International Journal of Advance Engineering and Research Development
Scientific Journal of Impact Factor (SJIF): 4.72
Special Issue SIEICON-2017,April -2017
e-ISSN : 2348-4470 p-ISSN : 2348-6406

Biclustering of web usage data using Genetic Algorithm

Pratiksha Raval[1], Mayuree Rathva [2], Nishant Khatri[3], Nimit Modi[4], Pragna Makwana[5]

[1] *Computer Engineering, Sigma Institute of Engineering*
[2] *Computer Engineering, Sigma Institute of Engineering*
[3] *Computer Engineering, Sigma Institute of Engineering*
[4] *Computer Engineering, Sigma Institute of Engineering*
[5] *Information Technologies, Sigma Institute of Engineering*

Abstract—Internet is a source of large amount of Data having large number of internet users on the web. Now days the users are facing many problems like information overload due to large number of internet users and rapid growth in the amount of information. The solution to this problem is to provide users with more exactly needed information. Mining is the process of extracting useful data from large database. 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 nontrivial process to discover valid, novel, potentially useful knowledge from web data using the data mining techniques or methods. It may give information that is useful for improving the services offered by web portals and information access and retrieval tools. In this study, we propose a novel biclustering algorithm based on genetic algorithms (GAs) to effectively segment the web usage data. In general, GAs is believed to be effective on NP-complete global optimization problems, and they can provide good near-optimal solutions in reasonable time. Thus, we believe that a biclustering technique with GA can provide a way of finding the relevant clusters more effectively. In this work Genetic Optimization technique is combined with biclustering approach to propose a recommendation system using GA based biclustering of Web Usage Data.

Keywords- Web Mining, Usage Mining, Biclustering, Genetic Algorithm.

I. INTRODUCTION

Web mining is the application of data mining techniques to the content, structure, and usage of Web resources. It is thus "the nontrivial process of identifying valid, previously unknown, and potentially useful patterns" in the huge amount of Web data. Like other data mining applications, Web mining can profit from given structure on data (as in database tables), but it can also be applied to semi structured or unstructured data like free-form text. [1]

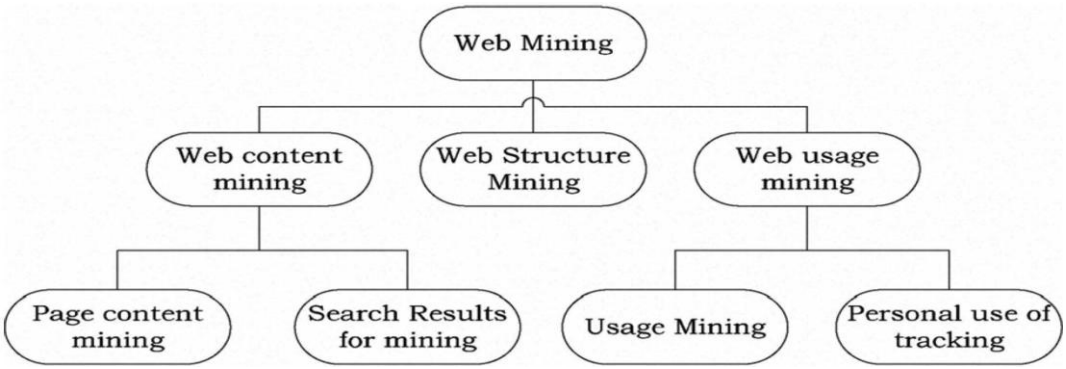


Figure 1: Classification of Web Mining

Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files; source data mainly consist of the (textual) logs, that are collected when users access web servers and might be represented in standard formats; typical applications are those based on user modeling techniques, such as web personalization, adaptive web sites, and user modeling. Recommender Systems (RSs) are software tools and techniques providing suggestions for items to be of use to a user. The suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read. "Item" is the general term used to denote what the system recommends to users. A RS normally focuses on a specific type of item (e.g., CDs, or news) and accordingly its design, its graphical user interface, and the core recommendation technique used to generate the recommendations are all customized to provide useful and effective suggestions for that specific type of item. RSs development initiated from a rather simple observation: individuals often rely on recommendations provided by others in making routine, daily decisions. For example it is common to rely on what one's peers recommend when selecting a book to read; employers

count on recommendation letters in their recruiting decisions; and when selecting a movie to watch, individuals tend to read and rely on the movie reviews that a film critic has written and which appear in the newspaper they read.

II.METHODS AND MATERIALS

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.[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**

Figure 2: Sample Bicluster

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. To overcome this problem Biclustering or Two-way clustering was introduced. Biclustering was first introduced by Hartigan and called it direct clustering. A bicluster of a web usage data is defined as a subset of users which exhibit similar interest or browsing patterns along a subset of pages.

