10] Robert.Cooley,Bamshed Mobasher and Jaideep Srinivastava,"Data Preparation for Mining World Wide Web Browsing Patterns," journal of knowledge and Information Systems,1999.

[11] Robert.Cooley,Bamshed Mobasher, and Jaideep Srinivastava, "Web mining:Information and Pattern Discovery on the World Wide Web,",In International conference on Tools with Artificial Intelligence, Newport Beach, IEEE,1997, pages 558-567.

[12] Suresh R.M. and Padmajavalli .R. ,"An Overview of Data Preprocessing in Data and Web usage Mining ," IEEE, 2006.

[13] Pierrakos. D, "Web usage mining as a tool for personalization: a survey", User Modeling and User-Adapted Interaction, 13(4), pp. 311-372.

[14] Tanasa. D and Trousse. B, "Advanced Data Preprocessing for Intersites Web Usage Mining", IEEE Intelligent Systems, 19(2), pp. 59-65, 2004.

[15] M. Y. Kan and H. O. N. Thi, "Fast webpage classification using URL features", Processing of CIKM, 2005.

[16] F. Masseglia, P. Poncelet and M. Teisseire, "Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure", ACM SigWeb Letters, vol. 8, pp. 13-19, 1999.

[17] C.H. Lee and Y.H. Fu, "Two Levels of Prediction Model for Users' Browsing Behavior, The 2008 IAENG International Conference on Internet Computing and Web Services (IMECS'08), Pp.751-756, 2008.

[18] UCI KDD archive, http://kdd.ics.uci.edu/.

R.Khanchana graduated with M.Sc., Computer Science in 2004 from J.K.K Nataraja college of Arts and Science for women, Erode, India and completed M. Phil., from Bharathiar University, India during 2006-2007. She has presented number of papers in National conferences and International seminarsand conferences. She is guiding research scholars and has published many papers in national and international conference and in many international journals. Her areas of Interest include Software Engineering and Data Mining. She has about 6 years of teaching experience. Currently she is working as a Lecturer in Sri Ramakrishna college of Arts & Science for Women, India.

Dr. M. Punithavalli is presently working as Director and Head of the Dept of Computer Science, SNS College of arts and science, India. She has published more than twenty papers in National/International Journals. Her areas of interest include E-Learning, Software Engineering, Data Mining, Networking and etc. She has about 19 years of teaching experience. She is guiding many research scholars and has published many papers in national and international conference and in many international journals.