IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-6, June2018 E-ISSN: 2347-2693

Consummate Approach for Classification and Pattern Matching for a Web usage based Recommendation System

Hutashan Vishal Bhagat[1*], Shashi Bhushan[2], Sachin Majithia [3]

[1] Information Technology Department, Chandigarh Engineering College, IK Gujral Punjab Technical University, Mohali, India
[2] Information Technology Department, Chandigarh Engineering College, IK Gujral Punjab Technical University, Mohali, India
[3] Information Technology Department, Chandigarh Engineering College, IK Gujral Punjab Technical University, Mohali, India

[*]*Corresponding Author: hutashan20@gmail.com, Tel.: +91-97188-35810*

Available online at: www.ijcseonline.org

Accepted: 17/Jun/2018, Published: 30/Jun/2018

***Abstract*—** Recommendation system is used to generate the recommendations on the basis of the input and processed data. This study develops a web usage data mining based recommendation system. To perform the classification and pattern matching is the major task of a recommendation system. For the purpose of classification, traditional recommendation system prefers the KNN classifiers but the issue was that the KNN performs the classification and pattern matching on the basis of the nearest neighbour and thus in this manner, it lacks the scalability and fails to perform the exact matches for the recommendation system. Therefore the Naïve Bayes classifier is implemented and analyzed in this study for the recommendation system and after simulation, it is found that the Naïve Bayes classifier generates the highly accurate and error-free recommendations for the users. The JAVA platform is used for simulation and the results are evaluated in the form of Accuracy, Error Rate, RMSE and Precision.

***Keywords*—**Data Mining, Web Usage Data mining, Classification, Pattern Matching, Naïve Bayes Classifiers.

I. INTRODUCTION

Web usage mining or web-based data mining falls under the category of data mining or web mining. The web usage mining is done to mine the web-based data such as information accessed from the web or web pages etc. The web usage data facilitate the leading access to the web pages available over the web. It is mandatory to gather the web usage based information for the purpose of web data mining and it can be done by accessing the logs through web servers. By gathering and mining the web usage data, the companies or organizations can enhance their profit, sales as the web usage data mining would make them capable to the produce more effective and qualitative results. Web usage data mining is also found to be helpful for creating marketing skills to outperform the rivals and to promote the business at a higher level. It is not embellished to state that the internet or World Wide Web has seen to have a great interference and impact on the human life. As it has changed out the way to learning, business, service providing etc. along with this the way of information gathering, assigning and communication means has been changed drastically.

II. RELATED WORK

In the recent internet-based era, web usage data mining act as a key tool for establishing and garnishing the e-commerce based businesses. Thus the web usage data mining has become the most prominent topic for research works and till now, a large number of surveys and researches have been conducted in this domain to introduce advanced mining techniques.

A. Data Classification

For the purpose of effective and qualitative data mining, the authors utilize the concepts such as classification, pattern matching etc. Among the classification, the KNN has become most popular and suitable classifier for this purpose. The KNN can also be employed for pattern matching as it matches the patterns on the basis of the factors or parameters that are defined by the users.

In [1], RSS (Really Simple Syndication) reader website was developed by simply considering the closed match of URLs to the user choice. But after having a review to these studies, it has been extracted that the traditional classification technique [1, 2] lacks at the point of accuracy due to poor selection criteria i.e. the selection was performed on the basis of the nearest neighbour which was detected on the behalf of the distance factor only. It did not bother whether the selected nearest neighbour is an exact match or not. Thus it leads to the less effective in the results of recommendation system. So the proposed model will first extract the information based on Keywords, Tags and Age of the user to

© 2018, IJCSE All Rights Reserved

classify users into distinct clusters. These clusters are then processed by the Naïve Bayes Classifier to provide the recommendation to the user. Thus the recommendations provided will not only be based on the distance of the nearest neighbour but it will also include closure sets of attributes in tags.

In [2], a two-stage algorithm using self-organizing map (SOM) and fuzzy k-means with an improved distance function to classify users into clusters was developed. Results from the combination of SOM and fuzzy K-means revealed better accuracy in identifying user-related classes or clusters.

B. Problem in traditional Classifiers

In [11], the author implements the KNN for web usage mining by using the fuzzy sets to it. After performing the simulation, the observations delineates that the fuzzy inference system with crisp values manages the low error rate in the system. Whereas in [12] the KNN was also implemented corresponding to the five other classification schemes in order to collaborate the web server based logs and web-based information. This was done with an objective to classify the user's surfing pattern and on this basis of this pattern, the user's future requests were predicted. The simulation proves the KNN as the most suitable classifier in comparison to other classifiers. But the study explained that the KNN worked on basis of nearest neighbour approach [1]. The fact was that this work lacks scalability and capability of selecting the exactly matched pattern for the recommendation system. Thus its recommendation quality and accuracy of some were doubtful since it performs the pattern matching on the basis of the distance or nearest neighbour only which is not a factor of matching the section as no information of content is there only a match count is there.

