WEBPAGE RECOMMENDATION IN WEB USAGE MINING USING GENETIC ALGORITHM

Panjwani Heena[1], Pooja Jardosh[2]

[1] *Research Scholar,Computer Department, Silver Oak College of Engineering & Technology, Gujarat, India*
[2] *Assistant Professor., IT Department, Silver Oak College of Engineering & Technology, Gujarat, India*

ABSTRACT

Web mining is an emerging field of data mining used to provide personalization on the web. The basic focus of web mining is to use data mining techniques and algorithms to extract useful and hidden patterns from unstructured and huge web data or resources. One of the applications of WUM is Recommendation system which is personalized information filtering technique used to either determine whether a certain user will approve a given item or to identify a list of items which can be of significant importance to the user. The purpose of using genetic algorithm is to find optimal sequential web pages. The practical implementation of this algorithm shows the more accurate prediction of user's intuition on web. The knowledge gained by the analysis is applied to target marketing and in the designing of web portals.

Keyword : - *Recommendation, Webpage Recommendation, Genetic Algorithm, Pattern Recognitions , Web Usage Mining*

---

1. INTRODUCTION

The growth of web site over the World Wide Web[1] (WWW) has not only raised many concerns but also opened a window of opportunity for organizations to analyze the lifetime value of their customers, and also improve their cross marketing strategies. The new strategies involve analyzing a large collection of data. Web mining is the withdrawal of notable and potentially valuable patterns[3] and implicit information from activity related to the web site .Web Usage Mining involves determining the frequency of the page access by the clients and then finding the common traversal paths of the users.

In web usage mining, clustering is based on the assumption that the entire related user behave similarly across all the pages of a web site and vice versa. Data mining is also one of the important application fields of genetic algorithms. In data mining, GA can be used to either optimize parameters for other kinds of data mining algorithms or discover knowledge by itself[3]. Generally, all the recommendation systems follows a framework for generating efficient

recommendations. Various recommendation systems use different approaches based on the sources of information they utilize.

1.1 Web Mining

Web mining is the strategy to group the pages and web clients by looking into the substance of the page and conduct of web client previously. Web mining[1] can be classified into content mining, structure mining or usage mining.

• **Web Usage Mining:** web analysis visitors usage habits can be provide significant clues about present market trends and help organizations to predict the future potential customers trends. Also, mined information can be used to provide preferred web content to visitors[1].

• **Web Structure Mining:** Web structure mining means[1] to mine the information from the linkage structure of the Webpage then this linkage information is used to capture the list of interesting patterns. It is the Procedure using in graph theory to explore the node and connection structure of a web site.

• **Web Content Mining:** Web content mining means that to mine[1] the info from record set or weblog. Website mining is additionally called Text mining. In this technique content like text, graphs and picture, are parsed or scanned to find out the significance of the content to the requested query.

2. Background Terminologies

The clustering techniques such as k-means clustering and hierarchical clustering 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. To overcome these problems bi-clustering technique is introduced in the literature. The bi-cluster [2] are defined to be a set of users and a set of pages where similar users are grouped under specific pages.

A. Bi-cluster Types

A bi-cluster of a web usage data is defined as a subset of users which exhibit similar interest or browsing patterns along a subset of pages[2]. . The different types of bi-clusters are Constant bi-cluster, Constant column bi-cluster and Coherent bi-cluster.

B. Coherent Bi-cluster

A bi-cluster 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[2].

C. Genetic algorithm

Genetic algorithms (GA) are search algorithms based on the principles of natural selection and genetics, introduced by John Holland in the 1970s and inspired by the biological evolution of living beings.

Working steps of Genetic Algorithm are:

1. Generate random population of n chromosomes i.e. suitable for the problem.

2. Evaluate the fitness f(x) of each chromosome x in the population[3].

3. Create a new population by repeating following steps until the new population is complete.

a) [SELECTION]: Reproduction is an operator that makes more copies of better strings in a new population. Reproduction is usually the first operator applied on a population [3].

b) [CROSSOVER]: A crossover operator is used to re- combine two strings/parents to get better new two strings/children. In the crossover operator, new strings are created by exchanging information.

c) [MUTATION]:.It is an operator that introduces diversity in the population whenever the population tends to become homogeneous due to repeated use of reproduction and crossover operators[3] .

4. [REPLACE] use new generated population for the further run of the algorithm.

5. [TEST] if the condition is satisfied then stops and re- turns the best solution in current population

6. [LOOP] Go back step 2.

3.LITERATURE REVIEW

Web usage mining is required to process web data for predicting and identifying accessed information. Web usage mining is the category of web mining that helps in automatically discovering user access pattern. Preprocessing of server access log data is very important task in mining process. Data is preprocessed in order to improve the quality of data, the efficiency and ease of the mining process.

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.

Hiral Y. Modi , Meera Narvekar [4] proposed involved two phases that work in conjunction with each other i.e. the online and offline phase.

