Indonesian Journal of Electrical Engineering and Computer Science
Vol. 11, No. 1, July 2018, pp. 333~340
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v11.i1.pp333-340

Analyzing and Predicting User Navigation Pattern from Weblogs using Modified Classification Algorithm

P.G.Om Prakash, A.Jaya
Department of Computer Science and Engineering, B.S.Abdur Rahman Cresent Institute of Science & Technology, Chennai 600048, India
Department of Computer Applications, B.S.Abdur Rahman Cresent Institute of Science & Technology, Chennai 600048, India

Article Info

Article history:

Received Sep 12, 2017
Revised Oct 16, 2017
Accepted Dec 21, 2017

Keywords:

Data mining
Prediction accuracy
Traversal pattern
User behavior
User navigation
Web mining

ABSTRACT

A Weblogs contains the history of User Navigation Pattern while user accessing the websites. The user navigation pattern can be analyzed based on the previous user navigation that is stored in weblog. The weblog comprises of various entries like IP address, status code and number of bytes transferred, categories and time stamp. The user interest can be classified based on categories and attributes and it is helpful in identifying user behavior. The aim of the research is to identifying the interested user behavior and not interested user behavior based on classification. The process of identifying user interest, it consists of Modified Span Algorithm and Personalization Algorithm based on the classification algorithm user prediction can be analyzed. The research work explores to analyze user prediction behavior based on user personalization that is captured from weblogs.

Corresponding Author:

P.G.Om Prakash,
Department of Computer Science and Engineering,
B.S.Abdur Rahman Cresent Institute of Science & Technology,
Chennai 600048, India.
Email: mmail2004@gmail.com

1. INTRODUCTION

In the current situation, the lot of products are available through online shopping. More number of users attracted towards the internet, so the users are increasing to access the websites day by day. The online shopping has the enormous growth. Identify the interested customer and not-interested customer is a difficult task, through weblog the user interest can be classified. The weblog consists of history of information while the user accessing the websites. The prediction of user behaviour can be identified only through logs. The weblog contain unstructured format from that user behaviour is analyzed.

The weblog consists of various entries like IP address, date, time, request method, protocol, categories and number of bytes transmitted and status code. The user interest can be classified based on the categories and attributes of a product.

Analyzing the user behaviour of user is helpful in understanding demands for manufacturing industry. Web mining algorithm helps to analyze the useful information from web log. Based on the kind of information, web mining can be classified as web structure mining (WSM), web usage mining (WUM) and web content mining (WCM). Web content mining is to analyze the contents from web documents such as audio, video, image and text etc. Web structure mining is to analyze the link of websites. Web usage mining is to analyze the user browsing behaviour activity.

The WUM consists of data from weblog. Whenever the user access the websites that information will stored as logs. The log contains series of user transactions that are frequently updated whenever the user will access the website. The prediction of user behaviour can be identified only through weblogs. The weblog contains unstructured format, so convert to raw weblog to processed weblog using data preprocessing, the data preprocessing includes data cleaning, user and session identification. The User Classification System consists of two phases Modified Spanning algorithm and Personalization algorithm to identify user behavior in short time.

The process of identifying user behavior is based on previous user navigation pattern that is taken from weblogs; through weblog web traversal pattern is discovered. The clustering is to group the similar web traversal pattern for classification. The classifier classifies the sequence as frequent sequence and infrequent sequence from that user behavior is analyzed.

Every organization will analyze the interested and not interested user, the manufacturing industry is interested in predicting the user preferences of each customers. The industry will identify the user preference based on classification algorithm. The various phases of classification are Modified Spanning Algorithm and Personalization Algorithm from that user prediction accuracy is increased.

The part of the paper contains, Section 2.0 Related Work, Section 3.0 Architecture Diagram, and Section 4.0 Implementation Results. Section 5.0 concludes the paper by giving to future directions of research in this area.

