Web Usage Data based Web Page Recommender System

[1]Prof. Mehul Barot, [2]Dr. Kalpesh H. Wandra, [3]Dr. Samir B. Patel
[1]Research Scholar, [2]Dean, C.U.Shah University, [3]Assitant Professor, Pandit Deendayal Petroleum University
[1]Computer Engineering Department,
[1]C.U.Shah University, Wadhwan City, India

ABSTRACT*-There is a concern that how to provide web users with the more accurate and exactly needed information. Web mining addresses this critical issue in web applications by extractingsuch interesting pattern or knowledge from web data. In this paper, we propose a new approach for web page recommendation along with user profile generation. The approach makes use of evolutionary biclustering technique for web page recommendation. Different datasets have been used for performing data mining operations, one is clickstream data and other used is web access log file of Kadi Sarva Vidhyalaya (KSV) University. The outcome of the approach using optimal biclusters and evolutionary biclustering techniques have been analyzed in this paper and the outcome shows that, almost all records of the database are used and accurate results are generated. This result is further useful for making good strategic decision in applications target marketing and direct marketing.*

KEYWORDS*-Web Mining, Usage Mining, Recommender system, Target Marketing, Clustering, Biclustering,Genetic Algorithm, Average Correlation Value, Mean Square Residue, Accuracy Evaluation.*

I. INTRODUCTION

With the rapid growth of World Wide Web (WWW), it becomes very important to find the useful information from this voluminous amount of data. The Web also contains the rich and dynamic collection of hyperlink information,web page access and its usage information which provides a valuable source for data mining. The web poses great challenges for effective knowledge discovery and data mining applications. Web mining is defined as an application of data mining techniques to automatically discover and extract information from 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 Weband available domain specific data 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 the primary goal of web usage mining is to guide web users to discover useful knowledge and to support them for making decisions.

A genetic algorithm (GA) is a technique that can solve both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. The algorithm repeatedly modifies a population of individual solutions.

Biclustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix and Genetic Algorithm (GA) takes these biclusters as initial population and generates optimal biclusters.

In this paper, we are focusing on recommendation system, which is one of the best applications of web usage mining based upon evolutionary biclustering for web page recommendation. The paper is organized in 6 sections. Section II provides a brief overviewof existing work in this domain. Section III describes the methods and material for GA based biclustering for web usage mining. Our proposed algorithm is described in Section IV. Experimental outcome is presented in Section V. Section VI provides conclusion and future direction of work in this domain.

II. RELATED WORK

Researchers have proposed evolutionary Biclustering method for clickstream data [1]. The authors 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 for web usage data by using Binary Particle Swarm Optimization (BPSO) and the biclustering technique. 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. This, interpretation of Co-cluster results are used by the companies for focused marketing campaigns for identification of interested target user clusters.

Biclustering framework using Genetic Algorithm for web usage mining could be referred at [4-7]. 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. Improved Fuzzy C-Means Clustering of Web Usage Data with Genetic Algorithm based approach could be referred at [5]. This 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 algorithm for online shopping market. GA K-means clustering approach 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.Using ant clustering algorithm we can discover web usage patterns and using linear genetic programming approach we can analyze the visitors trend(s). In [8], researchers have provided survey of recent developments in web usage mining. In [10], A Survey of Accuracy Evaluation Metrics of Recommendation tasks have been carried out. Based upon users surfing behaviour we get user information learned from user's web logs data to construct accurate comprehensive individual user profiles [11]. In [12], authors have given a generic framework that delivers "Contextual recommendations" that are based on the combination of previously gathered user feedback data (i.e. ratings and clickstream history),context data, and ontology based content categorization scheme. A web recommendation approach which is based on learning from web logs and recommends user a list of pages which are relevant to him by comparing with user's historic pattern could be obtained from [13]. A combined approach of content-based model and memory-based collaborative filtering is used in order to remove drawbacks of existing system and used feed forward back propagation neural network for training data [14]. An intelligent approach that explores the idea of applying a semantic recommender system in process plant design is discussed in [15]. In [16], authors have shown experiments based on Markov Logic Network, through which, one can do web page recommendation with very high accuracy. Tag Based Recommender System for Social Bookmarking sites is discussed in [17] and user's preference transition applied for Hotel Recommendation System is discussed in [18].

III. METHODOLOGY

3.1 Biclustering

Biclustering is a two way clustering approach which has a data matrix showing users and pages. Biclustering is widely used for data analysis of gene expression. The application of biclustering in web usage mining is when users have similar behaviour in the 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.

3.2Clickstream Data Pattern

Clickstream data [1] is defined as a sequence of Uniform Resource Locators (URLs) browsed by the user within a stipulated time frame. By analyzing the clickstream data, we can discover pattern of group of users with similar interest and motivation for visiting the particular website. It requires some pre-processing before it can be utilized to analyze.

3.3 Preprocessing of click stream data

The data matrix in biclustering has user and their respective visited page categories. So the rows of a data matrix will have users and the columns in the matrix will have the pages visited by all users. To generate these data matrix from the clickstream data, we need to pre-process the clickstream data. We generate the user access matrix from clickstream data using equation (1).

