Bisecting K-Means for Clustering Web Log data

Ruchika R. Patil
Department of Computer Technology
YCCE
Nagpur, India

Amreen Khan
Department of Computer Technology
YCCE
Nagpur, India

ABSTRACT
Web usage mining is the area of web mining which deals with extraction of useful knowledge from web log information produced by web servers. One of the most important tasks of Web Usage Mining (WUM) is web user clustering which forms groups of users exhibiting similar interests or similar browsing patterns. This paper presents results of clustering techniques for Web log data using K-means and Bisecting K-means algorithm. Clusters are formed with respect to similar IP address and packet combinations. The clustering framework is further used as an approach for intrusion detection from the log files. The system is trained first by labeling the classes and then tested to check for any intrusions. Recommendation output is generated which help in classifying the whether the input IP's are "safe" or "infected". Comparison of both algorithms is done and performance is evaluated with respect to time and accuracy. From the experimental results, it is found that Bisecting K-means overcomes the major drawbacks of basic K-means algorithm.

Keywords
Web mining, Clustering, Bisecting K-means, Intrusion detection

1. INTRODUCTION
In this internet era, the World Wide Web has become a major source of communication and vast source of information in everyday life. The growth of the internet over the last two decades has resulted in a large amount of data that is available for user access. The user interactions with the Web are recorded and stored in web access logs. Web mining [1][2] is the use of data mining techniques to automatically discover and extract information from Web documents and services.

Web mining can be categorized into three areas of interest based on which part of the Web to mine: 1) Web content mining: refers to discovery of useful information or knowledge from web page contents i.e. text, multimedia data like images, audio, video etc. 2) Web structure mining: aims at analyzing, discovering and modeling link structure of web pages and/or web site to generate structural summary. 3) Web usage mining deals with understanding user behavior while interacting with web site, by using various log files to extract knowledge from them[2][3]

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [1][2]. One of the most important tasks of Web Usage Mining (WUM) is web user clustering which forms groups of users exhibiting having common interests and behavior by analyzing the data collected in the web servers[5]. The K-means is most popular algorithm for clustering and well known for its simplicity and low time complexity. However, it has some major drawbacks like quality of the resulting clusters heavily depends on the selection of initial centroids, clusters produced are of varying sizes, hence unbalanced and may also lead to empty clusters. Bisecting k-means is modification over basic k-means algorithm. As Bisecting k-means is based on k-means, it keeps the merits of k-means and also has some advantages over k-means

In network security, an intrusion is defined as unauthorized action on a single host or a network to access data, use or modify the data and make the network unreliable and unsecured. Intrusion Detection System (IDS) defines a set of hardware or software systems that automate the process of monitoring different activities in a single host or in networks to detect intrusions or indication of intrusions from collected log data or a live network traffic. As the work aims with the web logs, there is also possibility of detecting intrusions in system. Thus, clustering technique can be extended as an approach for intrusion detection in the system.

2. PREVIOUS WORK
Various data mining techniques have been applied and implemented to Web server logs in order to fetch the useful information from abundant data. Amongst them, is clustering technique which is one of the most important task of WUM. Cluster analysis to mine web logs differ from classic clustering process, because the Web log files records various information such as user IP address, request time, requested URL, HTTP status code, referrer and so on. Hence specialized techniques are used for clustering analysis for Web log data.

Khaled Alsabti [3] presented a novel algorithm for performing k-means clustering. It organizes all the patterns in a k-d tree structure such that one can find all the patterns which are closest to a given prototype efficiently. Jin Hua Xu[4] presented vector analysis and K-Means based algorithms for mining user clusters. Clustering web users with K-Means algorithm based on web user log data is done. Experiment results show the feasibility and efficiency of such algorithm application. Bisecting K-means was studied in [9] for WordNet based documents clustering. Experimental results suggested that BKM is much efficient than basic K-means. Mitchael Steinbach[8] presents the results of an experimental study of some common document clustering techniques: agglomerative hierarchical clustering and K-means. The results indicate that the variants of K-means technique are better than the standard K-means and hierarchical approaches.

Also, the Data mining approaches has been known to aid the process of detecting an intrusion in network environment. Thus, clustering algorithms were widely used for intrusion detection. M. Jianlian[10] introduced the application on intrusion detection based on K- means clustering algorithm. K-means is used for intrusion detection to detect unknown attack. Lei Li, et al [11] introduced a novel rule-based intrusion detection system using data mining. They proposed an improvement over apriori algorithm by bringing the concept of length-decreasing support to detect intrusion.

