Analysis of Web Performance based on Navigation Pattern using Progressive Web Datasets

Sapana Kumari
Computer Science & Engineering
Lakshmi Narain College of
Technology Excellence
Bhopal, India

Vikram Garg
Computer Science & Engineering
Lakshmi Narain College of
Technology Excellence
Bhopal, India

ABSTRACT
Now a day's the e-commerce websites are facing the biggest challenges of the massive growth of website data. There is a need to find behavior of user so that there is need to find next page in advance on cache. Further they do not have a good policy for finding user behavior in website. They need to use good approach for improving the quality and accuracy in today's scenario. Closed sequential pattern mining is an important technique among the different types of sequential pattern mining, since it preserves the details of the full pattern set and it is more compact than sequential pattern mining. In this paper the clustering task is used to improve performance of website navigation pattern in advance. The main goal of this research is to find the extract the knowledge that can enhance web performance of associate items in sequential manner with the quality.

Keywords
e-commerce datasets, knowledge mining, Decision Making, Data Classification, Performance Prediction.

1. INTRODUCTION
Web mining is a current research area where some need to improve the performance of website. Web mining having three types- web content mining, web structure mining, and web usage mining. In web usage mining there is need to find the next page prediction in advance and reduce the database scan or partial scan.

In the association rule mining which follow sequential navigational pattern. It uses some parameter just like minimum support and no. of cluster for finding the frequent items. To reduce the generation of item sets here closed sequential pattern uses in two ways. The first approach is greedily finding the final closed sequential patterns; this approach is more complicated because it is hard to verify the closeness of pattern without checking with the previously discovered patterns. The second approach is to find the closed sequential pattern candidate set and conduct post pruning on it, this approach requires storing the discovered patterns but with recent advances in technology we can store million patterns in main memory. The second approach is follow in this research paper.

2. RELATED PREVIOUS WORK
Md. Hedayetul Islam Shovon, et. al. (2012) [1] give a presentation on 'Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm'. In this presentation data clustering technique named k-means clustering is applied to analyze student's learning behavior.

Kavita Nagar et. al. (2015) [2] presents a paper "Data Mining Clustering Methods:" in this paper gives Clustering is an unsupervised learning method which makes the cluster of objects or documents according to their similarity and dissimilarity bases. This paper gives review about various clustering methods.

Oyelade, O. J et. All (2010) [3]presents a paper"Application of k-Means Clustering algorithm for prediction of Students' Academic Performance" In this paper, it also implemented k-mean clustering algorithm for analyzing students' result data. It provided a simple and qualitative methodology to compare the predictive power of clustering algorithm and the Euclidean distance as a measure ofsimilarity distance.

Miss. Vrinda Khairnar and Miss. Sonal Patil, [2016] [11] proposed to improving the performance of the k-means clustering algorithm. In this paper discusses about the different techniques and improvements of K-means clustering algorithm based on different research papers referred. These methods include Refined initial cluster center's method, a parallel K-means algorithm, a parallel k-means clustering algorithm based on Map Reduce technique, determine the initial centroids of the clusters and assign each data point to the appropriate matching clusters.

3. PROBLEM DEFINITION
In web usage mining there is a some need to improve the web performance. For the improvement of the web performance there are two option, first put item in cache in advance for better response time and second thing use partial database scan not fully database scan.

In the FP-Tree there is also need to reduce the number of database scan. So that accuracy increases and response time are reducing.

In the current scenario there is also need to maintain threshold value for finding the frequent data in the dataset and finding frequent item sets in less time.