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

Where$Hits(U_i, P_j)$is the count/frequency of the user $U_i$accesses the page $P_j$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 work,we are using Two Bicluster Evaluation Functions: 1) Average Correlation Value (ACV) and 2) Mean Square Residue (MSR). 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) Average Correlation Value is used to measure the degree of coherence of the biclusters as shown in equation (2). It is used to evaluate the homogeneity of a bicluster.

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

Where, r_row_{ij} is the correlation between row i and row j, r_col_{kl} is the correlation between column k and Column l. High value of ACV suggests high similarities among the users and pages.

2) Another measure used in our work is Mean Square Residue whose equation is shown in (3).

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

Where,

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

- a_{ij} = Element in a sub-matrix A ij.
- 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 algorithm is simple and flexible. Applying K-means method on the web user access matrix A(U, P) along both dimensions separately. It generates k_u user clusters and k_p page clusters. We then combine the results to obtain biclusters from sub matrices ($k_u \times k_p$). These correlated biclusters are known as seeds. These combined biclusters are initial biclusters. These biclusters are enlarged and refined to generate potential bicluster in greedy search approach.

3.5Greedy 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 has proved to be mostly time efficient [2].

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 [9]. Usually, GA is initialized with the population of random solutions. In our approach, after we apply the greedy local search procedure, Genetic Algorithm an optimization technique is applied on biclusters to get the optimum bicluster. This has resulted in faster convergence compared to random initialization as discussed in Section V.

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.

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

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

$$\text{ACV Threshold } \delta = \frac{\sum_{p=1}^{p} ACV(p)}{|p|} \ldots\ldots\ldots (6)$$

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

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

0, Otherwise ..(7)

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

This Section discusses about the proposed algorithmic representation of our approach and overall system architecture is as shown in Figure [1].

Algorithm for Web Page Recommendation System based on web usage data.

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.

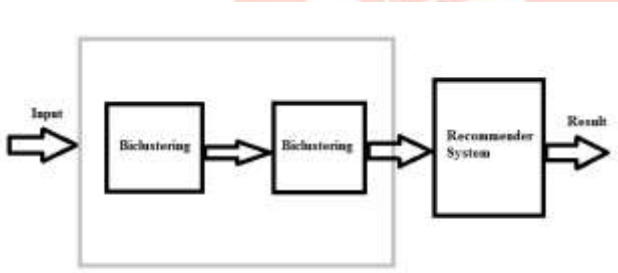

Figure 1. System Architecture

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. Size of msnbc.com dataset is 11.9MB. Page visit of users who visited msnbc.com on 28/9/1999 (Sources: David Hackerman (Heckerman@microsofot.com). 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. Size of Second dataset is 51.7MB (Sources: KSV University Access Log File).

We have shown results of the both dataset. The user access matrix is generated using equation (2). In the next biclustering step K_u User clusters and K_p Page clusters are generated from user access matrix and initial Biclusters K_u xK_p 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 which in turn generates 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:

$$R = \frac{1}{|U|*|P|}\sum_{i=1}^{|U|}\sum_{j=1}^{|P|} T_{ij} \quad \text{..... (8)}$$

Where

$$T_{ij} = \frac{1}{(N-1)} * \left(\sum_{k=1}^{N} W_k(a_{ij}) - 1\right) \text{ ..... (9)}$$

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 $W_k(a_{ij})$ is either 0 or 1. If the element (point) a_{ij} in A is present in the k[th] bicluster, then $W_k(a_{ij})$ = 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 Table [1]:

TABLE 1:- Bicluster Evaluation Function ACV after each step on msnbc.com

Parameters	Initial Bi-Cluster	After Applying Greedy Search Algorithm	After Applying Genetic 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	0.977778

TABLE 2:- Bicluster Evaluation Function MSR after each step on msnbc.com

Parameters	Initial Bi-Cluster	After Applying Greedy Search Algorithm	After Applying Genetic 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

Observing Table [1-4],we see that the average volume of biclusters is increasing after each step. Also we observe that, 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 obtain 0.23 overlapping degree for final biclusters in MSNBC.COM and get 0.21 overlapping degree for final biclusters in KSV Log's file.

TABLE 3:- Bicluster Evaluation Function ACV after each step on KSV Log's

Parameters	Initial Bi-Cluster	After Applying Greedy Search Algorithm	After Applying Genetic Algorithm
Seeds	10	10	10
Average Volume	5271.8	6233.8	6233.8
Overlapping Degree	0.0	0.0210045	0.2176000
ACV	0.5736842	0.5736842	0.6989713

TABLE 4:- Bicluster Evaluation Function MSR after each step on KSV Log's

Parameters	Initial Bi-Cluster	After Applying Greedy Search Algorithm	After Applying Genetic Algorithm
Seeds	10	10	10

Average Volume	5271.8	6233.8	6233.8
Overlapping Degree	0.0	0.0210	0.2176000
MSR	111.82604	28.335459	14.167724

Figure [2]& Figure [3] shows the final recommendation of the website www.msnbc.com and www.ksvuniversity.org respectively. The results are calculated from the optimal biclusters generated after the genetic algorithm. The X-axis of both the graph represents Page Categories and Y-axis of the graph represents the final percentage of the users, who have visited respective page categories of the website. It also represents how much percentage users of total users are viewing their web pages. It quantifies the relevant users and pages of a web site in high degree of homogeneity.