3. PROPOSED WORK
Web usage mining techniques can be applied for web log analysis. Analyzing the web access logs can help understand the user behavior and the web structure, thereby improving the website design. The proposed work is to cluster the web usage data using K-means and Bisecting K-means algorithm, evaluate their performance and then use this clustering technique as an approach for intrusion detection.

3.1 Algorithms and Design Modules

Bisecting K-means

Bisecting k-Means[8] is like a combination of k-Means and hierarchical clustering. Instead of partitioning the data into 'k' clusters in each iteration, Bisecting k-means splits one cluster into two sub clusters at each bisecting step(by using k-means) until k clusters are obtained.

As Bisecting k-means is based on k-means, it keeps the merits of k-means and also has some advantages over k-means.

First, Bisecting k-means is more efficient when 'k' is large. For the k-means algorithm, the computation involves every data point of the data set and k centroids. On the other hand, in each Bisecting step of Bisecting k-means, only the data points of one cluster and two centroids are involved in the computation. Thus, the computation time is reduced. Secondly, Bisecting k-means produce clusters of similar sizes, while k-means is known to produce clusters of widely different sizes

Basic Bisecting K-means Algorithm for finding *K* clusters [8]

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic K-means algorithm. (Bisecting step)
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

There are a number of different ways to choose which cluster is split. For example, we can choose the largest cluster at each step, the one with the least overall similarity, or use a criterion based on both size and overall similarity.

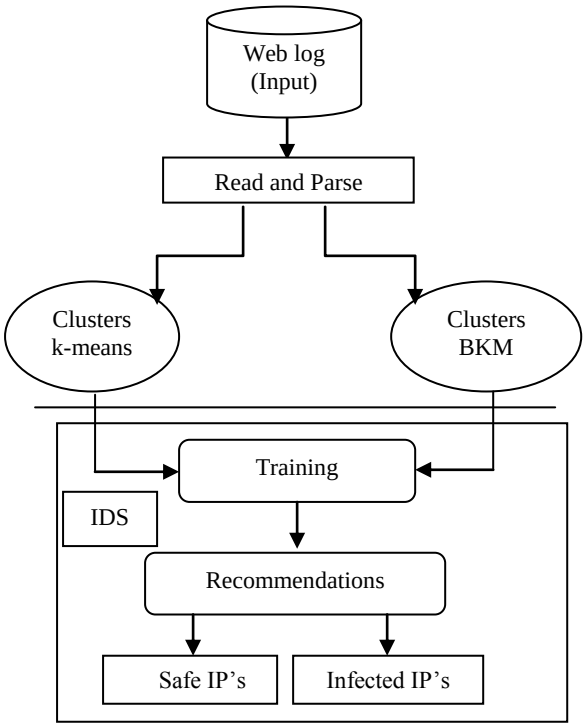


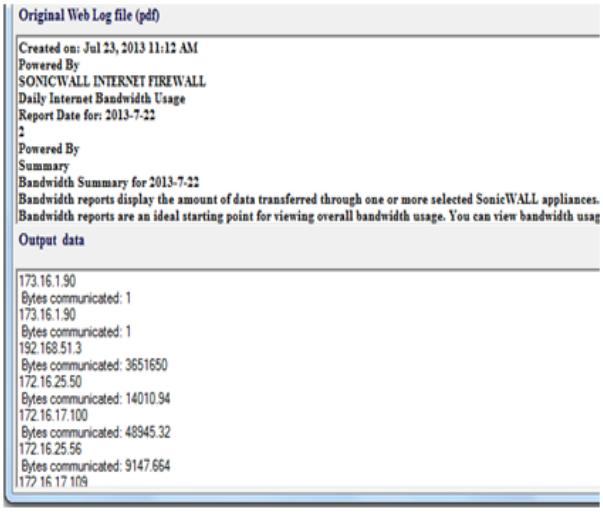
Fig 1: Flow of design Modules

Web log data set, in the form of PDF's are collected from college web site which consists of various reports and summaries. These files are in non standard format. Extraction of useful information is done, which includes taking in account bandwidth and web usage reports and summaries. Here the PDF files are first read and then parsed. Parsing means analyzing a text and converting it into useful form. It consists of displaying of IP address from the Bandwidth reports and the total bytes communicated by it. Application of the clustering technique by K-Means and its modification i.e. Bisecting K-Means is done in order to get similar IP addresses and Packet combinations together. We create a framework which will be able to detect number of intrusions present in the network environment. Training is essential because every recommendation system needs to have a database with which it can compare and recommend if input signatures (IP address, cluster no., packet size etc) are proper or not. Training helps to create a database having samples of values having IP signatures. Comparison of IP address occurrence with the IP address of similar signature in the database is done. We find the best matching IP signature from the database. The class of this best matched signature is given as the class of the input IP. Counts of each classes is done and a Recommendation output is generated.

