ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

A Review: Study of Various Clustering Techniques in Web Usage Mining

Miss. Aparna N. Gupta[1], Prof. Arti Karndikar[2]

M. Tech. Student, Department of Computer Science & Engineering, RCEOM, Nagpur, India [1]

Professor, Department of Computer Science & Engineering , RCEOM, Nagpur, India [2]

Abstract: The World Wide Web continuously growing repository of web pages and links at an exponential rate which makes exploiting all useful information a standing challenge. It has recently a wide range of applications in E-commerce web site and E-services such as building interactive marketing strategies, Web recommendation and Web personalization. The paper concerns Web server log file analysis to discover knowledge and by applying Clustering and optimization technique to get user interest which is helpful or useful for giving suggestion about specific user's interest.

Keywords: Web Mining, Pre-processing, FCM Clustering, ART1 –Neural Network Based Clustering, Partical Swarm Optimization.

I. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from Web data, Web log data is used in the mining process. Researchers have identified three different categories of Web mining[1,2,3].

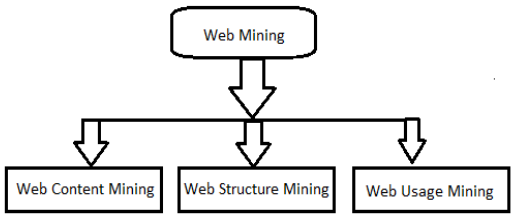

Fig.1 eb Mining Categories

1. *Web Content Mining* (Analyze the content of web pages as well as results of web Searching) Web content mining is a process of extracting up information from texts, images and other contents. The technologies that are mainly used in web content mining are NLP (Natural language processing) and IR (Information retrieval).
2. *Web Structure Mining* (Hyperlink Structure) Web structure mining is a process of extracting up information from linkages of web pages. Web structure mining is the process of using graph theory to analyse the node and connection structure of a web site. This graph structure can provide information about ranking and enhance search results of a page through filtering.
3. *Web Usage Mining* (analyzing user web navigation) Web usage mining is a process of extracting information from user how to navigate web sites. Web usage mining also known as web log mining, aims to discover interesting and frequent user access patterns from web browsing data that are stored in web server logs, proxy server logs or browser logs.

Application of Web Usage Mining:

1. *Personalization*: Reconstruct the website based on user's profile and usage behavior.
2. *System Improvement*: Provide help to understanding web traffic behavior. Web load balancing, data distribution or policies for web caching are benefits of such improvements.
3. *Adjustment of Website*: Understanding visitors' behavior in a web site provides hints for adequate design and update decision.
4. *Business intelligence* occupies the application of intelligent techniques in order to help certain businesses, mainly in marketing.
5. Valuing the *effectiveness of advertising* by analyzing large number of access behavior patterns.
6. Improving the design of e-commerce web site according to user's browsing behavior on site in order to better serve the needs of users.

II. VARIOUS WEB CLUSTERING TECHNIQUES

Clustering can identify user or data items with common characteristics. A group of users can be clustered that have similar navigation patterns on a web site. In e-commerce using cluster analysis technique, a group of customers can be clustered with similar browsing navigation and common characteristics of customers can be analysed. These can help the e-commerce users to get better understanding to their customers and customer-oriented service can be provided. Cluster analysis can help with marketing decisions, advertisement [4].

In this survey paper some clustering algorithms are suggesting for finding similar interest from the large datasets. Some clustering procedure are explaning given bellow:

A. Fuzzy C_Means Clustering

Clustering is the process of collecting similar object one another. In this paper, we can use object as user session as time generated by pre_processing stage .In clustering , grouping performed based on users having similar access sequences[5]. The data objects are represented by the feature vector. Given a set of data objects S =

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

{X1,X2,…Xn},where $X = (Xi1, Xi2, \ldots Xil)^P \in R^1$ is a feature vector and the similarity is calculated by the distance function D defined as D:S x S R such that for distinct X i , X j ε S. The distance between two data object is calculated as follow. Fuzzy C-means (FCM) clustering is of overlapped clustering which allows one data

$$D(x_i, x_j) = \sum_{k=1}^{i} (x_{ki} - x_{kj})^2 \quad \forall i, j = 1{:}\, n \text{ and } i \neq j$$

object to belong to two or more clusters. It is based on minimization of the following objective function:

$$j_m = \sum_{i=1}^{N}\sum_{k=1}^{c} U_{ij}^{m} \|X_{ki} - X_{kj}\|^2, \quad 1 <= m < infinity$$

