Improvement & Analysis of User Classification and Pattern Discovery in WUM

P.S.Shahu[1], D V Jamthe[2], A. A. Nikose[3]
Department of CSE
[1]PG Student, Priyadarshini Bhagwati College of Engineering, Nagpur.
[2, 3]Assistant Professor, Priyadarshini Bhagwati College of Engineering, Nagpur.

Abstract- *On web data, data mining has web mining as an application and web usage mining is an important component of web mining. The foremost objective of web usage mining is to understand the web site usage behaviour that can be achieved through the process of data mining of Web Data and web site user's behaviour through the process of data mining of Web Access Data. Facts obtain from the web using mining can be further used to improve web designs, introduce personalisation services and facilitate browsing effectively. An important application of web mining. The study has been done on this paper is to discuss various type of data preparation techniques to identify the unique user and user sessions. The paper shows the result of various clustering techniques for web log data using algorithm such as k-means and bisecting k-mean . With respect to the same IP address and packet combination clusters are formed. From the analysis it can be concluded that Bisecting K-means gives the better result than that of the K-means algorithm.*

I. INTRODUCTION

Web usage mining is application of data mining technique to discover automatic discover from particular web site. Web mining can be categories into three categories are shown in Fig.1, web content mining, web structure mining and web usage mining. Web content mining is referring to extract information of content of web pages. Web structure mining is used to analyze the links between web pages through the web structure to infer the knowledge. Web usage Mining is extracting the information from web log file which is accessed by users.The main aim of web usage mining is extract information from web pages which access by users. The K-means is very standard algorithm for the clustering process. But it has some disadvantages like cluster quality result depends on the selection of initial centroids, clusters produced are of different sizes, by using this method we also get empty clusters. Bisecting k-means is a enhancement over basic k-means algorithm. As it is based on k-means, it has the merits of k-means and also has some benefits over k-means.

II. RELATED WORK

In this paper [2] they explained the improved k-means clustering algorithm which can define number of clusters automatically and assign required cluster to un-clustered points. The main focus of this paper to overcome the drawbacks of the k-means clustering algorithm. The result of experiments depicts that with the improved k-means algorithm we get the better performance.

In this paper [3] they applied the improved k-means algorithm with respect to the various parameters like date, time, user agent and time taken on for the real data sets collected by the group of institution web servers. Also discussed that clustering is done group web users into same cluster based on users browsing behaviour during a specific time interval. The process cannot be completed until Pre-process is done properly.

In this paper [4] they presented a comparative analysis techniques such as support vector machine and artificial neural network with supervise learning technique. The Objective this is the Classification of user‟s web access pattern accurately after data cleaning concept for the future improvement. They propose data classification result based on accuracy.

III. PROPOSED WORK

For Web log analysis web usage mining techniques can be applied. By analysing the web access log can understand the user behaviour and the web structure. In this proposed work we cluster the web log data by using the techniques k-mean and bisecting k-mean algorithm and evaluate their performance with respect to the parameters like delay (execution time) and accuracy (infected).

Algorithms and Design Modules

3.1 Bisecting K-means

It is a combination of k-Means and hierarchical-clustering. Bisecting k-Means splits into two sub cluster at every bisecting stages instead of partitioning the data into 'k' clusters in each iteration. It is recurring until k clusters are found. Bisecting k-means is more effective when 'k' is large. At first, for the k-means algorithm, the computation includes

all data point of the data set and k centroids. Besides that in every Bisecting step of Bisecting k-means, only one cluster data point and two centroids are arises. Second, Cluster of similar sizes are being produced by the Bisecting k-means, while k-means produce different sizes clusters.

3.2 Algorithm for finding clusters by using k-Means

1. Choose a Cluster for splitting.
2. Discover two sub clusters by using k-means algorithm.
3. Repetition of second step until to get the best matching similar cluster.
4. To reach the desire clusters repeat steps first, second and third.

There are many ways to choose the clusters to make it split.

Web log of the dataset are collected in the pdf format of the college which consist of various details .The details are not in the proper format. In the first step extraction of useful information is done .So here first pdf file read and then parsed. Extraction of IP addresses is done by the regular expansion method. To get the same IP address and packet combination at same time bisection of k-means is done, which is the application of clustering technique by k-means.

