A Dynamic Web Mining Framework for E-Learning Recommendations using Rough Sets and Association Rule Mining

A.Anitha
Member, IEEE, UGC – Senior Research Fellow,
Centre for Information Tech& Engg,
Manonmaniam Sundaranar University,
Tirunelveli, Tamil Nadu – 627 012, INDIA

Dr.N.Krishnan
Senior Member, IEEE Professor & Head
Centre for Information Tech & Engg,
Manonmaniam Sundaranar University,
Tirunelveli, Tamil Nadu – 627 012, INDIA

ABSTRACT
World wide web is a huge information source, broadly used for learning now-a-days due to flexibility of time, sharing of learning resources and infrastructure etc., Most of web based learning system lacks expert-learner interaction, assessment of user activities and learners are getting drowned by huge number of web pages in the learning web site and they find difficulties in choosing suitable materials relevant to their interest. This work attempts to engage e-learners at an early stage of learning by providing navigation recommendations. E-learning personalization is done by mining the web usage data like recent browsing histories of learners of similar interest. The proposed method uses upper approximation based rough set clustering and dynamic all k^{th} order association rule mining using apriori for personalizing e-learners by providing learning shortcuts. The essence of combing association rule and clustering is that, using clustered access patterns can reduce the data set size for association rule mining task, and improves the recommendation accuracy.

Keywords
E-learning, Personalization, Rough sets, Association rule Mining.

1. INTRODUCTION
E-learning is the process of learning through web, a vast information resource. However e-learning is advantageous regarding huge subject content, presentation of materials, discussion facility through forums and chats etc., The main drawback is poor tutor-student interaction and choosing appropriate learning material from the enriched data source. This paper attempts to provide good e-learning personalization by web usage mining.

Data mining is the method of extracting implicit and useful patterns automatically from databases. Web mining is a task that extracts hidden information from web relevant information. Web mining is divided into web content mining, web structure mining and web usage mining. Web content mining deals with discovery of useful knowledge from actual content of the web pages, in the form of text, audio or video. Example - search engines. Web structure mining deals with analyzing link structure and topology of hyperlinks. Example - Adaptive web server. Web usage mining deals with analysis of various log file data including web server log, client log and proxy server log and discovering useful knowledge regarding web usage. Web usage mining can be applied to e-learning domain as the site records information recording learner profiles, web access information, academic details of students and evaluation results. Web usage mining can track learning activities and identifies web access patterns and user behaviors[8].

Web usage mining has lot of contributions in e-learning domain such as [6],

(i) Dynamic personalization like providing real time recommendations for e-learners
(ii) Commonly referenced web pages are cached in proxy servers.
(iii) Structuring or organizing the site structure according to learner's interest.
(iv) Creating access shortcuts for interested pages to enhance user friendliness.
(v) Updating course content of web site according to previous usage information.
(vi) Identifying groups of learners of similar interest and sending personalized course materials to interested groups.

2. BACKGROUND OF STUDY
Several works were done to improve e-learning effectiveness through web usage mining. Web usage mining using basic association rule proposed by Mei-Ling Shyu has drawbacks including generation of irrelevant rules, generation of too many rules leading to contradictory prediction resulting in reduction of overall recommendation accuracy. In [3] a summary of several data mining approaches are proposed for improving effectiveness of online teaching. A collaborative filtering using k nearest neighbor is proposed in [1], has serious drawback of time complexity in dynamically finding k nearest neighbor. Also, personalization through association rule mining is proposed by Mobasher [1] using multiple support and confidence levels. It is complex to fix and handle multiple support and confidence levels. In [6], recommendations are made for e-learners through k-means clustering and Apriori algorithm for association rule mining. It uses cosine distance measures. The cluster tightness is not good as distance is not updated dynamically when new object enters into clusters.

In the proposed work homogeneity of clusters is improved leading to determining highly similar user groups through simple calculations. Khalil proposed Markov model for web access prediction in which for analysis of every test session, all training data are considered, some of them can be less relevant to test

session that affects prediction accuracy [2].Moreover, Markov model uses only strict consecutive and sequential page access for matching session during prediction. It might lose some of the loosely connected but interesting sessions.

