http://dx.doi.org/10.21172/ijiet.123.11

Data Mining and Recommendation System

Sandeep Kaur[1]
[1]*Department of Computer Science and Engineering*
Punjab Technical University, Jalandhar(Punjab),India

Abstract— In this work, an investigation of automatic web usage data mining and recommendation framework is presented in light of current client conduct through his/her click stream information on the recently grew Really Simple Syndication (RSS) reader site, keeping in mind the end goal to give significant data to the person without explicitly requesting it. The hybridization of K-Nearest-Neighbor (KNN) characterization technique and Fuzzy Logic technique has been prepared to be utilized on-line and in Real-Time to distinguish customers/guests click stream information, coordinating it to a specific client gathering and recommend a modified browsing alternative that address the issue of the particular client at a specific time. To accomplish this, web clients RSS address document was extricated, cleansed, designed and gathered into significant session and information store was created. Our outcome demonstrates that the K-Nearest Neighbor classifier is straightforward, steady, direct, easy to understand, high inclination to have attractive characteristics and simple to execute than most other machine learning procedures particularly when there is almost no earlier information about information conveyance.
Keywords— Data Mining, K-Nearest-Neighbor, Knowledge Discovery in Database.

I. INTRODUCTION

Data mining is an automatic process which is used to remove the meaningful information from the data storage and further use this removed information for various purposes [1,2]. The meaningful data can be extracted using pattern matching techniques, which can be achieved by anomalies analysis, dependencies analysis, cluster analysis, Spatial indices are used to perform all above functions or processes [3,7]. The matched pattern is a form of brief summary of data stored in the data warehouse and these patterns are used for future prediction and various decision making systems to take right decision [4,6]. For example in case of machine learning systems this extracted information can be used for prediction analysis. Example, it is a procedure which examines different groups of linked data in the database which can be used for predictive analysis in near future. Data analysis, data collection and data compilation are involved in the process of KDD i.e. Knowledge Discovery [5,8,9]. These are the additional steps in KDD.

II. KNOWLEDGE DISCOVERY in DATA BASE

On the basis of KDD the data mining procedure is occurred. The KDD is Knowledge Discovery in Database. The KDD process includes various stages. The number of stages included in this is 5. These are as follows:
(1) Selection of data
(2) Pre-processing of selected data
(3) Transformation of data
(4) Data Mining
(5) Interpretation/Evaluation.

III. CLASSIFICATION ALGORITHMS

1) KNN :- KNN stands for K-Nearest Neighbor calculation which is generally utilized for categories [10, 12]. To classify the novel and classified data set the trained data set was saved by the KNN that is a sort of occasion based learning algorithm. The closest neighbor is found based on distance by utilizing the distance metrics. A distance metric is the operation in view of the real values for x, y and z coordinates.

$$d(x,y) \geq 0, \text{and } d(x,y) = 0 \text{ iff } x = 0 \ldots\ldots\ldots (6)$$

In other words the KNN is a non parametric paradigm. The word 'non parametric' means in this method no suspicions are required for distribution of information [11, 13]. Hence for practical implications this strategy is extremely effective as the hypothetical suppositions don't fit in reality.
Additionally, the non parametric KNN is additionally lazy algorithm that implies for any sort of generalization the training information points are not required. In KNN no explicit training stage is necessitated. Accordingly training stage in KNN algorithm executes rapidly [14]. No generalization implies every one of the information is kept by KNN. And all the training information is required while executing the testing stage. The lazy algorithm takes decision based on entire data set that experiences training.
The training stage in this method is very small yet testing stage is extremely expensive. Here the word cost depicts about time and capacity unit [15].

3.1 Assumptions in KNN

In this calculation it is viewed that the information as a characteristic or metric space. The data can be in any shape either scalar or vector frame. Where the points are considered as feature space therefore the idea of distance presented here.

All the training data sets are involved vectors and these vectors are coordinated with the class label. The class label can be either positive or negative. On differing various gatherings this algorithm can be executed.