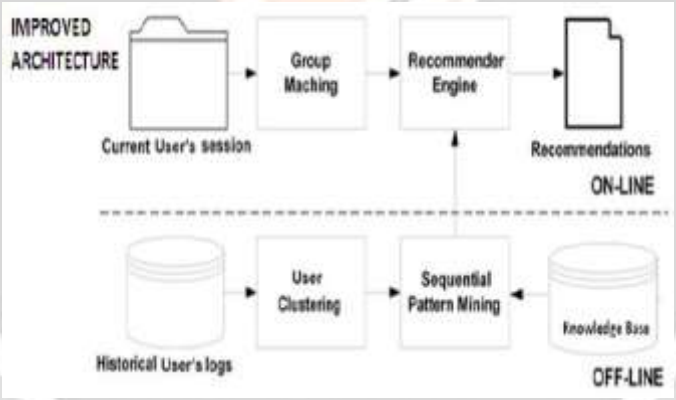

Fig -1. Online & Offline Phase Architecture[4]

4. EXPERIMENTAL RESULTS AND ANALYSIS

A. 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[6] for msnbc.com and news-related portions of msn.com for the entire day. 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".

The details of the data set are provided in Table-1.

Table -1: Dataset Used In The Experiment

Dataset	MSNBC
Total Number of Users	989818
Average number of visits per user	5.7
Number of URL for each categories	10-5000

The Average volume of bi-cluster is increasing after each step. As the value of ACV is increasing the value of MSR will be decrease. A high ACV and Low MSR value indicates that the bi-cluster is strongly coherent.

Table -2: Bi-cluster Evaluation Function ACV after each step.

Parameters	Initial Bi-clusters	After greedy search	After Genetic 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

Recommendation quality = ACV*100
So, our proposed system will gives the best quality and accurate results as compared to other recommendation functions like cosine similarity and hamming similarity.

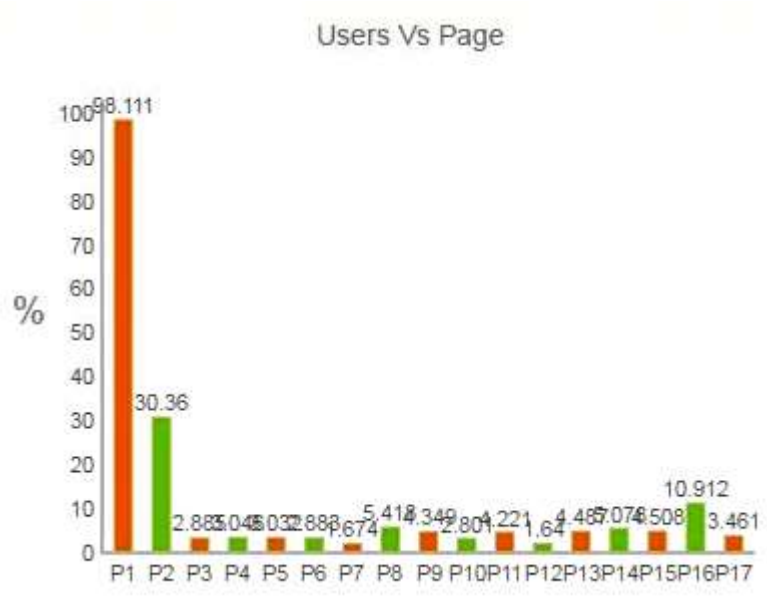

Chart -1: Recommendation Using ACV

5. CONCLUSIONS

The main contribution of this research is to development of recommender system using coherent bi-clustering framework with GA to identify coherent bi-clusters from the click-stream 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 behavior. The objective of this research is to find high volume bi-clusters with high degree of coherence between the users and pages.

6. REFERENCES

[1] Monika Dhandi,Rajesh kumar Chakrawarti "A Comprehensive Study of Web Usage Mining",Colossal Data Analysis and Networking (CDAN) ,978 -1-5090-0669-4/16/$31.00 ©2016 IEEE

[2] 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.

[3]Surbhi Bhatia, "New Improved Technique for initial cluster centers of K means Clustering Using Genetic Algorithm",International conference for convergence of technology-978-1-4799-3759-2/14/$31.00© 2014 IEEE,

[4] 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

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

[6]Deepti Sahu,Rishi Soni, "A New Method for Detecting Users Behavior from Web Access Logs",978 -1-5090-0076-0/15/$31.00 ©2015 IEEE

[7] Rosli Omar, Abu Osman Md Tap, Zainatul Shima Abdullah, "Web Usage Mining: A Review of Recent Works",date:11/8/2015 time:10:58 PM

[8] 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

[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] Ming Hu , Guannan Zheng , and Hongmei Wang, "Improvement and Research on AprioriAll Algorithm of Sequential Patterns Mining", 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, 978-1-4799-0245- 3/13/$31.00 ©2013 IEEE, pg no :158-161