A Comparison of ABK-Means Algorithm with Traditional Algorithms

Ms. H. N. Gangavane
Department of Computer Engineering
A.S. Polytechnic, Pipri, Wardha, Maharastra, India

ABSTRACT

Crime investigation has very difficult task for police. Department of police plays an important role for identifying the criminals and their related information. It is observable that there are so many amounts of increases in the crime rate due to the gap between the limited usages of investigation technologies. So, there are various new opportunities for the developing a new methodologies and techniques in this field for crime investigation. Using the methods like image processing, based on data mining, forensic, and social mining. Developing a good crime analysis tool to identify crime patterns quickly and efficiently for future crime pattern detection is required. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Data mining techniques are the result of a long process of research and product development. Data mining is the computer-assisted process to break up through and analyzing large amount of data. Then extracting the meaningful data. The proposed terminology provides combine approach of preprocessing by NLP clustering, outlier detection and rule engine to identify the criminals. To automatically group the retrieved data into a list of meaningful categories different clustering techniques can be used here we used the new approach to clustering i.e. combination of K-medoid and Bisecting K-means algorithm for clustering. Crime area somewhat helps to find out the criminals so in this work we focus on area wise analysis with require records. Those records having all information about criminals which helps to further investigation. In this paper we compare ABK-means algorithm with three basic clustering algorithms i.e. K-means K-medoid, and Bisecting K-means on crime Denver dataset on the basis of time and accuracy.

KEYWORD: Crime Dataset, NLP, Adaptive-Bisecting K-Means, Clustering, Rule Engine, Area-base and Cluster base graph

I. INTRODUCTION

Crime-domain is very sophisticated domain, proper input, data pre-processing and document clustering is very important. So many authors are used some traditional methods for clustering and data mining techniques which is available. But in our proposed work included the Natural Language for data Pre-Processing and providing the combined approach of rule engine and outlier detection. We will improve the efficiency and reduce the delay to identify crime. Practical databases slow down the performance. For such type of practical databases K-Mean algorithm is used. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-Mean algorithm is fully deterministic, once initial centroid is selected [10]. In this initial centroid plays a very important role, bad choice of initial centroid leads to poor cluster which may lead to poor clustering output. When considering association rule generation frequent item set or pattern discovery and searching interest in that is very important. From these issues we motivate to use the large amount of data set to be collection, which may be used in police department for further investigation.

Clustering is the means for achieving better organization of information [12]. In this the data space is partitioned into groups of entities with similar content. So our works focus on clustering on crime dataset. Here we are using ABK-Mean for clustering and clustered the selected dataset in a most efficient manner as well as combine approach of outlier

detection and rule engine. In this paper, document clustering for criminal identification is implemented. For clustering of the input document k-means clustering technique is used [17].

In this paper also use the data mining technique Association Rule Mining to improve the search result. Document clustering for criminal identification is implemented. The documents are collected from Denver-dataset and the Adaptive-Bisecting K-means Algorithm is used for clustering also applying part of speech tagging and chunking as part of NLP for data Pre-processing with Rule Engine and Outlier detection for finding area wise graph analysis. This tool also provide us cluster wise data analysis graph.

II. RELATED WORK

A wide body of research has been carried out in this particular the crime domain area. The some researchers are focused on extracting information from 'words' and 'terms', to indicate a certain crime by employing name entity, back of word, n-gram to improving document clustering better and more effectively. Concerning this, Zhiwei Li, Bin Wang, Mingjing Li, Wei-Ying Ma [12] conducted a study in which they ware compared back of words with name entity. They warefindings revealed that the results obtained through using the name-entity approach were better and more effective than those results generated from data using the back-of-word approach. In addition, Xiang-Ying Dai, Qing-Cai Chen, Xiao-Long Wang, and Jun Xu [13] Agglomerative Hierarchical Clustering was improved by taking into account the importance of the title part of a story. The higher weight was assigned if in cases where the occurrence of the term was found in the title.

Their findings showed that the proposed method was effective in clustering the documents of financial news. However, the focus of some other researchers addressed the clustering of topic or events, whereas current work focuses on the clustering of topics and events of crimes. Meanwhile, Sheng-Tun Li, Shu-ChingKuo, Fu-Ching Tsai [14], used a Fuzzy Self-Organizing Map (FSOM) network to detect and analyse the patterns of crime trends from temporal crime activity data. Other researchers, such as Christos Bouras and VassilisTsogkas [15], used clustering methodologies including single, maximum, linkage and centroid linkage hierarchical clustering, as well as regular k-means, k-medians and k-means++.