In KNN algorithm, the "k" chooses the quantity of neighbors and the estimation of 'k' influences the characterization in KNN calculation. Basically the estimation of k begins from 2 and if in case the estimation of k is 1 then it is alluded as NN algorithm.

3.2 KNN for Classification

In order to classify objects on the basis of closet training like in Feature space, KNN algorithm is used. It is basically a lazy learning where all computations are delayed until classifications and functions are approximated in the vicinity. It is basically said to be as a technique where there is very less or we can say has no knowledge of the data. A class represented by k-nearest neighbors in training set is assigned along with retainment of complete training set are accomplished by this technique. NN rule i.e. Nearest Neighbor rule is considered to be as the easiest rule of KMM where K=1. Each method is categorized same as that of its surrounding samples such that if this classification is unknown then it can be predicted through its neighbor sample. All the value of distance can be figured, if the value of unknown sample and all the samples in training set are given. The value of distance in training set having small difference is given to the unknown sample. Thus based on its nearest neighbor, the unknown sample can be classified. The figure 1 given divides the set of samples into 2 classes having value of k=1 and k=4 for KNN decision rule. Only one unknown sample is used for classification of an unknown sample in figure 4.1(a) whereas in figure 1(b) more than one sample is used. In last case value of K is set as 4 so that for classification of unknown sample, four samples can be kept under consideration where three samples are part of same class and remaining one sample is part of other class. Class which is at the left is taken under consideration for the classification of unknown sample in both of the above cases. Figure 2 gives the view of the KNN algorithm.

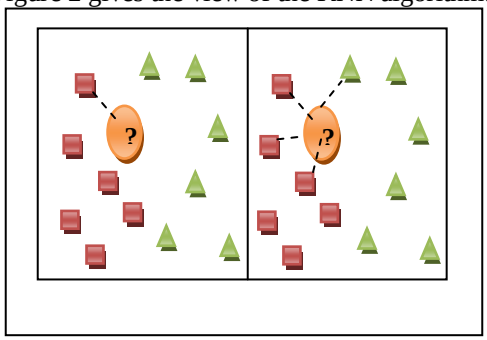

Figure 1 KNN Classifications

Distance applied and the choice of K is used for checking the performance of KNN classifier. As the radius of local region is evaluated by distance of kth nearest neighbor and its class probabilities along with selecting its neighbor size, it leads to affect in overall estimate. The local estimates lean to be poor and gets noisy and mislabeled point in regards to value of K if it is getting low. For making improvement in the estimate, the value of K needs to be increased. But the increase in value of K leads to decrease in overall classification performance and over smoothing along with introducing outliers from other classes. Various methods for improving the performance has been generated so as to overcome this problem. But selecting the required size of k is still an issue as it can affect the overall performance. Even the selection of neighbor size K can also be affected by training sample size which leads to reduction in performance. The distance of kth nearest neighbor is evaluated by the probability of query in local region that leads to overall performance of KNN. The change in distance of nearest neighbor and closer ones indicating class of query object leads to combining of class labels of KNN. Many new weighted methods are generated for KNN so as to reduce the issue of various choices of neighborhood size K.

IV. PROBLEM FORMULATION

The main issue of numerous on-line sites is the representation of numerous decisions to the customer at once; this generally results to strenuous and time consuming task in finding the right product or information on the site. In the traditional approach KNN based clustering techniques were proposed which were used for the recommendation process. But these have some major issue if the data is going to be varied the clustering approach that were used in

traditional work can only capable if the data variation was within the cluster information they are having if data goes out of bound it was difficult to perform classification. So there is need to add a classifier approach so can work in these conditions too.

V. PROPOSED WORK

The main motive of the data mining is to mine those part of the data warehouse which meaningful and helpful in different decision making processes. Sometimes the stored data in warehouse is not meaningful or say important for every purpose. It is knowledge getting process. Data mining also performs online updating, complexity consideration, pre-processing, visualization, data management.