3. RESEARCH HIGHLIGHTS

In this paper it is proposed to combine upper approximation based rough set clustering [3,4] and dynamic support pruned selective association rule mining using Apriori for e-learning recommendation.

The clustering of web sessions is the key aspect as it groups similar learning patterns. Hence in making e-learning recommendations, instead of considering all click stream sequences, it is necessary and sufficient to consider similar learning patterns. The highlights of this approach are as follows

- Considering learning activities of learners of similar interest (in a cluster) can improve recommendation accuracy.
- Clustering is done automatically, every learner is assigned to suitable cluster and most recent learning behaviors are considered for personalization.
- Using similarity measure instead of distance measure reduces computational complexity.
- For every prediction, only one cluster to which current user belongs to is considered. it reduces computational complexity and improves recommendation accuracy.

Association rule mining using Apriori is applied for personalization. The recommendation accuracy is improved in this stage in following ways, (i)All k th order approach is proposed by which the longest possible browsing sequence is considered for association rule mining, (ii) Contradictory predictions are reduced. As only small, highly similar data set (cluster) is considered, consequently the number of rules generated are very less, (iii) Ambiguous predictions if any, can be resolved by average number of bits per session and average access time user per page. The most referenced page is selected for recommendation. (iv)Minimum support and minimum confidence parameters are set in such a way to eliminate false discoveries. When minimum support is too small, every rule will get a chance to be true, leading to wrong recommendation and when minimum support is too large, for small data set, wrong predictions may occur due to poor coverage. But in the proposed approach, as clustering phase produces only similar patterns, the impact of minimum support and maximum confidence is not explosive regarding accuracy of prediction, (v) One of the major drawbacks of associations rule mining is that too many rules are generated and no guarantee for all generated rules to be relevant. In the proposed work, as clustering reduces the input data set to be small for Association rule mining, consequently the number of rules are reduced and the extracted rules are highly relevant and meaningful.

4. RECOMMENDATION FRAMEWORK

4.1 Log Data

The input data used for the proposed web usage mining model is the web server's access log. The web server records all learning activities carried out in the learning portal in the access log. It Includes IP address,URL,referrer URL, response code, size of files downloaded, date and timestamp etc., For experimentation of proposed work a sample log of www.e-learningcentre.com is used.

4.2 Formatting

In this paper mining is performed using only page view details. Hence it is sufficient to retrieve only page view details. The effect of web crawlers and web spiders which run indefinitely are totally irrelevant to mining task. Such items are filtered out. Similarly images and erroneous information such as noise are also removed. Only the IP address, page view and total time spent on every user is retained by preprocessing phase. The total time spent by user on all his reference to a particular page in a single session is determined. This parameter is retained in order to calculate average time spent by user in a page. This is used for resolving ambiguity when contradictions arise in association rule mining. Very short and very long transactions are eliminated as they affect cluster validity. Transactions containing only least referenced pages, that is, those with poor support are eliminated. Similarly transactions having only pages that appear in all transactions are eliminated, as they do not provide useful information. Hence a sample transaction after preprocessing phase is represented as,

128.5.16.102 </java;00:12> </loops;01:12> </if;02:23>, </while,03:20> </for,05:12>

Where 128.5.16.102 is the IP address of user who spent 01m12sec to learn loop,03 m,20 sec to learn while loop. A sample set of transactions extracted from web access log of e-learning website under analysis is given below.

