International Journal of Computational Intelligence and Informatics, Vol. 1 : No. 1, April - June 2011

Optimal Web Page Category for Web Personalization Using Biclustering Approach

P. S. Raja
Department of Computer Science,
Periyar University, Salem,
Tamil Nadu 636011, India.
psraja5@gmail.com

R. Rathipriya
Department of Computer Science,
Periyar University, Salem,
Tamil Nadu 636011, India.
rathipriyar@gmail.com

Abstract

In this paper, Biclustering of Web Usage Data using Genetic Algorithm is proposed to extract optimal web page category. Three different fitness functions based on Mean Squared Residue (MSR) score are used to study the performance of the proposed biclustering method. Experiment was conducted on the CTI dataset, and results of the different fitness functions are analyzed. The valuable outcome of the proposed biclustering method can be used to better understand behavioral characteristics of visitor or user segments, improve the organization and structure of the site, and create a personalized experience for visitors by providing recommendations.

Keywords: Biclustering, Genetic Algorithm, Mean Squared Residue (MSR) score, Web Personalization, Web Usage Mining.

1. Introduction

Web Mining is the generic term of applying data mining techniques to automatically discover and extract useful information and hidden patterns from the World Wide Web (WWW) documents and services. Based on the several research studies it can be broadly classified Web mining into three domains such as Web Content Mining, Web Structure Mining and Web Usage Mining.

Web Content Mining is the process of extracting knowledge from the content of the web documents and their descriptions. Web Structure Mining is the process of inferring knowledge from the structure of web data.

Web usage mining is the process of applying data mining techniques to the discovery of behavior patterns based on web usage data, for various applications. The overall process of web usage mining is generally divided into three main tasks: data preparation, pattern discovery and pattern analysis. The data preparation tasks build a server session file where each session is a sequence of requests of different types made by a single user during a single visit to a site. Pattern discovery converge the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. Pattern analysis is the most important process which analyzes the discovered patterns. Discovered patterns and statistics are further processed, filtered, possibly resulting in aggregate user models or user profiles that can be used as input to applications such as recommendation engines, visualization tools, and web analytics and report generation tools.

Personalization is a process of gathering and storing information about website visitors and analyzing the information. It helps us to deliver the right information or service to each visitor at the right time. A number of personalization techniques, with more on the way, can enable the website to target advertising, promote products, personalize news feeds and recommend documents , products, etc,. The major steps in the web personalization are collecting visitor information, filtering, and developing recommendations. Collecting the visitor's information are the fundamental step and its objective is to develop a profile that describes a site visitor's interests, role in an organization, entitlements, purchases, or some other set of descriptors important to the site owner.

This paper is focused on the biclustering approach to extract the optimal group of web pages based on usage behavior of the users of a web site. Clustering is the widely used technique to detect the browsing pattern from web usage data in the literature [1], [15]. The goal of clustering is to group a set of objects (web users or pages) into a number of more or less homogeneous clusters with respect to a suitable similarity measure. Objects that are similar are allocated in the same cluster, while the objects that are differ significantly are put in different clusters.

By analyzing the characteristics of web user, it is observed that some web users behave similarly only on a subset of pages (i.e. users do not show interest for all page of a web site). Traditional, clustering methods used for grouping the related users, it typically partitions users according to their similar browsing behavior under all pages of a web site. Therefore, it will fail to identify users groups whose behavior similar for a subset of pages [11].

To overcome this problem, the concept of biclustering is used in this paper. Biclustering was first introduced by Hartigan and called it direct clustering [7]. Biclustering is defined as the process of simultaneous clustering of rows and column of a data matrix. Biclustering is also known as coclustering, bidimensional clustering and subspace clustering in literature [6], [8]. A few works are available in the literature that has been applied biclustering for web usage data [11, 14]. The objective of this paper is to extract the optimized bicluster from the web usage data using genetic algorithm. In this work, three different

fitness functions based MSR is used to extract optimized biclusters. The evaluation index called coherence index is used to find the best fitness function to extract the optimized biclusters. These biclusters have a high degree of correlation between the web users and pages of a website.

Section II describes the methods and materials required for biclustering of web usage data. Section III describes the proposed work. Section IV contains experimental results of the proposed work and Section V concludes the paper with the possibility of future work.

2. Materials and Methods

2.1 Mean Squared Residue Score

Given A(U,P) be a web access matrix. For user subset U' ⊆ U and page subset P'⊆ P, A(U',P') denotes the submatrix of A called bicluster that contain only the elements a_{ij} satisfying U' ⊆U and P' ⊆ P. Cheng and Church defined a bicluster as a subset of rows and subset of columns which has low Mean Squared Residue (MSR) Score [4][13].

$$MSR(B) = \frac{1}{nm}\sum_I \sum_J \left(b_{ij} - \overline{b_I} - \overline{b_J} + \overline{b_{ij}}\right)^2 \quad (1)$$

Where $\overline{b_I}$ is the average value of row i, $\overline{b_J}$ is the average value for column j.

