Volume 8, No. 4, May 2017 (Special Issue)
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

# Recommendation system with Automated Web Usage data mining using K-Nearest Neighbor(KNN) classification

Er. Jyoti[1]
Deptt.of CSE[1]
Sant Baba Bhag Singh University[1]
Jalandhar, India[1]
jyot.paul@gmail.com[1]

Er.Amandeep Singh Walia[2]
Deptt.of CSE[2]
Sant Baba Bhag Singh University[2]
Jalandhar, India[2]
er.amanwalia@hotmail.com[2]

***Abstract*-**The major problem of many on-line web sites is the presentation of many choices to the various clients at a time. This usually results into time consumingtask in finding outthe right product or information on the site. The user's current interest depends upon the navigational behavior which helps the organizations to guideusers in their browsing activities and obtain some relevant information in a short span of time. Since, the resulting patterns which are obtained through data mining techniques did not perform well in the prediction of future browsing patterns because of the low matching rate of resulting rules and of user's browsing behavior. This paper focuses on the study of the automatic web usage data mining and recommendation system which is based on current user behavior through his/her click stream data. The K-Nearest-Neighbor (KNN) classification method has been trained to be used in real-time and on-line to identify clients and visitors click stream data, matching it to a particular user group and recommends a tailored browsing option that meet the needs of the specific user at particular time.

***Keywords:-Automated;*** Data Mining;K-Nearest Neighbor; Recommendation system ; Web Usage Mining.

## I. INTRODUCTION

A) *Data Mining*:Data mining is the process which comes under thecategory of computer science in order to investigate large data sets which belongs to the pattern. Here large data set stands for Big Data. Data mining is an automatic process which is used to extract meaningful information from the data storage and further use this information for various purposes [15,16].The extraction of meaningful data can be performed by matching patterns and it is achieved by cluster analysis, anomalies analysis, and dependencies analysis. Spatial indices are used toperform all above functions or processes. 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]

For example in case of machine learning systems this extracted information can be used for prediction analysis. Another example, data mining is a process which find or investigates various groups of correlated data in the database which further can be used for predictive analysis in near future [17]. Data analysis, data collection, compilation of data is not a connected to the data mining but still included in the process of KDD i.e. Knowledge Discovery.[1]

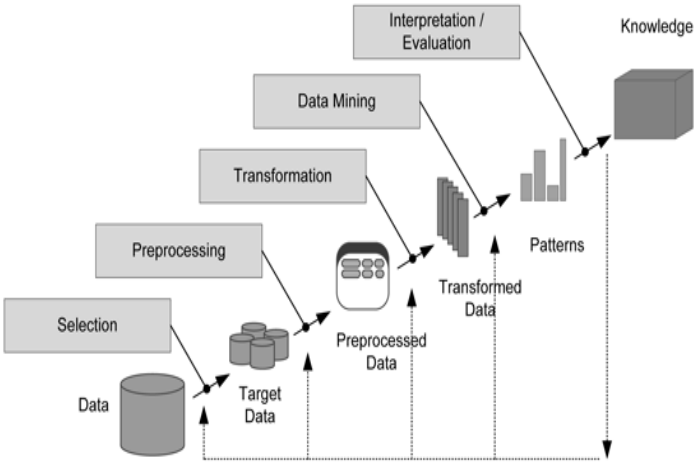


Figure 1 : Knowledge Discovery in Databases

Data mining is a process which is used to search large amount of data in order to find the useful data. The goal of this technique is to find patterns that were previously unknown. Once the patterns are found they can further be used to make certain decisions for the development of their businesses.

The iterative process consists of the following steps:

- **Data cleaning**: It is a phase in which noisy data and irrelevant data are removed from the collection.
- **Data integration**: At this stage, multiple data sources, mostly heterogeneous, may be combined in a common source.
- **Data selection**: At this step, the data which is relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation:**Data transformation is also known as Data Consolidation. It is a phase in which the selected data is transformed into appropriate forms for the mining procedure.
- **Data mining:** It is the crucial step in which clever techniques are applied to extract patterns which are potentially useful.



CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)

- **Pattern evaluation**: In this step, strictly interesting patterns representing knowledge are identified based on the given measures.
- **Knowledge representation:**It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help the users to understand and interpret the data mining results.[2]

*B) Web Usage Mining***:**It isone of the applications of the techniques of data mining to discoverand find out interesting patterns from the Web data. Usage data captures the origin or identity of Web users along with their browsing behaviour at the Web site. It usually focuses on the techniques which predict user behaviour while the user is interacting with the Web.[9,11,13]The potential strategic aims in each domain into mining goal as: prediction of the user's behaviour within the site, comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users. There are no such definite distinctions between Web usage mining and other two categories. In data preparation of Web usage mining, the Web content and Web site topology will be used as the information sources, which interacts Web usage mining with the Web content mining and Web structure mining.[10] Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining.[3]

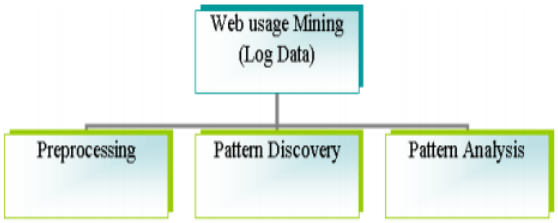


Figure 2: Contents of Web Usage Mining

*1) The Usage Of Mining On The Web*

It is one of the application of the techniques of data mining to discover and find out interesting patterns from the Web data. With the assistance of the diagram of the high-level Web usage mining process shown in Figure 1, may understand the architecture of the Web Usage Mining easily.

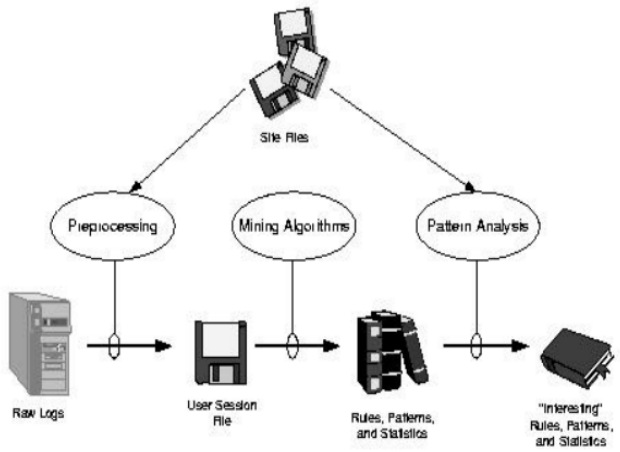


Figure 3: Web usage mining

## II. ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.[14]

These are explained as below:-

*A. Classification:*It is a technique in data mining, which employs set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and creditrisk applications are particularly well suited to this type of analysis. This approach sometimes employs decision tree and neural network-based classification algorithms. [8]The data classification process involves learning and classification. In Learning the training data are analysed by classification algorithm. In classification the test data are used for estimating the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. [5]The algorithm then encodes these parameters into a model called a classifier.
Different Types of classification models:

- Classification by decision tree induction
  - Bayesian Classification
  - Neural Networks
  - Support Vector Machines (SVM)
  - Classification Based on Associations

*B. Clustering:*Clustering can be defined as the identification of similar classes of objects. By using the clustering techniques we can also identify sparse and dense regions in the object space and can discover the overall pattern distribution and correlations among various data attributes. Classification approach can also be used for the effective means of distinguishing groups and classes of object but this

becomes costly. Thus clustering can be used as a pre-processing approach for the attribute subset classification and selection. [5]For example: - to form the group of customers which is based on purchasing patterns and to categories genes with similar functionality.

