I.J. Information Technology and Computer Science, **2018, 10, 31-37**
Published Online October 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.10.04

MF-NB Learning Based Approach for Recommendation System

Hutashan V. Bhagat
Chandigarh Engineering College/Department of I.T, Landran (Mohali), 140307, India
E-mail: hutashan20@gmail.com

Shashi B. and Sachin M.
Chandigarh Engineering College/Department of I.T, Landran (Mohali), 140307, India
E-mail: {cecm.infotech.shashi, cecm.infotech.sachinmajithia}@gmail.com

Received: 01 July 2018; Accepted: 12 August 2018; Published: 08 October 2018

***Abstract*—**The Multi Factor-Naive Bayes classifier based recommendation system is analyzed with respect to the traditional KNN classifier based recommendation system. The classification of the web usage data is done on the basis of the keyword name, keyword count, inbound links and age group of the users. Whereas, in traditional KNN the URL was the only factor that was considered for the purpose of classification. The performance evaluation is done in the terms of RMSE, Error Rate, Accuracy Rate and Precision. The MF-NB is observed to be outperforming the KNN classifier in all respective terms.

***Index Terms*—**Data Mining, Web Usage Data Mining, Classification, Naïve Bayes Classification, KNN Classifier.

I. Introduction

Data mining is a process in which important or meaningful values are extracted from the database. Data Warehouse is a kind of storage where the data is kept for the purpose of fetching it in near future when it is required. Data Warehouse can store any kind of data particularly the type of data depends upon the kind of industry for which it is being used [1]. Most of the industries keep the record of each and every kind of data whereas some companies only store that information which is beneficial and meaningful for them. The data stored in a warehouse is helpful in decision support system. On the basis of historical data, the decision regarding the future schemes can be taken easily or much effectively. The Web is a huge, explosive [2], diverse, dynamic and mostly unstructured data repository, which supplies an incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view – users, Web service providers, business analysts. The users want to have the effective search tools to find relevant information easily and precisely [3].

II. Related Works

A lot of researchers proposed different techniques and algorithms for web mining and use these to develop a system for recommendation. Several techniques that were found in the academic literature are described in this section.

D.A. Adeniyi et al. [1] in this paper the author had offered a study of automatic web usage data mining and recommendation mechanism on the basis of the present user behavior by his/her click stream data on the fresh generated Really Simple Syndication (RSS) reader website, to offer an appropriate knowledge to the user among explicitly asking no for it. To be utilized online and in actual time the K-Nearest-Neighbor (KNN) categorization mechanism has been instructed in order to recognize clients/visitors click stream data, matching it to a specific user class and propose a tailored browsing option that meet the requirement of the particular user at a specific time. The simulation results had demonstrated that the K-Nearest Neighbor classifier was transparent, constant, straightforward and easy.

A. Talakokkula et al. [3] in this paper the author had surveyed that the World Wide Web was a broad arrangement of unstructured web documents such as text, images, audio, video or Multimedia content. Because the web was rising quickly among millions of documents, mining the data from the web was a complicated job. In order to mine several patterns from the web was called as Web mining. The Web mining was categorized again as content mining, structure mining and web usage mining. The Web utilization mining extorts meaningful data from several web logs i.e. users utilization history. For the enhanced understanding this was helpful and serves the people for improved web applications. Web usage mining not only helpful for the people who access the documents from the World Wide Web, but also it helpful for various applications such as e-commerce to do personalized marketing, eservices, the government agencies to categorized threats and fight against terrorism, fraud detection, to recognize the criminal activities, the companies can introduce improved customer relationship and can enhance their businesses through examining the people buying strategies etc. In this paper the author had illustrated in detail regarding web utilization mining and how it was useful.

S.P. Singh et al. [5] in this paper the author had illustrated that the Web Usage Mining (WUM) was the part of Web Mining. There was data mining method to fetch and discover information from the web data. Web usage mining utilization data mining procedure for the research of the usage pattern from data fetched from the web log files. Web log was the collection of academic educational institute web server data was examined to help the institute for further enhancing the terms and policies of the service offered. Web usage was also helpful for enhancing or recognizing the visitor of website through auditing the log files of that site. The focus was on the data collection in web servers of academic educational institution and implement. The utilization of Web Log Expert-lite 9.3 tools is examined.