A Coherent bicluster is a bicluster with coherent values on both rows and columns. The degree of coherence of a bicluster is measured by using MSR score or hscore. It is the sum of the squared residue score. A submatrix B'_{ij} is called a δ bicluster if MSR(I,J)< δ for some δ >0. A high MSR value signifies that the data is uncorrelated. A low MSR value means that there is correlation in the matrix. The value of δ depends on the dataset.

2.2 Bicluster Seed Formation

In this work, Two-way K-Means algorithm is used to form the initial biclusters from web access matrix A[4]. The initial matrix A is partitioned into p*q submatrices where p is the number of clusters on row dimension and q is the number of clusters on column dimensions. From p*q submatrices, submatrices having hscore value below a certain limit called bicluster seeds. These bicluster seeds are used to initialize the initial population for GA.

2.3 Binary Encoding of Bicluster seeds

Each bicluster seed is encoded as a binary string [4]. The length of the string is the number of rows plus the number of columns of the web access matrix. A bit is set to one when the corresponding page or user is included in the bicluster otherwise zero. This representation is advantageous for node (i.e. row or column) addition and deletion.

2.4 Genetic Operators

This subsection gives a brief description of the genetic operators, since they play a key role in how to search is performed by the GA [2] [9] [10].

2.4.1 Selection

During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. Certain selection methods rate the fitness of each solution and preferentially select the best solutions. Other methods rate only a random sample of the population, as this process may be very time-consuming. Most functions are stochastic and designed so that a small proportion of less fit solutions are selected. This helps keep the diversity of the population large, preventing premature convergence on poor solutions. Popular and well studied selection methods include roulette wheel selection and tournament selection. In this paper, ***roulette wheel selection*** is used as a selection operator.

2.4.2 Crossover

This operator selects individuals from parent and creates a new offspring. The simplest way how to do this is to choose randomly some crossover point and everything before this point copy from a first parent and then everything after a crossover point copy from the second parent. In this work, **multipoint crossover operator** is used**.**

2.4.3 Mutation

After a crossover is performed, mutation takes place. This operator helps to prevent falling all solutions in population into a local optimum of solved problem. Mutation changes randomly the new offspring. The probability of mutation used in this work is **0.01** percentage.

3. Biclustering of Web Usage Data using Genetic Algorithm

The main objective of this paper is to introduce the biclustering concept for web usage data using GA and to extract the optimal page category based on the users' browsing behavior. In this work, GA is used as an optimization tool to extract global optimized biclusters from the given population to predict optimal web page category [2] [9] [10].

3.1Fitness Function

The fitness function, is normally used to transform the objective function value into a measure of relative fitness.

$$F(x) = g(f(x)) \quad (2)$$

where *f* is the objective function, *g* transforms the value of the objective function to a non-negative number and *F* is the resulting relative fitness. The main objective of this work is to

find maximal biclusters with low mean squared residue. The threshold value δ for the three fitness function is purely dependent on the dataset taken for the study. Three different fitness functions based on MSR score available in the literature are taken for the study to identify the optimized bicluster with low MSR score and high volume.

Algorithm 1: Biclustering Approach using Genetic Algorithm

1. [Start] Generate random population of *n* biclusters.
2. [Encoding] Encode the biclusters in to binary strings of size N x M, where N is the number of biclusters in population, M is the sum of number of rows and number of columns in the bicluster.
3. [Fitness] Evaluate the fitness *f(x)* of each bicluster *x* in the population
4. [New population] Create a new population by repeating following steps
   i. [Selection] Select the best individuals from a population according to their fitness
   ii. [Crossover] With a crossover probability crossover the biclusters to form a new offspring.
   iii. [Mutation] With a mutation probability mutate new offspring at each locus.
   iv. [Accepting] Place new offspring in a new population
5. [Replace] Use new generated population for a further run of algorithm
6. [Evaluate] Evaluate the fitness $f(x_n)$ of each bicluster x_n in the new population.
7. If convergence is met, stop, and return the best bicluster in current population as optimal bicluster.
8. [Loop] Go to step 2

Fitness Function I: Given the value of δ (δ>0), the following fitness function can be used to assess the quality of bicluster.

$$B(I,J) = \begin{cases} |I|.|J| & if\ MSR(I,J) \leq \delta \\ \delta / MSR(i,j), & Otherwise \end{cases} \quad (3)$$

Fitness Function II: The fitness function II used to evaluate a given bicluster is given as follows:

$$f(I,J) = R/\delta + w_c.\delta/|I| + w_r.\delta/|J| \quad (4)$$

where I and J are the set of rows and columns, respectively, in the bicluster, R is called the mean square residue of a bicluster. δ is a residue threshold (the maximum desired value for residue),w_c is the importance of the number of columns, and w_r the importance of the number of rows that denotes the number of elements on a given set.

