A Survey of GA based Clustering and Biclustering Approach in Web Usage Mining

Raval Pratiksha M.[1] Mehul Barot[2]
[1]Research Scholar [2]Professor
[1,2]Computer Engineering, KSV University, Gandhinagar, India

Abstract--- The World Wide Web contains an increasing amount of websites which in turn contains increasing number of web pages. When a user visits a new website he/she has to go through large number of web pages to meet their requirements. Web usage mining is the process of extracting useful knowledge from the server logs. This useful knowledge can be applied to target marketing and in the design of web portals. A Recommender system is one of the best web usage mining Application which reduces the difficulties faced by the users to meet their requirements .It recommends the pages of interest to the user. This survey paper includes the survey of different clustering and biclustering techniques. Also we will discuss the biclustering approach which has some advantages over the traditional clustering approach.

Keywords: Web Usage Mining, Recommender system, Target Marketing Clustering, Biclustering.

I. INTRODUCTION

The World Wide Web store, share, and distribute information in the large scale. There is large number of internet users on the web. They are facing many problems like information overload due to the significant and rapid growth in the amount of information and the number of users. As a result, how to provide web users with more exactly needed information is becoming a critical issue in web applications. Web mining extracts interesting pattern or knowledge from web data. It is classified into three types as web content mining, web structure, and web usage mining.

Web usage mining is the most important area of web mining which deals with the extraction of useful knowledge from the web usage data. There are different kinds of datasets on which web usage mining can be performed. They are in the form of log files. These log files can be stored at server side, proxy side and client side. Mostly the server side log files are used for web usage mining. Before the mining process various pre-processing techniques can be applied to the log files, for example, pre-processing, pattern discovery, pattern analysis. The data mining techniques like Association rule mining, Sequential pattern analysis; Classification and Clustering are used to mine the web usage data. The mined knowledge can be helpful in different web applications like personalization of web content, support for the design, E-commerce, and many other web applications.

In this paper we discuss clustering technique of data mining for web usage data. Clustering is one of the important data mining technique to discover usage pattern from the web usage data. The users with the same browsing pattern are clustered in the same group and the others are clustered in different groups. In this survey we consider a novel clustering and biclustering algorithm based on genetic algorithms (GAs) for effective clustering. In general, a genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a metaheuristic) is routinely used to generate useful solutions to optimization and search problems [10]. So, we believe that a clustering technique with Genetic algorithm can provide relevant clusters more effectively.

A traditional clustering method clusters users according to their similarity of browsing behaviour under all pages. However, it is often the case that some users have similar behaviour only on a subset of pages. For example consider below example user page matrix. [2]

	Page1	Page 2	Page 3	Page 4
User 1	0	5	3	6
User 2	1	2	4	7
User 3	1	1	2	6
User 4	5	0	8	11

Table 1: user page matrix

When all pages are considered users 1, 2, and 4 do not show similar behaviour since their hit count values are uncorrelated under page 2 ,while users 1 and 2 have an increased hit count value from page 1 to page 2, the hits of user 4 drops from page 1 to page 2. However, these users behave similarly under pages 1, 3, and 4 since all their hit count values increase from page 1 to page 3 and increase again for page 4. A traditional clustering method will fail to recognize such a cluster since the method requires the three users to behave similarly under all pages which are not the case [2]. To overcome this problem Biclustering or Two-way clustering was introduced. Biclustering was first introduced by Hartigan and called it direct clustering [1]. Following section describes some of the clustering and biclustering methods together with Genetic algorithm available in the literature.

II. RELATED WORK

In [3], researchers have proposed a recommendation system using GA k-means algorithm of online shopping market. The result showed that GA k-means clustering may improve segmentation performance in comparison to other typical clustering algorithms. In [4], the researchers have proposed an improved fussy C-means clustering of web usage data with genetic algorithm. The method is scalable and can be coupled with a scalable clustering algorithm to address the large-scale clustering problem in web data mining. In [5], researchers have developed a fast Genetic k-means Algorithm (GKA) proposed by Krishna and Murty in 1999. The Proposed FGKA runs much faster than GKA. In [6], researchers have proposed an ant clustering algorithm to

All rights reserved by www.ijsrd.com

discover Web usage patterns (data clusters) and a linear genetic programming approach to analyze the visitor trends. Empirical results clearly show that ant colony clustering performs well when compared to a self-organizing map even though the performance accuracy is not that efficient when compared to evolutionary-fuzzy clustering approach.