Yeqing Li et al. [6] in this paper the author had illustrated that among the quick generation of Internet, an era of information explosion was entered, and there was a lot of redundant information in the Network. This paper would be examined the realization of Web content mining and Web structure mining, their basic paradigm principles and their application fields.

V. Medvedev et al. [9] in this paper the author had projected a new mechanism for web-based solution called as DAMIS, motivated through the Cloud, was projected and executed. By constructing scientific workflows for data mining by applying a drag and drop interface for knowledge scientists and business intelligence professionals the constructed massive data mining easy, efficient, and understandable was allowed. The utilization of scientific workflows permitted the composing convenient tools for modeling data mining procedures and for simulation of actual-world time- and resource-utilizing data mining issues. The resolution was meaningful to resolve data categorization, clustering, and dimensionality decreasing issues. The DAMIS structural design was deliberated to assure simple availability, usability, scalability, and mobility of the resolution. The projected resolution has a broad array of applications and permitted to get deep insights into the data throughout the procedure of information detection.

Z. Balogh et al. [12] in this paper the author had illustrated that where entire the accessible data from the visitors of an online e-learning environment was saved and various data-mining techniques were performed on it to reveal hidden rules. As a result, intelligence and life satisfaction traits of the visitors can be predicted based on their behavior in the online environment. This information can be used to profile the visitor and targeted advertisements can be sent to them.

III. Objective

This study is the continuation of previous work that proposes a Naïve Bayes classifier for pattern matching and classification of web usage data for recommendation system. The previous study generates the automatic recommendation system by getting the motivation from traditional KNN classifier [1]. The motivation behind deriving this study is the shortcomings of traditional recommendation systems. Such as the KNN was not capable and scalable enough to find the exact match for the immediate neighbor and did not cover all the factors of web-based data for recommending the user were found in the previous research study of this chain. In previous work the Naïve bays classification based recommendation system was implemented in JAVA and analyzed to be effective and it was concluded that the proposed classification process did not get affected if the data goes out of bound because it will not only work on distance as a major factor, but it can be more refined by introducing the extraction of more information regarding the search by customer. Those can be the keywords matching, the tag matching, as major factors.

The objectives behind conducting this study is to analyze the existing recommendation system under different metrics of evaluation (precision rate, accuracy rate, error rate etc), to implement the proposed recommendation system by using Naïve Bayes classifier and to compare the performance with the existing recommendation system.

The advantage of proposed work over traditional recommendation system is that it is less complex method as it considers multiple parameters for deriving a recommendation for the users. And this feature also leads to the more accurate recommendations to the users. Along with this another advantage of proposed work is that it is highly reliable and efficient because it implements the Multi Factor Naïve Bayes (MF-NB) classifier instead of KNN classifier and MF-NB is found to be more prominent than KNN.

IV. Techniques Used

A. Traditional KNN Classifier

The traditional recommendation system implemented the KNN classifier for classification purpose. To perform the classification by using KNN classifier first of all a dataset was created. After creating the dataset, the dataset was divided into two parts i.e. one for training purpose and other for testing purpose. The dataset for training was passed as an input to the classifier whereas the testing dataset was used for testing purpose on the basis of the trained dataset. The mathematical model for KNN depicts that it only utilizes the local prior possibilities for classification purpose. For given query x_t the classifier works as follows:

$$y_t = \underset{c\in(c1,c2,\ldots\ldots cm)}{argmax} \sum x_i \in N(x_t,k)^{E(y_i,c)} \quad (1)$$

where, y_t denotes the predicted class corresponding to the x_t, m depicts the classes that are included in data. Also,

$$E(a,b) \;= \begin{cases} 1 \; if \; a = b \\ 0 \quad else \end{cases} \quad (2)$$

$$N(x,k) = set\ of\ k\ nearest\ neighbor\ of\ x.$$

The above defined equation (1) can also be formulated as follows:

$$y_t = argmax\left\{\sum x_i \in N(x_i,k)^{E(y_i,c1)}, \sum x_i \in N(x_i,k)^{E(y_i,c2)}, \ldots\ldots \sum x_i \in N(x_i,k)^{E(y_i,cm)}\right\} \quad (3)$$

$$y_t = argmax\left\{\sum x_i \in N(x_i,k)^{E\frac{(y_i,c1)}{k}}, \sum x_i \in N(x_i,k)^{E\frac{(y_i,c2)}{k}}, \ldots\ldots \sum x_i \in N(x_i,k)^{E\frac{(y_i,cm)}{k}}\right\} \quad (4)$$

It is known that,

$$p(c_j)_{(x_t,k)} = \sum x_i \in N(x_t,k)^{\frac{E(y_i,c_j)}{k}} \quad (5)$$

$p(c_j)_{(x_t,k)}$ denotes the probability of occurrence of j^{th} class in neighbor x_t.