In fitness function II, a ratio between two conflicting objectives minimizing the residue (variance of elements in the bicluster) and maximizing the size of the cluster is combined. Notice that, for a bicluster to be meaningful, it should contain
a reasonable number of elements so that some knowledge can be extracted. Also, it is important to maintain some correlation among its elements.

Fitness Function III: This fitness function is used to evaluate a given bicluster is given as follows:

$$F(I,J) = MSR(I,J)/\delta.(1 + var(I,J)) \quad (5)$$

where I and J are the set of rows and columns, respectively, in the bicluster, MSR is called the residue of a bicluster and is calculated as in the given equation above. δ is a residue threshold. var(I,J) is the variance of the rows and the columns of a bicluster.

4. Experimental Analysis

4.1 About Dataset

The dataset, ***CTI***, is from a university web site log and was made available by the authors of Mobasher (2004) and Zhang et al.[3], [14]. The data is based on a random collection of users visiting university site for a 2-week period during the month of April 2002. After data preprocessing, the filtered data consisted of 13745 sessions and 683 pages where root pages were considered as page view of a session. This preprocessing step resulted in total of 10 categories namely, search, programs, news, admissions, advising, courses, people, authenticate, cti, and miscellaneous. These pageviews are given numeric labels as 1 for search, 2 for programs and so on. These page views were given numeric labels as 1 for search, 2 for programs and so on. The session length in the dataset ranges from 2 to 68.Since comparing very long sessions with small sessions would not be meaningful, we considered only sessions of length between 3 and 7. Finally, 5915 user sessions are taken for the experimentation.

4.2 Validation of Biclusters

In this work, two types of validation index are used to evaluate the biclusters obtained from proposed biclustering approach using GA. They are Bicluster Index (BI) and Coherence Index (CI) [12].

Bicluster Index (BI) is the defined as the ratio of residue and row variance of the bicluster identified.

$$BI = \frac{MSR(B)}{(1+R)} \quad (6)$$

Coherence Index(CI) is the defined as the ratio of the mean square residue (MSR) score to the volume of the formed biclusters

$$CI = \min\left(\frac{MSR(B(k))}{Vol(B(k))}\right) \quad (7)$$

The k^{th} bicluster for k ∈ P is considered to be good, if it has minimum CI_k among all j ∈ P and j = k. A small Mean Square Residue (MSR) indicates that the corresponding user group has consistent browsing behavior value over the subset of pages.

Volume of bicluster is defined as the product of number of rows (users) and number of columns (pages) in the bicluster.
Vol(B(I,J))= |I|*|J|

Table 1: Parameter Setting for GA

Crossover Probability	**0.7**
Mutation Probability	0.01
Population Size	100
Generation	100
MSR Threshold	70

The parameter setting for this study is given in the Table 1and type operator for Genetic Algorithm is tabulated in Table 2. The characteristics of the optimal bicluster using Fitness function I, II and III are tabulated in the Table 3. In which, MSR and Row Variance are the quality measures and Volume is the quantity measure. The aim is to extract the optimal bicluster with minimal MSR and high volume. From this point of view, Fitness function I performs better than other two fitness function.

Table 2: Type of Genetic Operators used in this Study

GA Operators	Type
Crossover	Multi-point Crossover
Selection	Roulette Wheel Selection
Mutation	Bit mutation

Table 3: Characteristics of Optimal Biclusters using different Fitness Function

Type of Fitness Function	Fitness Value	MSR	Row Variance	Volume
Fitness Function I	0.42	103.23	69.861	10246
Fitness Function II	302.37	112.01	70.254	10152
Fitness Function III	0.033	143.49	70.037	10176

Mean Bicluster Index (BI) and Coherence Index (CI) are used validating indices for the biclustering algorithm. A small BI index and CI index indicate the subset users in the biclusters show highly correlated browsing behavior over the subset of pages in that biclusters.

Table 4: Validation Measures for Various Fitness Functions

	Mean Bicluster Index	Coherence Index
Fitness Function I	19.996	0.084
Fitness Function II	22.004	0.096
Fitness Function III	24.003	0.0908

From the values recorded in the Table 4, Function I has low BI and CI values. Therefore, the optimal page category is generated from the optimal bicluster extracted by biclustering approach using fitness function I and its page views are tabulated in the Table 5.
From the results, it is obvious that it correlates the relevant users and pages of a web site in high degree of correlation. Analyzing these results could be very beneficial for E-commerce applications such as target marketing and also web recommending system, web personalization systems and web usage categorization. This result is also used for focalized marketing campaigns to improve their performance of the business.