Where, m- is a real number greater than 1,
U ij- is the degree of membership of X i in the cluster j,
C is the total number of clusters,
N- is the total number of user sessions,
X i is the feature vector,
C j is the center of the cluster, and
||*|| is the any norm that expresses the similarity between any measured data and the center.
Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with update of membership U ij and the centers C j by:

$$U_{ij} = \frac{1}{\sum_{k=1}^{c} \frac{\|X_i - C_j\|}{\|X_i - C_k\|}^{\frac{2}{m-1}}} \quad - - - \quad (2)$$

$$C_j = \frac{\sum_{K=1}^{C} U_{ij}^{m} X_{ki}}{\sum_{K=1}^{C} U_{ij}^{m}} \quad - - - \quad (3)$$

The following steps explain the working of FCM:
Input : The feature vector X i that represent the navigational sequence of each user and the number of clusters.
Output: The clusters having users with similar access sequence.

Step 1: Start
Step 2: Initialize or update the fuzzy partition matrix Uij with equation (2)
Step3: Calculate the center vectors Cj using equation (3)
Step 4: Repeat step (2) and (3) until the termination criterion is satisfied.
Step 5: Stop
The fuzzy c-means procedure continues until the close criterion is satisfied. Termination criteria can be that the difference between updated and previous objective function value J< predefined minimum threshold[6,7].

B. ART1 –Neural Network Based Clustering

The ART1 algorithm[8] can use for clustering uses the concepts of competitive learning and interactive activation. Figure 2 illustrates the architecture for our ART1-based neural network for clustering usercommunities.1 It consists of Nu nodes in the L1 layer, with each node presented a 0 or 1 binary value.
The L1 layer presents the pattern vector U j(i) , which represents the access pattern of each user. The L2 layer consists of a variable number of nodes corresponding to the number of clusters. The nodes at L2 layer represents the clusters formed. The L1 and L2 layers are totally interconnected, that is the activation of each L1 unit is fed into all L2 unites and vice versa. This interlayer feedback structure is used to facilitate 'resonance' when a match between the encoded pattern and input pattern occurs. Following steps explains the working of ART1 [9] approach.

Input: The pattern vector Uji that represents the navigation patterns of each user .
Output: The prototype vector that forms for each cluster

Step1:Start
Step2: Select ε and ρ and initialize the interlayer connec tions as follows.
tij=1 and bij=1/(1+n)
Where ρ- represents the vigilance parameter that determines the error degree to be tolerated, tij- represents the top down weights and bij- represents the bottom up weights.
Step3: Present U j i the binary input pattern vector to the L1 layer .
Step 4: Using bij, determine the activations of the L2 layer.
Step 5: Determine the node with maximum activation in L2
Step6: The top down verification begins. Using the winner unit found in step 5, this result is then fed back to L1 via the top down weight tij. This is an attempted confirmation of the winning node found in step 4.
Step7: If the test in step 6 succeeds, the bij and tij Interconnections are updated to accommodate the results of input U j i using the discrete version of the slow learning dynamics.
If the step fails this node is ruled out and step 5 is repeated until a winner can be found or there are no Remaining candidates.
Step8: Stop
When the algorithm stabilizes, the nodes at layer L2 ent clusters with a generalized representation of the URLs most frequently requested by all members (hosts) of that cluster.

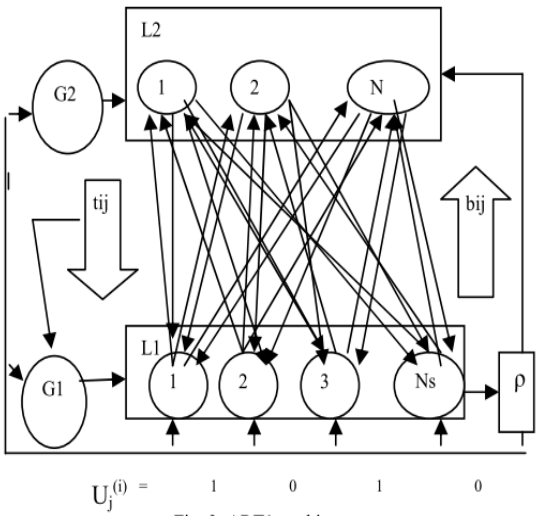

Fig.2 ART1 Architecture

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

C. Partical Swarm Optimization

PSO is an optimization technique originally proposed by [10] and is based upon the motivation from the swarm behavior of birds, fish and bees when they search for food or communicate with each other. Particles or agents represent an individual solution while the swarm is the assembly of particles which represent the solution space[11].