Thus the equation (4) can be written as

$$y_t = argmax\{p(c_1)_{(x_t,k)}, p(c_2)_{(x_t,k)}, \ldots\ldots p(c_m)_{(x_t,k)}\} \quad (6)$$

On the basis of above formulation, it is defined that the KNN performs the classification on the basis of the local possibilities. Following is the pseudo code for KNN that is implemented in proposed work.

```
1. Evaluate d(x,x_i) $i = 1,2,\ldots,n$
   where 'd' refers to the Euclidean Distance among the points. It is evaluated as follows:

   $d = \sqrt{(y1 - x1)^2 + (y2 - x2)^2}$

2. Let k a positive integer; consider the first k distance from d.
3. Evaluate k point corresponding to k distances.
4. Assume $k_i$ to define the number of points related to the $i^{th}$ class among k. $k \geq 0$
5. If $K_i > k_j \forall i \neq j$ then put x in class i.
```

B. Proposed MF-NB Classifier

The Naïve classifier is used for the purpose of classification of web usage information. This section gives a brief review of the proposed Multi Factor-Naïve Bayes (MF-NB) classifier. For the purpose of classification the following factors are considered:

- Keyword Name
- Inbound URLs
- Age group of the users

Then the Naïve Bayes classification technique is applied for generating the decision for recommendation system. The training is performed for evaluating the probability of recommendations decision over classes or categories of the dataset. Then the PT is obtained as a probability of recommended URL links. After performing the training the next step is to perform the testing of the dataset. The testing is performed on the part of the dataset that is defined by the user.

On the basis of the trained and tested dataset, the output of the recommendation system will be generated.

Naïve Bayes classification performs the classification on the basis of an equation that evaluates the probability for true decision firstly. After then the probability for input parameters are evaluated individually. In proposed recommendation system the Naïve Bayes classifier is used as follows:

$$(P_{TR}) = \frac{Count\ of\ positive\ recommendation}{count\ of\ total\ entries} \quad (7)$$

where, P_{TR} is the prior probability of true decision.

After evaluating the P_{TR} testing of the dataset is done by using the following formulations:

$$P_A = \frac{P(D/k)P(k)}{P(D)} \quad (8)$$

where *P(k)* refers to the prior probability of keyword name *k*, *P(D)* denotes the prior probability of training dataset D, P_A refers to the Likelihood of Keyword Name and $P(D/k)$ depicts the probability of D with respect to given *k*.

Similarly, the P_B, P_C, P_D, P_E is evaluated for URLs, Keyword Count, inbounds and age group with respect to the given dataset *D*.

$$P_F = \frac{P_A}{P_T} \times \frac{P_B}{P_T} \times \ldots\ldots \times \frac{P_N}{P_T} \quad (9)$$

where, P_B refers to the Likelihood of URLs, P_C refers to the Likelihood of Keyword Count, P_D refers to the Likelihood of inbounds, P_E refers to the Likelihood of Age Group, P_T refers to the total entries of the tested dataset and N is number of factors considered for the training purpose.

Final decision of Recommendation System:

$$Score = P_F \times P_{TR} \quad (10)$$

$$P_{F_{Mean}} = \frac{\sum_{i=1}^{N} P_{F_i}}{N} \tag{11}$$

After this, the score and $P_{F_{Mean}}$ is compared and if the score is found to be greater than $P_{F_{Mean}}$ then the recommendation system will generate "Yes" otherwise "No".

After training and testing of the data, the validation is done. The validation of the proposed work will be that the dataset which is created at an initial stage will have the actual recommendation results, once the traditional work having KNN approach and the Naïve Bayes approach with enhanced features extraction will get results of recommendation. The achieved recommendation of KNN and Proposed model will be matched with the actual results in the dataset. The number of recommendation those will get matched with the actual results will reflect the performance of the proposed model.