T.No-->0 catalog2/series/cprogramming/
T.No-->1 index/general-IT/submit/
T.No-->2 index/course/cprogramming/catalog1/
T.No-->3 index/course/cprogramming/catalog2/
T.No-->4 index/general-it/cprogramming/
T.No-->5 index/general-IT/course/clientsr/
T.No-->6 index/general-IT/course/datawh/
T.No-->7 index/webdesign/cgiperl/features/
T.No-->8 index/window2000/register/special/
T.No-->9 index/course/cprogramming/catalog1/
T.No-->10 index/general-IT/microsoft-cert/register/

4.3 Learning Pattern Discovery

The patterns of learners of similar interest are discovered by clustering. Clustering is the process of grouping data objects such that, objects within a cluster are highly similar and objects in different clusters are dissimilar to each other. A roughest is defined by pair of sets which give lower approximation and upper approximation of the set. Lower approximation includes elements that definitely belonging to a concept where as upper approximation of a set includes elements that possibly belonging to a concept [13].

Roughness is the term deals with uncertainty and vagueness or fuzziness. A rough cluster is a cluster whose elements can be the member of more than one cluster. The rough clusters are upper approximated until it resulted in mutually disjoint equivalence classes. As upper approximation successively adds data objects, this type of clustering is termed as upper approximation based rough agglomerative clustering. If U is the universe of discourse,

a binary relation R between two transactions t and s is defined as tRs=1 when certain condition like >th is met. In this work a similarity threshold >0.4 is suggested. Any two transactions whose similarity value>0.4 are similar learning patterns.