4. EXPERIMENTAL RESULTS

Experiment was carried out using a log retrieved. The web log files (Log files of Sonicwall) in the form of PDF are collected from reputed Engineering college, YCCE Nagpur.

STEP 1: Reading and Parsing


Original Web Log file (pdf)

Created on: Jul 23, 2013 11:12 AM
Powered By
SONICWALL INTERNET FIREWALL
Daily Internet Bandwidth Usage
Report Date for: 2013-7-22
2
Powered By
Summary
Bandwidth Summary for 2013-7-22
Bandwidth reports display the amount of data transferred through one or more selected SonicWALL appliances.
Bandwidth reports are an ideal starting point for viewing overall bandwidth usage. You can view bandwidth usag

Output data

173.16.1.90
Bytes communicated: 1
173.16.1.90
Bytes communicated: 1
192.168.51.3
Bytes communicated: 3651650
172.16.25.50
Bytes communicated: 14010.94
172.16.17.100
Bytes communicated: 48945.32
172.16.25.56
Bytes communicated: 9147.664
172.16.17.109

Fig 2: Reading and Parsing the file

The input web log file is the PDF file, which is first read by the system. The useful text is extracted from a large data and then parsed. Parsing involves displaying of various IP address and total bytes communicated by them, as shown in Fig 2.

STEP 2: Data Clustering using K-means

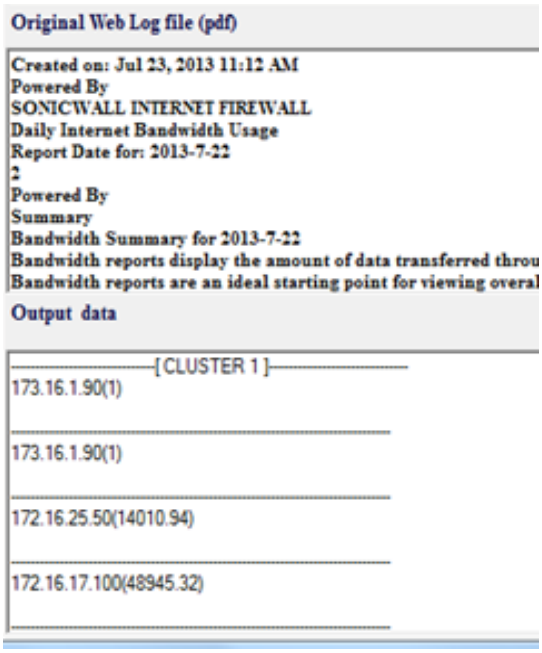


Fig 3(a): Clustering using K-means

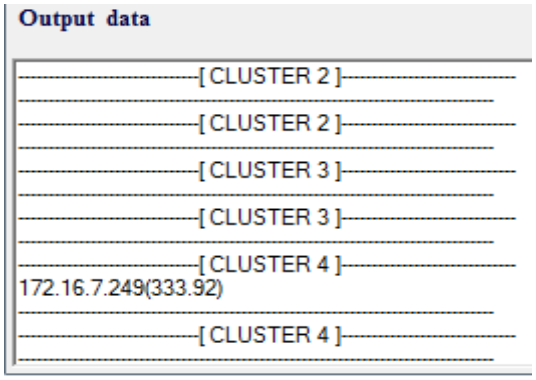
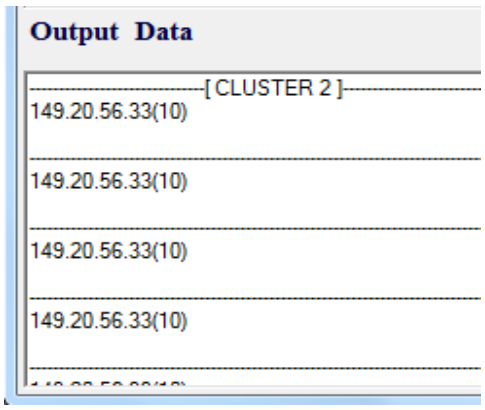
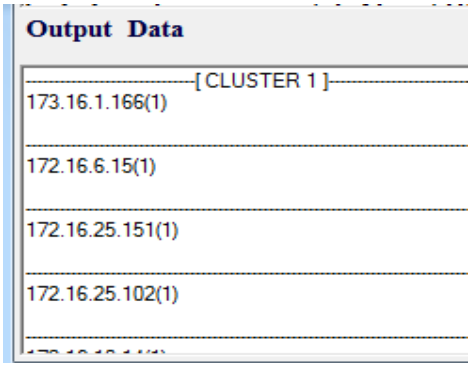