V. Methodology

The proposed system can be represented as a prototype by taking data with validating the results with actual results. The model is being designed in modular approach as shown in the Fig. 1. Various steps involved are:

1) Gathers the Dataset in the form of information from online sources. The information related to the world, entertainment, sports, business, politics etc. is gathered.
2) After gathering the data next step is to categorize the collected dataset as per its related domain. In proposed work, the dataset is categorized in following categories:

   a) Sports
   b) Entertainment
   c) News
   d) World
   e) Politics

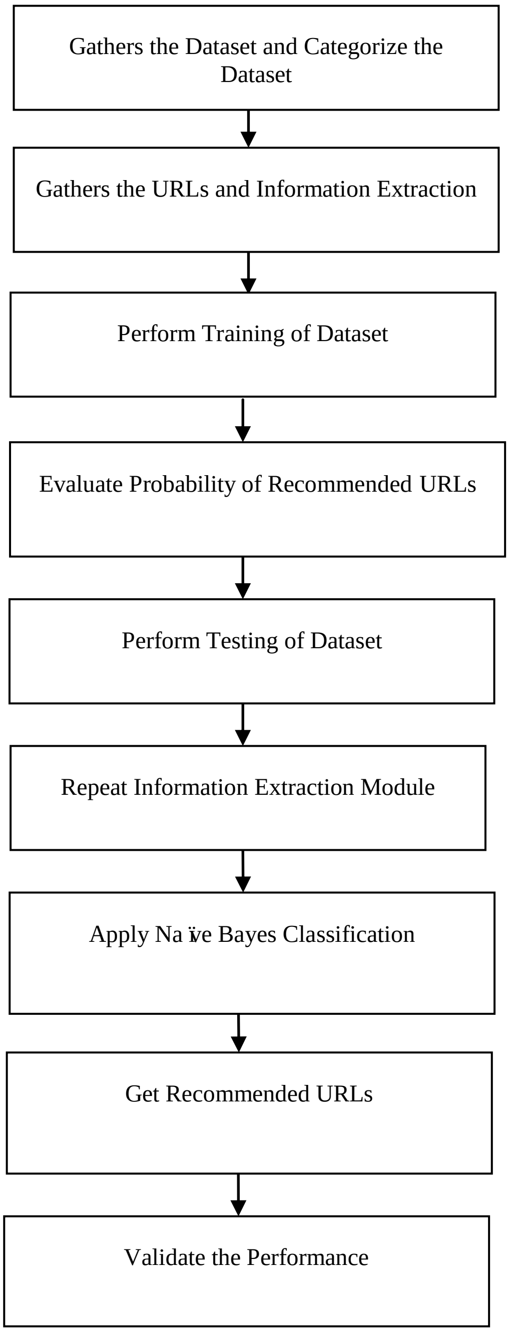
Gathers the Dataset and Categorize the Dataset
Gathers the URLs and Information Extraction
Perform Training of Dataset
Evaluate Probability of Recommended URLs
Perform Testing of Dataset
Repeat Information Extraction Module
Apply Naïve Bayes Classification
Get Recommended URLs
Validate the Performance

Fig.1. Flow Chart of the Proposed System.

3) After categorizing the data, the URLs for the related category is collected from online available sources.
4) On the basis of the collected URLs, the information extraction is done. The information such as keywords, inbound links, keyword count and age group of the users is extracted so that it can be further utilized for training purpose.
5) The training of dataset in proposed work is done by using equation (1) and the prior probability of true decision is evaluated.
6) After this, the testing is performed on various portions of the dataset such as 70%, 80%, 90% and 100%. For the purpose of testing the equations (2), (3), (4) and (5) are implemented.
7) The information extraction is repeated for keyword name, keyword count, inbounds and age group.
8) The Na ïve Bayes classification is applied for classifying the extracted trained dataset.

VI. Results and Performance Analysis

The proposed MF-NB based recommendation system is compared to the traditional KNN based recommendation system under this section. The recommended URLs are observed and the validation of the proposed work is done by generating a comparative analysis study among MF-NB and KNN classifiers in the terms of Accuracy, Error Rate, RMSE and Precision.