2. RELATED WORK

WUM is to predict the user behavior pattern based on navigation preferences in the website. Wangshu Liu et al. [1] suggested that two stage preprocessing eliminates irrelevant and redundant datasets, whereas ranking removes irrelevant and features and clustering remove redundant data. Xin ruan et al. [2] classified online social networks (OSN) in to extroversive behavior and introversive behavior. Extroversive behaviors directly reflect how a user interacts with online friends, introversive behavior gather and consume social information. Guoshuai Zhao et al. [3] suggested user rating prediction service approach consists of user personal interest. Ruili Geng et al. [4] showed user navigation pattern is to identify the actual usage pattern and anticipated usage patterns. The actual usage patterns can be extracted from web logs to identify user sessions and user transactions. The anticipated usage contains user path and user session. Surbhi Huriaet al. [5] suggested that Back Navigation Approach utilizes forward and backward path to identify frequent pattern. Indre Zliobaite et al [6] suggested adaptive preprocessing mechanisms for data streaming, due to data streaming the accuracy of prediction will change.

Chin teng Lin et al. [7] suggested that support vector based fuzzy neural network is a pattern classification algorithm to identify user buying behavior that helps to improve the accuracy of classification. Dezhi wu et al. [8] suggested that individual user behavior can be analyzed based on user personalization, using personalization the customization was under user preference. Zhang et al. [9] suggested that the user personalization focus is a valid for text information retrieval. Based on user behavior the personalization recommendation is analyzed. Santhosh K. et al. [10] suggested that the degree of personalization to organize the site structure of different user groups. Shen Hui-zhangl et al. [11] suggested that web personalization is based on dynamic clustering. It is to cluster the similar group of data where as markov model utilizes the previous results from that user web personalization is analyzed. LI Weil et al. [12] suggested that web personalization is to cluster and classify the data based on user profiles. Cluster is to grouping the discovering users from access log, classifier is to classify the log based on user preference. Gang Fang et al. [13] suggested that double algorithm; it will work based on sequence number by mining the session patterns to increase the efficiency that will generate the frequent item set.

Santral A.K. et al. [14] suggested that Naïve Bayesian algorithm classifies the log in to relevant and irrelevant item sets that will helps to classify interested user and not interested user behavior. Suneetha K.R et al. [15] suggested that Decision trees are useful for classifying the log data that helps to predict the user prediction. Decision trees is one of the way to represent knowledge representation, this results shows the improvement in time and memory utilization.

Mobasher et al. [16] suggested that Web personalize needs dynamic recommendation from a list of links to the users. The web personalizer shows the usage of web data and also shows the hyperlinks of the website. Weblogs are preprocessed from raw log, from that navigation path is analyzed and generated a sequence. The similar sequence is grouped as cluster, from that clustering the log can be classified in to frequent sequence, semi-frequent sequence and infrequent sequences will helps to understand the user's behavior.

Joachims et al. [17] and Pazzani et al. [18] showed the examples of visitors suggested links of individual user. The user access the server based on his interest. The browser will follow the query based on user request, if more number of users follow the same request then it is frequent viewed items.

Fu and Perkowitz et al. [19] showed the current and past sessions whenever the user visits the web server. The algorithm suggest that first top m pages based on user's most visited current sessions, the system should analyze the maxpath and minpath of user navigated sessions, then only the visitor should easily identify the user behavior. The algorithm should select the best path from the list of frequent path.

The user prediction can be analyzed based on the previous user behavior that is stored in the weblog. The following techniques help to identify the user behavior such as data preprocessing, pattern discovery and classification.

Cadez et al. [20] suggested the markov model to analyze user behavior, the markov model analyze various results based on previous navigated user and reduces the number of path that is helpful in identifying user behavior. Om Prakash et al. [21] suggested that user log can be classified as Successful and Unsuccessful log. The success log can be classified based on the user interest from that user behavior can be analyzed. Addanki Ramya et al. [22] showed that cluster the similar path based on back propagation approach, it contains both forward and backword approach. Lutfi Fanani et al. [23] suggested reducing the waiting time for passengers in peak time by normal distribution of random travel time. Yahya zare khafri et al. [24] showed a customized tracking system is for autonomous navigation for high speed vehicles. The proposed algorithm helps to reduce the spped by light weight navigational algorithm.

