International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

"A Review on Recommendation System and Web Usage Data Mining using K-Nearest Neighbor(KNN) method"

Er.Jyoti[1], Er.Amandeep Singh Walia[2]

[1]*M.Tech Scholar, Department of Computer Science &Engineering, S.B.B.S.U, Jalandhar, Punjab, India*
[2]*Assistant Professor, Department of Computer Science & Engineering, S.B.B.S.U, Jalandhar, Punjab, India*

Abstract - *Data Mining is a extraction of knowledge from large amount of Observational datasets. In Web world, there is immense of information available on the internet but user is not capable to find relevant information in short period of time. Therefore, a system called recommendation system developed to assist user in their browsing activities. It analyzes users need and provides relevant information in shorter span. K-NN algorithm helps to know the users behavior and his interest. This paper focuses on recommendation systems based on the user's navigational patterns and provides suitable recommendations to cater to the current needs of the user.*

Key Words: ***Automated , Data Mining, K-Nearest Neighbor, Recommendation system, Web usage Data Mining.***

1. INTRODUCTION

1.1 Data Mining

It is the process of analyzing data from different perspectives and generally summarizing it into useful information. Data mining is a knowledge discovery process. It performs the data management and pre-processing, visualization, complexity consideration and online updating.

1.2 Web Usage Mining

It is one of the application of the techniques of data mining to discover and find out interesting patterns from the Web data. Usage data captures the origin or identity of Web users along with their browsing behavior at the Web site. It usually focuses on the techniques which predicts user behaviour while the user is interacting with the Web. [16,17,18]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 the process of 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.[20] Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining.[19]

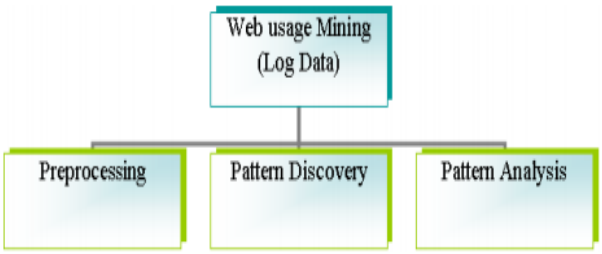

Fig 1: Contents of Web Usage Mining

K-NN Classification Method - The K-Nearest-Neighbor (KNN) method has been used to on-line and real-time to identify clients/visitors click stream data, matching it to a particular user group. K-NN algorithm is mainly used for Pattern recognition.

2. K-NN ALGORITHM

According to Leif a 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. 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 neighbors taken into account in determining the class. 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 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.

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

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

3. LITERATURE REVIEW

Zdravko and Daniel [3], described web data mining as application of data mining techniques to discover patterns in web content, structure and usage. It is a branch of applied artificial intelligence that deals with storage, retrieval and analysis of web log files in order to discover users accessing and usage pattern of web pages [2].

The use of classification and regression tree (CART) was adopted by Amartya and Kundan [4] in their work. In constructing a decision tree, they applied both the gini index(g) and entropy value (e_i) as the splitting indexes, the model was experimented with a given set of values, different sets of results were obtained for both the outlook, humidity, windy, Temp, and Time for execution. The result of the experiment shows that the best splitting attribute in each case was found to be outlook with the same order of splitting attributes for both indices.

Resul and Ibrahim [2], in their work used the path analysis method to investigate the URL information of access to the Firat University web server, web log file so as to discover user accessing pattern of the web pages, in order to improve the impressiveness of the web site. They explain further that, the application of path analysis method provides a count of number of time a link occur in the data set, together with the list of association rules which help to understand the path that users follow as they navigate through the Firat University web site.

Decision rule and Bayesian network, support vector machine and classification tree techniques were used by Rivas et al. [5], to model accidents and incidents in two companies in order to identify the cause of accident. Data were collected through interview and modeled. The experimental result was compared with statistics techniques, which shows that the Bayesian network and the other methods applied are more superior than the statistics technique.
Rivas et al. [5], stated further that the Bayesian/K2 network is of advantage as it allows what-if analysis on data, which make the data to be deeply explored.
Self Organizing Map (SOM) or Kohonen neural network model was explored by Xuejuu et al. [6], in their work, to model customers navigation behavior. The model was used to create clusters of queries based on user session as extracted from web log with each cluster representing a class of users with similar characteristics, in order to find the web links or product of interest to a current user on a Real-Time basis. The experimental result of the SOM model performance was compared with that of K-Means model, and the SOM model was found to outperform the K-Means model with value of correlation co-efficient of SOM model scoring twice that of K-means result.

Many researchers have attempted to use K-Nearest Neighbor classifier for pattern recognition and classification in which a specific test tuple is compared with a set of training tuples that are similar to it. [7], in their own work introduced the theory of fuzzy set into K-Nearest Neighbor technique to develop a fuzzy version of the algorithm. The result of comparing the fuzzy version with the Crisp version shows that the fuzzy algorithm dominates its counterpart in terms of low error rate

The K-Nearest Neighbor algorithm was used alongside with five other classification methods to combine mining of web server logs and web contents for classifying users' navigation pattern and predict users' future request. The result shows that the KNN outperformed three of the other algorithms, while two of them performed uniformly. It was also observed that KNN archives the highest F-Score and A(c) on the training set among the six algorithms. [8], as well adopted the KNN classifier to predict protein cellular localization site. The result of the test using stratified cross validation shows the KNN classifier to perform better than the other methods which includes binary decision tree classifier and the naïve Bayesian classifiers.