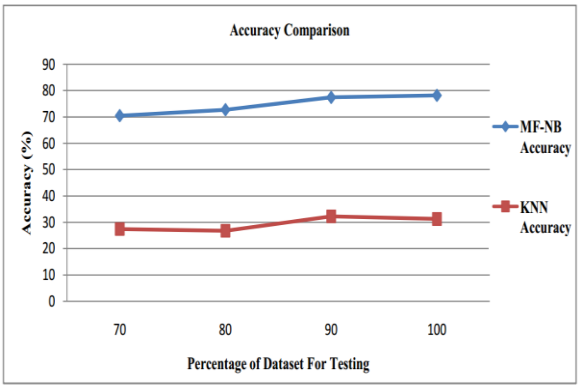

Fig.2. Accuracy Analysis

A. Accuracy Analysis

The graph in Fig. 2 depicts the comparative analysis of MF-NB and KNN on the basis of the accuracy rate. The accuracy is the statistical measure that is used to measure the exactness of the observed results. On the basis of the graph, it is observed that the accuracy rate of MF-NB is higher for all respective category of the tested dataset.

B. Error Analysis

The graph in Fig. 3 delineates the contrast of MF-NB and KNN classifier based recommendation system in the terms of Error Rate. The x-axis in the graph calibrates the data in the form of the tested portion of the dataset and y-axis in the graph defines the error rate that ranges from 0 to 80. The graph proves that the error rate of MF-NB is lower in comparison to the KNN classifiers.

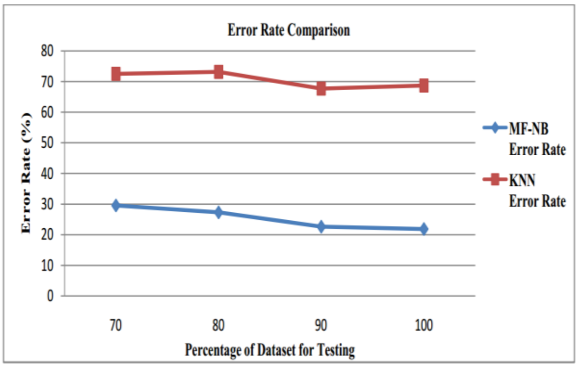

Fig.3. Error Rate Analysis

C. RMSE Analysis

Similarly, the Fig. 4 shows the RMSE of MF-NB and KNN classifier. The respective factors are evaluated on the basis of the different testing dataset. The RMSE is Root Mean Square Error and it is essential to have a lower RMSE rate for an ideal recommendation system. As per the observation from the graph, it can be said that the RMSE of proposed work is quite effective and better than the RMSE of KNN based recommendation system.

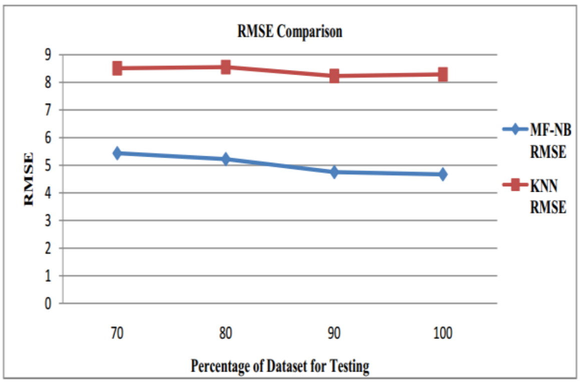

Fig.4. RMSE Analysis

D. Precision Analysis

The Precision is a performance metric that is used to evaluate the rate of positively predicted values by the recommendation system. Thus on the basis of definition, it should be beneficial to have a higher precision rate. The graph in Fig. 5 defines that the precision of MF-NB is evaluated to be high for all respective tested datasets.

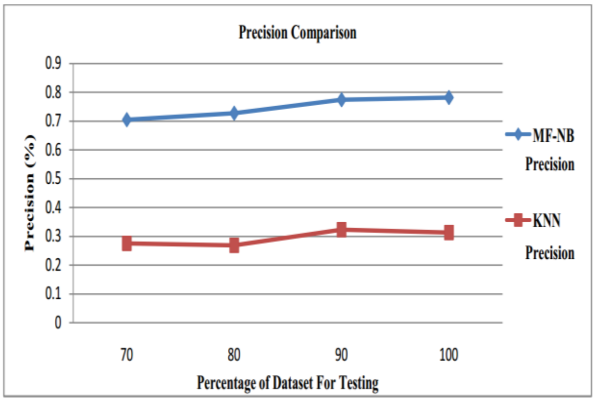

Fig.5. Precision Analysis

Table 1 depicts the Accuracy, Error, RMSE, and Precision for MF-NB and KNN. The observations are being carried out at different values of data set. Initially results are evaluated at 70% of the total data in the data set, afterwards at 80% and 90% of the total data, and finally at 100% of the data in the data set. The facts observed from the results are gathered in the table below.

