Hybrid Modelling Of Web Usage Mining For Web Page Recommendation Based On Click Stream Data

Darshan K Prajapati[1], Dr. Shyamal M Tanna[2]

[1] *M.E. Student, Computer Engineering, L.J. Institute of Engineering & Technology, Gujarat, India*
[2] *Assistant Prof., Computer Engineering, L.J. Institute of Engineering & Technology, Gujarat, India*

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. 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. With the rapid development of biclustering, more researchers have applied to the web usage data it automatically captures the hidden browsing patterns from it in the form of biclusters. In this study, we propose a novel biclustering algorithm based on genetic algorithms (GAs) to effectively segment the web usage data. In this work Genetic Optimization technique is combined with biclustering approach to propose a recommendation system using GA based biclustering of Web Usage Data. The main objective of this Algorithm is to retrieve the global optimal bicluster from the web usage data.

Keyword: - *W*ebpage *Recommendation, Clickstream Data, Gentic Algorithm, Biclustering*

1. INTRODUCTION

World Wide Web (WWW) is very popular and interactive with the information available over the Internet, it becomes a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Web contains the rich and dynamic collection of hyperlink information and web page access and usage information, providing source for data mining.

Web mining is defined as an application of data mining technique to automatically discover and extract information from web documents and service. Web mining is categorized into three areas which are Web usage mining, Web content mining, Web structure mining.

Web Usage Mining (WUM) is the process of extracting knowledge from Web user's access logs or usage data, by making use of data mining technologies. The primary source of data in web usage mining is the server log captured by web servers. Web usage mining involves three phases are data preparation and transformation, pattern discovery and recommendation. The main goal of the recommender system is to improve the web site usability by knowing the interest of the users.

A web recommendation process consists of two components namely online and off-line components with respect to web server activity. Offline component builds the knowledge base by analysing historical data, such as server access log file or web logs which are captured from the server. Then these web logs are used in the online component for capturing the intuition list of the user so as to recommend page views to the user whenever user comes online for the next time[9].

In this paper, a framework is generated for capturing recommendations in the form of recommendation list using biclustering techniques. A recommendation list consists of pages visited by User as well as list of pages visited by other users of having similar usage profile. Biclustering is extension of clustering technique which allows simultaneously clustering of columns and rows of a matrix. The goal of biclustering is to identify maximum sub groups of users and sub group of pages such that users express highly correlated activities over a range of pages. The greedy search improves the result of bicluster and then genetic algorithm takes these biclusters as initial population and generates optimal biclusters and gives final recommendation with accuracy.

2. RELATED WORK

Mr.M.Saravanan and,Dr.V.L.Jyothi [1] proposed the optimal sequence of pages in log file. Log files has huge amount of data so to obtain sequential pages first data preprocessing is need to be done and then genetic algorithm has to be applied to get the optimal sequence of pages. Genetic algorithm helps in large amount of data input. The sequential pages of visited user is presented which enhances the pre-processing steps of web log usage data in data mining. Firstly, user identification and session identification and session identification is applied on the log files. This pre-processing step filters the number of users and the number of unique users. Sequential access table is the combination of both user and session identification. All the sequential pages visited by the users can be seen in this table. For getting the best sequential pages, genetic algorithm is used. And lastly best sequential pages is obtained.

R.Rathipriya, Dr. K.Thangavel, J.Bagyamani [2] proposed a bi-clustering approach for web data, which identifies groups of related web users and pages using spectral clustering method on both row and column dimensions. biclustering algorithms are widely applied to the gene expression data. Most of these algorithms are failed to extract the coherent pattern from the data matrix. In web mining, there is no related work that has been applied specific biclustering algorithms for discovering the coherent browsing patterns. In this paper, Greedy Search Procedure and evolutionary approach namely Genetic Algorithm (GA) is introduced to obtain the optimal coherent browsing patterns. The results show that GA outperforms the greedy procedure by identifying coherent browsing patterns. These patterns are very useful in the decision making for target marketing.

Hiral Y. Modi , Meera Narvekar [3] proposed an Online recommendation System. The system involved two phases that work in conjunction with each other i.e. the online and offline phase. Data pretreatment and navigation pattern mining is carried out in offline phase while predictions are generated in the online phase. They also proposed the online and offline phase architecture.

Ravi Bhushan and Rajender Nath [4] proposed architecture can be divided into two main phases Back end and Front end. In the back end phase, there are main two modules: Data pre-processing and sequential pattern mining. The block diagram of recommendation system is given below. Back End Phase consists of two modules and modules are data preprocessing and sequential pattern mining. In the front end phase, URL request of the user is processed by search engine and captures the recommended list of web pages relevant to user query and then rank updating algorithm is applied on them.