2.1 Problem definition

In the current situation more number of users attracted towards the internet, so there is growth of users accessing the internet are increases day by day and reduces the shopping time, so data size will increase. Identifying the interested user and not interested user is difficult, based on weblog user interest can be classified. The weblog consists of history of information while the user accessing the websites. The prediction of user behaviour is a difficult task.

3. FRAMEWORK FOR USER NAVIGATION SYSTEM

The framework for navigator system is shown in Figure 1.

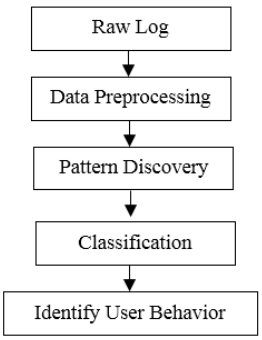

Figure 1. Framework for Analyzing User Behavior

3.1. Raw Log

While user accessing the web server, the user transaction is generated in web server as log, it contains unstructured format.The weblog consists of various entries like IP address, date, time, request method, protocol, categories and number of bytes transmitted and status code. The above attributes helps to identify user navigation and the pattern can be classified after pre-processing.

3.2. Data Preprocessing

The weblog contains unstructured format of user navigation, the conversion of unstructured format to structured format by data preprocessing techniques.

1) Data cleaning
2) User identification
3) Session Identification

3.2.1. Data Cleaning

Let us consider the portal of Electronic prediction sale. The users were browsing the information with their own interest. These logs have been recorded from the period of 10-03-2016 18:23:41 to 03-02-2017 17:15:10, logs details were acquired and preprocessed for further navigation prediction. The log file contains 20400 records, in that each record having the status code. The status code decides the success and failure of webpage. The status code in between 200 to 400 is valid. The data cleaning process removes the records such as gif, jpeg etc., that is easily to identify user navigation process. After the cleaning process in log 1800 records are obtained.

E.g.: Figure 2 shows "192.168.1.103- [10-03-2016 18:23:41] "http://localhost:8080/ws/mob.jsp" 200 441" the stats code 200 is considered.

192.168.2.11	14-03-2016 11:26:23	http://localhost:8080/ws/mob2.jsp	den
192.168.2.11	14-03-2016 11:26:50	http://localhost:8080/ws/mob2.jsp	den
192.168.2.11	14-03-2016 11:33:41	http://localhost:8080/ws/mob.jsp	den
192.168.2.11	14-03-2016 11:33:46	http://localhost:8080/ws/mob2.jsp	den
192.168.2.11	14-03-2016 11:35:27	http://localhost:8080/ws/mob2.jsp	den
192.168.2.11	14-03-2016 11:37:15	http://localhost:8080/ws/mob2.jsp	den
192.168.2.11	14-03-2016 11:39:07	http://localhost:8080/ws/mob2.jsp	den
192.168.2.11	14-03-2016 11:45:58	http://localhost:8080/ws/product.jsp	den
192.168.2.11	14-03-2016 11:46:19	http://localhost:8080/ws/mob.jsp	den
192.168.2.11	14-03-2016 11:46:25	http://localhost:8080/ws/mob2.jsp	den
192.168.2.11	14-03-2016 12:19:15	http://localhost:8080/ws/single.jsp	den
192.168.2.11	14-03-2016 12:19:22	http://localhost:8080/ws/single.jsp	den
192.168.2.11	14-03-2016 12:26:23	http://localhost:8080/ws/single.jsp	den
192.168.2.11	14-03-2016 12:36:37	http://localhost:8080/ws/LGO	guest
192.168.2.11	14-03-2016 12:37:48	http://localhost:8080/ws/index.jsp	guest
192.168.2.11	14-03-2016 12:40:57	http://localhost:8080/ws/index.jsp	guest
192.168.2.11	14-03-2016 12:42:15	http://localhost:8080/ws/index.jsp	guest
192.168.2.11	14-03-2016 12:42:36	http://localhost:8080/ws/mob2.jsp	guest
192.168.2.11	14-03-2016 12:42:41	http://localhost:8080/ws/single.jsp	guest
192.168.2.11	14-03-2016 12:47:08	http://localhost:8080/ws/index.jsp	guest
192.168.2.11	14-03-2016 12:47:13	http://localhost:8080/ws/index1.jsp	guest
192.168.2.11	14-03-2016 13:00:07	http://localhost:8080/ws/index.jsp	guest
192.168.2.11	14-03-2016 13:00:15	http://localhost:8080/ws/index1.jsp	guest
192.168.2.11	14-03-2016 13:01:30	http://localhost:8080/ws/index.jsp	guest
192.168.2.11	14-03-2016 13:01:34	http://localhost:8080/ws/index1.jsp	guest
192.168.2.11	14-03-2016 13:01:53	http://localhost:8080/ws/US	sam
192.168.2.11	14-03-2016 13:02:01	http://localhost:8080/ws/product.jsp	sam
192.168.2.11	14-03-2016 13:02:06	http://localhost:8080/ws/mob.jsp	sam

