Accuracy in Data Extraction and Predicting Performance of Mining Algorithms Using Weka

Prathibha. G[1], Sankara Subbulakshmi. A[1] and Kavitha. R[2]
[1]Student, [2]AP, Department of IT,
Parisutham Institute of Technology and Science, Thanjavur, Tamilnadu

***Abstract* -- Data mining finds valuable information hidden in large volume of data. The basic principles of data mining is to analyze the data from different angle, categorize it and finally to summarize it. In today's world data mining have increasingly become very interesting and popular in terms of all application. WEKA is a data mining tool allows user to analyze the data . It introduces the key principle of data pre-processing, classification , clustering , etc. It provides the facility to classify the data through various algorithms . Here we are going to predict the performances of data mining algorithms by using sample dataset. We are going to create a framework , to estimate the preferences of users over the articles based on browsing behaviour using hybrid collaborative filtering. The framework produced result will be utilized by WEKA tool for generating the preference ratios of different users.**

Keywords: Data mining; data preprocessing, classification, cluster analysis, Weka tool etc.

I INTRODUCTION

Data mining is the concept of analyzing the data through various algorithm such as data preprocessing , pattern recognition, clustering , classification, association rule mining etc.
These algorithms have been developed and implemented to extract information and discover knowledge patterns that may be useful for decision support. Data mining also known as KDD(knowledge discovery in databases). Normally, all data mining algorithms have been processed by developing code manually and then the datasets are applied. In this case, there may be a chance of human or programmer can make a mistake .When the data are applied either in small or medium volume, it is easy to access as many times based on users requirements and also possible to achieve efficiency. But in case of large volume data it is difficult to achieve that much efficiency while processing and also it is not applicable for processing more time to satisfy user needs. In this paper , we will discuss about performance of algorithms which is already inbuilt in the tool by using sample dataset. Based on the efficiency, we are going to choose one algorithm for processing the data generated from the framework which we are creating for achieving effective browsing . The data processed in order to predict the preference ratio of different users over the articles that published in the internet.

This paper is organized as followings. In Chapter 2, introduction to WEKA are referred. Chapter 3 introduces the background and approaches of the research. Chapter 4 describes the related works . Chapter 5 and Chapter 6 describes the relationship between the interest of a user on an article and the behavior of the user for browsing it. The evaluation method and experiments on the value of article estimation are shown in Chapter 7. Finally, a summary of this paper and future works are shown in Chapter 8.

II INTRODUCTION TO WEKA

The aim of this section is to give a brief description of algorithms inbuilt in WEKA, overview of processing of default dataset and introduction of WEKA.
WEKA consists of Explorer, Experimenter, Knowledge Flow, Simple Command line Interface, Java Interface.

a) Explorer:

WEKA is a main graphical user interface. Each of the major weka packages are Filters, Classifiers, Clusterers, Associations, and Attribute Selection is represented along with a Visualization tools. Refer Fig.1. the tool which we are going to use.

Fig. 1. WEKA tool

b) Experimenter:

Comparing different learning algorithms on different datasets with various parameter setting and analyzing the performance statistics. Experimenter makes it easy to compare the performance of different learning schemes and for classification and regression problems. Results can be written into file or database. Evaluation options are cross-validation, learning curve, hold-out. It can also iterate over different parameter settings. It has the major significance of testing process to be built in the tool itself.

c) Knowledge Flow:

The Knowledge Flow provides an alternative to the Explorer as a graphical front end to Weka's core algorithms. The Knowledge Flow is a work in progress so some of the functionality from the Explorer is not yet available.
New graphical user interface for WEKA are as follows:
• Java-Beans-based interface for setting up and running machine learning experiments
• Data sources, classifiers, etc. are beans and can be connected graphically
• Data "flows" through components: e.g.,"data source" -> "filter" -> "classifier" -> "evaluator"
• Layouts can be saved and loaded again later

d) Simple Command Line Interface:

All implementations of the algorithms have a uniform command line interface.