Diviya Prabha, R. Rathipriya [5] proposed Gravitational Search Algorithm (GSA) is used to propose a new biclustering algorithm to extract the highly correlated pattern from the optimal bicluster. This GSA is based on the Newtonian gravity: ‘‘Every particle in the universe attracts every other particle with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them”. The clustering techniques work good for small dataset value but work poor for large data sets and if web data is huge that groups similar users under all pages. Besides, algorithm cannot overlap for clusters that are generated i.e. user belong to one cluster may participate in many other clusters with different conditions. To overcome these problems biclustering technique is introduced in the literature. The bicluster are defined to be a set of users and a set of pages where similar users are grouped under specific page.

Thiyagarajan et al. [6] a new recommendation system is proposed to predict the user’s navigational behaviour. The practical implementation of this algorithm shows that the prediction of user intuition capturing is more accurate. This paper has paid an attention to group the similar usage behaviour of users using K-Means algorithm and new validating measure called MSR is applied to evaluate the cluster’s quality. It was observed that the experimental results of the proposed approach using Hamming similarity measure improved the quality of recommendation for the case of binary data. In real life a user may be placed in more than one clusters and this has not been taken care in the proposed approach.

3. METHODS AND ALGORITHMS

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.

Bicluster Types [10]

Different biclustering algorithms have different definitions of Bicluster.

1) Bicluster with constant values.

2) Bicluster with constant values on rows or columns.

3) Bicluster with coherent values

3.2. Clickstream Data Pattern

Clickstream data [5] is one of the web usage data format. It 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 analysing the clickstream data. It cannot use as such, it requires the some pre-processing before it is taken for analyse.

3.3. Pre-processing of clickstream data [5]

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.

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

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

3.4. 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.

3.5. 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. 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.

i) ACV: The following fitness function F (I, J) is used to extract optimal bicluster

$$F(I, J) = \begin{cases} |I|*|J|, \text{ if ACV (bicluster)} \geq \delta \\ 0, \text{ Otherwise} \end{cases} \quad (2)$$

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

$$\text{ACV Threshold } \delta = \frac{\sum_{p=1}^{p} \text{ACV (p)}}{|P|} \quad (3)$$

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

$$F(I, J) = \begin{cases} |I|*|J|, \text{ if } MSR(\text{bicluster}) \leq \delta \\ 0, \text{ Otherwise} \end{cases} \quad (4)$$

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.

4. PROPOSED WORK

Step 1: Load data set.

Step 2: Pre-process data and generate user matrix A.

Step 3: Generate initial Bicluster using two way K-means clustering from user matrix A.

Step 4: Improve the quality and quantity of the initial Biclusters using Greedy search procedure with two Bicluster evaluation function ACV & MSR.

Step 5: Apply genetic algorithm.

Step 6: Evaluate the fitness of individuals.

Step 7: For i=1 to max_iteration

Selection ()

Crossover ()

Mutation () Evaluate the fitness

End (For)

Step 8: Return the optimal Bicluster.

Step 9: Generate Recommendation for website.

Step 10: Stop

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Data Set:

A real dataset is used for this experiment. The data set is taken from the UCI dataset repository (http://kdd.ics.uci.edu/) that consists of Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 28, 1999 (Pacific Standard Time). Visits are recorded at the level of URL category and are recorded in time order. Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail that is, at the level of URL, but rather, they are recorded at the level of page category. The categories are "front page", "news, "tech", "local", "opinion", "on-air", "misc",

"weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". Any page requests served via a caching mechanism were not recorded in the server logs and, hence, not present in the data. This dataset is slightly changed flattering to our experiment, if the user visit only the "front page" then 1 is recorded on the first position of the matrix and other 16 column (category) are filled by 0 [6].

TABLE I: Dataset Used In The Experiment

Dataset	MSNBC
Total Number of Users	989818
Average Number Per visit	5.7
Number of URL for each categories	10-5000

We have shown results of the MSNBC dataset. 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

$$Tij = \frac{1}{(N-1)} * \left(\sum_{k=1}^{N} W_k(a_{ij}) - 1 \right) \quad \ldots (5)$$

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 $0 \leq R \leq 1$.

Fig 1 describe the biclustering details after the each and every algorithm . Result can see that ACV is increased after applying the genetic algorithm and MSR value is decreased after applying the genetic algorithm. That means optimized bicluster is generated after genetic algorithm.

Parameters	Initial Bi-Cluster	After Applying Greedy Search Algorithm	After Applying Genetic Algo Algorithm
Seeds	114	114	114
Average Volume	16263.28	1938.0	351575.28
Overlapping Degree	0.0	0.0190045	0.21350000000000002
ACV	0.42643172	0.58894	0.922269
MSR	8281359.0	7906758.5	50768.105

Fig-1 Details Of Bi-Cluster After Each Steps

Chart 1 & 2 describes the recommended pages to the user. Result displayed all the percentage of the web pages. Front ,News and BBS pages hits frequently by the all users. Chart-1 shows the recommended web pages to the user using ACV ad chart-2 show the recommended web pages to the user using MSR.