III. PROPOSED WORK

After discussing the bottlenecks of the traditional work in the previous section, it has been concluded that there is a requirement to develop a new recommendation system in order to overcome the backlogs of traditional one. Thus, this study is organized with an objective to introduce an enhanced recommendation system for web usage data mining. The proposed or current system performs the pattern matching on the basis of the following factors:

- Keywords matching.
- URLs extraction.
- Age of user matching.

These factors are introduced to overcome the shortcoming of traditional recommendation system which only considered the distance as a factor to select the nearest neighbour. Another amendment is done by introducing the Naïve Bayes classifiers for the purpose of classification. The advantage of Naïve Bayes classifiers over other classifiers is that it is highly scalable and it is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Thus it leads to the generation of exact patching and is capable induces the capability of the recommendation system in the terms of accuracy, precision rate etc.

A. Data Acquisition

The proposed recommendation system recommends the online data on the basis of the age of the users and category of the dataset. The age group is divided into two categories where the first category is related to the Sports and Entertainment and the second category is related to the Politics, Business and World. The proposed Naïve Bayes Classification based recommendation system initiates the implementation firstly by creating a dataset. The proposed dataset is categorized on the basis of the following categories:

- Sports.
- Politics.
- Entertainment.
- Business.
- World.

Under above-defined categories, the data is gathered from online resources in the form of inbound links and keywords. The following table shows the respective URLs for considered categories.

Table 1. Sample Dataset Classes

Category	URLs
Sports	http://www.cricbuzz.com/
Politics	http://www.business-standard.com/politics
Entertainment	http://www.rediff.com/movies
Business	https://www.businesstoday.in/
World	https://www.nbcnews.com/news/world

B. Feature Extraction

After creating the dataset, next step is to extract the information from created dataset. The information such as URLs, keywords and links are extracted from the dataset. On the basis of the extracted dataset, the training of the dataset is performed.

C. Classification and Pattern Matching

The Naïve Bayes classifier is found to be prominent to overcome the shortcomings of traditional KNN classifier. As the Naïve Bayes classifier works on the basis of the Bayesian

© 2018, IJCSE All Rights Reserved

theory and it assigns the most likely class to an available instance of the class. In Naïve Bayes classifier, it is assumed that the impact of an attribute on a class is statistically independent to rest of the attributes in the class. Regardless of this assumption, the Naïve Bayes classifier is proved to be more efficient from the point of computations, simple to implement, scalable for real-world applications the following formulation delineates the Bayes rule that is used for the classification purpose.

$$P(Y|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|Y)P(Y)}{P(X_1, \ldots, X_n)}$$

Likelihood, Prior, Normalization Constant

Initially, the prior probability is evaluated and then the probability for respective factors that is keyword name, inbound links, age group and keyword count is evaluated. After this, on the basis of the probability of individual factors the final probability is evaluated. The scoring factor is evaluated by combining the final probability and initial prior probability. After this if the score is evaluated to be greater than the mean of final probability then the recommendation system will generate the recommendation otherwise it will not.

IV. RESULTS AND DISCUSSION

This section depicts the results that are observed by implementing the proposed recommendation system. The Naïve Bayes classifier is implemented to generate the decision from recommendation system. After training the testing is done by considering the different portions of the dataset.

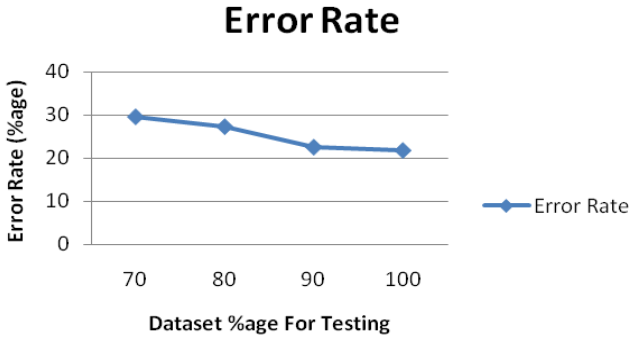

Figure 1. Error Rate of proposed recommendation system

The graph in figure 1 shows the Error Rate of the tested dataset. The error rate is depicted with respect to the different portion of the tested dataset as shown by the x-axis of the graph. First of all the testing is done on 70% dataset, then 80%, 90% and 100%. The y-axis in the graph calibrates the data in the form of the error rate (%) that ranges from 0 to 40.

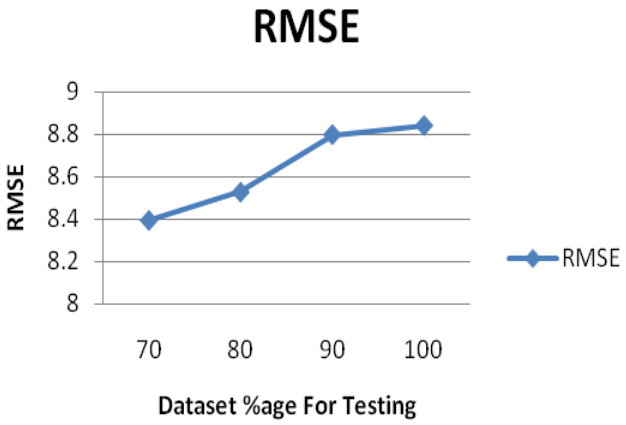

Figure 2. RMSE of proposed recommendation system

The graph in figure 2 depicts the RMSE of proposed recommendation system. The RMSE refers to the Root Mean Square Error. The following formulation is used for evaluating the RMSE. It refers to the larger absolute error thus it is mandatory to have lower RMSE for an ideal recommendation system.

