IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Prediction of User Browsing Behavior Using Web Log Data

Abhishek Chauhan[1], Dr.Sandhya Tarar[2]

Master of Technology in Computer Science & Engineering, Gautam Buddha University,Uttar Pradesh, India[1]

Research Associate, Department of Information & Communication Technology, Gautam Buddha University, Uttar Pradesh, India[2]

Abstract: Web usage mining is one of the areas of data mining where it not only deals with the analysis of web data but also discovers frequent patterns from them to improve the web based applications. It consists of three phases preprocessing, pattern discovery and pattern analysis. After the completion of these phases, the user can find interesting and frequent patterns whatever he/she wants to be need with the use of these information for specific needs. In other words, one can say that it deals with the extracting of useful information and find interesting usage patterns from web server, proxy server or web clients. In this paper, combined approach of Ant Colony Optimization and DBSCAN is used to get effective Association Rules to predict the customer's browsing behavior .Moreover, the aim is to find the user's behavior with less error rate in prediction.

Keywords: Web mining, Web log, Pattern discovery, Clustering, User's browsing behavior.

I. INTRODUCTION

With the rapid growth of explosive Internet information, the user's always find themselves in the 'ocean' of information as they didn't find desirable and accurate information which they actually wants due to which there are mainly two problem arises i.e. low recall and low precision when they interacting with the web. In the 21 st century, technologies grow day by day and without having Internet we couldn't imagine our life. To overcome these web problems, one of the applications of the data mining technique is used. Web mining is the process of extracting the useful information and find the frequent patterns from the web. It consists of three types namely as web content mining, web structure mining and web usage mining.

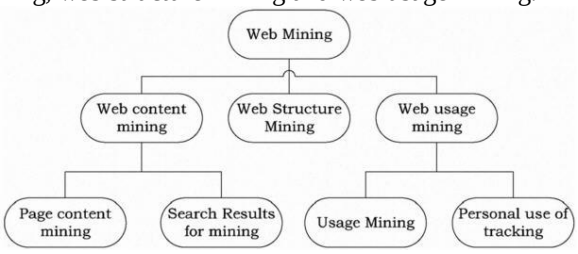


Fig. 1: Classification of Web Mining.

A. Web content mining

Web content mining is also called text mining as it is the process of finding extracting, integration of useful data and information from the web contents which consists of text, audio, images and hyperlinks etc. Moreover, the technologies used for web content mining are NLP and IR i.e. Natural Language Processing & Information Retrieval respectively.

B. Web structure mining

Web structure mining is the mining process by analyzing and mining the information or direct link connection to provide the relationship between the web pages linked. As according to the type of web structure data provided, it is basically divided into two kinds namely as: extracting patterns from hyperlinks in the web and mining the document structure.

C. Web usage mining

Web usage mining is the process to find interesting and frequent access patterns of the user's browsing behavior from the web log. It deals the navigation pattern of the customer to serve them better what they actually wants. All the e- commerce shopping sites used web usage mining technique in order to help designer not only to improve the web-site & to attract the visitors but also to give regular user's a personalized and adaptive service. The main goal of the WUM is to understand, capture and model the browsing behavior of the user's while they interacting with the web.

II. LITERATURE SURVEY

Mathias Grey, Hatem Huddad [2] proposed a framework for a recommender system that from web log data, it will predicts the user's next request based on their behavior discovered. They had combined three web usage mining techniques i.e. association rules, frequent sequence and frequent generalized sequence to predict user's next request web page. All the results show after performing the experiments that frequent sequence gives better accuracy than association and frequent generalized sequence. The main disadvantage of association mining is that it can be conclude not only inherent use the notion of temporal distance but also frequent sequences cannot predict navigation patterns for data sets. Vedpriya Dongre, Jagdish Raikwal [3] proposed a system architecture using

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

web log for finding the hidden navigational patterns data. Mehradad Jalali, Narwat Mustapha, Md. Nasir Sulaiman, Ali Mannat [7] proposed an online and offline phase architecture. Both of these architecture works simultaneously. For improving the quality of the recommendation, the semantic knowledge about underlying domain can be used. Megha P. Jarkad ,Prof.Mansi, Bhonsle [8] proposed a system architecture which contain five steps using classification, clustering and background algorithm.. By using classification, clustering and backtracking algorithm not only improves the performance and decreases the time complexity of the proposed system. The author used graph partitioned clustering algorithm instead of calculating weights of the web pages direct frequencies can be used for prediction. Dilpreetkaur, A.P.sukhpreetkaur [9] proposed a system architecture using Fuzzy Clustering i.e. fuzzy c-means and kernelized fuzzy c-means algorithm. Authors concluded that KFCM is not only more robust than FCM but also creates better clusters for prediction.