Table 1. Performance Analysis

Dataset (%)	Accuracy		Error		RMSE		Precision	
	MF-NB	KNN	MF-NB	KNN	MF-NB	KNN	MF-NB	KNN
70	70.4	27.4	29.5	72.5	5.4	8.5	0.70	0.2
80	72.7	26.8	27.2	73.1	5.2	8.5	0.73	0.2
90	77.4	32.2	22.5	67.7	4.7	8.2	0.77	0.3
100	78.1	31.2	21.8	68.7	4.6	8.2	0.78	0.3

VII. Contribution and Conclusion

A. Contribution

Taken into consideration both the shortcomings of the existing KNN based recommendation system and research objectives; a system is proposed to develop a data mining technique by using the Naïve Bayes classifier. The proposed technique has not only provided results based on the distance of the nearest neighbour but also consider closure set of attributes in tags. Because of this, the proposed classifier is named as MF-NB Classifier that stands for Multi Factor-Naïve Bayes Classifier. Results have been derived by using the MF-NB classifier over different performance metrics that are analyzed and compared. Experimental values proved that the proposed MF-NB classifier outperforms the existing KNN based systems. The main contribution of this work is to provide a recommendation system that overcomes the limitations of the existing traditional KNN based recommendation systems in terms of different performance metrics such as Accuracy, RMSE, Error Rate and Precision. Based on the experimental values, the proposed recommendation system has outperformed with respect to all the performance metrics.

B. Conclusion

Many authors have been done research in the domain of recommendation system to generate the best decision on the basis of the web generated information. This study introduces a recommendation system that generates the results on the basis of the keyword names, inbounds links and age group of the users by using the Naïve Bayes classifier for training and testing purpose. The results conclude that the proposed recommendation system generates the highly accurate and error free decision. The results have been calculated over various performance parameters. The experimental results drawn by using the KNN classifier gives an accuracy of 31.27%, error rate of 21.81%, average RMSE value for whole data set is 8.29% and average precision is found to be 0.312%. Similarly, for the same data set the results drawn by using the proposed MF-NB classifier gives an accuracy of 78.18%, error rate of 21.81%, average RMSE is 4.67% and average precision is 0.78%. Hence, the testing of the proposed recommendation system done on various datasets observed to be effective in terms of performance.

To sum up, it is concluded that the web usage data mining has gathered the popularity among the researchers due to the introduction of the advancements in the recent technology as each and every domain utilizes the internet services for various purposes such as education, entertainment, communication etc. therefore, the web service providers perform the data mining on the web-based data to generate the recommendation of the online content to the users. In previous work of the same study, a Naïve Bayes classification based recommendation system was proposed and that work is continued under this work by comparing the performance of the MF-NB based recommendation system and traditional KNN based recommendation system. On the basis of the analysis, it is concluded that the proposed MF-NB based recommendation system generates the recommendations of the URLs that are more accurate, error-free and more positive than the recommendations of KNN classifiers.

References

[1] D.A. Adeniyi, Z. Wei, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", *ELSEVIER*, Vol. 12, pp. 90-108, 2014.

[2] Ngoc Nhu Van, J. Rokne, "Integrating SOM and Fuzzy K-means Clustering for Customer Classification in Personalized Recommendation System for Non-Text based Transactional Data", *International Conference on Information Technology*, Amman, Jordan, 2017.

[3] Anitha Talakokkula, "A Survey on Web Usage Mining, Applications and Tools", *Computer Engineering and Intelligent System*, Vol. 6, No.2, pp. 22-30, 2015.

[4] Bo Cheng, Shuai Zhao, Changbao Li, Junliang Chen, "A Web Services Discovery Approach Based on Mining Underlying Interface Semantics", *IEEE*, Vol. 29, pp 950-962, 2017.

[5] Satya Prakash Singh , Meenu, "Analysis of web site using web log expert tool based on web data mining", *IEEE*, 2017.

[6] Yeqing Li, “Research on Technology, Algorithm and Application of Web Mining”, *IEEE*, Vol. 1, pp. 772-775, 2017.