4. PROPOSED MODEL FOR CLUSTERING OF WEB DATA

4.1 Proposed Framework

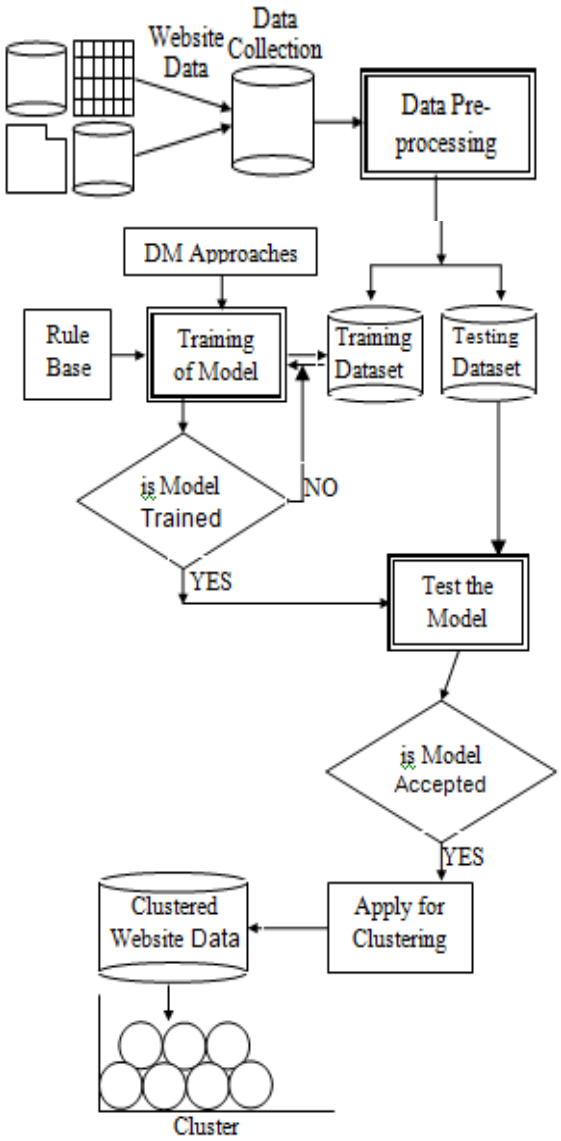

Fig. 1: Approach for Clustered Website Dataset

4.2: Algorithms for proposed model.

Step 1. Collection of web data from website.

Step 2. Apply pre-processing techniques to remove noise from data.

Step 3. Divide the data into two parts one is training data (80%) and testing data (20%).

Step 4. Choose appropriate data mining approach to design a model.

Step 5. Trained the model as per given rule base.

Step 6. If the model is not trained, go to step 5.

Step 7. Test the model using Test data.

Step8. If model is accepted, use it for clustering & prediction purpose.

5. IMPLEMENTATION AND RESULT

All experiments were conducted on a 2GHz Intel Core2 Duo processor PC with 2GB main memory running Microsoft Windows XP. The algorithms were implemented in Asp.Net with C# and were executed using 10% support value. In this experiments a real world dataset BMS Web View of KDD CUP 2000 is used, which having click stream data from an e-commerce web store named Gazelle and it has been used widely to assess the performance of frequent pattern mining. This dataset contains sequences of 59601 customers with a total of 146000 purchases in 497 distinct product categories.

In data mining, data scientists use algorithms to identify previously unrecognized patterns and trends hidden within vast amounts of structured and unstructured information. These patterns are used to create predictive models that try to forecast future behavior.

These models have many practical business applications—they help banks decide which customers to approve for loans, and marketers use them to determine which leads to target with campaigns.

a) Accuracy: Accuracy is an evaluation metrics on how a model perform.

ACCURACY = (TP+TN)/ (TP+TN+FP+FN)

TN is the number of true negative cases

FP is the number of false positive cases

FN is the number of false negative cases

TP is the number of true positive cases

b) Precision: Prediction is a calculation of positive predicted values precision, which is the fraction of retrieved documents that are relevant. The precision is calculated using the formula as:

Precision = (TP) / (TP + FP)

c) Recall: Recall in information retrieval is the fraction of the documents that are relevant to the query and that are successfully retrieved.

Precision=(TP) / (TP + FN)

The experiments are done with 1000, 2000, 3000, 4000 and 5000 records to Calculate Accuracy Precision, and Recall which is shown by Fig. 2, fig. 3 & Fig. 4 respectively.

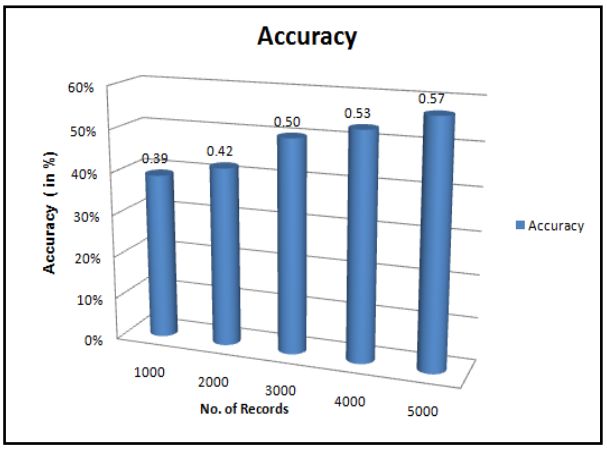

Fig. 2: Accuracy obtained for website by Naïve Bayes method using different no. of records

Fig. 2 showing the accuracy of the preprocessed web dataset by Naïve Bayes Method using different no. of records. If records are 1000, 2000, 3000, 4000 and 5000 then accuracy (in %) are 39%, 42%, 50%, 53% and 57%. So when number of records are increases accuracy also maintain.