Their findings revealed that using k-means generated the best results, not only at the level of internal measurement of clustering index function, but also on real users' experimentation. Furthermore, when comparing k-means, single pass clustering and other approaches of clustering topics of news, Taeho Jo [16] revealed that k-means was better than single pass clustering. As suggested by Zhiwei Li, Bin Wang, Mingjing Li, Wei-Ying Ma [12], estimation of the initial number of events depends, or is based on, the article count-time distribution in their probabilistic model, where the estimation of events number represents the initial (K) clusters. However, in this current study, k-means and single pass clustering were compared in terms of their effectiveness or better results generated from analyzing the events of crime documents, and thus, evaluating k-means when being used in a number of topics larger than the initial number of clusters, and when it was used in a number of themes smaller than the initial number. It was therefore expected that this method's result would often be suboptimal [17].

The researchers were carried out compare its performance in the correct number of initial number of clusters, where the benefits of the initial number of clusters were grouped documents based on this initial number, in which it was difficult to decide the initial number of clusters and the required groups or sets of data of crime. The performance of k-means clustering highly depended on the initial seed centroids [18].Jingke Xi (2008)[19]ware worked onto attempt and bring to get various outlier detection techniques, in a structured and generic description with some exercise. In this research field who could then pick up the links to different areas of applications in details.
The researcher ware discusses as well as compares approach of different outlier detection techniques data mining perspective that could be categorized into two categories i.e. Classic outlier and spatial outlier approach respectively [20]. The first approach analyzes outlier based on transaction dataset it could be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach. The second approach analyzes outlier based on spatial dataset that non-spatial and spatial data are significantly different from transaction data that could be grouped into space-based approach and

graph-based approach. Finally, this paper concludes some advances in outlier detection recently.

III. TRADITIONAL METHOD FOR CLUSTERING

Clustering is used to create group the relevant retrieved documents into meaningful categories. Descriptors are nothing but sets of words that describe the contents of the cluster. Document cluster is generally considered as a centralized process [12]. Figure 2 shows example of document clustering is web document clustering.

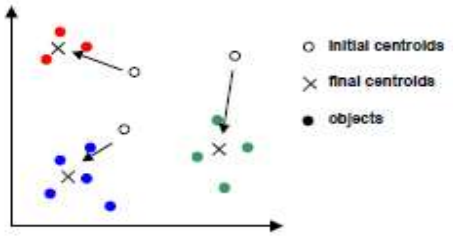

Fig 2: Example of Clustering

- **K-Mean Algorithm**

The K-Means algorithm used numerical, unsupervised, non-deterministic and iterative method.K-Mean Properties are there are always K number of cluster, always at least one item in each cluster i.e. it create non empty cluster, clusters are non-hierarchical and they do not overlap, every member of a cluster is near to its cluster than any other cluster because nearness does not always involve the centroid of clusters. Due to this approximation, these methods are usually efficient, with computational requirements ranging in the order of O (N) to O (N logN) for clusteringN documents [12]. The k-means algorithm is one of the most commonly used partition clustering methods. But this algorithm unable to handle noisy data and outliers, as well as very sensitive with respect to initial choices of clusters.

- **Bisecting K-Mean Algorithm**

Scalability problem of K-means algorithm has been tries to solve by Bisecting K-means by improving quality over K-Means. It chooses one large cluster of all the data points and divides this data-set into two clusters. K-Means algorithm is run multiple times to find a large cluster and then split it, that will produce maximum intra cluster similarity. Then again the cluster with largest size is picked to split further. This cluster can be chosen based upon minimum intra cluster similarity also.

This algorithm is run k - 1 time to get k clusters. This algorithm performs better than regular K Means because bisecting K Means produces almost uniform sized clusters. While in regular K Means there can be notable difference between sizes of the clusters. As small cluster tends to have high intra cluster similarity, large clusters have very low intra cluster similarity and overall intra cluster similarity decreases.

Document clustering is very important so, in third step we developing a clustering algorithm for creating the cluster from input data i.e. ABK-means algorithm. This algorithm is nothing but the combine approach of two algorithms K-medoid and Bisecting K-means.

IV. PROPOSED WORK

The proposed work is done on the crime data set. The data is extract and fast information retrieval or filtering with Related to on the basis of data clustering. This is done by using NLP method of part-of-speech tagging and grouping of those words.