Fig 3(b): Formation of various clusters-by K-means

Once information is obtained, the web log data is clustered using K-means first. Fig 3(a) shows the clustering process, while Fig 3(b) shows various empty clusters generated using K-means. Here, we can see a single cluster(cluster1) consisting of large portion of the dataset.

STEP 3: Data Clustering using Bisecting K-means





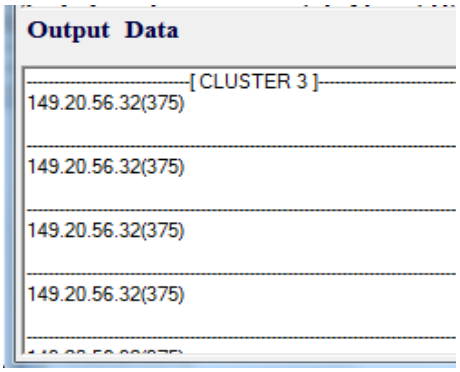


Fig 4: Formation of various clusters- Bisecting K-means

Fig 4 shows outputs of clustering using Bisecting K-means algorithm. The results concludes that BKM does not produce any empty cluster and the clusters are uniform and balanced.

Now this clustering framework is further used to detect intrusions in system. We first train the data and then apply recommendations on them

STEP 4: Training the system

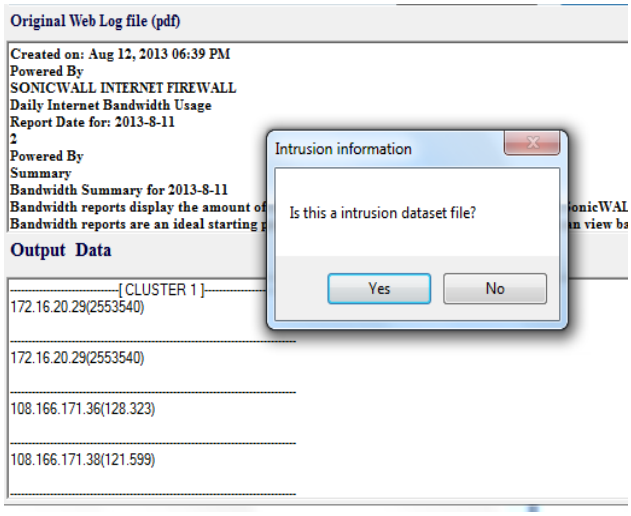


Fig 5(a): Training the system

id	cluster	ip	packets	type
14913	1	173.16.1.95	1	1
14914	1	173.16.1.95	1	1
14915	1	172.16.20.29	2553540	1
14916	1	172.16.20.29	2553540	1
14917	1	205.196.123.114	728.112	1
14918	1	108.166.171.36	128.323	1
14919	1	108.166.171.38	121.599	1
14920	1	108.166.171.37	205.254	1

Fig 5(b): Infected packets

id	cluster	ip	packets	type
21455	1	172.16.25.80	4	0
21456	1	172.16.17.122	13	0
21457	1	172.16.17.101	96048	0
21458	1	172.16.25.70	5	0
21459	1	172.16.25.5	14010	0
21460	1	172.16.20.38	38	0
21461	1	172.16.17.99	423	0
21462	1	172.16.20.41	8	0

Fig 5(c): Normal packets

The system asks the end user if the data set is an intrusion data, as shown in Fig 5(a). If user the replies "yes", then the type value in the field is labeled as "1" otherwise it is labeled as "0".A database is created in MYSQL having fields id, cluster, IP address, packets and type. We have defined 1 for infected packet and 0 for normal packet in the database. Fig. 5(b) shows the infected packets and Fig 5(c) shows the normal packets stored in the database while training

