International Journal of Advance Engineering and Research Development (IJAERD)
Volume 1,Issue 5,May 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

A Web Page Recommendation system using GA based biclustering of web usage data

Raval Pratiksha M. [1], Mehul Barot [2]

[1] Computer Engineering, LDRP-ITR,Gandhinagar,cepratiksha.2011@gmail.com

[2] Computer Engineering, LDRP-ITR,Gandhinagar,mpbarot@ldrp.ac.in

Abstract--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 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. It may give information that is useful for improving the services offered by web portals and information access and retrieval tools. In this paper we are introducing a new approach for web page recommendation and user profile generation. This approach makes use of evolutionary biclustering technique for web page recommendation. We have applied it on two different datasets. One is clickstream data and other is web access log file of KSV University. The final results are generated from optimal biclusters obtained from evolutionary biclustering.

Keywords- Web Mining, Usage Mining, Recommender system, Target Marketing, Biclustering.

I. INTRODUCTION

With the rapid growth of WWW, it becomes very important to find the useful information from this huge amount of data. The Web also contains the rich and dynamic collection of hyperlink information and Web page access and usage information, providing sources for data mining. The Web poses great challenges for effective knowledge discovery and data mining application. Web mining is defined as application of data mining techniques to automatically discover and extract information form Web documents and services. In general, Web mining is a common term for three knowledge discovery domains that are concerned with mining different parts of web: Web Structure Mining, Web Content Mining and Web Usage Mining.

While Web Structure and Content mining utilize real or primary data on the Web, Web usage mining works on secondary data such as Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries and bookmark data. The continuous growth of World Wide Web and available data in that domain imposes new design and development of efficient Web Usage Mining process. Web Usage Mining refers to the application of data mining technique to discover usage patterns in order to understand and better serve the needs of Web based applications. As Web data is unstructured it becomes more difficult to find relevant and useful information for Web users. Thus one of the goal of Web Usage Mining is to guide Web users to discover useful knowledge and to support them for decision making.

In this paper we are focusing on recommendation system, one of the best applications of web usage mining. We are using a new approach of evolutionary biclustering for web page recommendation. Biclustering is a data mining technique which allows simultaneous clustering of the rows and columns

of a matrix. The genetic algorithm takes these biclusters as initial population and generates optimal biclusters. The remaining part of the paper is organized as follows. Section 2 provides a brief overview of existing work in the literature. Section 3 describes the methods and material for GA based biclustering for web usage mining. The proposed algorithm is described in the Section 4. Experiment carried out in this paper is described and presents the results in the Section 5. Summary of the paper and suggestion for the future work is given Section 6.

II. RELATED WORK

In [1], researchers have proposed evolutionary Biclustering method for clickstream data. They have Developed a coherent biclustering framework using GA to identify overlapped coherent biclusters from the clickstream data patterns and a coherence quality measure ACV .In [2], researchers have proposed an 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 [3], 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. In [4], 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 [5], researchers have developed Improved Fuzzy 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 problems in web data mining. In [6], researchers have proposed recommender system using GA K-means clustering for online shopping market. GA K-means clustering improves segmentation performance in comparison to other typical clustering algorithms. In [7], researchers have proposed Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming. An ant clustering algorithm discovers Web usage patterns and a linear genetic programming approach analyze the visitor trends. In [8], researchers have given survey of recent developments in web usage mining. Following Section describes our proposed biclustering Framework with Genetic algorithm for web usage mining.

III. METHODS AND MATERIALS

3.1 Biclustering

Biclustering is a two way clustering of a data matrix. Biclustering is mostly used for gene expression data analysis. The application of biclustering in web usage mining is when users have similar behaviour in subset of pages. It is used for clickstream data generated from web logs. The traditional clustering algorithm will try to identify users who have similar behaviour in similar set of pages but biclustering extracts users who have similar behaviour over subset of pages.

3.2 Clickstream Data Pattern [1]

Clickstream data is defined as a sequence of Uniform Resource Locators (URLs) browsed by the user within a particular period of time. To discover pattern of group of users with similar interest and

motivation for visiting the particular website can be found by analyzing the clickstream data. It requires the some pre-processing before it is taken for analyze.