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i} |p_{u,i} - r_{u,i}|} \qquad (1)$$

Where $p_{u,i}$ denotes the predicted rating for the user, $r_{u,i}$ depicts the actual rating and N refers to the total number of rating on the available dataset.

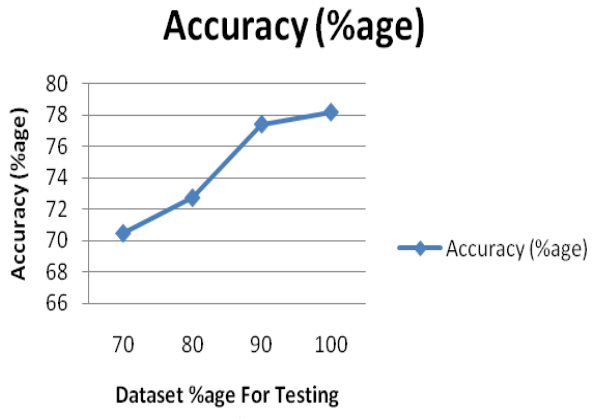

Figure 3. Accuracy of proposed recommendation system

The graph in figure 3 shows the accuracy rate of proposed work. Accuracy matrix is used for the purpose of evaluating the exact accuracy rate of predicted decision with the actually recommended decision.

© 2018, IJCSE All Rights Reserved

The graph in figure 4 represents the precision of proposed recommendation system after simulating it. The Precision is known as positive predictive value.

$$Precision = \frac{Correctly\ Recommended\ Items}{Total\ \ recommended\ Items} \qquad (2)$$

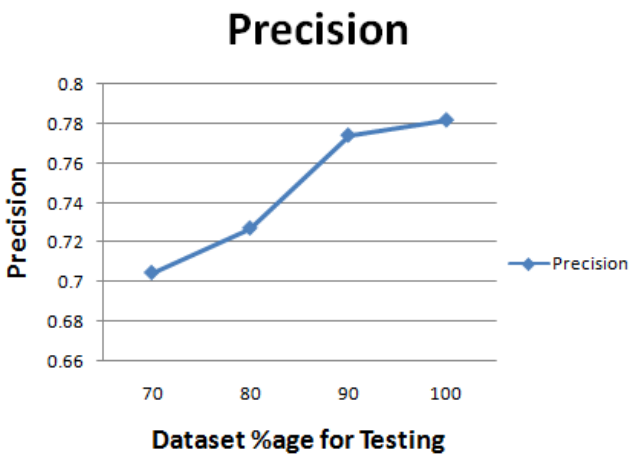

Figure 4. Precision of proposed recommendation system

Table 2 comprised of facts and figures corresponding to the performance metrics that are evaluated in above section. On the basis of the values that are shown in the table below, it is observed that the all of the performance parameters are evaluated to be efficient and better in all of the test cases of the work. The testing with 100% data produces the best results as the error rate, in this case, is 22.58065%, Accuracy is 78.18182%, the RMSE is 8.842049 and Precision is 0.721818%.

Table 2. Performance Analysis

Testing Data %age	Accuracy (%age)	Error Rate	RMSE	Precision
70	70.46632	29.53368	5.43449	0.704663
80	72.72727	27.27273	5.22233	0.727273
90	77.41935	22.58065	4.75191	0.774194
100	78.18182	21.81818	4.670993	0.781818

V. CONCLUSION

A large number of 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 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 testing of the proposed recommendation system is done on various datasets as shown in table 2 and observed to be effective.

As this study is organized as a proposal work and also found to be effective individually, thus, in future to assure the efficiency and effectiveness of the proposed recommendation system, a comparative analysis could be generated with respect to the traditional KNN classifier based recommendation system.

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.

© 2018, IJCSE All Rights Reserved

[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.
[22] K.Reka, T.N.Ravi, *"Hybrid Recommender Systems: Process, Challenges, Approaches and Metrics"*, International Journal of Computer Sciences and Engineering, Vol. 6, No. 5, pp. 1024-1033, 2018.
[23] S.N. Patil, S.M. Deshpande, Amol D. Potgantwar, *"Product Recommendation using Multiple Filtering Mechanisms on Apache Spark"*, International Journal of Scientific Research in Network Security and Communication, Vol. 5, No. 3, pp. 76-83, 2017

Authors Profile

Mr. Hutashan Vishal Bhagat pursed 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.

Dr. Shashi Bhushan 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.

Mr Sachin Majithia 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 & is pursuing 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.

© 2018, IJCSE All Rights Reserved