It search the large amount of data, performs the pattern matching. Data mining also becomes a term which is repeatedly use at the place of large scale data processing or information processing and where the data processing includes the processes like collection of data, extraction of information, analysis of data warehouse etc. it is also substituted with the process decision making systems, AI, machine learning etc. the jargons used for data mining like data analysis on large scale database or data analytics or machine learning and AI is much appropriate terms.

As it was studied from the literature that for the data mining the KNN was proposed and have some problem in case of classification if the data is going to be changed of values are out of bound in the cluster so there will some problems that can be faced during the decision time so there is need to hybridized the proposed work of paper with a classifier which will be capable to take decision in the worst condition also so in this proposed model an Fuzzy Logic and KNN clustering approach hybridization is taken as the proposal work so can be capable to take decision in the data variation cases.

VI. METHODOLOGY

Initial step is to take the dataset on which the recommendation decision is to took
Calculate the KNN based clustering information from the dataset
Train the system with applying the KNN input to for the training purpose
User inputs for the recommendation purpose
Pass the user input to Fuzzy Logic for the classification purpose
Fuzzy Logic and KNN hybrid model will perform the data mining and do recommendation.

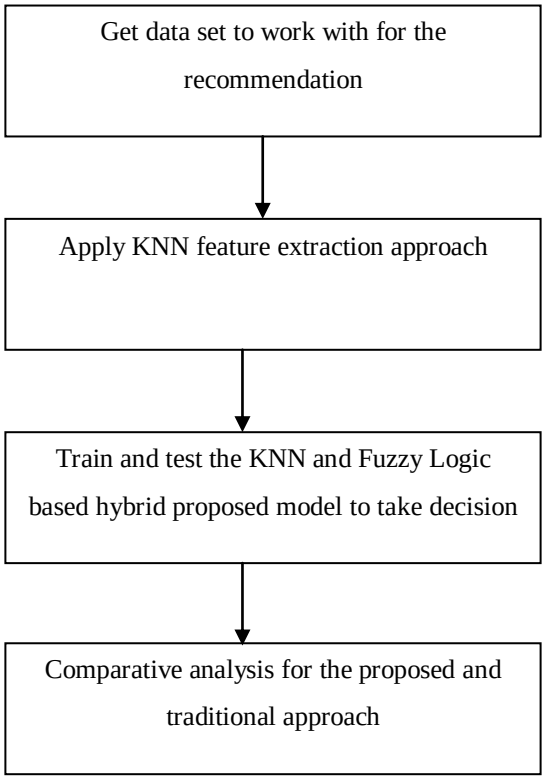

Figure 2 Methodology of the proposed work

http://dx.doi.org/10.21172/ijiet.123.11

VII. RESULTS

The main problem observed from conventional recommendation framework is identified with the characterization and number of features utilized for classifying the recommendations. This investigation means to build up a proficient web-based recommendation framework by utilizing some significant parameters, for example, URLs, keywords, inbound links and outbound links. With the end goal of characterization, the Mamdani fuzzy inference model (FIS) is utilized. Alongside this, the idea of user authorization and verification is additionally represented to estimate whether a user is authorized for the recommendation framework or not. The user's authorization for recommendation framework is characterized based on the News type. This segment portrays the outcomes that are acquired after implementing the proposed recommendation framework in MATLAB.

The graph of Figure 3 shows the Different Rules for the Different Users. The Decision Fuzzy technique is used to offer better decisions for the different users in order to provide more accuracy. The Decision Fuzzy technique decides that what output or recommendation should be related to which user. Here in this graph there are 5 rules for the 5 users for which the recommendations or outputs are also 5.