This knowledge can be used by the company for focused marketing campaigns to improve their performance of the business, web personalization systems, web usage-categorization and user profiling. It is generated using the optimal biclusters which follows Genetic Algorithm.

Figure 2 outcome shows that FrontPage and news have been hit maximum number of times by the users and from Figure 3 outcome shows that Indexand Result have been hit maximum number of times by the users.

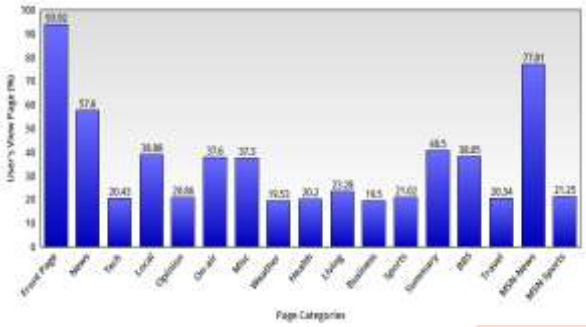

Figure 2. Final Recommendation Graph for MSNBC.Com

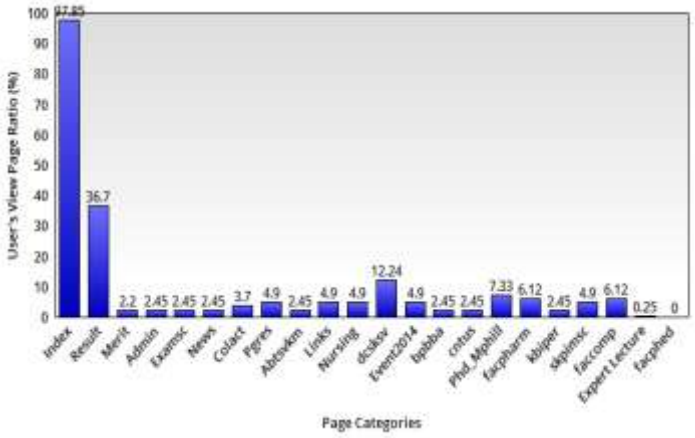

Figure 3. Final Recommendation Graph for KSV Log's File

VI. CONCLUSION

In this paper,we 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 gives most visited pages by different users from the website. It also provides the information of the users having similar behaviour on subset of pages. The results of biclustering approach can be widely used for market strategy like target marketing and direct marketing which are data mining applications. The final optimal biclustering outcomes can also be used towards improving the website's design, information availability and to improve the quality of services provided by the website.

Future work aims at extending this framework by enriching clustering process for enhanced cluster's quality. We have worked with web access log file format to get the clickstream data; one can work with other formats of web log file, also. In this approach, we have taken two parameters from the log file,user IP address and page Categories visited by the users. One can also consider other parameter like the total session time of a user for particular web page which in-turn could improve the recommendation by considering the time factor.

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 inWeb Usage Mining Research Federico Michele Facca and Pier Luca Lanzi.

[9] http://en.wikipedia.org/wiki/Biclustering.

[10] AselaGunawardana, Guy Shani, "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks" Journal of Machine Learning Research 10 (2009) 2935-2962, published 12/09.

[11] SutheeraPuntheeranurak ,Hidekazu Tsuji, " Mining Web Logs for Personalized Recommender System", 2005 IEEE.

[12] Christian Rick, Stefan Arbanowski, Stephen Steglich, " A Generic Multipurpose Recommender System for Contextual Recommendations", Eighth International Symposium on Autonomous Decentralized System (ISADS'07), IEEE.

[13] Ravi Bhusan, Rajendra Nath, "Recommendation of Optimized Web Pages to Users Web Log Mining Techniques",2012 IEEE.

[14] Anant Gupta, Dr. B.K.Tripathy,"A Generic Hybrid Recommender System based on Neural Networks", 2014 IEEE.

[15] MahsaMehrpoor, Andreas Gjaerde, Ole IvarSivertsen, "Intelligent Services: A Semantic RecommenderSystem for Knowledge Representation in Industry", 2014 IEEE.

[16] Wang Ping," Web Page Recommendation Based on Markov Logic Network" 2010 IEEE.

[17] FatemehGhiyafehDavoodi, OmidFatemi, " Tag based Recommender System for Social Bookmarking sites", 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[18] Ryosuke Saga, Yoshihiro Hayashi and Hiroshi Tsuji, "Hotel Recommender System Based on User's Preference Transition", 2008 IEEE.