In Mobasher et al[9] present Web personalizer a system which provides dynamic recommendations, as a list of hypertext links, to users. This analysis is based upon

anonymous usage data combined with the structure formed by hyperlinks of the site.

WUM system which is called SUGGEST,as proposed by Baraglia and Palmerini, which provides very useful information to make easier the web user navigation and to optimize the web server performance [10,11]. A two level architecture was adopted by SUGGEST which is composed of offline creation of historical knowledge and online engine that understands user's behavior. Since a request arrives on the system module, it incrementally updates a graph representation of web site based on the active user sessions and classifies the active session using a graph partitioning algorithm.
The Potential limitation of the architecture may be:
a) the memory required to store Web server pages in quadratic in the number of pages .This might be severe limitation in larger sites made of million pages;
b)it does not permit us to manage web sites made up of pages dynamically generated. All the work attempts to find the architecture and algorithm to improve accuracy of personalized recommendation, but accuracy still does not meet satisfaction.

D.A. Adeniyi et al, proposed that 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 this work, a study of automatic web usage data mining and recommendation system based on current user behavior through his/her click stream data on the newly developed Really Simple Syndication (RSS) reader website, in order to provide relevant information to the individual without explicitly asking for it. 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. To achieve this, web users RSS address file was extracted, cleansed, formatted and grouped into meaningful session and data mart was developed. Our result shows that the K-Nearest Neighbor classifier is transparent, consistent, straightforward, simple.

4. PROPOSED WORK

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 ANN and KNN clustering approach hybridization is taken as the proposal work so can be capable to take decision in the data variation cases.

5. CONCLUSION

This paper gives the brief literature review about the designing and developing a recommendation system. In most of the cases KNN algorithm was used . K-NN Classification Method is very efficient and reliable method to know users behavior and interest at a particular session. Thus hybridization of KNN with ANN can be done. Thus by comparing these two techniques we can get more accurate result. Thus, we can increase the efficiency of the system.

REFERENCES

[1] Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method/ D.A. Adeniyi, Z. Wai, Y. Yongquan/ Science direct 2014.
[2] D. Resul, T. Ibrahim.Creating meaningful data from web log for improving the impressiveness of a web site by using path analysis method.Journal of expert system with applications, 36 (2008) (2008), pp.66356644 http://dx.doi.org/10.1016/j.eswa.2008.08.067
[3] M. Zdravko, T.L. Daniel.Data mining the Web, Uncovering patterns in Web content, structure, and usage.John Wiley & sons Inc., New Jersey, USA (2007) p. 115–132.
[4] S. Amartya, K.D. Kundan, Application of Data mining Techniques in Bioinformatics, B.Tech Computer Science Engineering thesis, National Institute of Technology, (Deemed University), Rourkela, 2007.
[5] T. Rivas, M. Paz, J.E. Martins, J.M. Matias, J.F. Gracia, J. Taboadas.Explaining and predicting workplace accidents using data-mining Techniques.Journal of Reliable Engineering and System safety,96(7)(2011),pp.739747 http://dx.doi.org/10.1016/j.ress.2011.03.006.
[6] Z. Xuejuu, E. John, H. Jenny.Personalised online sales using web usage data mining.J. Comput. Ind., 58 (2007) (2007), pp.772782 http://dx.doi.org/10.1016/j.compind.2007.02.004.
[7] M.K. James, R.G. Michael, A.G. James, A fuzzy K-Nearest Neighbor Algorithm. IEEE Transactions on System Man and Cybernetics, vol. SMC-15 No4.[0018-9472/85/0700-0580$01.00], 1985.
[8] H. Paul, N. Kenta, Better Prediction of Protein Cellular Localization Sites with the K-Nearest Neighbor Classifier, ISMB-97, Proceeding of America Association for Artificial Intelligence, USA, 1997, pp. 147–152.
[9] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on Web usage mining" Communications of the ACM, vol. 43, pp. 142-151, 2000.

[10]M. Perkowitz and O. Etzioni, "Towards adaptive Web sites: Conceptual framework and case study," Artificial Intelligence, vol. 118, pp. 245-275, 2000.

[11]M. Jalali, N. Mustapha, A. Mamat, Md N. Sulaiman, "OPWUMP An architecture for online predicting in WUM-based personalization system", In 13th International CSI Computer Science, Springer Verlag,2008. 307

[12]Amartya, S., Kundan, K.D., 2007. Application of Data mining Techniques in Bioinformatics. B.Tech Computer Science Engineering thesis, National Institute of Technology, (Deemed University), Rourkela.

[13]Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In EMNLP, pages 79–86.

[14] A. Balahur, R. Steinberger, M. A. Kabadjov,V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news." in LREC, 2010.

[15] J. Ali, R. Khan, N. Ahmad, and I. Maqsood,"Random forests and decision trees." nternational Journal of Computer Science Issues (IJCSI), vol. 9, no. 5, 2012.

[16]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.

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

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

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

[20].Himangni Rathore, Hemant Verma, "Analysis on Recommended System for Web Information Retrieval Using HMM", International Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, November 2014.