Types of clustering methods:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

*C. Prediction:*For prediction, Regression technique can be adapted. Regression analysis can be used to model the relationship between one or more dependent variables and independent variables. In data mining, independent variables are the attributes which are already known and response variables are those we want to predict. Unfortunately, many of the real-world problems are not simply prediction. For instance, stock prices, sales volumes and product failure rates are all very difficult to predict because they may depends on the complex interaction of multiple predictor variables. Thus, more complex techniques (e.g. Decision trees, logistic regression and neural nets) may be necessary to forecast the future values. The same model types can mostly be used for both classification and regression. For instance, the CART (Classification and Regression Trees) decision tree algorithm can be used to building up both the classification trees (to classify the categorical response variables) and regression trees (to forecast continuous response variables). Neural networks can also create both classification and regression models. [5]

*1) Types of regression methods:*

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

*D Association Rule:*Basically the Association and correlation are used to find out the frequent item set findings among large data sets. These types of finding helps in business and to make various decisions such as catalogue design, cross marketing and customer shopping behaviour analysis. Association Rule algorithms need to be able to generate the rules with confidence value less than one. Since, the number of possible Association Rules for the given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. [5,7]

Types of association rule:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

*E. Neural networks:* Neural network is a set of connected input/output units and each of the connection has a weight present with it. During the learning phase, network learns by adjusting the weights so as to predict the correct class labels of the input tuples. Neural networks have the ability to derive meaning from thecomplicated and imprecise data and this is used to extract patterns and to detect the trends which are very complex to be noticed either by humans or by other computer techniques. These are suited for the continuous valued inputs and outputs. For instance, handwritten character reorganization, for training a computer to pronounce English text and many real world business problems have already been applied in many of the industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of neural networks

- Back Propagation

*F. K-NN Algorithm:*According to Leifa non-parametric method of pattern classification popularly known as K-Nearest Neighbor rule was believed to have been first introduced by Fix and Hodges in 1951, in an unpublished US Air Force School of Aviation Medicine report[6,12] .The method however, did not gain popularity until the 1960's with the availability of more computing power, since then it has become widely used in pattern recognition and classification .K-Nearest Neighbor could be described as learning by analogy, it learns by comparing a specific test tuple with a set of training tuples that are similar to it. It classifies based on the class of their closest neighbors, most often, more than one neighbor is taken into consideration hence, the name K-Nearest Neighbor (K-NN), the ''K'' indicates the number of neighborstaken into account in determining the class.[11] The K-Nearest-Neighbor (KNN) classification method has been trained to be used on-line and in real-time to identify clients/visitors click stream data, matching it to a particular user group and recommend a tailored browsing option that meet the need of the specific user at a particular time.[1]

K-Nearest Neighbor classifier is used for pattern recognition and classification in which a specific test tuple is compared with a set of training tuples that are similar to it. The K-Nearest Neighbor (K-NN) algorithm is one of the simplest methods for solving classification problems; it often yields competitive results and has significant advantages over several other data mining methods. [18]

(1) Providing a faster and more accurate recommendation to the client with desirable qualities as a result of straightforward application of similarity or distance for the purpose of classification.
(2) Our recommendation engine collects the active users' click stream data, match it to a particular user's group in order to generate a set of recommendation to the client at a faster rate.

The K-Nearest Neighbor classifier usually applies the Euclidean distance between the training tuples and the test tuple.

$$d(x_i, x_t) = \sqrt{(x_{i1} - x_{t1})^2 + (x_{i2} - x_{t2})^2 + \ldots + (x_{ip} - x_{tp})^2}$$

In general term, the Euclidean distance between two Tuples for instance





CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)

X1 = (x11, x12, . . .. . .. . .. . . x1n) and X2 = (x21, x22, . . .. . .. . .. . ...x2n) will be

$$\text{dist}(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

## III. PROBLEMS

The major problem of many on-line web sites is the presentation of many choices to the client at a time; this usually 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.