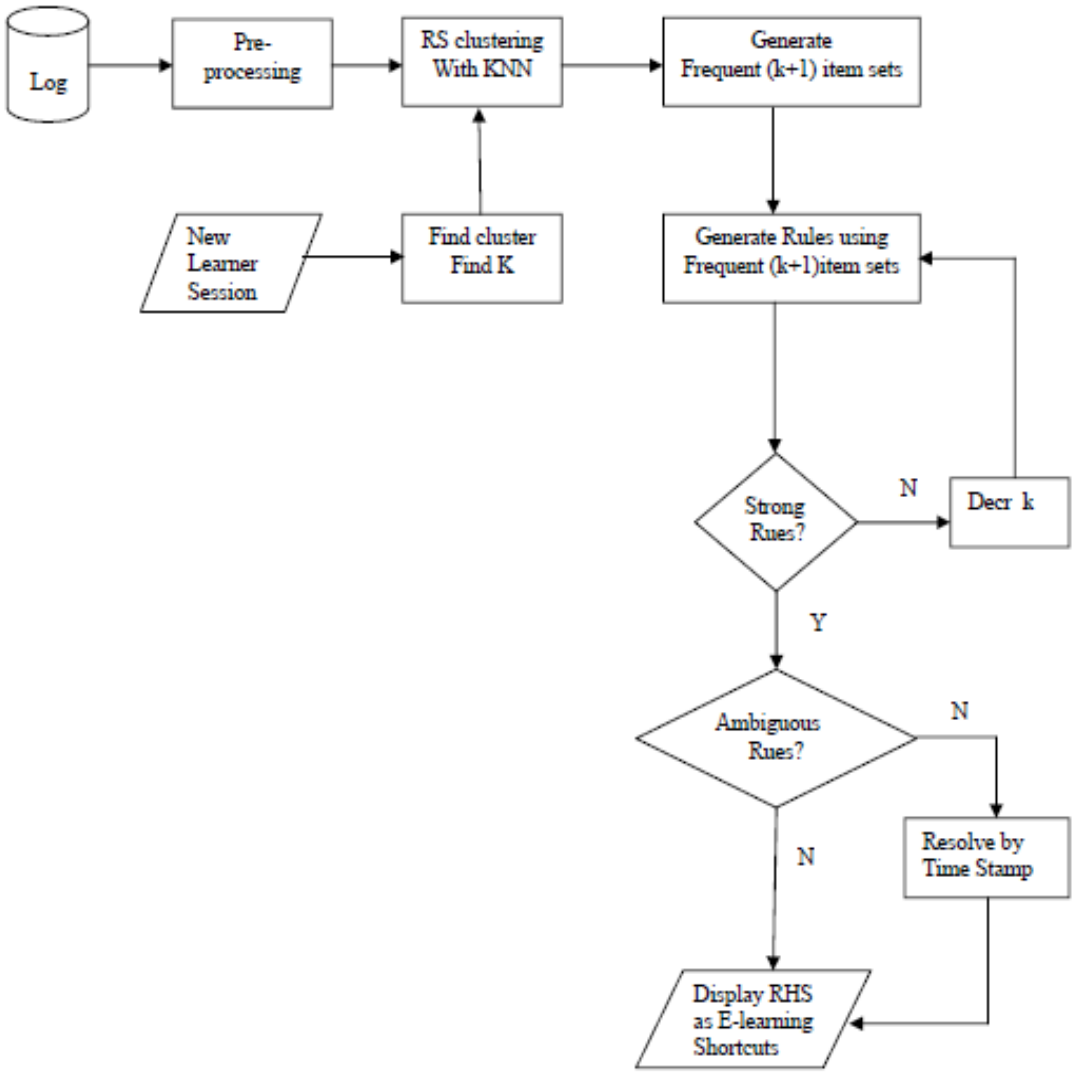

Figure 1. Flow diagram of Proposed Recommendation System for e-learning

An equivalence class is defined by a binary relation whose value is 1 for all members of the class. The relation $R \subseteq UXU$ partitions U into equivalence classes $\{ E_1, E_2 \ldots E_n\}$ where every pair of classes$\{E_i, E_j\}$ are either equal or disjoint and U E_i constitute U [5,12].

The pseudo code for rough set clustering is stated as follows:
For all transactions obtained in module2 from Fig.1,

1. Construct similarity matrix
2. Find similarity class for each transaction
3. Perform upper approximation by considering objects in k neighborhood, by which all transactions whose intersection with class considered is not equal to Ø are combined. Repeat step (3) until result of two successive iterations are same.

The similarity value is determined by Jaccord co-efficient as

$$Sim(a,b)=a \cap b/aUb \quad [1]$$

The proposed algorithm resulted in dense clusters with less computational complexity. As only nearest neighborhood is considered in each iteration only learners with similar learning behaviors are combined at each step and outliers are eliminated from rough cluster. Also, as two or more learners are merged at every step, agglomeration occurs faster by the proposed approach. As possibility factor is responsible for reducing cluster tightness, the most possible learning behaviors are added first by upper

approximation, than the least possible one. The threshold value eliminates odd behavior learners. Hence, this module discovers groups of learners having similar learning habits. This is applied to Association rule mining phase.

The similarity matrix for transactions identified in section 4.2 is as given in Table 1.

Table 1. Similarity Matrix

1	0	0.17	0.4	0.2	0	0	0	0	0.17	0
0	1	0.17	0.17	0.2	0.4	0.4	0.17	0.17	0.17	0.4
0.17	0.17	1	0.6	0.4	0.33	0.33	0.14	0.14	1	0.14
0.4	0.17	0.6	1	0.4	0.33	0.33	0.14	0.14	0.6	0.14
0.2	0.2	0.4	0.4	1	0.17	0.17	0.17	0.17	0.4	0.17
0	0.4	0.33	0.33	0.17	1	0.6	0.14	0.14	0.33	0.33
0	0.4	0.33	0.33	0.17	0.6	1	0.14	0.14	0.33	0.33
0	0.17	0.14	0.14	0.17	0.14	0.14	1	0.14	0.14	0.14
0	0.17	0.14	0.14	0.17	0.14	0.14	0.14	1	0.14	0.33
0.17	0.17	1	0.6	0.4	0.33	0.33	0.14	0.14	1	0.14
0	0.4	0.14	0.14	0.17	0.33	0.33	0.14	0.33	0.14	1

Confidence (P=>Q) =supcnt(PUQ) / Supcnt (P) [3]

By applying proposed rough set clustering with k=3 and th=0.4, following learning patterns are generated.

cluster 0:9,2,4,3,0
cluster 1:10,1,6,5
cluster 2:7
cluster 3:8

4.4 Association Rule Mining

This phase finds relationships or association between set of pages viewed by similar type of learners. It involves two phases namely frequent item set mining and rule generation. For every Active learning session, first, its corresponding cluster is identified. The number of pages in browsing history is recorded. Only the learner session in its corresponding cluster is input to Association rule mining. For k available page visits of active session, dynamic association rule mining generates frequent (k+1) items using apriori. Only the frequent (k+1) item sets are used for generating association rules. The cluster of user sessions is a set of page views. It is essential to identify associations or correlations between those pages.