Table 5: Optimal Page Category of CTI Dataset

Page View Index	**1**	**2**	**7**	**10**
Page View Category	Search	Program	People	Miscellaneous
Weights	0.5033	0.2091	0.8374	0.3017

5. Conclusion

In this paper, a biclustering algorithm using Genetic Algorithm is proposed for web usage data, for identifying an optimal page category of a web site. With this technique, optimized bicluster is obtained with their most representative pages that correspond to users in a real web site. This information can be also used for marketing purpose, customizing the web site once one knows the kind of user through the navigation characteristics.

6. References

[1] Amos Tanay, Roded Sharan, Ron Shamir, "Biclustering Algorithms: A Survey", Oxford University Press, 2002.

[2] Anupam Chakraborty and Hitashyam Maka "Biclustering of Gene Expression Data Using Genetic Algorithm" Proceedings of Computation Intelligence in Bioinformatics and Computational Biology CIBCB, pp 1-8, 2005.

[3] B. Mobasher, "In The Adaptive Web: Methods and Strategies of Web Personalization", Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.), Lecture Notes in Computer Science, Vol. 4321, pp. 90-135, Springer, 2007

[4] Baiyi Xie, Shihong Chen, Feng Liu, "Biclustering of Gene Expression data using PSO-GA hybrid", Proc. of the First International Conference on Bioinformatics and Biomedical Engineering, pp.302-305, 2007.

[5] Cao Y.J and Wu Q.H., "Teaching Genetic Algorithm Using Matlab", Int. J. Elect. Enging. Educ., Vol. 36, pp. 139–153, 1999.
[6] Chakraborty Aand Maka H, "Biclustering of gene expression data by simulated annealing", HPCASIA '05, pp 627-632, 2005.
[7] Hartigan J.A., "Direct clustering of Data Matrix", Journal of the American Statistical Association, Vol. 67, pp. 123- 129, 1972.
[8] Madeira S. C, Oliveira A. L., "Biclustering algorithms for Biological Data analysis: a survey" IEEE Transactions on computational biology and bioinformatics, pp. 24-45, 2004.
[9] Maheswara Rao V.V.R., Valli Kumari V. and Raju, "An Advanced Optimal Web Intelligent Model for Mining Web User Usage Behavior using Genetic Algorithm", In Proc. of Int. Conf. on Advances in Computer Science 2010.
[10] Manoj Kumar, Mohammad Husian, Naveen Upreti & Deepti Gupta, "Genetic Algorithm: Review and Application", International Journal of Information Technology and Knowledge Management, Vol. 2, pp. 451-454, 2010.
[11] R. Rathipriya, K. Thangavel, J. Bagyamani: Evolutionary Biclustering of Clickstream Data ,International Journal Computer Science Issue, Vol. 8, pp:32-38, 2011.
[12] Santamaría, R., Quintales, L., and Therón, R. "Methods to Bicluster Validation and Comparison in Microarray Data", Intelligent Data Engineering and Automated Learning, Yin, H., Tino, P., Corchado, E., Byrne, W., and Yao, X., (eds), Lecture Notes in Computer Science, Vol. 4881, pp. 780–789, 2007.
[13] Yizong Cheng and George M. Church, "Biclustering of Gene expression data", Proc Int Conf Intell Syst MolBiol, Vol. 8, pp. 93-103,2000.
[14] Zhang. Y., Xu, G., & Zhou, X. "A latent usage approach for clustering Web transaction and building user profile". In Proc. of First International Conference on Advanced data mining and applications, pp. 31-42, 2005.
[15] Zong Y, Xu G, Dolog P and Zhang Y, Co-Clustering for Weblogs in Semantic Space. In Proc. of 11th International Conference on Web Information Systems Engineering (WISE'10), Lecture Notes in Computer Science, Vol. 6488, 120-127, 2010.

P.S.Raja was born in 1984 at Rasipuram in Nammakal, Tamilnadu, India. He is received the Master of Computer Application in 2010, from Anna University, Coimbatore. He obtained his M.Phil (Computer Science) Degree from Periyar University, Salem, India in 2011. His area of interests includes, Data Mining, Biclustering, Web Mining, Image processing, Gene Clustering Analysis.

Rathipriya R, working as Assistant Professor in Periyar University, Salem, India. She received her Bachelor of Science and Master of Science degrees in Computer Science from the Periyar University. Currently, she pursuing her Ph.D in Bharathiyar University, India. Her research interests are in several areas of data mining, web mining, Optimization techniques (in particular, optimization of biclusters in web mining area), and Bio-Informatics.