Clickstream Data Pattern [5]

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.

3.1.3 Preprocessing of Clickstream Data Pattern [5]

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.

$$a_{ij} = \begin{cases} \text{Hits}(U_i, P_j), \text{ if } P_j \text{ is visited by } U_i \\ 0 \text{ , otherwise} \end{cases} \quad (1)$$

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

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.

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

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.

$$H(I,J) = \frac{1}{|I||J|}\sum_{i\in I, j\in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \quad (3)$$

Where,

$$a_{iJ} = \frac{1}{|J|}\sum_{j\in J} a_{ij},\ a_{Ij} = \frac{1}{|I|}\sum_{i\in I} a_{ij}\ and\ a_{IJ} = \frac{1}{|I||J|}\sum_{i\in I, j\in J} a_{ij} \quad (4)$$

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.

Initial Biclusters [5]

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.

Greedy Local Search Procedure [5]

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 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.

Structure of Greedy Search Procedure

Step 1: Start with initial bicluster.

Step 2: For every iteration.

Add/ remove the element (user/page) to/from the bicluster which maximize the objective function.

End for

The Objective function is to maximize ACV of a bicluster.

Coherent Biclustering Framework using Genetic Algorithm (GA) [5]

The GA is a stochastic global search method that mimics the metaphor of natural biological evolution. GA operates on a population of potential solutions applying the principle of survival of the fittest to produce better and better approximations to a solution. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators borrowed from natural genetics. This process leads to the evolution of populations of individuals that are better suited to their environment than the individuals that they were created from, just as in natural adaptation. Biclustering approach is viewed as optimization problem with the objective of discovering overlapping coherent biclusters with high ACV and high volume. In this paper, Genetic Algorithm (GA) is used for optimization of bicluster. The important feature of GA is that it provides a number of potential solutions to a given problem and the choice of final solution is left to the user. Usually, GA is initialized with the population of random solutions. In order to avoid random interference, biclusters obtained from greedy search procedure are used to initialize GA. This will result in faster convergence compared to random initialization.

Fitness Functions

1) ACV:-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.

$$F(I,J) = \begin{cases} |I|*|J|, \text{ if ACV (bicluster)} \geq \delta \\ 0, \text{ Otherwise} \end{cases} \quad (5)$$

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

ACV threshold δ = Max (ACV (P))

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

$$F(I,J) = \begin{cases} |I|*|J|, \text{ if MSR (bicluster)} \leq \delta \\ 0, \text{ Otherwise} \end{cases} \quad (6)$$

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.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.

III. PROPOSED ALGORITHM

We have proposed following algorithm for web page recommendation:

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.
5. Initialize the population with improved initial biclusters.
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.

PROPOSED SYSTEMFLOWCHART

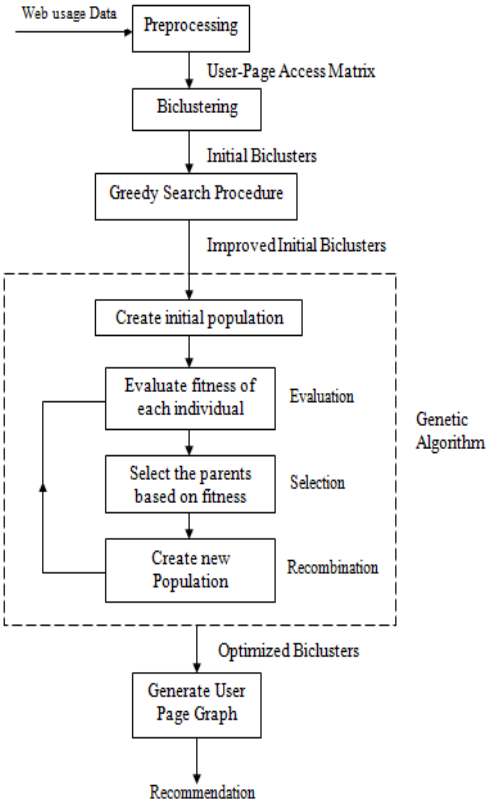


Figure 3: System Flowchart

IV. 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.

Step 1 : Load Dataset

Sr. No.	Dataset	Size	Users	Page Categories
1.	msnbc.com anonymous web data	11.9 MB	9,89,974	17
2.	KSV access log file	51.7MB	4592	22

Figure 4: Dataset

MSR(Mean Square Residue)