Association rule mining is a technique of identifying hidden associations and correlations of items transactional databases. The association rule mining technique produces association in the form of if-then rules. The common application areas of association rule mining are market based analysis, E-Commerce, E-Learning,Web-Caching Etc., An implication P=>Q is called an Association rule if it satisfies two important statistical parameters namely minimum support and minimum confidence. Support is the percentage of transactions in database that contain all the items in left and right hand side of the implication. Confidence in the percentage of transaction in the database containing items in left hand side, that also contain items on right hand side of rule.

$$\text{Support } (P=>Q)=P(PUQ) = \text{supcnt}(PUQ) / |\mathbf{D}| \quad [2]$$

where, |D| is the number of objects in cluster. If the minimum support is very less, it results in generation of too many rules leading to contradictory prediction. If minimum support is too high, very few rules are generated due to poor coverage. Thus many interesting rules are filtered out. Hence proper domain knowledge is essential for setting minimum support and minimum confidence. The value of minimum support is 30 % and minimum confidence=75% for cleaning database. For a learning pattern to be considered as frequent learning pattern, the number of minimum occurrences of the pattern is 30% of total number of learning sequences considered. The set of pages viewed is called item set. The learning sequence with k number of pages is called k item set. The support of k pages, that satisfies minimum support is called frequent k item set.

The proposed e-learning recommendation work involve page that takes Boolean values as viewed or not viewed. Hence it requires generation of Boolean association rule. The variant of apriori algorithm for finding frequent k item set is proposed to find frequent learning sub sequence of pages. The proposed approach is advantageous in two ways.

(i) instead of finding longest possible frequent item set, frequent(k+1) item set is generated. Value of k is determined by active learner sessions browsing hit.
(ii) instead of using whole data set, only one cluster is used.

4.4.1 Apriori Algorithm

Apriori algorithm finds frequent k- item sets using apriori property that all subsets of frequent item sets are also frequent and an anti monotonic property that all supersets of infrequent item sets are also infrequent. For every learner session, the user ID, each page viewed and time spent on each page is available. At very first step, all pages are placed as candidate 1-itemset, and their support count is recorded. From fig.1, the large (L_1) item sets are generated by eliminating those pages that do not have minimum support or those referred less frequently as they are not useful for mining risk. The large items are also called frequent 1-itemset. In the subsequent stages candidates (C_k) are generated by joining L_{k-1x} L_{k-1} in lexicographic order. The candidate item sets

whose subsets are infrequent are pruned. And the remaining C_k item sets are used for finding frequent k+1 item sets (large item sets).This process is repeated until k+1 frequent item sets are generated, for active learner session containing k previous browsing history pages [6].

4.4.2 Rule generation

The rules that satisfy minimum support and minimum confidence are called strong rules. Using frequent k+1 item sets determined by apriori algorithm, association rules are generated as follows
For frequent (k+1) item set, calculate

$$Conf(I_1,I_2\ldots\ldots\ldots I_k \Rightarrow I_{k+1}) = \frac{supcnt(I_1,I_2..I_{k+1})}{Supcnt(I_1,I_2..I_k)} \quad [4]$$

If this confidence > minimum confidence, represent the rule $I_1,I_2,I_3\ldots\ldots\ldots\ldots I_{k+1}$ as a strong association rule. It implies such a rule is considered as recommendation for viewing page I_{k+1} after k page views by considering all learners of similar interest. If no matching rule for active learners are generated from frequent (k+1) item set, due to poor average, then freq k item sets are used for rule generation and the process is repeated by considering latest (k-i) visits and so on, until unambiguous recommendations are made. Recommendations are sorted in the order of decreasing confidence and the one with highest confidence is selected as recommended page. If any ambiguity arises, it is resolved by selecting page with high average viewing time / user. It is determined as,

$$\text{Avg time spent on a page} = \frac{\text{total time spent by all user in cluster}}{\text{Sup-cnt}} \quad [5]$$

Among the two conflicting recommendations the page on which more time in spent by user on an average is considered as recommended page.