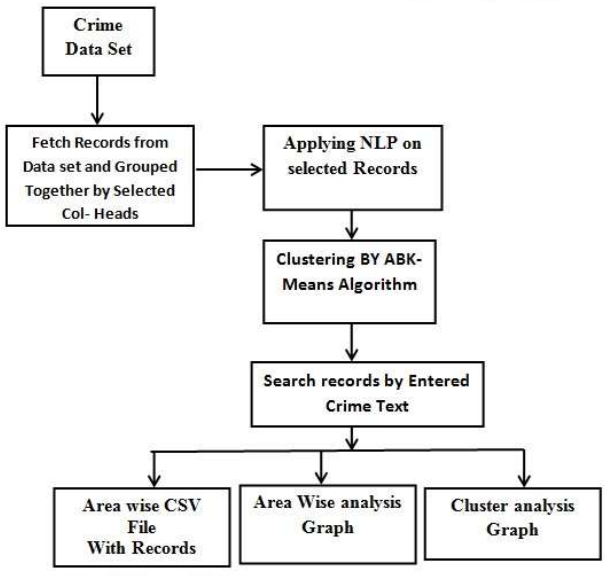

Fig 1: Block diagram of the overall proposed method.

Criminal document i.e. crime dataset csv file is given as the input to the system. NLP is applied on the input document to find out the nouns. Now ABK-means

clustering algorithm is used to find the similar categories from the input criminal document. Then applying the outlier detection method and Rule engine by entering the crime text which will give us the following output.

- **Area Wise Excel Sheet**

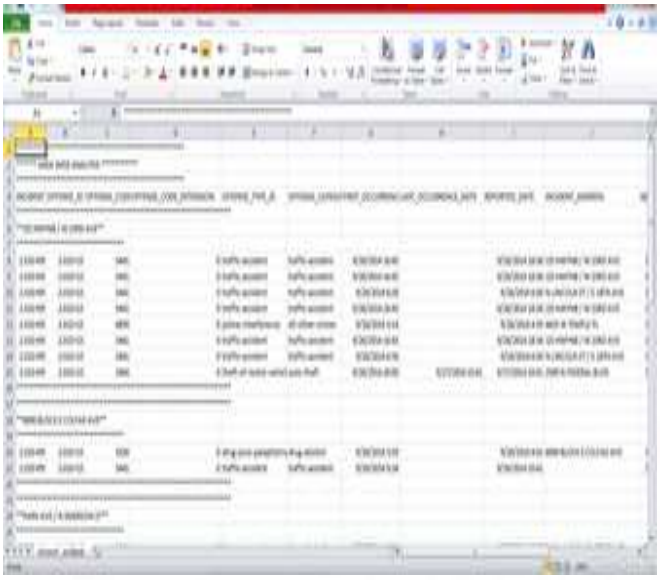

Fig 1: Excel Sheet of crime records

- **Area Wise Criminal Records**

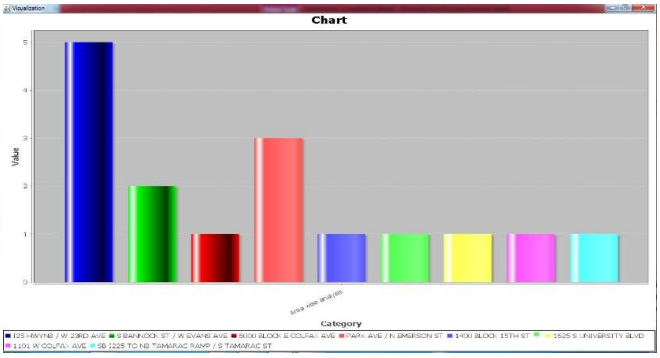

Fig 2: Shows Area Wise Criminal record in graphical format.

- **Cluster Wise Criminal Records**

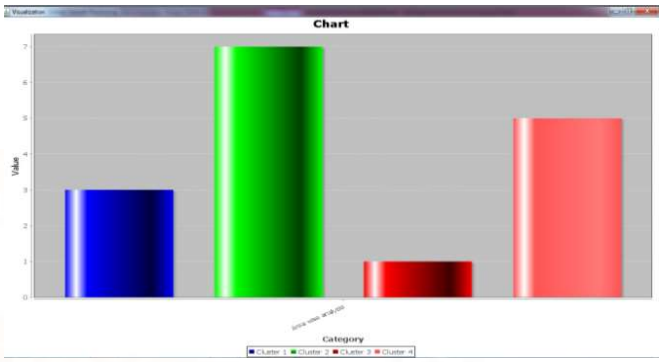

Fig 3: Shows Cluster Wise Criminal record in graphical format.