1. Java Interface:

WEKA is fully java based data mining tool. All algorithms that present in this tool have been developed using java codes.

```
Instances data = new Instances( "data.arff"); // loading data
data.setClassIndex(position); // setting class attribute
Remove remove = new Remove(); // new instance of filter
remove.setOptions("-R"); //set options
remove.setInputFormat(data); // to inform filter about dataset
Instances newData = Filter.useFilter(data, remove); // apply filter
J48 tree = new J48(); // new instance of tree
tree.setOptions("-U"); // set the options
tree.buildClassifier(data); // build classifier // using 10 times 10-fold cross-validation. Evaluation eval=new Evaluation(newData); eval.crossValidateModel(tree, newData, 10, newData.getRandomNumberGenerator(1));
Instances unlabeled = new Instances( "unlabeled.arff" ); // unlabeled data unlabeled.setClassIndex(position); // set class attribute
Instances labeled = new Instances(unlabeled); // create copy // label instances
for (int i = 0; i < unlabeled.numInstances(); i++)
{
clsLabel = tree.classifyInstance(unlabeled.instance(i));
labeled.instance(i).setClassValue(clsLabel);
}
```

A. DATA PREPROCESSING

Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary. Data can also be read from a URL or from an SQL database (using JDBC). Pre-processing tools in WEKA are called "filters". WEKA contains filters for: Discretization, normalization, resampling, attribute selection, transforming and combining attributes. Data that given to the weka tool may be unstructured or no quality data so data pre-processing is important. In every field ,the data collection is the important factor but if the information is irrelevant then the huge problem may occur. Those problems are missing values, impossible data combination, out of range values. Due to these problems, it may produce unexpected fault result. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

B. CLASSIFICATION

Classifiers in WEKA are models for predicting nominal or numeric quantities. Implemented learning schemes include: Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets. Classification is the process to identify the set of models that developed using datasets and it involves in the process of differentiating the data classes and concepts.
The process involves following steps:

a. Create training dataset.
b. Identify class attributes and classes.
c. Identify useful attributes for classification
d. Learn a model using training example in training set.
e. Use the model to classify the unknown data samples.

C. CLUSTERING

WEKA contains "clusterers" for finding groups of similar instances in a dataset. Implemented schemes are: *k*-Means, EM, Cobweb, *X*-means, FarthestFirst. Clusters can be visualized and compared to"true" clusters (if given). Evaluation is based on loglikelihood, if clustering scheme produces a probability distribution.

D. ASSOCIATION RULEMINING:

WEKA contains an implementation of the Apriori algorithm for learning association rules. It can work only with discrete datasets. It can identify statistical dependencies between groups of attributes. Apriori can

compute all rules that have given minimum support and exceed a given confidence.

E. ATTRIBUTE SELECTION:

Panel that can be used to investigate which (subsets of) attributes are the most predictive ones. Attribute selection methods contain two parts:
– A search method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
– An evaluation method: correlation-based, wrapper, information gain, chi-squared. Very flexible: WEKA allows (almost) arbitrary combinations of these two methods.

F. VISUALIZATION:

Visualization is very useful in practice: e.g. helps to determine difficulty of the learning problem. WEKA can visualize single attributes (1-d) and pairs of attributes (2-d). To do rotating 3-d visualizations (Xgobi-style) contains Color-coded class values. "Jitter" option is to deal with nominal attributes (and to detect "hidden" data points). "Zoom-in" function is involved.

III EXISTING SYSTEM:

In recent years , the news websites like abc.com, cnn.com , bbc.com etc. are growing in popular. There are huge contents that delivered in websites over several categories includes education, political , sports, economic, cinema, etc. Therefore various types of news are delivered in real time as a stream. It is not easy to browse all the content of the news because users would sometimes skip the valuable contents. Under this background , the processing of big stream data to discover and use contents effectively attracted more and more attention. At present, for obtaining necessary information from big data, the keyword-based search engine method is used to provide the information estimated valuable and interesting to users. In internet the huge contents will be published . The users do not completely read the entire content to understand, instead of that they pick out the certain content as on headlines by skip reading and judge the segment of the news usually.

IV RELATED WORKS:

Up to now, many researches about discovering and recommending valuable contents for individual users from a large amount of stored contents have been done actively, and many recommender systems have been proposed. Recommender systems have changed the way people find products, information, and even other people. They study patterns of behavior to know what someone will prefer from among a collection of things that has never experienced. Fig.2. shows different recommendation systems.