STEP 5: Applying Recommendations

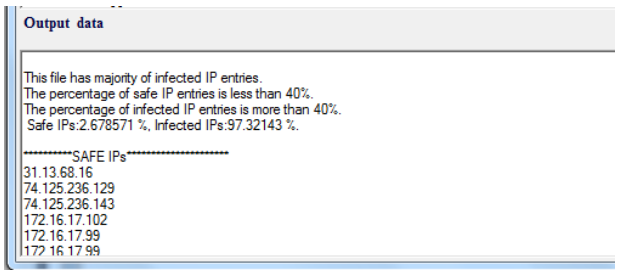


Fig 6(a) : Recommendations-List of Safe IP's

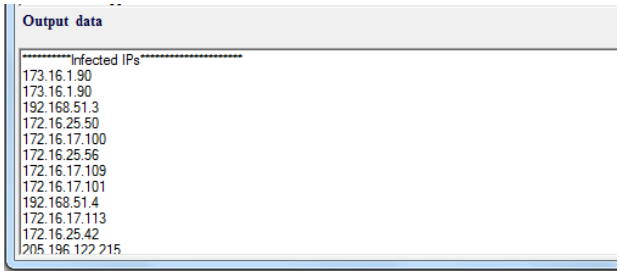


Fig 6(b): Recommendations-List of Intrusion IP's

Each IP address is checked to find the value of its signature. Comparison of the IP address occurrence is done with the IP address of similar signature in the database. If the IP address and cluster match, the difference in packet size is calculated as packet size – packet of matching entry. If the difference in the sizes is minimum (as per the threshold set), we use this packet type to decide whether the particular IP categorized as safe IP or an infected IP. Fig 6(a) shows the List of safe IP's and Fig 6(b) shows list of Infected IP's detected by the system.

5.1 Comparing algorithms

We compare the algorithms with respect to two parameters: Time and Accuracy

Table 1: Comparison of algorithms with respect to time

1) TIME (in seconds)

For k=4

PDF No.	K-means	Bisecting k-means
1	52.1	33.7
2	37.3	22.7
3	15.9	4.6
4	60	35.2
5	11.5	3.8

For k=5

PDF No.	K-means	Bisecting k-means
1	67	31.5
2	39.8	22.4
3	35	5.7
4	63	33.1
5	17.3	4.4

For k=6

PDF No.	K-means	Bisecting k-means
1	55.5	30.9
2	47.6	21.1
3	30.5	4.9
4	66	32.8
5	16.9	5.4

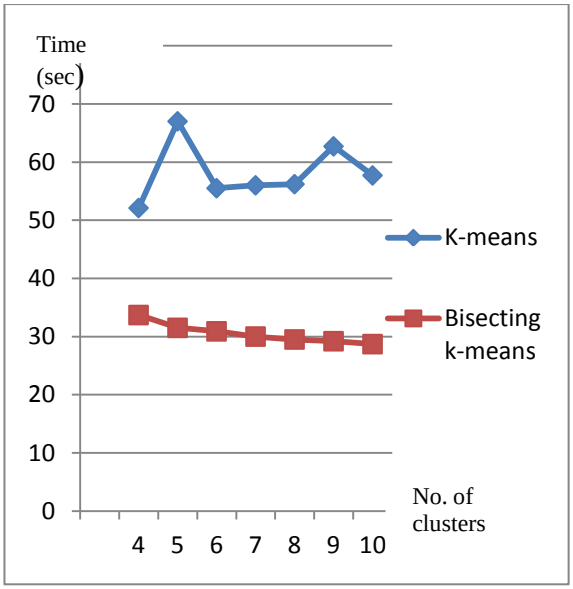


Fig. 7: Comparison of K-means and BKM

Fig. 7 shows the comparison of K-means and Bisecting K-means with respect to time for a given PDF file. From the graph, it can be concluded that as the number of cluster increases, Bisecting k-means is more efficient i.e. it is more efficient when 'k' is large.

2) ACCURACY

To evaluate the accuracy of a system, we use **Detection Rate (DR)** which is equal to the number of intrusions divided by the total number of intrusions in the data set.

In our system we find Detection Rate as, the number of counts produced as "Infected IP's" by the IDS divided by the number of intrusions found in the database. We manually checked the IP's by putting it in the college PC's browser to check whether IP's are opening correctly or not. The ones which opened were the "Safe IP's" and the ones which did not were categorized as "Infected".