A. Collecting criminal data set from.

We collect the criminal's Data set from various sources like NationalArchive of Criminal justice data (NACJD). They provide facilitate to researchin criminals justice and criminology, through the preservation, enhancement, andsharing of computerized data resources; through the production of original research based on archived data, Denver Open data catalog and other on-line research. Also we can collect the data from Communities and Crime Data Set.

INCIDENT_ID	OFFENSE_ID	OFFENSE_CODE	OFFENSE_CODE_	OFFENSE_TYPE_ID	OFFENSE_CATEGORY_I	FIRST_OCCURRE	LAST_OCCURRENCE_R
2014510780	2.01E+15	5441	0	traffic-accident	traffic-accident	9/26/2014 16:45	
2014513767	2.01E+15	4899	0	police-interference	all-other-crimes	9/28/2014 2:04	
2014514476	2.01E+15	1316	0	threats-to-injure	public-disorder	9/28/2014 11:19	
2014514116	2.01E+15	5401	0	traffic-accident-hit-i	traffic-accident	9/28/2014 6:46	
2014514026	2.01E+15	2203	0	burglary-business-b	burglary	9/27/2014 17:00	9/28/2014 4:34
2014513634	2.01E+15	1315	0	aggravated-assault	aggravated-assault	9/27/2014 23:57	9/28/2014 0:04
2014510725	2.01E+15	5707	0	criminal-trespassing	all-other-crimes	9/26/2014 16:14	
2014510177	2.01E+15	5203	0	weapon-carrying-pr	all-other-crimes	9/26/2014 9:59	9/26/2014 10:30
2014513751	2.01E+15	1313	2	assault-dv	other-crimes-against-p	9/28/2014 0:53	
2014513989	2.01E+15	3550	0	drug-poss-paraphern	drug-alcohol	9/28/2014 3:59	
2014510756	2.01E+15	2399	0	theft-other	larceny	9/26/2014 9:59	9/26/2014 15:07
2014514003	2.01E+15	2999	0	criminal-mischief-ot	public-disorder	9/28/2014 3:00	
2014509870	2.01E+15	5441	0	traffic-accident	traffic-accident	9/26/2014 6:56	
2014513967	2.01E+15	4899	0	police-interference	all-other-crimes	9/28/2014 3:14	
2014510729	2.01E+15	4104	0	liquor-possession	drug-alcohol	9/26/2014 16:15	
2014510899	2.01E+15	2399	1	theft-bicycle	larceny	9/26/2014 17:39	
20146009226	2.01E+16	2305	0	theft-items-from-ve	theft-from-motor-vehi	9/28/2014 1:05	9/28/2014 1:09
20146009125	2.01E+16	2605	0	theft-unauth-use-of	white-collar-crime	8/29/2014 18:40	8/29/2014 18:42
2014510091	2.01E+15	5441	0	traffic-accident	traffic-accident	9/26/2014 9:24	
2014515466	2.01E+15	2404	0	theft-of-motor-vehi	auto-theft	9/28/2014 23:15	9/28/2014 23:29
2014510968	2.01E+15	2204	0	burglary-residence-l	burglary	9/26/2014 14:15	9/26/2014 14:45
2014514955	2.01E+15	5441	0	traffic-accident	traffic-accident	9/28/2014 17:39	

Fig 4: Crime Data Set

The precipitate violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery. Here we used Denver criminal dataset which having near about 3.5 lack records.

B. Developing preprocessing technique.

After collection of criminal data set we have to remove unwanted words. In this step the preprocessing of given data is done by using Natural Language Processing (NLP) that will first perform part of the speech tagging then applies to chunking

technique in order to filter out only action words [21]. Here this technique is used on the crime data set which reduces the words form crime data set fetching records which will reduce the time for processing. After that this data is group together. For that we are used RiWordNet API.