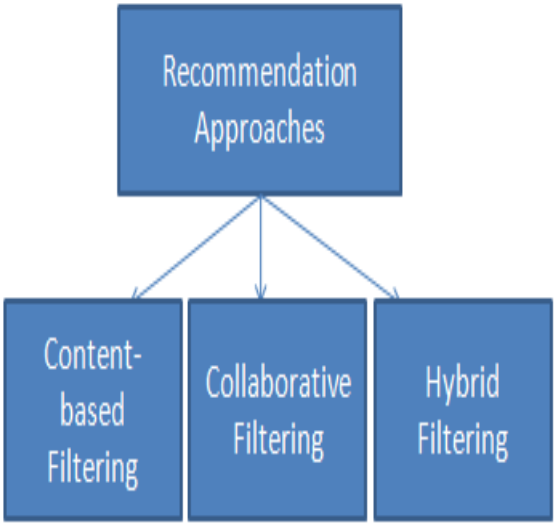

Fig.2. Different recommendation systems

Recommender systems typically produce a list of recommendations in one of two ways - through collaborative or content-based filtering. Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviors, activities or preferences and predicting what users will like based on their similarity to other users. Content-based filtering methods are based on a description of the item and a profile of the user's preference.In a content-based recommender system, keywords are used to describe the items; beside, a user profile is built to indicate the type of item this user likes. It can often suffer from the following problem :

- Cold Start: These systems often require a large amount of existing data on a user in order to make accurate recommendations.
- Scalability: In many of the environments that the system make recommendations , there are millions of users and products. Thus, a large amount of computation power is often necessary to calculate recommendations.
- Sparsity: The number of items sold on major e-commerce sites is extremely large. The most active users will only have rated a small subset of the overall database. Thus, even the most popular items have very few ratings.

In those recommendation systems one of the most widely used technique is the hybrid-collaborative filtering method. The hybrid filtering is the combination of both the content- based and collaborative filter in which both technique process simultaneously . Here we are going to apply collaboration first and then content- based is applied. Hence we are mentioning as hybrid collaborative filter here. It is proposed to use the personalization for

recommending new articles which is building a profile for the user's genuine interests. In order to handle the cold-start problem, the system includes the opinions of chosen experts (who uses the social networks have significant influence on new users). So that the system can make recommendations to a new user. By using the TF-IDF method new items can also be recommended. It includes another approach that is grouping the users as an old or a new user according to the number of articles they read. For each group of users, a separate model is trained for predicting the most interesting news category. And the top new article from each filtered category is recommended to the user. So that cold start problem can be eliminated by using the current contextual information for first recommendations. When a user starts to read a news item the system generates recommendations and during the session of the user the system updates the model and makes better recommendations.

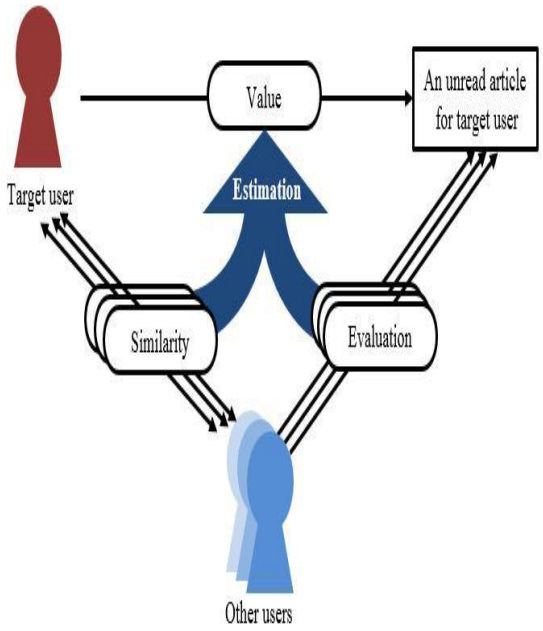

Fig. 3. Idea for estimating unread article's value.

V. FRAMEWORK:

The increasing amount of data on the internet makes harder to find what we are really looking for. Even though the technologies like search engines , it is still hard to find the information we really want to get. On the other hand, we are not always sure about what we want to get. We can only search for what we know and we try to find some connections to the new information. This research aims to develop a method to support a user to browse a news websites. Our method can improve the efficiency of browsing unread news articles. In our proposal method, a hybrid- collaborative filtering technique for information recommendation is utilized to filter valuable articles effectively based on estimating the similarities of behaviors between users. Figure 3 illustrates how to estimate the value of the unread articles for the target user in our method.