In this study the Decision making Fuzzy Logic technique is used to offer more accuracy in the work. There are some rules defined by the Decision Fuzzy in order that different rules for the different users so that every user can attain their exact recommendation. The simulation results obtained by the proposed method are shown below as:

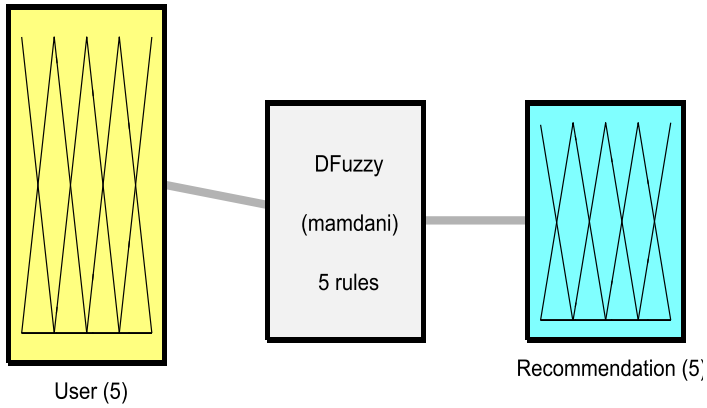

Figure 3 Different Rules for the Different Users.

The graph of Figure 4 shows the Euclidean Distance of the other users to the unknown users that are X1, X2, X3, X4, X5 and X6. The graph shows that if the recommendation is obtained by the user then the distance is 0 and if the recommendation is not obtained by the user then the distance is 1. The Formula of Euclidean Distance is used to evaluate the distance that is shown in the equation below as:

$$distt(x_m, x_n) = \sqrt{\sum_{i=1}^{j} (x_{mi} - x_{ni})^2} \ldots\ldots (1)$$

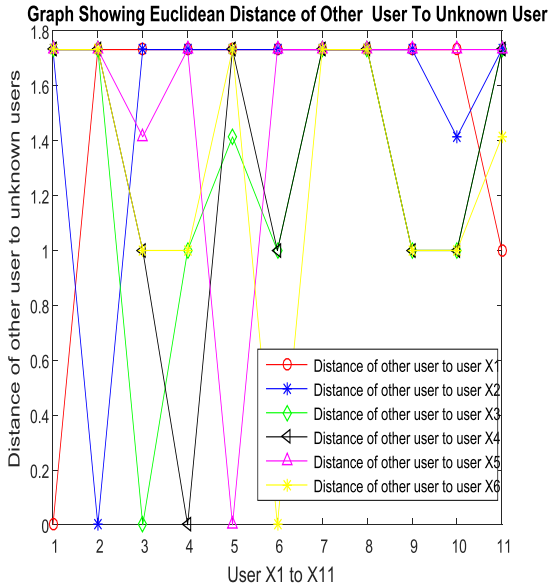

Figure 4: Distance between the users from X1 to X6

The graph of Figure 5 shows the Fuzzy Classification. In this graph the users are shown on the x-axis that ranges from 1 to 11. On the y-axis the recommendation field is shown that ranges from 0 to 12. The concept of Fuzzy offers more accuracy in the recommendation.

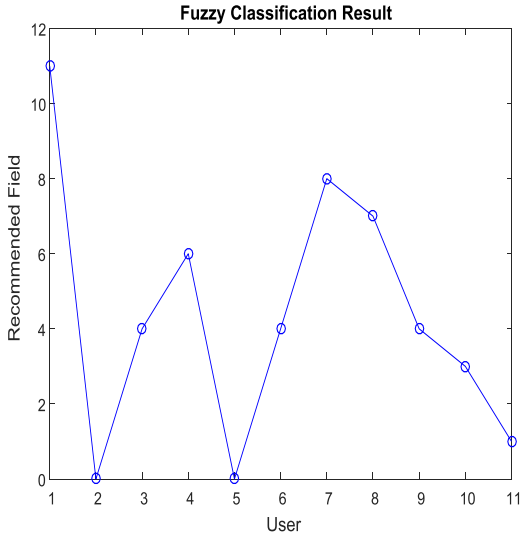

Figure 5 Fuzzy Classification

The graph of Figure 6 shows the Euclidean Distance of the other users to the unknown users. The number of users is shown on the x-axis that ranges from 1 to 11 and the distance of the other users to the unknown users are shown on the y-axis that ranges from 0 to 1.8. With respect to the recommended URL the distance is evaluated. The user has o Distance to whom the URL is recommended and the other users have a distance nearby 1. This graph shows the distance of other users to the user X7 to X11.