Table 2. Frequent Item set Generation using dynamic apriori for cluster 0

C_1	L_1	C_2	L_2	C_3	L_3
{index}-(4)	{index}-(4)	{index, course}-(3)	{index, course}-(3)	{index, course,cpgm}-(3)	{index, course, cpgm}-(3)
(course}-(3)	{course}-(3)	{index, cpgm}-(4)	{index, cpgm}-(4)	{index, course,catalog1}-(2)	
{cpgm}-(4)	{cpgm}-(4)	{index, catalog1}-(2)	{index, catalog1}-(2)	{index, cpgm,catalog1}-(2)	{index, course, catalog1}-(2)
{catalog1}-(2)	{catalog1}-(2)	{index, catalog2}-(1)	{course, cpgm}-(3)	{course, cpgm,catalog1}-(2)	
{generalIT}-(1)	{catalog2}-(2)	{course, cpgm}-(3)	{course, catalog1}-(2)		{index, cpgm, catalog1}-(2)
{catalog2}-(2)		{course, catalog1}-(2)	{cpgm,catalog1}-(2)		
{series}-(1)		{course, catalog2}-(1)			course, cpgm, catalog1}-(2)
		{cpgm,catalog1}-(2)			
		{cpgm,catalog2}-(1)			
		{catalog1,catalog2}-(1)			

For an active learning sequence belonging to cluster 0, the frequent item sets are generated as shown in table 2. For example, for the first page click "index/?", initially frequent 2-item sets (C_2,L_2) are generated and the recommended pages are

index->cpgm (conf=4/4=>100%)
index->course(conf=3/4=>75%)
index->catalog1(conf=2/4=50%)

The recommendations are provided in the decreasing order of confidence. No ambiguous predictions occurred. If the learner followed "index/course/?", then frequent 3itemsets(C_3,L_3) are generated and next set of recommendations are provided in decreasing order of confidence as follows.

{index, course}=>{cpgm}(conf=3/3=>100%)
{index, course}=>{ catalog1}(conf=2/3=>66.7%)

The active learner can make use of these recommendations for further learning. If no matching antecedent found, then next lower order predictions using only latest 1-page "course" are used. In this way, proposed approach improved the effectiveness of learning.

5. RESULTS

For an e-learning portal under analysis, 100 learning sessions are tested. Recommendation accuracy of 84% is achieved. It means, out of test sessions considered, 84% of learners followed the recommendations made by the proposed approach.

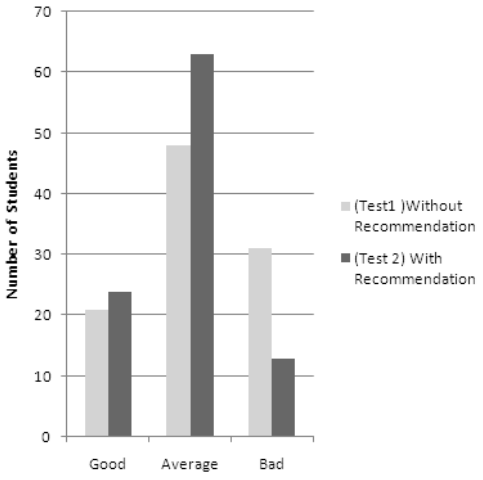

Figure 2. Learning Effectiveness Analysis Before and after recommendations

However, it is not necessary that new learners should follow the recommendations. But, in this experimentation, out of 100 learning sessions considered, 84 learners viewed recommended pages. Thus, recommendation accuracy represents the usefulness of recommendations while learning. In order to evaluate the effectiveness, two online tests were conducted for same set of

students. First test after self-learning and second test after learning through recommendations.