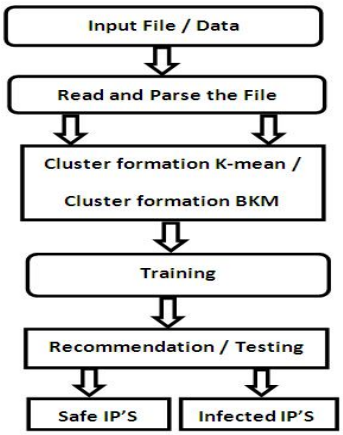

Fig 3: Flow Graph of design Module

IV. EXPERIMENTAL RESULT

Experiment is done by the web log file retrieved. The web log files are in the form of PDF format which is collected from the college.

Step1: Reading and Parsing

The code is executing on the Microsoft Visual Studio. In this c# languages is used. In the first step we have to collect dataset for the pre-processing. For this we maintain a pdf file to collect all the records of the dataset. By using the regular expression we extract the IP address.

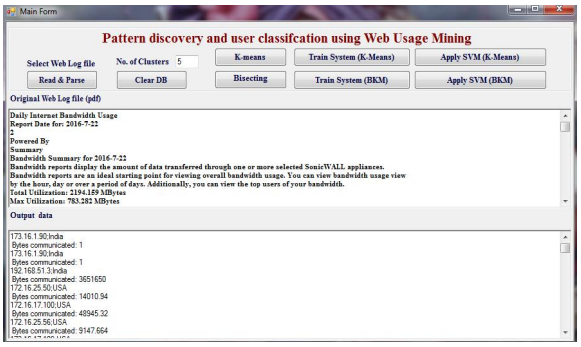

Fig 4.1: Reading and parsing the pdf file

Step 2: Clustering using K-means

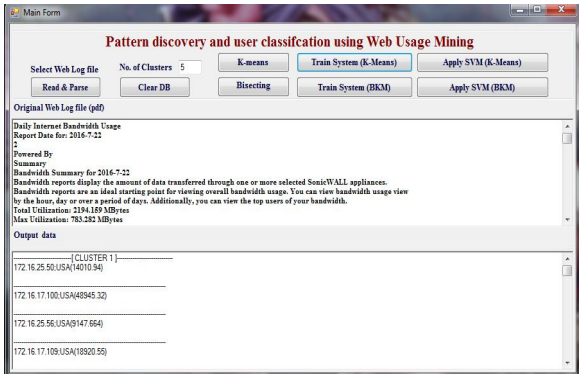

Fig 4.2(a): Formation of clusters using k-means

Fig 4.2(a) shows the result of the k-means clustering algorithm that forms the clusters.

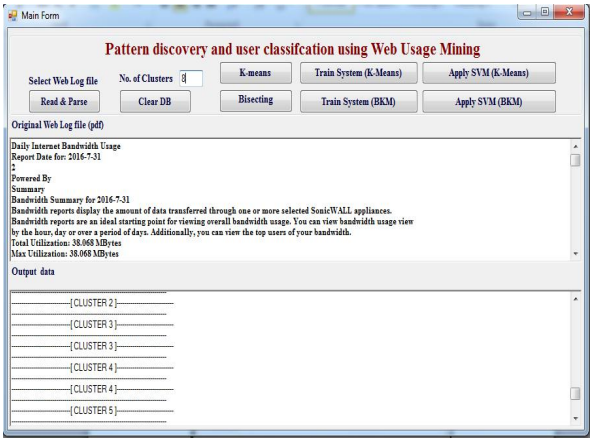

Fig 4.2(b): Formation of various blank clusters using k-means

Fig 4.2(b) shows the output of the k-means clustering. Here we can see various blank clusters are formed by the algorithm.