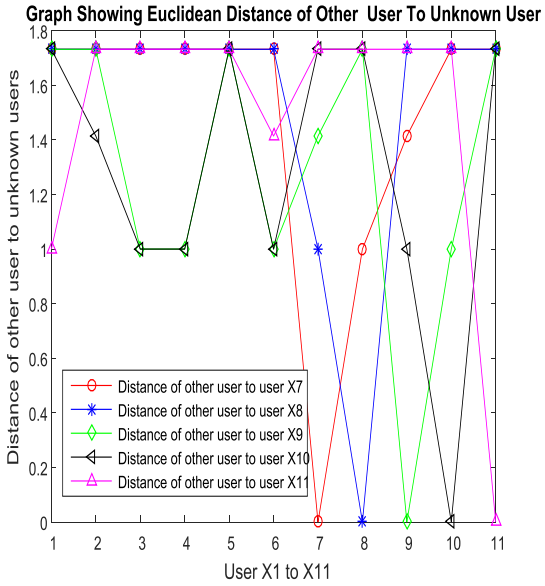

Figure 6: Distance between the users from X1 to X11.

The graph of Figure 7 shows the accuracy proposed by the concept of Fuzzy. The proposed fuzzy is shown on the x-axis of the graph and the percentage of the accuracy is shown on the y-axis that ranges from 0 to 100%. The proposed Fuzzy concept offers the accuracy upto 90% as shown in the graph.

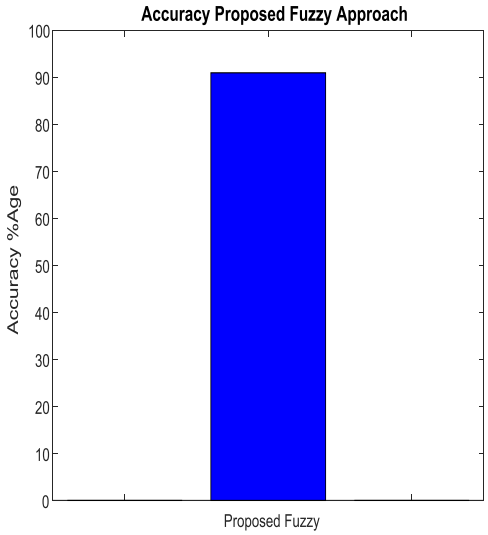

Figure 7 Accuracy proposed Fuzzy method

The graph of Figure 8 depicts the comparison of the accuracy offered by the conventional method that is KNN and the proposed method that is Fuzzy approach. The traditional KNN offers accuracy upto 75 % and the proposed Fuzzy method offers accuracy upto 90 %. The graph shows that the Proposed Fuzzy mechanism offers more accuracy than the traditional method.