The performance of e-learners after providing recommendations is good. From Fig 2, the number of poor students is reduced. Most of them are moved to average performance. As recommendations are more useful to slow learners, the overall learning effectiveness is improved. The accuracy of learning recommendations is defined as percentage of learners followed the recommended page.

6. CONCLUSION

The proposed work achieves good recommendation accuracy, with less computational complexity. The recommendations are more useful to beginners in web based learning systems. In this work association rule mining is tried in different approach and overall learning effectiveness is improved through recommendations. Clustering played a vital role in reducing complexity of mining process and its contribution on making recommendations is also good. For future researchers, it is left out that, this work can be extended on different portals with different mining strategies.

REFERENCES

[1] Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa, "Effective Personalization based on Association Rule Discovery from web usage data",ACM workshop on Web Information and Data management , Nov 2001

[2] Faten Khalil Jiuyong LiHua Wang," Integrating Recommendation Models for Improved Web Page Prediction Accuracy, Conferences in Research and Practice in Information Technology (CRPIT), 2008,Vol. 74.

[3] Ioannis Kazanidis, Stavros Valsamidis, "Proposed Framework for Data Mining in E-learning: The case of Open E-class", ISBN:978-972-8924-97-3,2009 , pp 254-258

[4] Jaideep Srivastava, Robert Cooley et al.,"Web Usage Mining:Discovery and Applications of Usage patterns from Web Data", ACM SIGKDD, Vol 1, Issue 2, Jan 2000, pp 12-22

[5] Ms.Jyoti," A Novel Approach for clustering web user sessions using RST", International Journal on Computer Science and Engineering Vol.2(1), 2009, pp.56-61

[6] Khribi, M.K., Jemni M., Nasraoui O ,"Automatic Recommendation for E-learning Personalization based on Web Usage mining techniques and information retrieval", Educational Technology and Society, I2(4), 30-42

[7] Mei-Ling Shyu, Choochart Haruechaiyasak, "Collaborative Filtering by Mining Association Rules from User Access Sequences",Proceedings of 2005 International workshop on challenges in web information retrieval and integration, 0-7695-2414-1/05,IEEE, 2005

[8] Nasraoui, O. Soliman, M. Saka, E. Badia, A.Germain, R. "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites",IEEE transaction on Knowledge and dataengineering,Volume 20,Issue 2, Feb 2008 pp. 202-215

[9] Pasi Franti,Olli Virmajoki, and Ville Hautamaki "Fast Agglomerative Clustering Using a k Nearest Neighbor graph",IEEE transaction on pattern analysis and machine intelligence.Vol 28,No11. November 2006, pp 1875-1881

[10] Sathiya Moorthi V, Murali Bhaskaran V, "Data preparation Techniques for Web Usage Mining in World Wide Web – an approach", International Journal of Recent Trends in Engineering, Vol 2, No 4, November 2009

[11] Sen Guo, Yongshen Liang et al.,"Association Rule Retrieved from Web log based on Rough Set Theory", Fourth International conference on Fuzzy systems and Knowledge discovery, IEEE, 2007

[12] Siripom chimphlee,Naomie Salim,Mohd Salihin Bin Ngadiman, Witcha,Surat ,"Rough Sets Clustering and Markov Model for Web Access Prediction" ,Proceedings of post graduate annual seminar 2006, pp. 470-474

[13] Ying Cong, Changxu Ji, "Application of Web-based Data Mining in Personalized online learning system", Proceedings of Wuhan International Conference on E-Business, pp.150-156

[14] A.Anitha," A New Web Usage Mining Approach for Next Page Access Prediction". International Journal of Computer Applications, doi:10.5120/1252-1700 October 2010, pp 7-10

[15] A.Anitha," A Web Recommendation Model for E-Commerce Using Web Usage Mining Techniques", Advances in Computational Sciences and Technology, Volume 3 Number 4 (2010),pp.507–512