3.3 Preprocessing of clickstream data

Biclustering is performed on a data matrix. In our case this data matrix is of user and their respective visited page categories. So the rows of a data matrix will be users and the columns will be the pages visited by all users. To generate these data matrix from the clickstream data we need to pre-process the clickstream data. We can generate the user access matrix A from clickstream data using following equation.

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

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

3.3 Bicluster Evaluation Functions.

An Evaluation Function is the measure of coherence degree of a bicluster in a data matrix. There are several Bicluster evaluation functions available. In our research we are using Two Bicluster Evaluation Functions: 1) ACV(Average Correlation Value and 2) MSR (Mean Square Residue). 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.

1) 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.

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

Where, 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.

2) The Second and the most popular Evaluation function is Mean Square Residue.

$$H(I,J) = \frac{1}{|I||J|}\sum_{i\in I, j\in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

Where,

$$a_{iJ} = \frac{1}{|J|}\sum_{j\in J} a_{ij},\ a_{Ij} = \frac{1}{|I|}\sum_{i\in I} a_{ij} \text{ and } a_{IJ} = \frac{1}{|I||J|}\sum_{i\in I, j\in J} a_{ij}$$

- aij= Element in a sub-matrix Aij.
- aiJ= mean of ith row of bicluster (I,J).
- aIj=Mean of the j-th column of (I,J).
- aIJ=Mean of all the elements in bicluster.

A Low MSR value indicates that the bicluster is strongly coherent.

3.4 K- means Clustering

K-means clustering is a simple and flexible. The K- means algorithm is easy to understand and implement. This 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. These combined biclusters are initial biclusters. These biclusters are enlarged and refined to generate potential bicluster in greedy search procedure.

3.5 Greedy Search Procedure

A greedy algorithm repeatedly executes a search procedure which tries to maximize the bicluster based on examining local conditions, with the hope that the outcome will lead to a desired outcome for the global problem. ACV and MSR are used as merit function to grow the bicluster. With ACV 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. With MSR it Insert/Remove the user/pages to/from the bicluster if it decreases MSR of the bicluster. Our objective function is to minimize MSR of a bicluster. The greedy approach is easy to implement and mostly time efficient.

3.6 Genetic Algorithm (GA)

Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover[wiki].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.

Fitness Functions

The main objective of this work is to discover high volume biclusters with high ACV and low MSR.

1) AVC: - The following fitness function F (I, J) is used to extract optimal bicluster.

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

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

$$\text{ACV Threshold } \delta = \frac{\sum_{p=1}^{P} \text{ACV (p)}}{|P|}$$

2) MSR:- The following fitness function F (I, J) is used to extract optimal bicluster.

$$F(I,J) = \begin{cases} |I|*|J|, \text{ if MSR (bicluster)} \leq \delta \\ 0, \text{ Otherwise} \end{cases}$$

Where |I| and |J| are number of rows and columns of bicluster and δ is defined same as ACV Threshold but using MSR value in it.
The Roulette Wheel Selection (RWS) is used for selection process. One point and two point crossover is used for crossover of selected parents and to generate new offspring.

IV. PROPOSED ALGORITHM

We have proposed following algorithm for web page recommendation:

4.1 Proposed Algorithm

1. Load data set.
2. Preprocess data and generate user access matrix A.
3. Generate initial biclusters using Two-Way K-Means clustering from user access matrix A.
4. Improve the quality and quantity of the initial biclusters using Greedy Search procedure with two Bicluster Evaluation function ACV and MSR.
5.Apply Genetic Algorithm.
6. Evaluate the fitness of individuals.
7. For *i* =1 to max_iteration.
Selection ()
Crossover ()
Mutation () Evaluate the fitness
End (For)
8. Return the optimal bicluster.
9. Generate Recommendation for website.
10. Stop.

4.2 Proposed System Architecture

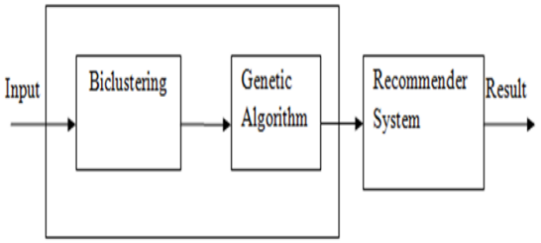

Figure 1. System Architecture

V. EXPERIMENTAL RESULTS AND ANALYSIS

The Experiments are conducted on two different datasets. One is the clickstream dataset collected from MSNBC.com. This dataset is collected from UCI repository. It contains 9, 89,818 users and 17 distinct page categories. Second dataset is a web access log file of KSV University, Gandhinagar. After converting it to clickstream data we got 4592 total users and 22 distinct page categories.