III. WEB LOG FILES

The files that contain information about the user's activity on the websites are called web log files. These log files are created automatically. Each time the user visits the websites and requesting any file such as images, pages, videos etc. are being recorded in web log file. It has range 1kb to 100 mb and also has text format i.e. all log files are in the form of as line of text. It may consists of IP address, host, User ID, date, time, requesting URL, HTTP status, size of recoded file etc. It is one of the data source of web usage mining and also it is located in three different locations namely as web server, proxy server & web clients or client browser.

IV. INTRODUCTION TO ANT COLONY OPTIMIZATION, DB SCAN CLUSTERING ALGORITHM AND ASSOCIATION RULE MINING

Ant colony optimization is one of the algorithms which based on behavior of ants searching for food i.e. for finding the optimal paths. In other words, it is a technique which is used for computational problem for finding the good paths from the graph. Basically, just like ants wander randomly and after finding the food return to their colony by leaving pheromone trails for the other ants, if they finding foods so that they don't wander here and there. The idea of ant colony algorithm is to 'mimic the behavior' of the ants to find the good path from the graph. From this optimization, one can get not only best iterative cost but also true and false behavior of the customer's i.e. high prediction rate at low iterative cost. DBSCAN (density based spatial clustering of applications with noise) is one of the clustering algorithm to determine the clusters of arbitrary shape. Basically, it depends upon two 2 parameters i.eeps and minpts. By setting these parameters, formation of clusters would be formed having no noises. Besides its formation of clusters i.e. it takes less area to form clusters, it has some disadvantage that it is less sensitive to the parameters that is by decreasing or increasing the value of eps , we get more duplicate of data having more outliers. Whereas, association rule mining is used to discover those pages that are visited together. It is very common in grocery stores i.e. if a customer buys cheese and butter so what's the probability of him to buy bread also. It depends upon two factors i.e. support and confidence. As the database is very huge, the customers is only concern to those items whose threshold values are predefined i.e. minimum support and minimum confidence parameters.

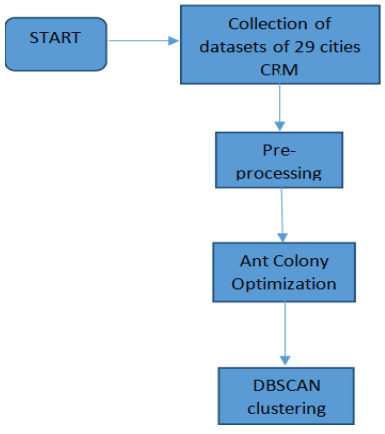
Support (XY) = Support Count of XY / Total number of transaction.

And Confidence can be calculated by; Confidence ($X_j Y$) = Support (XY)/ Support(X). If one can specify these parameters correctly, then get precise and accurate rules.

V. PROPOSED WORK

The proposed work for web usage mining is shown in figure below:

1. First of all, synthetic data sets of 29 cities CRM has been collected of the customers.
2. Then, the datasets are going under pre-processing phase where all unwanted entries are removed.
3. In this step, Ant Colony Optimization has been applied to determine the best possible paths from the graph i.e.to find the best iterative cost of the customers. Basically, after the preprocessed datasets are being going under cross mutation in various customer levels and through this optimization, one can conclude that the customer's true and false behavior.
4. In the clustering phase with the help of DBSCAN (Density Based Spatial Clustering Of Application With Noise) to reduce the outliers or noises and make the noises to be part of any clusters by adjusting values of eps and minpts.
5. At last in the next step, association rule mining is applied to make rules without any duplicity of data. Its main issue is that give irrelevant rules which leads reduction to less accuracy in prediction that is why one can apply clustering algorithm to make clusters of user's having similar behavior to get efficient and accurate rules for prediction of user's behavior.