By, our estimation method is based on the assumption that articles that have not read by the target user and have been judged as valuable by other users who are similar with the target user would be also valuable for target user too. Of course, it is impossible to make some users evaluate each article explicitly. Therefore browsing behaviors on an article would be utilized for estimating the value of the article. However, the relevance between the value of an article and browsing behaviors of the article is needed to verify experimentally. There are many types of browsing behaviors that we can observe. In this paper, we focused on the reading speed and flicking speed of an article in any website.

Consider the reading speed rv_i, is defined as the number of characters that are processed in a unit time for an article *i*. The following formula (1) gives the formal definition.

$$rv_i = length_i / time_i \quad \text{-------(1)}$$

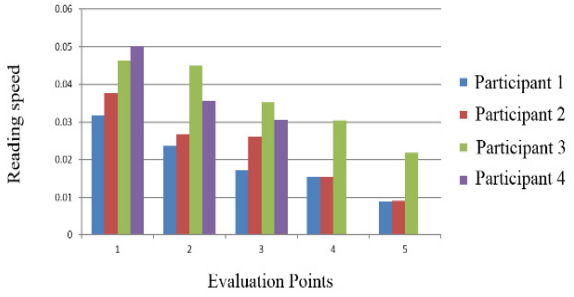

Fig. 4. Experimental results of reading speed (unit: word/ms).

Here, $length_i$ represents the number of characters of an article *i*, and $time_i$ represents the display time (milliseconds) of article *i*.

Higher reading speed for an article would represent higher possibility to skip it in user browsing.

The flicking speed fv_i is defined as the length of the trajectory of the finger on the touch screen of the smart phone when the user changes the articles for reading. The following formula (2) gives the formal definition.

$$fv_i = dist_i / dur_i \quad \text{--------(2)}$$

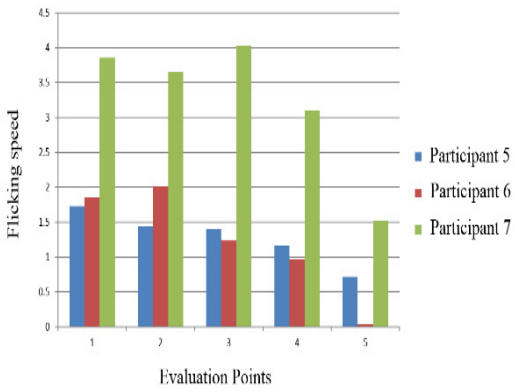

Fig. 5. Experimental results of flicking speed (unit: pixel/ms).

Here, $dist_i$ represents the length of (pixels) of the finger during the flick operation for displaying the next content from the article *i*, and dur_i represents moving time (milliseconds).

VI. BEHAVIOURAL HYBRID COLLABORATIVE FILTERING ON NEWS WEBSITES:

In this section, we introduce a hybrid collaborative filtering method based on user behaviors. Our proposal method aims to estimate the value of unread articles based on user behaviors. Firstly, our system collects users who have read the same articles that the target user has read. Secondly, the score for each article of each user is estimated based on the reading speed and flicking speed of the article. Thirdly, the score for each article that the target user have not read is estimated using hybrid collaborative filtering technique.

A. HYBRID APPROACH CF ALGORITHM

Since hybrid covers information deduction made according to both users and items, hybrid approach combines CF-U (user based) and CF-I (item based) techniques.

TABLE I
USER-ITEM MATRIX ON THE SCORES OF THE SAMPLE TABLE

User/ Item	I1 Education	I2 Sports	I3 Political
U1	3	1	2
U2	3	1	1
U3	3	1	5
U4	1	5	5
U5	2	5	5
U6	3	5	3
U7	3	3	2

We define our method in the followings.

a) User contribution to an article

The sets of articles that are distributed to each user are different. In order to estimate the value of an unread article for the target user, our system have to select sufficient other users who have read the article. In order to evaluate the sufficiency of a user, we introduce the contribution. The contribution of a user u_k to the target user u is defined as following where c_u and c_{uk} represents the set of articles that the target user has read, and the set of articles that the user u_k has read respectively.

$$Contribution(u,u_k)=|C_u \cap C_{uk}| \div |C_u \cup C_{uk} \quad ---(3)$$

b) User similarity

The similarities between the target user and other users are calculated by the cosine similarity measure, which is one of classic similarity calculation measures. As the ordinary cosine of the trigonometric function, the cosine similarity represents the proximity of the angle between user's vectors. The vectors are more similar when the cosine similarity closer to 1, and to the contrast, not similar to closer to 0. Figure 6 shows the image of our similarity function between users which is defined where n represents the number of articles that both u and u_k have read and r represents a parameter.