We have shown results of the first dataset only. The user access matrix is generated from the first equation. In the next biclustering step Ku User clusters and Kp Page clusters are generated from user access matrix and initial Biclusters Ku * Kp are generated. These biclusters are enlarged and refined using Greedy search procedure. In this step the volume of biclusters is higher than initial biclusters. The Enlarged and refined biclusters are set as initial population to the Genetic Algorithm. It will generate optimal biclusters.

The measure R is used to evaluate the overlapping degree between biclusters. It calculates the amount of overlapping among biclusters. The degree of overlapping of biclusters is defined as follows:

$$R = \frac{1}{|U| * |P|} \sum_{i=1}^{|U|} \sum_{j=1}^{|P|} T_{ij}$$

where

$$Tij = \frac{1}{(N-1)} * \left(\sum_{k=1}^{N} W_k(a_{ij}) - 1 \right)$$

Where,

N is the total number of biclusters,

|U| represents the total number of users,

|P| represents the total number of pages in the data matrix A.The value of wk(aij) is either 0 or 1. If the element (point) aij in A is present in the kth bicluster, then wk(aij) = 1, otherwise 0. If R index value is higher, then degree of overlapping of the generated biclusters would be high. The range of R index is $0 \leq R \leq 1$.

The results generated after each step are shown in the following tables:

Parameters	Initial Bi-Cluster	After Applying Greedy Search Algorithm	After Applying Genetic Algo Algorithm
Seeds	114	114	114
Average Volume	463.693	1938.0	13543.091
Overlapping Degree	0.0	0.0390045	0.2335
ACV	0.52643174	0.91118836	0.97778

Table 1:- Bicluster Evaluation Function ACV after each step.

Parameters	Initial Bi-Cluster	After Applying Greedy Search Algorithm	After Applying Genetic Algo Algorithm
Seeds	114	114	114
Average Volume	463.693	1938.0	13543.091
Overlapping Degree	0.0	0.02	0.2335
MSR	605.36957	452.0117	159.6281

Table 2 :- Bicluster Evaluation Function MSR after each step.

From table 1 and 2 we can see that the Average volume of biclusters is increasing after each step. Also the value of ACV is increasing and the value of MSR is decreasing after each step. A high ACV and Low MSR value indicates that the bicluster is strongly coherent. We get 0.23 overlapping degree for final biclusters.

The following Graph shows the final recommendation to the website. It is generated from the optimal biclusters we got after Genetic Algorithm.

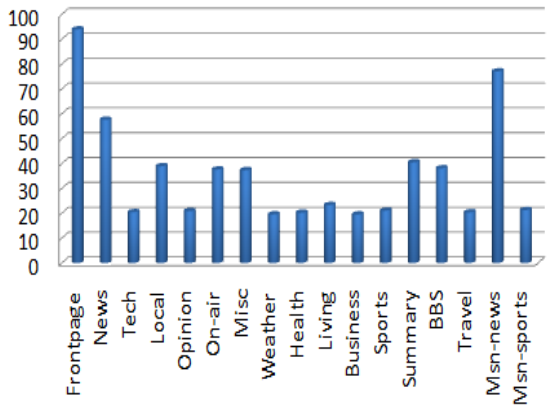

Figure 2. Final Recommendation Graph

CONCLUSION

This paper proposed a recommendation system using evolutionary biclustering Technique. The objective of this algorithm is to find high volume biclusters with high degree of coherence between the users and pages. The final recommendation step will give the website its most visited pages by all its users. Also it gives the information of the users having similar behavior on subset of pages. The results of biclustering can be used in market strategy like target marketing and direct marketing. The final optimal biclustering results can also be used towards improving the website's design, information availability and quality of provided services.

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] 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.
[4] 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.
[5] 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.
[6] 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.
[7] Ajith Abraham, Vitorino Ramos, "Web Usage mining using artificial ant colony clustering and genetic programming".
[8] Recent Developments in Web Usage Mining Research Federico Michele Facca and Pier Luca Lanzi.
[9] http://en.wikipedia.org/wiki/Biclustering.