## IV.CONCLUSION

As all information is available on the Internet but it is not easy for every user to find relevant information in short span of time. In order to overcome this problem recommendation system introduced in Web world. In this paper, the problem and various techniques are explained for recommendation models. In this work, knn algorithm is used. Since the data doesn't remain within the bound so if we use the hybridisation of KNN(K-Nearest Neighbor classification) and ANN (Artificial Neural Network) then recommendation system can be improved. In future, the implementation of this proposed model will be provided. Their performance will be compared in terms of error rate, memory required, time consumed.

## REFERENCES


[1] Jiawei Han and MichelineKamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[2] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases".http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996Fayyad.pdf Retrieved 2008-12-17.

[3] Dr.R.Lakshmipathy, V.Mohanraj, J.Senthilkumar, Y.Suresh, "Capturing Intuition of Online Users using a Web Usage Mining" Proceedings of 2009 IEEE International Advance Computing Conference (IACC 2009)Patiala, India, 6-7 March 2009

[4] PhridviRaj MSB., GuruRao CV (2013) Data mining – past, present and future – a typical survey on data streams. INTER-ENG Procedia Technology 12:255 – 263

[5] Srivastava S (2014) Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. International Journal of Computer Applications (0975 – 8887) 88:.10

[6] L. L. Tang, J. S. Pan, X. Guo, S. C. Chu, and J. F. Roddick, "A novelapproachonbehaviorofsleepylizardsbasedonK-nearest neighbor algorithm," in Social Networks: A Framework of ComputationalIntelligence,vol.526ofStudiesinComputational Intelligence,pp.287–311,Springer,Cham,Switzerland,2014.

[7] A.Gosainand M.Bhugra,"A comprehensive surveyof association rules on quantitative data in data mining," in Proceedings of the IEEE Conference on Information & Communication Technologies (ICT '13), pp. 1003–1008, JeJu Island, Republic of Korea,April2013.

[8] Luca, C., Paolo, G., 2013. Improving classification models with taxonomy information. J. Data Knowledge Eng. 86, 85–101. http://dx.doi.org/10.1016/j. datak.2013.01.005.

[9] Hitesh Hasija and Deepak Chaurasia;" Recommender System with Web Usage Mining basedon Fuzzy C Means and Neural Networks"; NGCT-2015

[10] Prajyoti Lopes and BidishaRoy;"Dynamic recommendation system using web usage mining for E-commerce users";ICACTA-2015

[11] D.A. Adeniyi, Z. Wei,and Y. Yongquan;" Automated web usage data miningand recommendation system usingK-Nearest Neighbor (KNN)classification method";Applied Computing and Informatics (2016).

[12] Hiral Y Modi, MeeraNarvekar ."Enhancement of Online web Recommendation System Using A Hybrid Clustering And Pattern Matching Approach": 2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2015).

[13] HimangniRathore, HemantVerma, "Analysis on Recommended System for Web Information Retrieval Using HMM", International Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, November 2014.

[14] S.Sharma, P.Sharma,"Use of Data Mining in Various Fields", In IOSR Journal of Computer Engineering, IOSR-JCE Volume 16, Issue 3, Ver. V May-Jun. 2014 .

[15] S.Kaviarasan,K.Hemapriya,K.Gopinath,Semantic Web Usage Mining Techniques for Predicting Users' Navigation Requests, International Journal of Innovative Research in Computer Science and Communication Engineering,Vol. 3,Issue 5,[ISSN:2320- 9801],2015.

[16] B.Lalithadevi, A. Mary Ida,W.Ancy Breen, A New Approach for Improving World Wide Web Techniques in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 3,Issue 1,[ISSN:2277 128X],2015.

[17] Shu-Hsien, C. Pei-Hui, H. Pei-Yuan, Data mining techniques and applications- A decade review from 2000 to 2011, Journal of expert system with applications 39 (2012) (2012).

[18] ShihuaCai,LiangxiaoJiang,DianhongWang.Survey of Improving K-NN for Classification. International Journal of Advanced Research in Computer Science and Software Engineering.




CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)