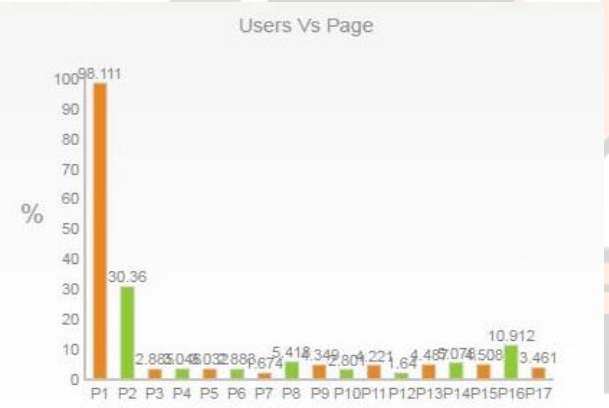

Chart-1 Recommended Pages With ACV

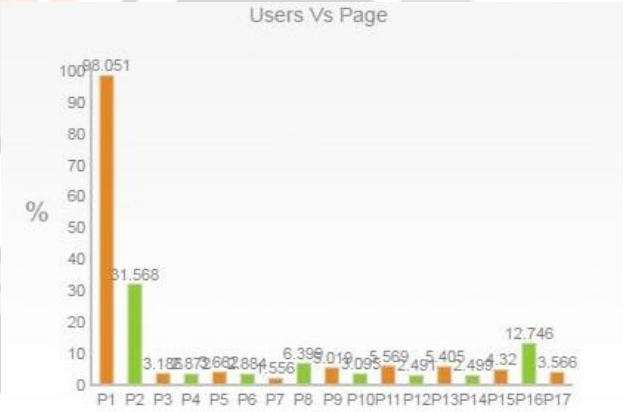

Chart-2 Recommended Pages With MSR

6. CONCLUSIONS

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. The next step recommender system will take as input the optimum bicluster. The combination of the biclustering and GA will be treated as the pre-processing tool for our proposed recommender system. That is the two dimensional optimal user X page matrix. The Last step will generate the nearest users which are having same browsing pattern and also recommend the user their page of interest which they have visited mostly in the past. Hence in the last step web pages are recommended to the user.

7. REFERENCES

[1] Mr.M.Saravanan, Dr.V.L.Jyothi, " A Novel Approach for Sequential Pattern Mining By Using Genetic Algorithm " , International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) ,978 -1-4799-4190-2/14/$31.00 ©2014 IEEE , pg no :284-288

[2] 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 ,ISSN (Online): 1694-0814 , pg no: 341-347

[3] Hiral Y. Modi, Meera Narvekar, "Enhancement Of Online Web Recommendation System Using A Hybrid Clustering And Pattern Matching Approach" , 2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2015), 978-1-4799-7263-0/15/$31.00 ©2015 IEEE

[4] Ravi Bhushan and Rajender Nath, "Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques" , 978-1-4673-4529-3/12/$31.00c 2012 IEEE, pg no :1030-1033

[5] V. Diviya Prabha, R. Rathipriya, "Biclustering of Web Usage Data Using Gravitational Search Algorithm", Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, 978-1-4673-5845-3/13/$31.00©2013 IEEE, pg no: 500-505

[6] R. Thiyagarajan, K. Thangavel, R. Rathipriya, "Usage Profile based Recommendation system", IEEE, International Conference on Intelligent Computing Applications (ICICA), 2014, Page(s): 382–386, ISBN: 978-1-4799-3966-4.

[7] Pablo A. D. de Castro, Fabrício O. de França Hamilton M. Ferreira and Fernando J. Von Zuben, "Evaluating the Performance of a Biclustering Algorithm Applied to Collaborative Filtering – A Comparative Analysis", Seventh International Conference on Hybrid Intelligent Systems, 0-7695-2946-1/07 $25.00 © 2007 IEEE DOI 10.1109/HIS.2007.55, pg no: 65-70

[8] Zhongyun Ying, Zhurong Zhou, Fengjiao Han and Guofeng Zhu, "Research on Personalized Web Page Recommendation Algorithm Based on User Context and Collaborative Filtering", 978-1-4673-5000-6/13/$31.00 ©2013 IEEE, pg no :220-224

[9] Ruimei Lian, "The Construction of Personalized Web Page Recommendation System in E-commerce", 978-1-4244-9763-8/11/$26.00 ©2011 IEEE, pg no :2687-2690

[10] http://en.wikipedia.org/wiki/Recommender _system, Date: 21/12/2014, Time:19:06:00.