Figure 2. Acquisition of Weblog

3.2.2. User Identification

The user can be classified from the obtained records in the data cleaning. The user identification is based on user IP address, user name, user request time, user requested URL, date & time, server IP, bytes sent and receive, server name, service instance, HTTP request and status code etc. The Internet Information Service (IIS) log file format from that user is identified.

E.g.: from figure 2 shows the user IP address 192.168.1.103 is identified from log.

3.2.3. Session Identification

After the data cleaning and user identification, the navigation pattern mining classifies the each page of user browse consider as a session. The important process of session identification is to cluster the session.

E.g.: from figure 2 shows the time stamp session, for e.g. 11:26:41 to 11:37:46 is considered as one session.

3.3. Framework for Pattern Discovery

After the preprocessing of server log file, data mining techniques are applied. The Pattern Discovery consists of Path Analysis, Sequential Pattern and Clustering. The sequence of pattern is generated based on Path analysis. From that Sequential pattern the sub sequence pattern is generated by maximum forward algorithm. The maximum forward algorithm considers both forward and backward reference from that sequence of log can be clustered.

The Pattern discovery is a clustering technique to cluster the browsing pattern of users.

3.4. User Classification System

Figure 3, shows the framework for analysis of user classification pattern from clustering. It consists of two phases

- Modified Spanning Algorithm
- Personalization Algorithm

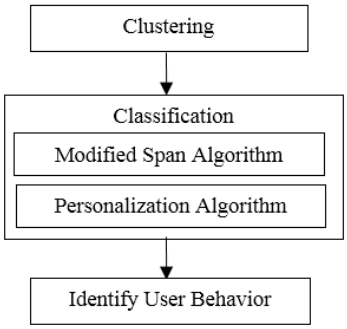

Figure 3. Framework for User Classification

3.4.1. Modified Spanning Algorithm

Modified Spanning algorithm classifies the user preference based on clustering. If the clustering size is above the threshold point then it is interested user. If the cluster size is below the threshold value then it is not interested user, from the clustering threshold value user behavior is analyzed.

step 1 : add the new sequence in F'
step 2 :if F' in IU_SEQ then goto step 3 else step 9
step 3 :cmp [F' in Seq_DB_His]
step 4 :if F' is Greaterthan Succ_count[i] Seq_DB_His
step 5 :then it is "Interested User"
step 6 : IU_SEQ=F'
step 7:else "Not Interested User"
step 8:NIU_SEQ=F'
step 9:if F' in NIU_SEQ then goto step 3
step 10:Seq_DB_His = F'