[7] Z. A. Usmani, Saiqa Khan, Mustafa Kazi, Aadil Bhatkar, Shuaib Shaikh, “ZAIMUS: A department automation system using data mining and web technology”, *IEEE*, pp 1-6, 2017.

[8] Martin Lnenicka, Jan Hovad , Jitka Komarkova , Miroslav Pasler, “A proposal of web data mining application for mapping crime areas in the Czech Republic”, *IEEE*, 2016.

[9] Viktor Medvedev, Olga Kurasova, Gintautas Dzemyda, “A new web-based solution for modelling data mining processes”, *ELSEVIER*, Vol. 76, pp. 34-46, 2016.

[10] Petar Ristoski, Heiko Paulheim, “Semantic Web in data mining and knowledge discovery: A comprehensive survey”, *ELSEVIER*, Vol. 36, pp. 1-22, 2016.

[11] Venkata Subba Reddy Poli, “Fuzzy data mining and web intelligence”, *IEEE*, 2016.

[12] Zoltán Balogh, “Data-mining behavioural data from the web”, *IEEE*, Vol.1, pp. 122-127, 2016.

[13] Suvarn Sharma, Amit Bhagat, “Data preprocessing algorithm for Web Structure Mining”, *IEEE*, pp. 94-98, 2016.

[14] Wang Lei, Liu Chong, “Implementation and Application of Web Data Mining Based on Cloud Computing”, *IEEE*, 2016.

[15] D. Bavarva Bhaskar, Dheeraj Kumar Singh, “Multimedia questions and answering using web data mining”, *IEEE*, 2015.

[16] Ying Han, Kejian Xia, “Data Preprocessing Method Based on User Characteristic of Interests for Web Log Mining”, *IEEE*, 2014.

[17] Quang yang, “10 Challenging problems in Data Mining research”, *World Scientific*, Vol. 5, No. 4, pp 597-604, 2006.

[18] L. Habin, K. Vlado, “Combining mining of web server logs and web content for classifying users’ navigation pattern and predicting users future request”, *J. Data Knowledge Eng.*, Vol. 61, pp. 304–330, 2014.

[19] Dhanashree S. medhekar, “Heart Disease prediction System using Naïve Bayes”, *IJERSTE*, Vol. 2, No.3, pp. 1-5, 2013.

[20] Arno J. Knobbe, “Multi-Relational Data Mining”, *SIKS*, pp 1-130, 2015.

[21] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, “Recommendation Systems: Principles, methods and evaluation”, *ELSEVIER*, Vol. 16, pp. 261-273, 2015.

Authors’ Profiles

Hutashan V. Bhagat pursued Bachelor of Technology in Information Technology from National Institite of Technology Srinagar, India in 2012. He has worked for three years as a Software Engineer in Samsung Research Institute Noida, India. He is currently pursuing Master of Technology in Information Technology from Chandigarh Engineering College, Mohali, Punjab. His main research work focuses on Big Data Analytics, Data Mining, IoT and Wireless Sensors.

Shashi B. did his Ph.D from NIT, Kurukshetra, India in 2015. Dr. Bhushan is presently the Head of the Department (Information Technology Department) at CEC, Landran since April 2011. He is having more than 20 years of academic and administrative experience. Dr. Bhushan has published more than 20 research papers in various National/International Journals of repute. He had also filed two patents under Intellectual Property Right (IPR) entitled ―System and Method of Self Destructive Program on Privacy‖ and ―Advance Server Protection Framework (ASPF)‖. He had haired the technical sessions in Technical Seminars and in National/International Conferences. He had also delivered the expert lecture in various Workshops and Faculty development Program. His areas of interest are Peer to Peer Networks, Mobile Computing and Databases.

Sachin M. is presently working as an Assistant Professor, CEC Landran Mohali, in Information Technology Department. He has done his B.Tech in Information Technology in year 2003 and M.Tech in computer science and engineering. His area of interest is in Mobile Communication & wireless networks. He has published more than 50 papers in various journals and conferences of international and national repute. He is member of Computer Society of India.

How to cite this paper: Hutashan V. Bhagat, Shashi B. and Sachin M., "MF-NB Learning Based Approach for Recommendation System", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.10, pp.31-37, 2018. DOI: 10.5815/ijitcs.2018.10.04