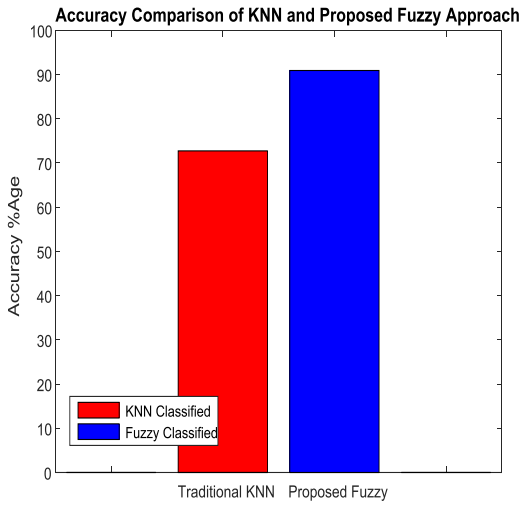

Figure 8 Comparison of Accuracy

VIII. CONCLUSION

In broadening recommendation parameter model, the search won't reliant on the URL coordinate yet in addition the elements those will give the inside data. The final decision is arrived at on the basis of URL matching and the proposed recommendation parameters. This examination builds up an automatic recommendation system for web usage mining. The proposed recommendation framework deals with the premise of the Fuzzy Inference Model and Euclidean Distance. The execution of proposed recommendation framework is observed to be exceptional than conventional suggestion framework. The proposed work of paper is hybridized with a classifier to take decision in the worst condition also so in this proposed model the Fuzzy Logic and KNN clustering approach hybridization is taken as the proposal work so can be capable to take decision in the data variation cases.

As the proposed method offers better results but more amendments can be done in future. In future the proposed suggestion framework can be improved by expanding the number of data sets. Alongside this, the idea of clustering technique is also used for the feature extraction purpose.

IX. REFERENCES

[1] Shengsheng Shi, Chengfei Liu, Yi Shen, Chunfeng Yuan, Yihua Huang , "AutoRM: An effective approach for automatic Web data record mining", Elsevier, vol 89, pp 314-331, 2015.
[2] Petar Ristoski, Heiko Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey", vol 36, pp 1-22, 2016.
[3] Petar Ristoski, Christian Bizer, Heiko Paulheim, "Mining the Web of Linked Data with RapidMiner", Elsevier, vol 35, pp 142-151, 2015.
[4] Viktor Medvedev, Olga Kurasova, "A new web-based solution for modelling data mining processes", Elsevier, vol 76, pp 34-46, 2017.
[5] Bhukya shankarnayak, K. Venkatesh sharma, Betala rakesh, "Mining the data using aggregator from the dynamic web page", Elsevier, vol 5, pp 980-993, 2018.
[6] Sumaiya Kabir, Shamim Ripon, Mamunur Rahman, Tanjim Rahman, "Knowledge-based Data Mining Using Semantic Web", Elsevier, vol 7,pp 113-119, 2014.
[7] Christopher Newman, Zach Agioutantis, Nathaniel Schaefer, "Development of a web-platform for mining applications", Elsevier, vol 28, pp 95-99, 2018.
[8] D.A. Adeniyi, Z. Wei, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", ELSEVIER, Journal of Applied Computing and Informatics, Vol. 12, pp. 90-108, 2014.
[9] Anitha Talakokkula, "A Survey on Web Usage Mining, Applications and Tools", Computer Engineering and Intelligent System, IJSTE, Computer Engineering and Intelligent System, Vol. 6,Issue 2, pp. 22-30, 2015.
[10] Satya Prakash Singh , Meenu, "Analysis of web site using web log expert tool based on web data mining", IEEE, International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) , 2017.
[11] Yeqing Li, "Research on Technology, Algorithm and Application of Web Mining", IEEE, International Conference on Computational Science and Engineering (CSE) and International Conference on Embedded and Ubiquitous Computing (EUC) , Vol. 1, pp. 772-775, 2017.
[12] Viktor Medvedev, Olga Kurasova, Gintautas Dzemyda, "A new web-based solution for modelling data mining processes", ELSEVIER, Simulation Modeling Practice and Theory, Vol. 76, pp 34-46, 2016.
[13] Petar Ristoski, Heiko Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey", ELSEVIER, Vol. 36, pp. 1-22, 2016.
[14] Venkata Subba Reddy Poli, "Fuzzy data mining and web intelligence", IEEE, International Conference on Fuzzy Theory and Its Applications (iFUZZY), 2016.
[15] Zoltán Balogh, "Data-mining behavioral data from the web", IEEE, International Conference on Software, Knowledge, Information Management & Applications (SKIMA), Vol.1, pp. 122-127, 2016.