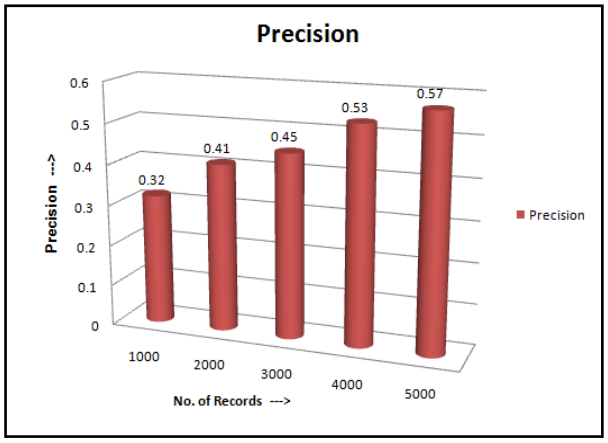

Fig. 3:Precision obtained for website by Naïve Bayes method using different no. of records

Fig. 3 showing the precision of the preprocessed web dataset by Naïve Bayes Method using different no. of records. If records are 1000, 2000, 3000, 4000 and 5000 then precision are 32, 41, 45, 53 and 57. So when number of records are increases accuracy also maintain.

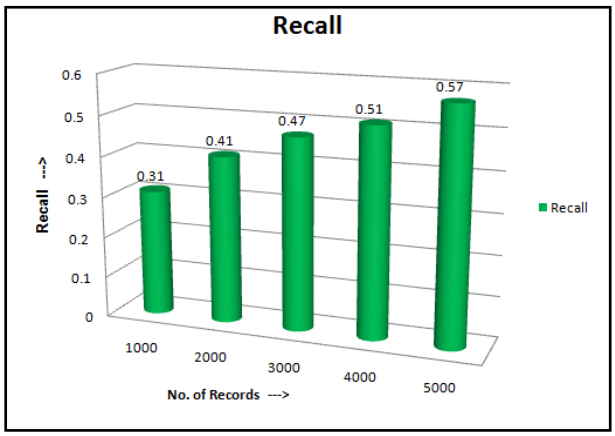

Fig. 4: Recall obtained for website by Naïve Bayes method using different no. of records

Fig. 4 showing the recall of the preprocessed web dataset by Naïve Bayes Method using different no. of records. If records are 1000, 2000, 3000, 4000 and 5000 then recall are 31, 41, 47, 51, and 57. So when number of records are increases recall also maintain.

6. CONCLUSION

This work proposed a framework to classify a knowledge in web dataset. The Naïve Bayes algorithm is used for finding centroid of data, which is used for grouping of data and generate the perfect cluster. The presented work insists the better utilizing of web data in term of getting more insight knowledge which will be helpful to find next item in advance in cache. This framework is used to identify and group the items through their occurrences for used clustering approach to accomplish the web performance. In future this research will prefer to use distributed data from various websites.

7. REFERENCES

[1] Md. Hedayetul Islam Shovon, Mahfuza Haque "Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 2, Issue 7, ISSN: 2277 128X, pp 353-355, July 2012,

[2] Kavita Nagar "Data Mining Clustering Methods: A Review" International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Volume 5, Issue 4, ISS: 2277 128X, pp 575-579, 2015, www.ijarcsse.com.

[3] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C "Application of k-Means Clustering Algorithm for Prediction of Student's Academic Performance", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 1, 2010, pp 292-295.

[4] K. C. Kandpal and R. Agnihotril, Sblock – A closed sequential pattern mining algorithm, International Journal of Computer Applications in Engineering Sciences, vol.1, no.3, pp.296-299, 2011.

[5] S.-Y. Yang, C.-M. Chao, P.-Z. Chen and C.-H. Sun, Incremental mining of across-streams sequential patterns in multiple data streams, Journal of Computers, vol.6, no.3, pp.449-457, 2011.

[6] S.-Y. Yang, C.-M. Chao, P.-Z. Chen and C.-H. Sun, Incremental mining of closed sequential patterns in multiple data streams, Journal of Networks, vol.6, no.5, pp.728-735, 2011.

[7] G. Lee, K.-C. Hung and Y.-C. Chen, Path tree: Mining sequential patterns efficiently in data streams environments, Intelligent Systems and Applications, SIST, vol.20, pp.261-268, 2013.

[8] T. Guyet and R. Quiniou, Incremental mining of frequent sequences from a window sliding over a stream of itemsets, Actes IAF, 2012.

[9] A. Koper and H. S. Nguyen, Sequential pattern mining from stream data, Advanced Data Mining and Applications, vol.7121, pp.278-291, 2011.

[10] L. Vinceslas, J.-E. Symphor, A. Mancheron and P. Poncelet, SPAMS: A novel incremental approach for sequential pattern mining in data streams, Advances in Knowledge Discovery and Management, vol.292, pp.201-216, 2010.

[11] Miss. Vrinda Khairnar and Miss. Sonal Patil, "Efficient clustering of data using improved K-means algorithm", Imperial Journal of Interdisciplinary Research (IJIR), Vol.2, Issue-1, 2016.