Particle Swarm Optimization (PSO) [7] is calculation technique introduced by Kennedy and Eberhart in 1995.This algorithm is modified with a population of random solutions known as particles. Each particle changes in feature space with a rate that is dynamically adjusted. These dynamic modifications are based on the historical behaviors of itself and other particles in the population. Each particle keeps track of its coordinates in the feature space which is associated with the best solution (fitness) it has achieved known as pbest. Another best value is tracked called gbest value. Fitness of all instances of generated clusters is calculated aslbest.

$$Fi = \frac{1}{\sum_{k=1}^{N} IIxi - xkII^2}$$

Particle Swarm Optimization Algorithm
P x: Dataset to be clustered
N: Number of clusters
S: Small positive valued constant
V: Random Speed

Step 1: For dataset x, set initial random cluster vector <C1 , C2,…..Cn >and random speed V=<V1,V2, ..Vn >
Step 2: Determine Euclidian distance from all clusters to all instances of Dataset.
Step 3: Determine fitness of all instances (Fi) of clusters.
Step 4: Choose instance having maximum fitness in each cluster is chosen as gbest of that cluster. Create n number of gbest.
Step 5: Compute new speed, lbest, gbest.
Step 6: Update position of all cluster centers with new speed V new and generate C new .
Step 7: if Euclidian distance <= S then repeat from step 3 otherwise display result with final clusters.

III. CONCLUSION

This paper describes Comparative Study of FCM clustering algorithm, Ant Cluster Track algorithm, Particle Swarm Optimization Algorithm. A good clustering technique may yield clusters thus have high inter cluster and low intra cluster distance. The fuzzy c-mean algorithm is sensitive to initialization. By using **FCM** clustering technique, give better result on overlapped clustering which allows one object to belong to two or more clusters and web log data but a traditional clustering technique has lot of limitations(for accurate results to overcome the limitations of traditional clustering techniques). The **ART1** algorithm uses the concepts of competitive learning and interactive activation. On other hand the Swarm intelligence techniques , **PSO** solves various function optimization problems.

REFERENCES

[1] Shaily G. Langhnoja, Mehul P. Barot, Darshak B. Mehta , "Web Usage Mining to Discover Visitor Group with Common Behavior Using DBSCAN Clustering Algorithm", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.
[2] Monika Yadav, Mr. Pradeep Mittal, "Web Mining: An Introduction",
[3] International Journal of Advanced Research in Computer Science and Software Engineering, March 2013 .
[4] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations. Copyright c
2000 ACM SIGKDD, July 2000.
[5] Sung-Shun Weng, Mei-Ju Liu, "Personalized product recommendation in e-commerce", IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004.
[6] Nayana Mariya VargheseJomina John, "Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic", IEEE Std. 2012.
[7] K.Suresh, R.Madana Mohana, A.RamaMohanReddy, "Improved FCM algorithm for Clustering on Web Usage Mining", IJCSI International Journal of Computer Science ,January 2011.
[8] Soniya P. Chaudhari, Prof. Hitesh Gupta, Prof. S. J. Patil, "Web Log Clustering using FCM and Swarm Intelligence Based Algorithms", International Journal of Innovative Research in Science, Engineering and Technology , Janauary 2013 .
[9] Anna Alphy 1 , S. Prabakaran , "Cluster Optimization for Improved Web Usage Mining using Ant Nestmate Approach", IEEE-International Conference on Recent Trends in Information Technology, MIT, Anna University, Chennai. June 3-5, 2011.
[10] Robert Schalkoff, "pattern Recognition: statistical, structural and neural approaches, John wiley & sons , Inc,1992..
[11] J. Kennedy, and R. C. Eberhart, "Particle Swarm Optimization", Proc. Of IEEE ICNN, Vol. IV, Perth, Australia (1995) 1942-1948.
[12] Shafiq Alam, Gillian Dobbie, Patricia Riddle, "Particle Swarm Optimization Based Clustering Of Web Usage Data", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.