Step 3: Clustering using Bisecting K-means

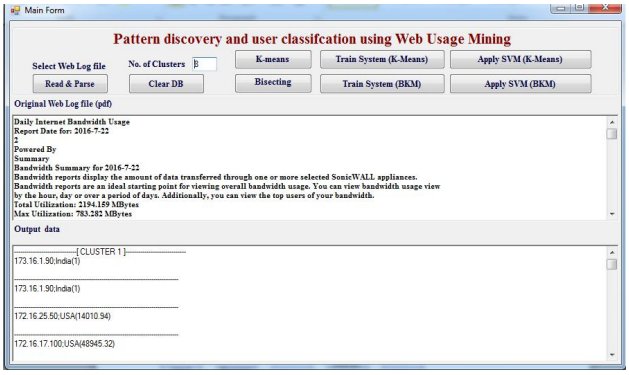

Fig 4 .3(a) : Formation of clusters using BKM

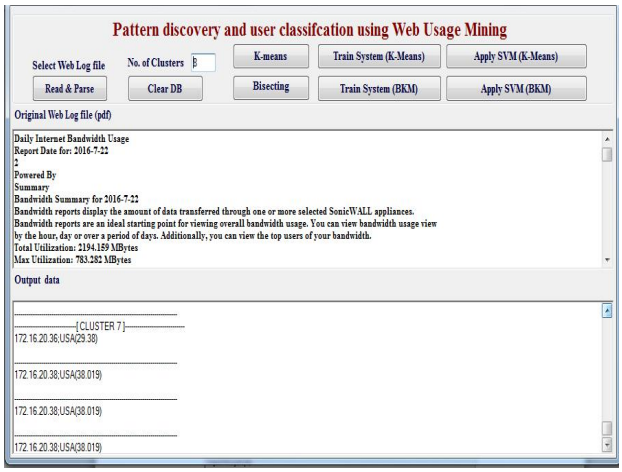

Fig 4 .3 (b) : Formation of clusters using BKM

Fig 4 .3(c): Formation of clusters using BKM

Fig 4.3(a), Fig 4.3(b) and Fig 4.3(c) shows the output executed by the bisecting k-means algorithm. In this we can clearly see that the formations of clusters are uniform. And here we not get any empty clusters.

Step4: Training the system

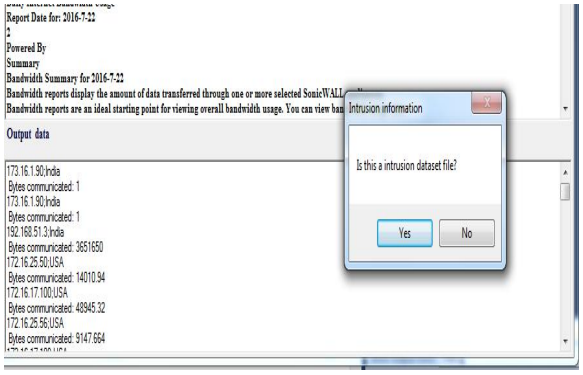

Fig 4.4(a): Training the system

For intrusion data set system asks the end user, as shown in Fig 4.4(a). If user selection is "yes", then the type value is labelled as "1" otherwise it is labelled as "0". Fig. 4.4(b) and Fig 4.4(c) shows respectively, the packets are infected and the normal packets stored in the database while training.

4.4(b) Infected Packets

Fig4.4(c): Normal Packets

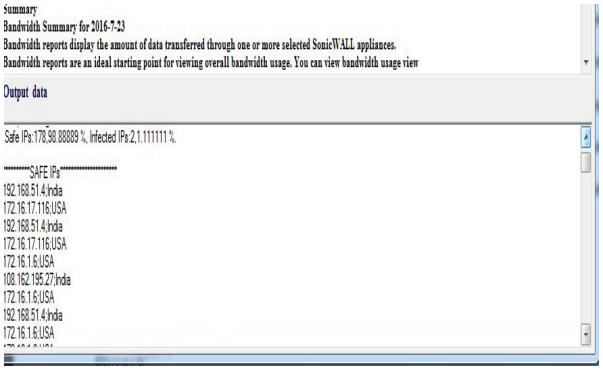

Fig4.4 (d): List of Safe IP'S

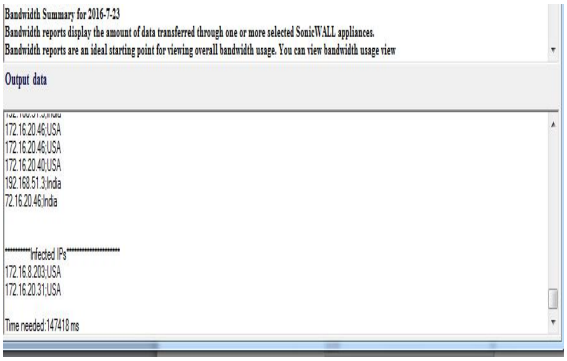

Fig 4.4(e): List of Infected IP's

Fig 4.4(d) and Fig 4.4(e) computes the results of the infected and the safe IP's detected by the system.