Part-Of-Speech Tagging
In this the given sentences are determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. Chinese is prone to such ambiguity because it is a tonal language during verbalization. Such inflection is not readily conveyed via the entities employed within the orthography to convey intended meaning. But in our project we used this methodology on crime data set for reducing unwanted words.
Chunking Grouped all extracted records

C. Developing clustering Algorithm.

Create a algorithm which can work best in case of all kinds of datasets. Document clustering is being studied from many decades but still it is far from a trivial and solved problem by applying combine approach of Bisecting K-Means and K-Medoid. The most intuitive and frequently used criterion function in partition clustering techniques is the squared error criterion, which tends to work well with compact and isolated clusters. The squared error for a clustering y of a pattern set x (containing K clusters) is

$$e^2(x,y) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} ||x_i^{(j)} - c_j||^2 ,$$

Where $x_i^{(j)}$ is the j^{th} pattern belonging to the j^{th}cluster and $c_j$is the centroid ofthe j^{th}cluster.

- **Algorithm**

Step 1: Initialize: randomly select k of the n data points as the medoids
Step 2: Assignment step: Associate each data point to the closest medoid.
Step 3: Update step: For each medoid m and each data point o associated to m swap m and o and compute the total cost of the configuration (that is, the average dissimilarity of o to all the data points associated to m). Select the medoid o with the lowest cost of the configuration.
Step 4: Pick a cluster to split
Step 5: Find 2 sub-clusters using the basic k-Means algorithm (Bisecting step)
Step 6: **Repeat Step 2,** in the bisecting step, for iterative times and take the split that produces the clustering with the highest overall similarity.
Step 7: Repeat Steps 4, 5 and 6 until the desired number of clusters is reached.

D. Perform outlier detection if any

Outlier detection is an irrelevant attributes can be termed as noisy attributes and such attributes further magnify the challenge of working with data streams [21]. This is the post processing techniques in which the clustered data is processes by using semantic unigram approach. This technique like as template matching. Here we used Semantic unigram approach for detecting outliers.

E. Developing a rule engine.

The described work presents the possibility to provide a common interface for rule-driven components in distributed system. The authors' approach leverages on a set of discovery protocol, rule interchange and user interface to alleviate the environment's complexity. In this step we create a rule engine which apply if - else type rule to the clustered data so, that the criminal should be identified.

V. COMPARISONOF ABK-MEANS ALGORITHM

Here we compare our proposed algorithm with three mostly used algorithms with respective time required for clustering and accuracy. This is calculated by simple Date function with end time minus from start time of the processing and accuracy based ontotal outputs obtained divided by number of desired outputs. For this analysis we taken200 to 500 records from crime data set with crime text search is used Public road accident.

I. K-Means algorithm having time and accuracy

Here we used standard k-Means algorithm on our crime data set with 200 to 500 records from crime data set with crime text search is used Public road

accident and it gives following accuracy and time required for the processing.

Sr. No.	Number of Records	Number of K(Cluster)	Time required byBisecting K-Means	Accuracy%By BsectingK-means
1	200	24	209670.8	92
2	250	24	226101.9	94
3	300	24	238377.1	94.5
4	350	24	241386.9	94.9
5	400	24	251528.5	96
6	450	24	278164.1	96.4
7	500	24	294568.9	97

Table1: Time and accuracy by K-Means.

II. Bisecting K-Means algorithm having time and Accuracy

Here we used standard Bisecting k-Means algorithm on our crime data set with 200 to500 records from crime data set with crime text search is used Public road accident and it gives following accuracy and time required for the processing.

Sr. No.	Number of Records	Number of K(Cluster)	Time required byBisecting K-Means	Accuracy%By BsectingK-means
1	200	24	209670.8	92
2	250	24	226101.9	94
3	300	24	238377.1	94.5
4	350	24	241386.9	94.9
5	400	24	251528.5	96
6	450	24	278164.1	96.4
7	500	24	294568.9	97

Table2: Time and accuracy by Bisecting K-Means

III. K-Medoid algorithm having time and accuracy

Here we used standard K-Medoid algorithm on our crime data set with 200 to 500 records from crime data set with crime text search is used Public road accident and it gives following accuracy and time required for the processing.

Sr. No.	Number of Records	Number of K(Cluster)	Time required by K-Medoid	Accuracy%By K-Medoids
1	200	24	182001.2	92
2	250	24	185934.9	94
3	300	24	197339.7	95
4	350	24	201043.5	95
5	400	24	241478.3	96
6	450	24	258315.8	97.1
7	500	24	274456.7	98

Table3: Time and accuracy by K-Medoid

IV. ABK-Means algorithm having time and accuracy

Here we used standard ABK-Means algorithm on our crime data set with 200 to500 records from crime data set with crime text search is used Public road accident and it gives following accuracy and time required for the processing.