In [1], researchers have proposed evolutionary Biclustering method for clickstream data. They proposed biclustering approach for web usage data using a combination of k-means, Greedy search procedure and Genetic algorithms to identify the coherent browsing pattern. In [2], researchers have proposed a combination of optimization technique, Binary Practical Swarm Optimization and the biclustering technique and developed a BPSO based biclustering of web usage data. The Objective of this algorithm is to find high volume of biclusters with high degree of coherence between the users and pages. In [7], researchers have proposed Biclustering approach with genetic algorithm for optimal web page category. Three different fitness functions based on Mean squared residue score are used to study the performance of the proposed biclustering method. In [8], researchers have proposed a fuzzy Co-clustering approach for clickstream data Pattern. The results proved its efficiency in correlating the relevant users and web pages of a web site. Thus, interpretation of Co- Cluster results are used by the company for focalized marketing campaigns to an interesting target user cluster. Following section describes the biclustering framework using Genetic Algorithm for web usage mining.

III. METHODS AND MATERIALS

A. Bicluster Types [9]

Different biclustering algorithms have different definitions of bicluster.

1) Bicluster with constant values (a),
2) Bicluster with constant values on rows (b) or columns (c),
3) Bicluster with coherent values (d, e).

a	a	a	a
a	a	a	a
a	a	a	a
a	a	a	a

(a)

a	a	a	a
a+i	a+i	a+i	a+i
a+j	a+j	a+j	a+j
a+k	a+k	a+k	a+k

(b)

a	a+i	a+j	a+k
a	a+i	a+j	a+k
a	a+i	a+j	a+k
a	a+i	a+j	a+k

(c)

a	b	c	d
a+i	b+i	c+i	d+i
a+j	b+j	c+j	d+j
a+k	b+k	c+k	d+k

(d)

B. Clickstream Data Pattern [1]

Clickstream data is a sequence of Uniform Resource Locators (URLs) browsed by the user within a particular period of time. By analyzing these data we can discover web users having similar browsing pattern. It requires some pre-processing before it is taken for analyze.

C. Preprocessing of Clickstream Data Pattern [1]

Clickstream data pattern is converted into web user access matrix A by using equation (1.1) in which rows represent users and columns represent pages of web sites. Let A (U, P) be an „n x m" user access matrix where U be a set of users , P be a set of pages of a web site, n" be the number of web user and „m" be the number of web pages. It is used to describe the relationship between web pages and users who access these web pages. The element aij of A(U,P) represents frequency of the user Ui of U
visit the page Pj of P during a given period of time.

$$a_{ij} = \begin{cases} \text{Hits}(U_i, P_j), \text{ if } P_j \text{ is visited by } U_i \\ 0, \quad \text{otherwise} \end{cases} \quad (1.1)$$

where Hits(Ui ,Pj) is the count/frequency of the user Ui accesses the page Pj during a given period of time.

D. Coherent Bicluster [1]

A bicluster with coherent values is defined as the subset of users and subsets of pages with coherent values on both dimensions of the user access matrix A. A measure called Average Correlation Value (ACV) is used to measure the degree of coherence of the biclusters. It is used to evaluate the homogeneity of a bicluster.

$$ACV(B) = \max\left\{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|row_{ij}| - n}{n^2 - n}, \frac{\sum_{k=1}^{m}\sum_{l=1}^{m}|col_{kl}| - m}{m^2 - m}\right\} \quad (1.2)$$

r_rowij is the correlation between row i and row j,
r_colkl is the correlation between column k and
Column l. A high ACV suggests high similarities among the users or pages.

E. Initial Biclusters [1]

K-Means clustering method is applied on the web user access matrix A(U, P) along both dimensions separately to generate k_u user clusters and k_p page clusters .And then combine the results to obtain small co-regulated sub matrices ($k_u \times k_p$) called biclusters. These correlated biclusters are also called seeds.

F. Encoding of Biclusters [1]

Each initial bicluster is encoded as a binary string. The length of the string is the number of rows plus the number of columns of the user access matrix A i.e. n + m. where n and m are the number of rows (users) and of columns (pages) of the user access matrix, respectively.

u_1	u_2	...	u_{n-1}	u_n	p_1	p_2	...	p_m

These binary encoded biclusters are used as initial population for genetic algorithm.

G. Greedy Local Search Procedure to enlarge and refine biclusters [1]

A greedy algorithm repeatedly executes a search
Procedure which tries to maximize the bicluster based on examining local conditions. Here ACV is used as merit

All rights reserved by www.ijsrd.com

function to grow the biclusters. It Insert/Remove the user/pages to/from the bicluster if it increases ACV of the bicluster. Our objective function is to maximize ACV of a bicluster. This approach employs simple strategies that are easy to implement and most of the time quite efficient.