DOI 10.17148/IJARCCE.2016.5512

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

VI. EXPERIMENTAL RESULTS

In the research, we take synthetic datasets of 29 cities of the customers then we apply preprocessing step on it and removing the unwanted entries.

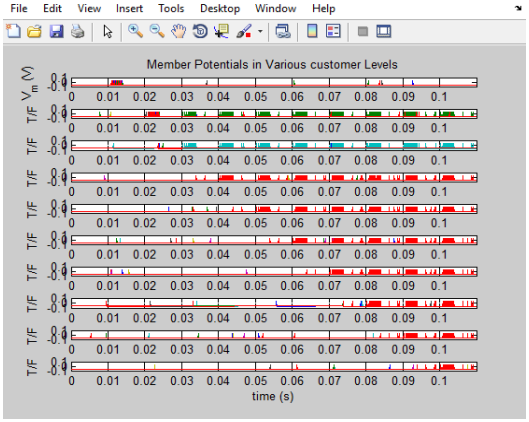


Figure 2. Cross over mutation will happen in various customer levels to find out best iterative cost and also it will do iterations at customer level for finding the test values.

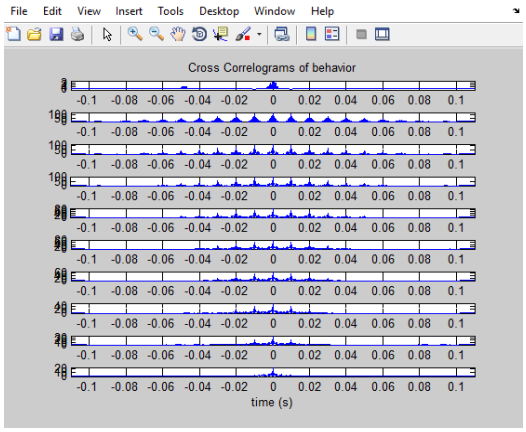


Figure 3. Through the test values, one can find cross correlograms of customer behavior which shows the true and false behavior of the customers.

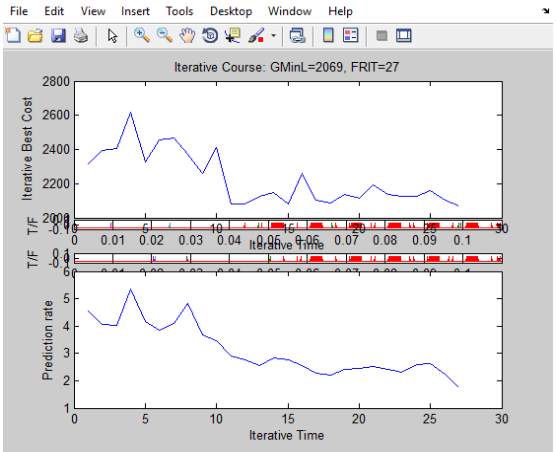


Figure 4.Graph between iterative time and prediction rate. Through this one can come to know that at the highest point of the prediction rate we get the iterative best cost at low point.

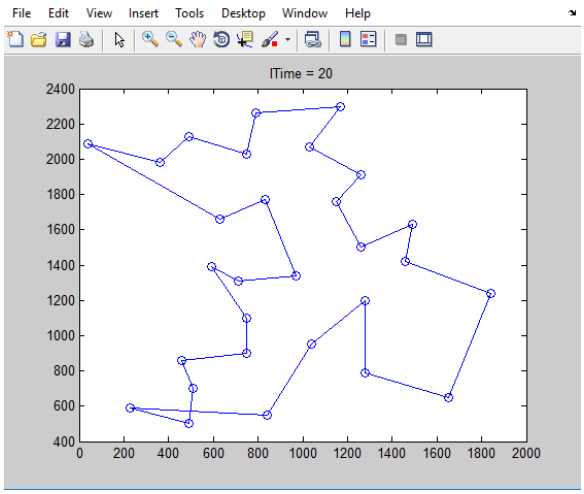


Figure 5.We plot the graph having 29 nodes and traverse the data by using the test values coming from the cross over mutation and optimization.