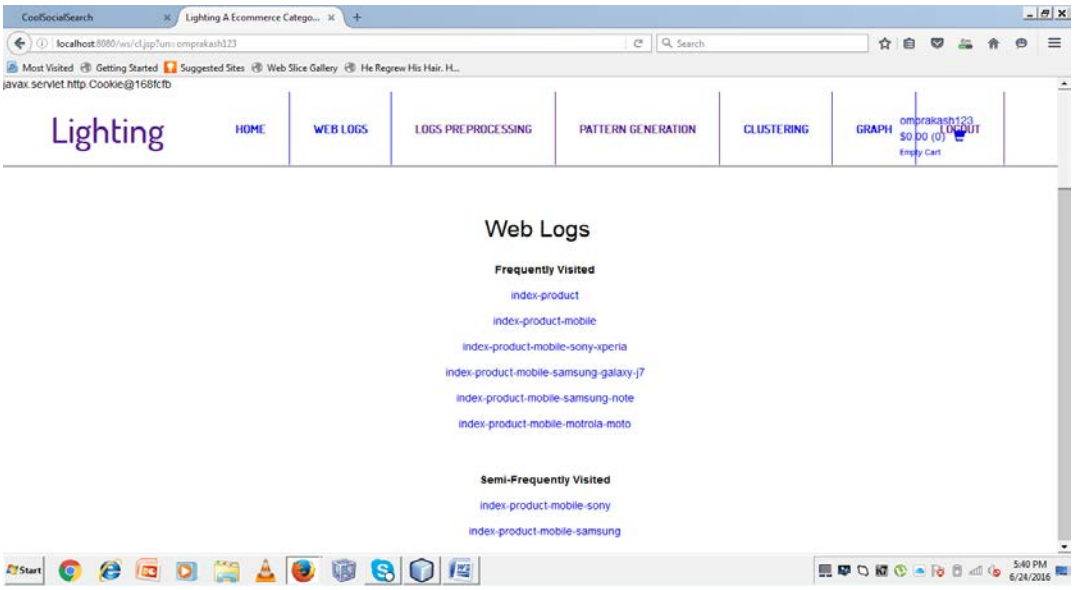

Figure 4. Modified Spanning Algorithm

3.4.2. User Personalization

The user personalization algorithm determines the success count of the user transaction, if the user count is greater than the Seq_DB then the user has interested user, otherwise the user has not interested user.

step 1 : get new sequence F' from user
step 2 :cmp [U_ID of F' in Seq_DB_His]
step 3 :if F' of U_ID is Greaterthan Succ_count[i]
step 4 :then it is "Interested User"
step 5 : IU_SEQ=F'
step 6: else "Not Interested User"
step 7: NIU_SEQ=F'
step 8: Seq_DB_His = F'

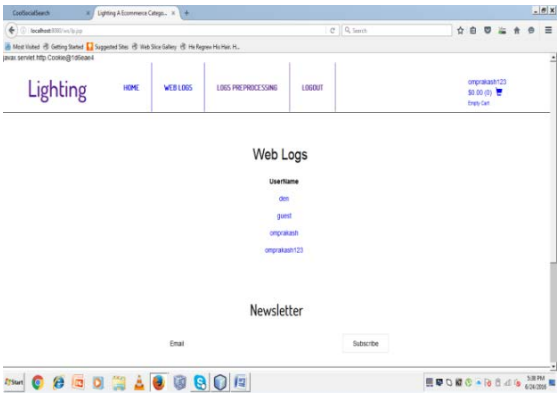

Figure.5 User Personalization

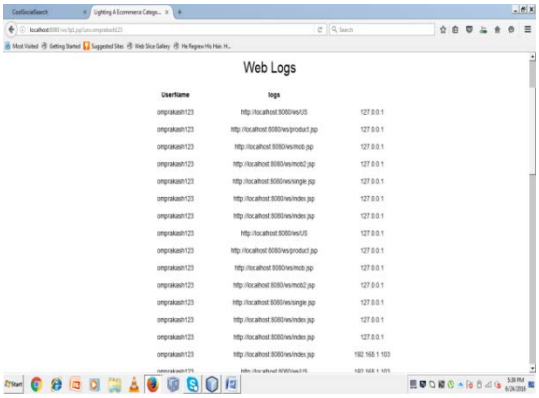

Figure.6 User Navigation of Weblog

4. IMPLEMENTATION

The proposed architecture is implemented by using Modified Span algorithm and User Personalization algorithm, the logs are recorded from the period 10-03-206 to 03-02-2017, it contains 20400 records. After preprocessing the log contains 1800 records, the logs are preprocessed based on status code, if the status code in-between 200 to 400, it will be success response, and otherwise it will be failure response. The knowledge base has 1800 instance in log, in that 25 attributes generated. The prediction accuracy is calculated by total number of instances in log by total number of user and attributes.
Figure 7, 8, 9, 10 & 11 shows the sample output screen for Search Result.