H. Coherent Biclustering Framework using Genetic Algorithm (GA) [1]

Usually, GA is initialized with the population of random solutions. In our case, after the greedy local search procedure the optimization technique genetic algorithm is applied on biclusters to get the optimum bicluster. This will result in faster convergence compared to random initialization.

1) Fitness Function

The main objective of this work is to discover high volume biclusters with high ACV. The following fitness function F (I, J) is used to extract optimal bicluster.

$$F(I,J) = \begin{cases} |I|*|J| \text{ , if ACV(bicluster)} \geq \delta \\ 0, \text{ Otherwise} \end{cases} \quad (1.3)$$

Where |I| and |J| are number of rows and columns of bicluster and δ is defined as follows

ACV threshold δ = Max(ACV(P))

Here, the objective function should be maximized. P is the set of biclusters in each population, mp is the probability of mutation, r is the fraction of the population to be replaced by crossover in each population, cp is the fraction of the population to be replaced by crossover in each population, n is the number of biclusters in each population. The biclustering framework using genetic algorithm is given below.

Algorithm: Evolutionary Biclustering Algorithm [1]
Input: Set enlarged and refined seed
Output: Optimal Bicluster

Step 1. Initialize the population.
Step 2. Evaluate the fitness of individuals
Step 3. For i =1 to max_iteration
Selection()
Crossover()
Mutation() Evaluate the fitness
End(For)
Step 4. Return the optimal bicluster

Using the above algorithm we can generate optimum biclusters from web usage data which exhibits high coherence between the web user and the pages visited by them. Analyzing these overlapping coherent biclusters could be very beneficial for direct marketing, target marketing and also useful for recommending system, web personalization systems, web usage categorization and user profiling. The interpretation of biclustering results is also used by the company for focalized marketing campaigns to improve their performance of the business [1].

IV. CONCLUSION

In this paper we compared the traditional clustering and the biclustering approach with the genetic optimization technique in the context of web usage mining. The biclustering approach overcomes the problem associated with traditional clustering methods by showing the higher coherence between the web user and the subset of pages visited by them. This method has potential to identify the coherent patterns automatically from the clickstream data. The biclustering results can be used in the focalized marketing strategy like direct marketing and target marketing. Future work aims at extending this framework by using it as a pre-processing tool for the web page recommendation system.

ACKNOWLEDGMENT

I would like to express my special appreciation and thanks to my Guide Professor Mehul Barot, you have been a tremendous mentor for me. I would like to thank you for encouraging my research.

REFERENCES

[1] R.Rathipriya , Dr. K.Thangavel , J.Bagyamani "Evolutionary Biclustering of Clickstream Data" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.
[2] R.Rathipriya , Dr. K.Thangavel , J.Bagyamani "Binary Practical swarm Optimization based Biclustering of web usage data" International Journal of Computer Applications (0975 – 8887)Volume 25– No.2, July 2011.
[3] Kyoung-jae Kim a, Hyunchul Ahn,,A Recommender system using GA K-means clustering in an online shopping market", Expert Systems with Applications (2007), doi:10.1016/j.eswa.2006.12.025.
[4] N. Sujatha and Dr. K. Iyakutti, "Improved fuzzy C-Means clustering of web usage data with Genetic Algorithm", CiiT International Journal of Data Mining and Knowledge Engineering, Vol 1, No 7, October 2009.
[5] Yi Lu, Shiyong Lu, Farshad Fotouhi, Youping Deng, Susan J. Brown , "FGKA: A Fast Genetic K-means Clustering Algorithm", SAC'04, March 14-17, 2004, Nicosia, Cyprus.
[6] Ajith Abraham, Vitorino Ramos, "Web Usage mining using artificial ant colony clustering and genetic programming".
[7] P.S.Raja, R.Rathipriya, "Optimal web page category for web personalization using biclustering approach". International Journal of computational intelligence and informatics, vol. 1:No. 1,April-June 2011.
[8] R.Rathipriya, Dr. K.Thangavel ,"A Fuzzy Co-Clustering approach for Clickstream Data Pattern", Global Journal of Computer Science and Technology Vol. 10 Issue 6 Ver. 1.0 July 2010 Page.
[9] http://en.wikipedia.org/wiki/Biclustering
[10] http://en.wikipedia.org/wiki/Genetic_algorithm
[11] Federico Michele Facca and Pier Luca Lanzi,"Recent Developments in Web Usage Mining Research", DaWaK 2003, LNCS 2737, pp. 140–150, 2003.

All rights reserved by www.ijsrd.com