For example for BKM,

Counts of infected IP's shown by system = 25

Counts of infected IP's actually in database=31

Therefore, DR= 25/31 = 0.8064

Accuracy % = 0.8064 * 100 = 80.64%

Similarly, we find the accuracy for different files using both the algorithms:

Table 2: Comparison of algorithms with respect to accuracy

PDF No.	K-means	Bisecting k-means
1	74.19%	86.64%
2	75.78%	83.18%
3	77.85%	85.37%
4	77.28%	84.74%
5	76.75%	85.43%

Thus, from the research carried out, it is seen that Bisecting K-means has higher percentage of accuracy than basic K-means algorithm.

5.2 Comparison of K-means and Bisecting K-means:

Table 3: Comparison of K-means and Bisecting K-means

Criteria	K-means	Bisecting K-means
Working	Partitions data into k clusters in each iteration	Splits one cluster into two subclusters at each Bisecting step (using k-means)
Computational Time	The computation of each iteration involves every data point of data set and k-centroids.	Only data points of one cluster and two centroids are involved at each bisecting step. Hence, computational time is less.
Degeneracy	May or may not generate empty clusters	Does not generate Empty clusters
Shape of clusters	Clusters formed are unbalanced, not uniform is in size.	Uniform clusters are produced having similar sizes
Accuracy	Lesser compared to BKM	Greater compared to k-means
Overall Performance	Depends on selection of initial centroids which is random, hence not efficient.	No initialization of centroids. Hence, more efficient than K-means

5. CONCLUSIONS

This paper presented application of clustering algorithms on Web Log data. This clustering technique was used able to detect number of intrusions present in the network environment, which helped to classify IP's as "Safe IP's" or "Infected IP's". We used two algorithms for clustering the log data: K-means and its modification- Bisecting K-means. Comparing the two algorithms and from the experimental results it was found that Bisecting K-means, outperforms simple K-means algorithm. BKM produced uniform clusters and did not generate any empty clusters. Also, it took less time for computation, provided with more accuracy and was more efficient when number of clusters were increased.

Tough the Bisecting k-means is relatively efficient than K-means, but even it is dependent on initial number of cluster 'k' given by the user. The future scope would be overcoming the initial cluster dependency. Also, we need to train the data first for labeling the packets as "Normal" or "Intrusion data". Thus, the proposed work can be extended for detection of intrusions with unlabelled data using clustering

6. REFERENCES

[1] Oren Etzioni "*The world wide Web: Quagmire or gold mine*" Communications of the ACM, 39(11):65-68, 1996

[2] J. Srivastava, R.Cooley, M. Deshpande and P. N. Tan, "*Web usage mining: discovery and applications of usage patterns from Web data*", ACM SIGKDD Explorations, Volume 1 Issue 2, January 2000.

[3] Bamshad Mobasher, Chapter: 12, "Web Usage Mining in Data Collection and Pre-Processing", ACM SIGKKD 2007 Pages 450-483.

[4] K. Alsabti, S. Ranka, and V. Singh, "*An Efficient k-means Clustering Algorithm*", Proc. First Workshop High Performance Data Mining, Mar. 1998.

[5] JinHuaXu and HongLiu, "*Web User Clustering Analysis based on K-Means Algorithm*", IEEE International Conference on Information, Networking and Automation, 2010.

[6] Natheer Khasawneh and Hien-Chung Chan, "*Active User-Based and Ontology-Based Weblog data preprocessing for Web Usage Mining*", IEEE International Conference on Web Intelligence, 2006.

[7] Peilin Shi, "An *Efficient Approach for Clustering Web Access Patterns from Web Logs*", International Journal of Advanced Science and Technology, 2009

[8] K. Poongothai, M.Parimala and Dr. S. Sathiyabama," *Efficient Web Usage Mining with Clustering*", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

[9] M. Steinbach, G. Karypis , V. Kumar, "*A comparison of document clustering techniques*", In KDD Workshop on Text Mining, 2000

[10] B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R., "*Comparative Study of K-means and Bisecting k-means Techniques in Wordnet Based Document Clustering*", International Journal of Engineering and Advanced Technology, Volume-1, Issue-6, August 2012

[11] M.Jianliang, S.Haikun and B.Ling, *"The Application on Intrusion Detection based on K-Means Cluster Algorithm*", IEEE International Conference on Information Technology and Applications, 2009.

[12] Lei Li, De-Zhang, Fang-Cheng Shen, " *A novel rule-based Intrusion Detection System using data mining*", IEEE International conference on Computer Science and Information Technology, 2010.