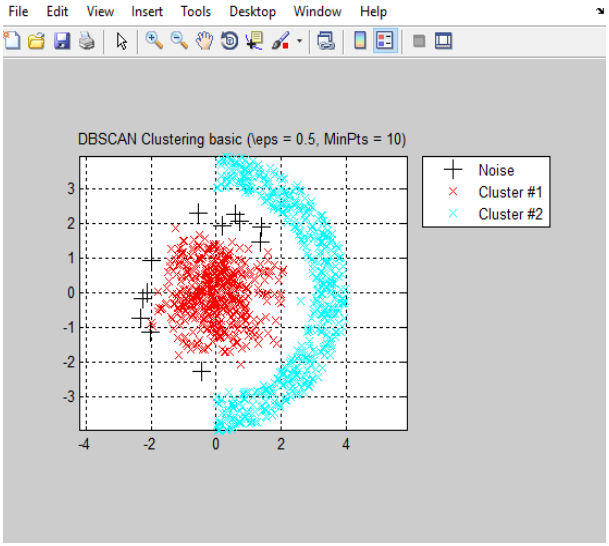


Figure 6.Clusters would be formed with some noisy data.

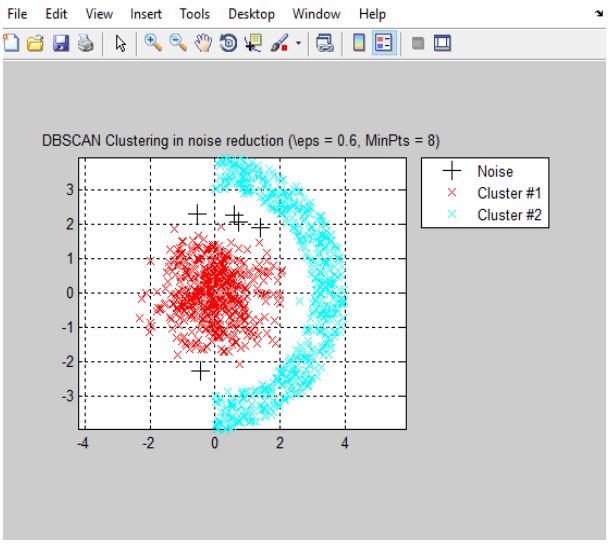


Figure 7. By increasing the eps value, clusters would be formed having less noise or should say that those noises would be belong to any of the cluster.

DOI 10.17148/IJARCCE.2016.5512

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

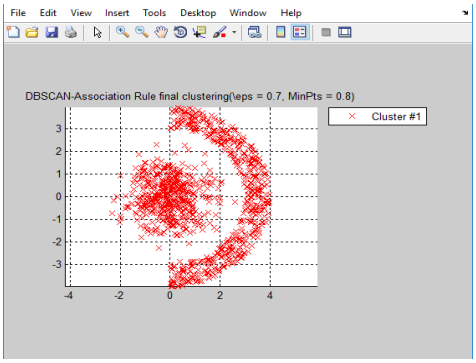


Figure 8. Applying Association rule on the clustered data without having outliers or noises.

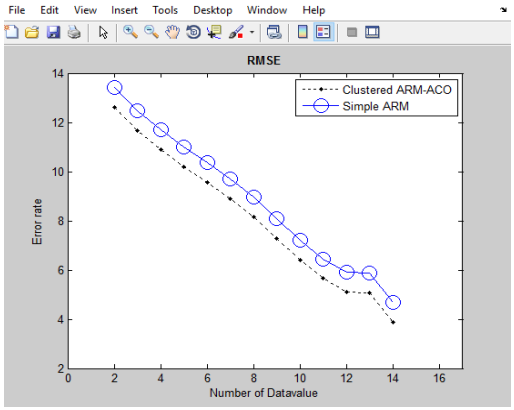


Figure 9. Graph between Simple ARM & Clustered ARM which showing that applying Association rules on clustered data, we get more accurate results having less rate as compared to apply it on preprocessed datasets.