Sr. No.	Number of Records	Number of K(Cluster)	Time required by Proposed method /ABK-Means	Accuracy % by ABK-Means
1	200	24	96027.4	97
2	250	24	97232.2	98
3	300	24	98392.2	98
4	350	24	99585.2	98
5	400	24	107220.4	98
6	450	24	112632.2	98.5
7	500	24	137864.9	99

Table 6.4: Time and accuracy by ABK-Means

Graphical representation of comparison of all algorithms with respective time required for processing with 200 to 500 records having 24 clusters. Where x-Axis belongs to number of records and y-Axis belongs to time in ms.

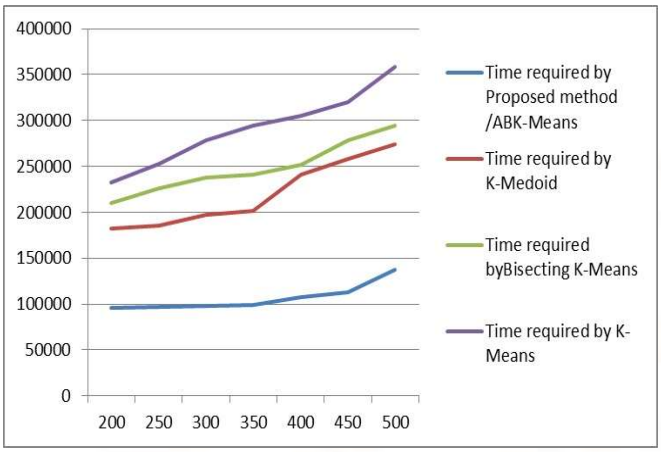

Fig4: Analysis with respective time.

Graphical representation of comparison of all algorithms with respective accuracy. Where x-Axis belongs to number of records and y-Axis belongs to time in accuracy in percentage.

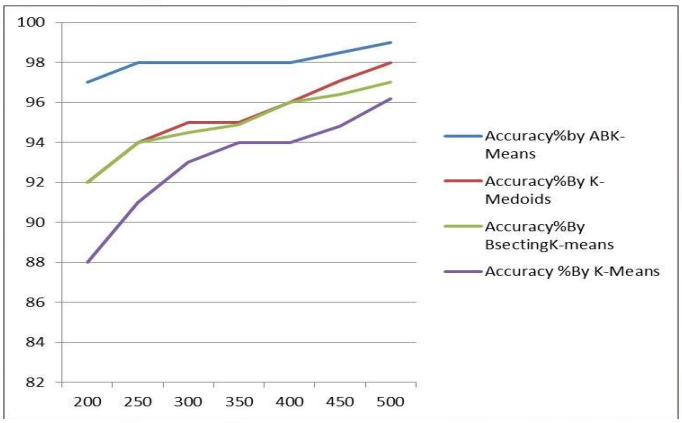

Fig 5: Analysis with respective accuracy.

According to graph we analysis the processing time is minimum as compare to other standard algorithm as well as high accuracy given by ABK-Means algorithm.

V. CONCLUSION

This paper describes a novel methodology to identifying a criminal. Crime data is a sensitive domain so, efficient clustering techniques play important role for crime analysts and law-enforcers to predict the crime. In the investigation this tool will help to solving unsolved crimes faster. Clustering is very important that's why, we used new approach for clustering i.e. Adaptive Bisecting K-Means algorithm and preprocessing by using NLP on crime data set. By developing new approach for document clustering and applying the rule engine for identifying the criminal become quite easy.

In this paper we compare ABK-means algorithm with three basic clustering algorithms i.e. K-means, K-medoid, and Bisecting K-means on crime Denver dataset on the basis of time and accuracy. The result shows that the ABK-means algorithm gives better result as compare to those three algorithms. This algorithm takes less time with higher accuracy as compare to k-means while Bisecting K-means gives the result nearer to ABK-means but it doesn't work on outlier detection. The tool provides us area wise and cluster wise graph of criminal data basing on the type crime. But the limitation of this tool it is used only one crime data set i.e. Denver dataset but we can change the dataset. In future we can apply this tool on various crime dataset.