Figure.7 Comparison of Mobile Configuration

The Figure 7, shows the comparison of mobile configuration of various manufacturing industries, the various features of the product are Camera, Processor, memory, weight, rating and cost. Through online the user can aware of each and every features of the products. The user give the feedback of the product

purchase, from the above feedback the product rating will incresae, that will increase the sale of the product. The rating describe the product quality and reliability of mobiles, from that user behavior is analysed.

Figure.8 Prediction Statistics of Samsung & Sony

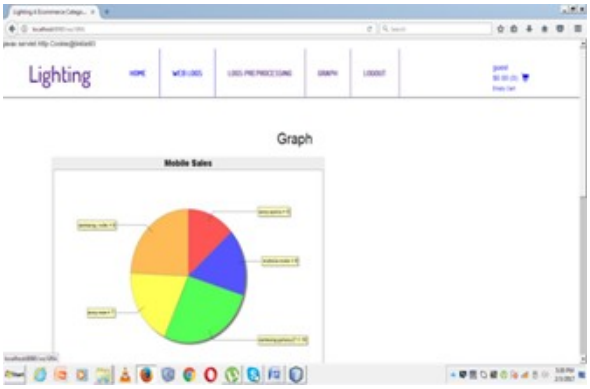

Figure.9 Prediction Statistics of Mobile Sales

The Figure 8 shows the prediction statistics of the products. The prediction statistics of the product is to classify feedback and sale. Modified spanning is to classify the interested and not interested user. The personalization is to classify the individual user interest. Modified spanning algorithm and User personalization algorithm is to increase the prediction accuracy.

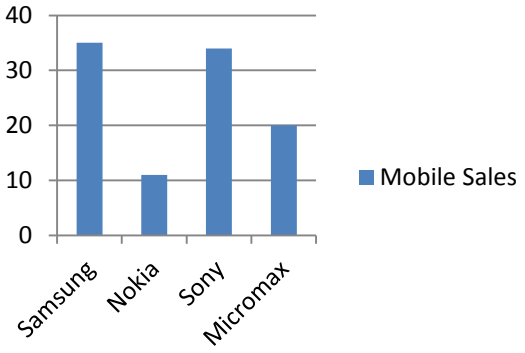

Figure.10. Prediction Statistics of User

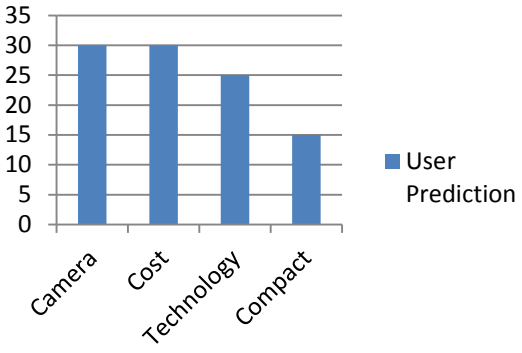

Figure 11. User Prediction

Figure 11, shows the Prediction Statistics of User by mobile sales, the major categories of Mobile prediction is taken from Figure 9. It shows the percentage of mobile sales that is helpful to understanding the demand of manufacture industry, from that the sales can be improved.

The Figure 11, shows the prediction of Mobile sales will decide by various class of people, by comparing the major attributes such as Camera, Color, Cost, Technology and Compact from that prediction level were generated. The major four attributes are calculated by ratio between the number of user instance generated by the system and the total number of attributes. The user prediction can be generated using the majority of attributes; with the help of above attributes user prediction is analyzed.