SNO.	No.of data value	Error rate for Simple ARM	Clustered ARM
1.	2	14	13
2.	4	12	11
3.	6	10	9
4.	8	9	8
5.	10	7	6
6.	12	6	5

Table 1:Differentiate between Simple ARM & Clustered ARM.

VII. CONCLUSION AND FUTURE WORK

There are lots of research have been done for prediction of user's browsing behavior. Due to vast amount of data available on the Internet, which would be turn the result in delay in response of the user. To overcome this issue prediction will take a step forward. All the e- commerce sites i.e. jabong , myntra, ebay, snap deal etc. used for more understanding what their customer's demands and requests to increase the productivity .With the help of ACO i.e ant colony optimization, one can not only get the best iterative cost but also find the true and false behavior of the user. Through the results, one can conclude that applying association rules on the clustered data gives result having with less error rate as compared to the simple ARM which results into precise and accurate results. The performance of the proposed work is more accurate and efficient with less error rate in predicting of user's browsing behavior. In future, the work can be extended by applying different methods and methodology to increase the performance of the architecture and applied it on large datasets.

REFERNCES

[1]. Vrishali P. Sonavane,"Study And Implementation Of LCS Algorithm For Web Mining", International Journal of Computer Science Issues, Vol. 9, Issue 2, No 3, March 2012.

[2]. Mathias Gery, Hatem Haddad ,"Evaluation of Web Usage Mining Approaches for User's Next Request Prediction" WIDM"03 Proceedings of the 5th ACM international workshop on web information and data management p.74-81, November 7-8,2003.

[3]. VedpriyaDongre, JagdishRaikwal ,"An improved user browsing behavior prediction using web log analysis", International Journal of Advanced Research in Computer Engineering and technology (IJARCET), Vol. 4, Issue 5, , May 2015.

[4]. R.Khanchana and M. Punithavalli, "Web Usage Mining for Predicting Users' Browsing Behaviors by using FPCM Clustering", IACSIT International Journal of Engineering and Technology, Vol. 3, No. 5, October 2011.

[5]. M. Jalali, N. Mustapha et al ," WebPUM: A Web-based recommendation system to predict user future movements", in international journal Expert Systems with Applications 37 (2010) 6201–6212 .

[6]. YanRong Zhang1 and ZhiJie Zhao," Study on Consumer Behavior Predict in E-commerce Based on Multi-Agent ",International Journal of u- and e- Service, Science and Technology Vol. 7, No. 6 (2014), pp. 403-412.

[7]. MehrdadJalali,Norwati Mustapha, Ali Mamat , Md. Nasir B Sulaiman "A new classification model for online predicting users" future movements", Information Technology, 2008. ITSim 2008. International Symposium on Information Technology, p.1-7, 26-28 Aug. 2008.

[8]. Megha P. Jarkad, Prof. Mansi, Bhonsle," Improved Web Prediction Algorithm Using Web Log Data", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 5, May 2015.

[9]. Alexandros Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos "Effective prediction of web-user accesses: A data mining approach," in Proc. Of the Workshop WEBKDD, 2001.

[10]. A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications, Volume 8– No.11, October 2010.

[11]. Dilpreet Kaur, A.P. Sukhpreet Kaur," User Future Request Prediction Using KFCM in Web Usage Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013.

[12]. Yogesh Rajaram Bhalerao1, Prof. P. P. Rokade," User Navigation Pattern Prediction using Statistical Classifier and Modern Techniques", International Journal of Emerging Technology and Advanced Engineering ,Volume 5, Issue 5, May 2015.

[13]. R.Khanchana and M. Punithavalli.," Web Usage Mining for Predicting Users' Browsing Behaviors by using FPCM Clustering", IACSIT International Journal of Engineering and Technology, Vol. 3, No. 5, October 2011.

[14]. Ujwala Patil, Sachin Pardeshi," A Survey on User Future Request Prediction: Web Usage Mining ",International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 3, March 2012.

[15]. Ketul B. Patel, Jignesh A. Chauhan, Jigar D. Patel," Web Mining in E-Commerce: Pattern Discovery, Issues and Applications", International Journal of P2P Network Trends and Technology- Volume1Issue3- 2011.