REFERENCES

[1] Dhanabhakyam, M and Punithavalli, M. A Survey on Data MiningAlgorithm for Market Basket Analysis.Global Journal of ComputerScience and Technology, Vol. 11 issue 11, version 1.0, 2011.

[2] Agrawal, R., Imielinski, T. and Swami, A. Mining association rulesbetween sets of items in large databases. In Proceedings of theACM SIGMOD International Conference on Management of Data,207-216, 1993.

[3] Qiankun Zhao and Sourav S. Bhowmick. Association rule mining :A Survey. Technical Report, CAIS, NanyangTechnologicalUniversity, Singapore, No. 2003116, 2003.

[4] Agrawal, R. and Srikant, R. Fast Algorithms for MiningAssociation rules. Proc. 20th VLDB conference, Santiago, Chile,1994.

[5] Viet PhanLuong and Lif.Mining normal and abonormalclassassociationrules.IEEE 27th International Conference on AdvancedInformation Networking and Applications. 968-975, Barcelona, 25-28 Mar 2013.

[6] Bourbakis, N., Mills M., Converting Natural Language Text Sentences Into SPN Representations for Associating Events, International Journal of Semantic Computing, Vol. 6, No. 3 (2012) pp. 353-370. World Scientific Publishing Company

[7] Mich, L., Franch, M. and Novi Inverardi, P., "Requirements Analysis using Linguistic Tools: Results of an On-line Survey", Requirements Engineering Journal, Technical Report 66,

Department of Computer and Management Sciences, University of Trento, Itlay, 2003.

[8] Cristiaximiliano and Plss, Brian (2010). Generating natural language descriptions of Z test cases. In: Proceedings of the 6th International Natural Language Generation Conference, 7-9 July 2010, Dublin Ireland.

[9] Miss. Aparna N. Gupta , Prof. ArtiKarndikar, "A Review : Study of Various clustering Technique in web usage mining" , International Journal of Advance Research In Computer And Communication Engineering, Vol. 3 , Issue 3 , March 2014.

[10] Michael Steinbach, George Karypis, Vipin Kumar, "A Comparison of Document Clustering Techniques" at In KDD Workshop on Text Mining

[11]B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R "Comparative Study of K-meansand Bisecting k-means Techniques in Wordnet Based Document Clustering".International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-1, Issue-6, August 2012

[12] Zhiwei Li, Bin Wang, Mingjing Li, Wei-Ying Ma."A Probabilistic Model for Retrospective News Event Detection",in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 106-113, 2005.

[13] Dai X, Chen Q, Wang X,Xu J. "Online topic detection and tracking of financial news based on hierarchical clustering," in Proceeding of International Conference on Machine Learning and Cybernetics, pp. 3341-3346, 2010.

[14] Sheng-Tun Li, Shu-ChingKuo, Fu-ChingTsai."An intelligent decision-support model using FSOM and rule extraction for crime prevention", Expert Systems with Applications, Elsevier, Vol (37), no. 10, PP. 7108–7119, 2010.

[15] Aouf M, Lyanage L, Hansen S. "Review of data mining clustering techniques toanalyse data with high dimensionality as applied in gene expression data (June2008)" in Proceeding of International Conference on Service Systems and ServiceManagement, pp. 1-5, 2008.

[16] TaehoJo."Clustering News Groups using Inverted Index based NTSO,"NDT, First International Conference on Networked Digital Technologies, PP. 1-7, 2009.

[17] Aouf M, Lyanage L, Hansen S. "Review of data mining clustering techniques to analyzedata with high dimensionality as applied in gene expression data (June 2008)" in Proceeding of International Conference on Service Systems and Service Management, pp. 1-5, 2008.

[18] SubhashTatale, SachinSakhare" IntellectualCrimerecognition System ."IOSR Journal of Computer Science(IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 40-

[19] Jingke Xi "Outlier Detection Algorithms in Data Mining" Intelligent InformationTechnology Application, 2008.IITA '08. Second International Symposium on(Volume:1) 20-22 Dec. 2008

[20] Thakran, Yogita, and DurgaToshniwal. (2012). "Unsupervised outlier detectionin streaming data using weighted clustering." Intelligent Systems Design andApplications (ISDA), 2012 12th International Conference on.IEEE.

[21] Sheng-Tun Li, Shu-ChingKuo, Fu-ChingTsai."An intelligent decision-supportmodel using FSOM and rule extraction for crime prevention", Expert Systems withApplications, Elsevier, Vol (37), no. 10, PP. 7108-7119, 2010.