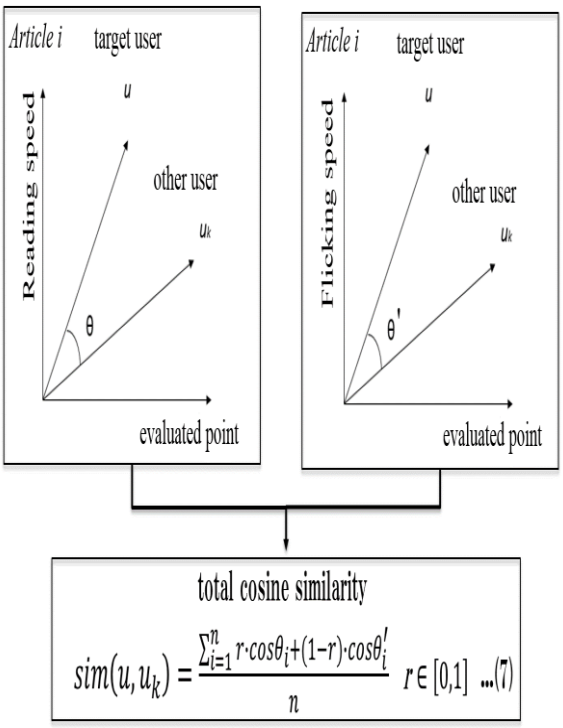

Fig. 6. The similarity measure of users.

VII. EVALUATION:

In order to evaluate the effectiveness of the proposed method, we conducted some experiments. The information or data of user browsing behavior from the framework analysis will be given to the data mining tool WEKA for processing. After the data has been processed, it will produce the result of user preference level with minimum error and accuracy of the estimation will be proved to be higher.

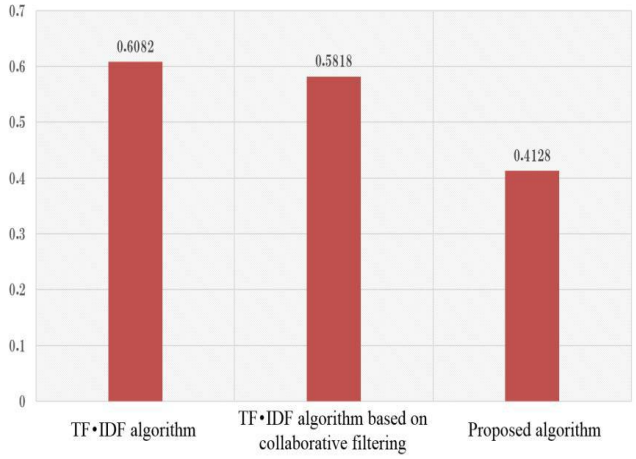

Fig. 7. Proposed method comparison with baselines.

VIII. CONCLUSION:

In this paper , we have proposed a hybrid collaborative filtering method for browsing news articles efficiently from the internet. Our method predict's the user behavior's of browsing news articles in the internet for estimating the skipped content of the target user . In order to make the accurate estimation of user preference ratio , the data from application which has been developed is given to tool for displaying the variation of different users preferances.

REFERENCES:

[1] J.R Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufman, 1993.
[2] P. Arabie, and Y. Wind, "Marketing and Social Networks." In Stanley Wasserman and Joseph Galaskiewicz, Advances in Social Network Analysis: Research in the Social and Behavioral Sciences. Thousand Oaks, Calif.: Sage Publications, pp. 254–273, 1994.
[3] M. Balabanović, "An adaptive web page recommendation service," Proceedings of the first International Conference on Autonomous Agents. 1997: 378-385.
[4] C. Basu, H. Hirsh, and W. Cohen, "Recommendation as classification: using social and content based information in recommendation," Proceedings of the 1998 workshop on Recommender Systems, 1998: 11-15.
[5] Li Y., Lu L., Xuefeng L., "A hybrid collaborative method for multiple-interest and multiple-content recommendation in E-Commerce, " *International Journal of Expert System with Application*, pp.67-77,2005.
[6] Koren Y., "Collaborative Fİltering with Temporal Dynamics," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.447-456, 2009.
[7] Sarwar B., Karypis G., Konstan J. and Rield J. , "Item-based Collaborative Filtering Recommendation Algorithms," *in WWW'01:International Conference on World Wide Web*, pp.285-295, 2001.
[8] Choi K., Yoo D., Kim G. ve Suh Y., "A hybrid online-product recommendation systems: Combining implicit rating-based collaborative filtering and sequential pattern analysis," *Elseiver Elektronic Commerce Research and Applications*, pp.309-317,2012.
[9] D. R. Liu, and Y. Y. Shih, "Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences," Journal of Systems and Software, vol.77, no.2, pp:181-191, 2005.
[10] George Lekakos, George M. Giaglis, Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors, Interacting with Computers 18 (2006) 410–431.
[11] Gong J.G., (2010). "A Collaborative Filtring Recommendation Algorithm Based on User Clustering and Item Clustering," *Academy Publisher*,pp.745-752,2010.
[12] L. H. Ungar and D. P. Foster. Clustering Methods for Collaborative Filtering. In Proc. Workshop on Recommendation Systems at the 15th National Conf. on Artificial Intelligence. Menlo Park, CA: AAAI Press.1998
[13] M. O. Conner and J. Herlocker. Clustering Items for Collaborative Filtering. In Proceedings of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, August 1999.
[14] A. Kohrs and B. Merialdo. Clustering for Collaborative Filtering Applications. In Proceedings of CIMCA'99. IOS Press, 1999.
[15] Lee, WS. Online clustering for collaborative filtering. School of Computing Technical Report TRA8/00. 2000
[16] Cantador, I., Castells, P. Multilayered Semantic Social Networks Modelling by Ontologybased User Profiles Clustering: Application to Collaborative Filtering. EKAW 2006, pp. 334-349.
[17] Gao Fengrong, Xing Chunxiao, Du Xiaoyong, Wang Shan, Personalized Service System Based on Hybrid Filtering for Digital Library, Tsinghua Science and Technology, Volume 12, Number 1, February 2007,1-8.
[18] Huang qin-hua, Ouyang wei-min, Fuzzy collaborative filtering with multiple agents, Journal of Shanghai University (English Edition), 2007,11(3):290-295.
[19] Songjie Gong, Chongben Huang, Employing Fuzzy Clustering to Alleviate the Sparsity Issue in Collaborative Filtering Recommendation Algorithms, In: Proceeding of 2008 International Pre-Olympic Congress on Computer Science, World Academic Press, 2008, pp.449-454.
[20] Jia D., Zhang F., Liu S., "A Robust Collaborative Filtering Recommendation Algorithm on Multidimensional Trust Model, "*Journal of Software*, vol 8,no.1,pp.11-18, 2013.
[21] Zhang Z., Lin H., Liu K., "A hybrid fuzzy-based personalized recommender system for telecom products/services, "*Information Sciences*,vol.235,pp.117-129, 2013.
[22] Said A., Fields B., Jain B.J, Albayarak S., "User-Centric Evaluation of a K-Further Neighbor Collaborative Filtering Recommender Algorithm,"*CSCW 2013*, San Antonio, USA, 2013.
[23] Wu J., Chen L., Feng Y., Zheng Z., "Prediction Quality of Service for Selection by Neighborhood-Based Collaborative Filtering, "*IEEE Transactions on Systems, Man, And Cybernetics:Systems*, vol.43, issue 2, 2013.
[24] M. Toki and T. Ushiama, "A Method for Composing a User Profile Based on Browsing Behaviors on Social Streams," IPSJ Transactions on Databases,vol.6, no.4, pp: 35-45, 2013.
[25] Hong Yan and Taketoshi Ushiama, " Effective browsing technique based on behavioural collaborative filtering on social streams", International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014.
[26] Çiğdem Bakır,"Hybrid Based Collaborative Filtering with Temporal Dynamics", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 11
[27] Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar and Nidhi Gupta, ―A Comparative Study of Various Clustering Algorithms in Data Mining‖, International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384
[28] Swasti Singhal, Monika Jena," A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013