V. COMPARISON

The objective of this is to analyze the web usage data by applying techniques such as K-means and Bisecting K-Means clustering algorithms. Here we consider the data file with the related cluster to perform the calculations. We compare the algorithm with respect to the parameter: time and accuracy.

Table 5.1: Comparison of algorithms with respect to time

K-mean(time in ms)	BKM(time in ms)
265045	254634
154288	151488
41622	13160
257444	257784
154818	150618
50722	12490
269765	279405
295326	222492
70654	44472

Fig 5.1: Comparison of K-means and Bisecting K-means (Delay)

Fig shows the comparison of both k-means and bisecting k-means algorithm with respect to the execution time required by them for the given input file. From the graph it can be concluded that the bisecting k-means algorithm give more efficient result then that of the k-means algorithm.

Table 5.2: Comparison of algorithms with respect to infected values (Accuracy)

K –mean (Accuracy in %)	BKM (Accuracy in %)
66.5	81.6
89.5	98.9
94.7	100
61.6	70.53
97.3	93.9
94.7	93.9
61.6	70.53
91.7	99.9
94.7	96.46

Fig 5.2: Comparison of K-means and Bisecting K-means (Accuracy)

From the above graph we can see the result of k-mean and bisecting k-means algorithm in terms of accuracy. We know the accuracy by how much safe IP's we get during the execution in the system. Hence the bisecting gives better accuracy as comparison with the k-mean algorithm.

VI. CONCLUSION

For the data mining, huge volumes of raw data explore with pattern. It refers to the techniques that help us to identify the web content and retrieve user's interest and their needs. This paper focus on the mining algorithms comparison. For the pattern discovery, web usage mining and various criteria are considered. Quality information required to satisfy upcoming needs of user which is allowed by adding huge data in repository. Two clustering algorithm k-means and bisecting k-means used for the formation of clusters. By the result of these two algorithms we conclude that the Bisecting k-means give better performance than that of k-means algorithm. And the execution time required less hence the accuracy get increased.

REFERENCES

[1] Manoj Kumar, Mrs. Meenu "A Survey on Pattern Discovery of Web Usage Mining" 2017IJARIIT

[2] Arpit Bansal, Mayur Sharma,Shalini goel "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining" 2017 IJCA

[3] S.Padmaja ,Dr.Ananthi Sheshasaayee "Clustering of User Behaviour based on Web Log data using Improved K-Means Clustering Algorithm"2016 IJET

[4] Ramandeep Kaur, Asst. Prof.Gauravdeep "To Study Web Pattern Discovery with Web Usage Mining " 2015 IJIACS

[5] K.Dharmarajan, M. A. Dorairangaswamy Discovering User Pattern Analysis from Web Log Data using Weblog Expert" November 2016 IJST

[6] Virendra R. Rathod, Govind V. Patel "Prediction of User Behavior using Web log in Web Usage Mining" 2016 IJCA

[7] M. Aldekhail "Application and Significance of Web Usage Mining in the 21st Century: A LiteratureReview" 2016 IJCTE

[8] M.SANTHANAKUMAR,C.CHRISTOPHER COLUMBUS "Web Usage Based Analysis of Web Pages Using RapidMiner"2015

[9] R. Suganya "Analysis of Implementating Web Usage Mining and I-Miner in E-Business" 2015 IJARCSSE

[10] Yukai Yao, Yang Liu, Yongqing Yu, Hong Xu, Weiming Lv, Zhao Li, Xiaoyun Chen"K-SVM: An Effective SVM Algorithm Based on K-means Clustering"2013 JCP

[11] Dhanamma Jagli, Sangeeta Oswal "Web Usage Mining: Pattern Discovery and Forecasting"2012IFRSA

[12] NeetuAnand, Prof(Dr.) SabaHilal "Identifying the User Access Pattern in Web Log Data"2012 IJCSIT

[13] K.Poongothai ,M.Parimala, Dr. S.Sathiyabama "Efficient Web Usage Mining with Clustering" 2011 IJCSI

[14] Jiaqi Wang, Xindong Wu, Chengqi Zhang "Support vector machines based on K-means clustering for real-time business intelligence systems"2005.