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

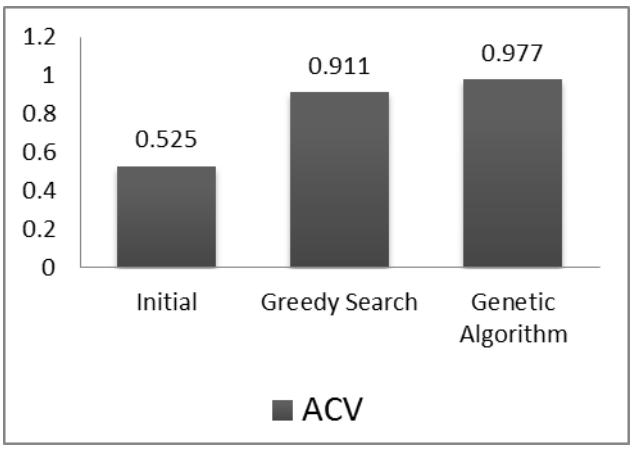
$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

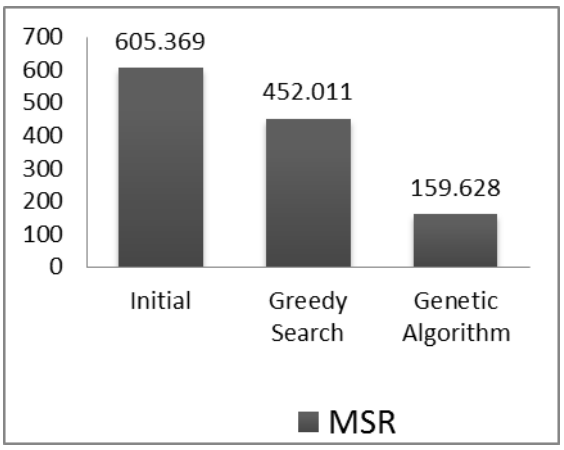
Where,

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

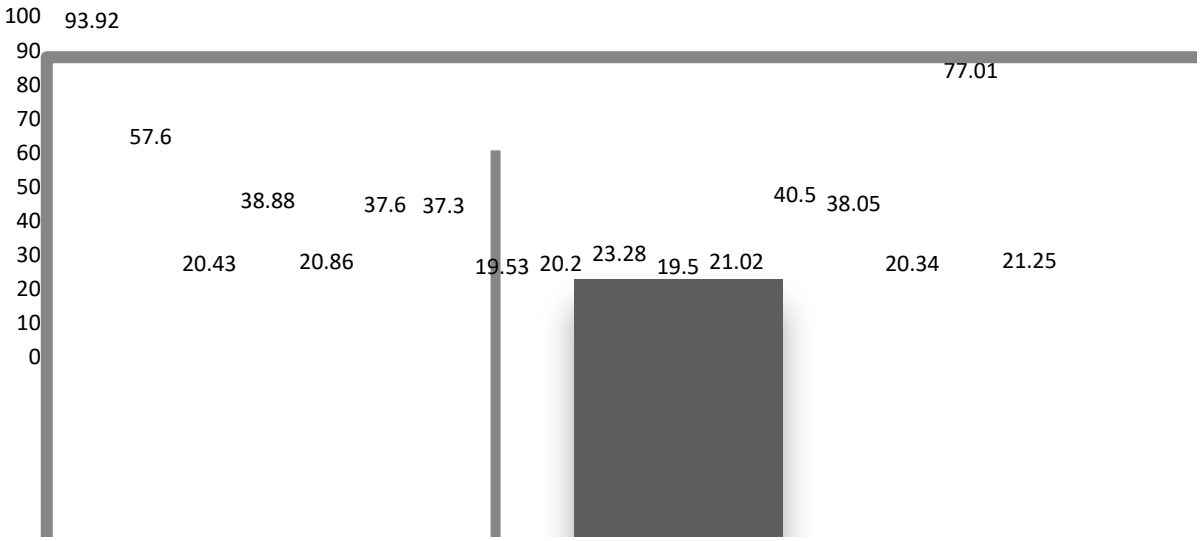
- aij= Element in a sub-matrix Aij.
- aiJ= mean of ith row of bicluster (I,J).
- aIj=Mean of the j-th column of (I,J).
- aIJ=Mean of all the elements in bicluster.
-

Comparison Graph(msnbc.com dataset)





Final Recommendation Graph



V. CONCLUSION

The main contribution of this research is to development of recommender system using coherent biclustering framework with GA to identify overlapped coherent biclusters from the clickstream data patterns. The interpretation of the recommender system can be used towards improving the website's design, information availability and quality of provided services. It is also useful in learning the user behaviour. The objective of this research is to find high volume biclusters with high degree of coherence between the users and pages. This method has potential to identify the coherent patterns automatically from the clickstream data.

VI. REFERENCES

[1] Semantic Web Mining and its application in Human Resource Management..IJCSMS International Journal of Computer Science & Management Studies, Vol. 11, Issue 02, August 2011 ..Ridhika Malik1, Kunjana Vasudev2 and Udayan Ghose3

[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 Know ledge 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.