5. CONCLUSION

Web service systems used by weblog based model is low-cost. The proposed architecture is implemented by using Modified Span algorithm and User Personalization algorithm, the system will show the mobile purchase based on user interest and improves prediction accuracy, it utilizes the previous user navigation results and classify the frequent item set, in-frequent item set with user preference and reduces the

number of path from the weblogs. The system will concentrate to extend with Fuzzy algorithm to identify the buying prediction behavior.

References

[1] Wangshu Liu et al,"Empirical studies of two stage data preprocessing approach for software fault prediction" *IEEE Transactions on Reliability*, Vol. 65, March 2016.
[2] Xin Ruan et al, "Profiling online social behaviors for compromised account detection", *IEEE transactions on information forensics and security*, Vol. 11, January 2016.
[3] Guoshuai Zhao et al, "User service rating prediction by exploring social users' rating behaviors", *IEEE transactions on multimedia*, Vol. 18, March 2016.
[4] Ruili Geng et al, "Improving web navigation usability by comparing actual and anticipated usage", *IEEE transactions on human-machine systems*, Vol. 45, February 2015.
[5] Surbhi Huria et al, "Implementation of dynamic association rule mining using back navigation approach", fifth international conference on communication systems and network technologies 2015.
[6] Indr E Zliobaite et al, "Adaptive preprocessing for streaming data", *IEEE transactions on knowledge and data engineering*, Vol. 26, February 2014.
[7] Chin-Teng Lin et al, "Support vector based fuzzy neural network for pattern classification", *IEEE transactions on fuzzy systems*, Vol. 14, February 2006.
[8] Dezhi et al, "A framework for classifying personalization scheme used on e-commerce websites", *IEEE Proceedings of 36 Hawaii international conference on system sciences* 2002.
[9] Zhang et al, "Enabling personalization recommendation with weighted for text information retrieval based on user focus", *IEEE proceedings of the international conference on information technology*, 2004.
[10] Santosh et al, "Adaptive neural network clustering of web users", Published By *the IEEE Computer Society*, 2004.
[11] Shen Hui-Zhangi et al, "A Web Mining Model For Real-Time Webpage Personalization", *Icmse 06*. 2006.
[12] Li Wei1 et al," Clustering of web users based on competitive agglomeration", 2008 *IEEE*.
[13] Gang Fang et al, "A double algorithm of web usage mining based on sequence number ", 2009 *IEEE*.
[14] A. K. Santra et al,"Classification of web log data to identify interested users using naïve bayesian classification", *International Journal of Computer Science Issues* 2012.
[15] K. R. Suneetha et al, "Classification of web log data to identify interested users using decision trees".
[16] B. Mobasher et al, "Automatic personalization based on web usage mining", *Communications of ACM,* Vol. 43, 2000.
[17] T. Joachims et al,"A tour guide for the world wide web", *15th International conference*, 1997.
[18] M.J. Pazzani et al,"Adaptive web site agent", *3rd international conference on autonomous agents*, 1999.
[19] M. Perkowitz et al,"Adaptive Websites an AI challenge". *15th International conference*, 1997.
[20] I. V. Cadez et al,"Visualization of navigation patterns on a web site using model based clustering" *Proc. 6 International conference on knowledge discovery and data mining*, 2000.
[21] Prakash, P. O., and A. Jaya. "Analyzing and predicting user behavior pattern from weblogs", *International Journal of Applied Engineering Research* 11.9 (2016): 6278-6283.
[22] Addanki Ramya et al, "Preprocessing and Unsupervised Approach For Web Usage Mining". *International Journal of Social Networking and Virtual Communities*, Vol 1 No 2, 2012.
[23] Lutfi Fanani et al, "Bus Arrival Prediction – to Ensure Users not to Miss the Bus". *International Journal of Electrical and Computer Engineering (IJECE)* Vol 5 No 2, 2015 pages 333-339.
[24] Yahya Zare Khafri et al, "Improved Line Tracking System for Autonomous Navigation of High-Speed Vehicle" , *IAES International Journal of Robotics and Automation (IJRA)*, Vol 1 No 3, 2012 pages 163-174.