Journal of Theoretical and Applied Information Technology
15th November 2012. Vol. 45 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMAL INITIAL CENTROID IN K-MEANS FOR CRIME TOPIC

[1]MASNIZAH MOHD, [2]QUSAY WALID BSOUL, [3]NAZLENA MOHAMAD ALI, [4]SHAHRUL AZMAN MOHD NOAH, [5]SAIDAH SAAD, [6]NAZLIA OMAR AND [7]MOHD JUZAIDDIN AB. AZIZ
[1,2,4,5,6,7] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

[3] Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Malaysia

E-mail: [1]mas@ftsm.ukm.my

ABSTRACT

A wide number of different clustering method applications and their effectiveness in crime topics have been examined in this paper. Several works have investigated the optimal initial centroid of clustering crime topics. In this paper, wehave compared the effectiveness of single pass clustering and k-means in detecting crime topics and aiding in the identification of events or crimes. We have also experimentedon enhanced k-means clustering, in order to select the optimal initial centroid to be automatically compared with regular k-means, to choose the initial centroid randomly. Based on the main findings of this study, it was revealed that the experimental method, which was based on k-means, was proved to be better and more effective than single pass clustering in detecting and identifying events or crimes. For the initial number of centroids, it was found that the proposed method was more effective when used in selecting terms that were more than the number of topics, than when they were less. However, the best result was obtained when choosing a number of topics equal to the number of original topics. This implies that the optimal accuracy of clustering is achieved when selecting a large number of documents that have termsbetter than randomly chosen documents as a centroid.

Keywords: *Crime clustering, Single pass, K-means, Crime topic*

1. INTRODUCTION

An increasingly large number of daily reports and news on crimes has resulted in making the detection of crimes complex and more difficult. Therefore, the need for detecting and identifying crime patterns from the news has emerged. In recent years, there has been an increasing interest from researchers in the detecting and tracking of crime news stories. This increasing interest is attributed to the social dilemma and epidemic disease represented and reflected by the occurrence of social crimes, which poses tremendous threats to societies [1, 2]. Since most news concerns stories in general, and those related to crime in particular, are being increasingly accumulated like a flood over the web; many challenges are encountered by decision-makers in law enforcement departments in detecting, identifying and tracing or tracking crime events [1, 3]. Therefore, tracking social crimes or events according to their time line is becoming a tedious task. These difficult challenges and complexities, in organizing the news of crime stories, are generated from a huge dimensionality of crime data, which usually refers to highly diverse embedded modalities, such as criminal data and weapon data [4]. In other words, law enforcement officers are provided through these modalities with justified explanations of international or worldwide views of crime patterns, by carrying out identification of the relations between local patterns [4, 5].

In organizing this current paper, Section 2 is concerned with presenting a review of related clustering crime research and Section 3 addresses the clustering technique. In Section 4, we reporton the experimental methodsused and Section 5 analyses the results. Finally, Section 6 presents the conclusion and future work.

2. CRIME-RELATED CLUSTERING

One of the most complex challenges faced by researchers interested in this domain, which still presents an interesting dilemma for them, is the data of crimes. Although some data remains top secret and private, some is made accessible to the public as public information. However, restrictions

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of data about crimes concerning narcotics or juvenile cases,distinguishessuch crime data from other data for other crimes. Similarly, in crimes such as sex offences, information related to the sex offenders is publicized for the purpose of warning others about the seriousness of such crimes in an area, but information pertaining to the victim's identity is not often permitted to be made publiclyaccessible. Therefore, this implies that when playing the role of a data miner, the analyst has to deal with issues related to the public versus private data issues; so that the process of text mining modelling does not violate such legal boundaries. Recently, electronic crime reporting systems have taken the place of traditional manuallywritten reports in the majority of police departments. These more advanced or technology-based crime reports include information categorized into types of crime, date/times, locations and others.

2.1. Crime Domain

From a Malaysian context, as far as crime information systems are concerned, it is evident that Malaysia has not yet applied such information systems into the crime domain. Therefore, the most challenging problem being faced by the researchers of this current study has been the lack of such data in this particular context. As one way of solving this problem, the researchers have made efforts to combine and collect information about such crimes from reports, news and articles, published in several Malaysian newspapers. The difficulty of accessing official reports or narratives from the policejustifies why the researchers have proposed the exploitation of newspapers to solve this problem. Moreover, the type of information about crimes in newspaper articles can be similar to the information containedwithin the police reports. Therefore, one of the most important steps in pursuing this current research is that it is expected that such data will provide the researchers with a better understanding of the crime domain and the nature of data that our system will have to deal with.

2.2. Related Work

A wide body of research has been carried out in this particular area of the crime domain. The focus of some researchers has been placed on extracting information from 'terms or words', indicating a certain crime by employing name entity, back of word, n-gram to improve document clustering better and more effectively. Concerning this, Zhiwei Li, Bin Wang, Mingjing Li, Wei-Ying Ma [6] conducted a study in which they compared back of words with name entity. Their findings 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 [7] improved Agglomerative Hierarchical Clustering by taking into account the importance of the title part of a story. In cases where the occurrence of the term was found in the title, that word was assigned as higher weight. 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-Ching Kuo, Fu-Ching Tsai [8], 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 Vassilis Tsogkas [9], used clustering methodologies includingsingle, 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 [10] revealed that k-means was better than single pass clustering. As suggested by Zhiwei Li, Bin Wang, Mingjing Li, Wei-Ying Ma [6], 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 analysing 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. This was carried out to 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. It was

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

therefore expected that this method's result would often be suboptimal [11].

2.3. Event Crime Description

Concerning the description of events in the crime domain, in the majority of Malaysian newspaper reports, such reports appeared to share the same structure in relation to writing style. In other words, such reports are usually initiated with a sentence having the title of the report, followed by the date, the author's name and details describing the crime. These crime description details usually provided the public with information regarding the type of crime committed and the criminal who committed it. Next, the reports moved on to support further details about the victims and other information related to the crime. In interpreting such reports, it is evident that Malaysian journalists and reporters follow a formulaic approach; since they are involved in a specialist domain that is characterized by its own language, known as the language of crime. According to Almas and Ahmad [12], each special language used in a specific domain for specific purposes is seen as having a limited amount of vocabulary and idiosyncratic syntactic structures. Therefore, such a restricted language's use of this vocabulary and structures becomes common for most journalists who are specialized in such a domain. In other words, the use of restricted and specific language items and structures, involves almost the same behaviour. Figure 1is an example of a document structure for a crime topic.

SMS ON SHARLINIE NOT TRUE
Date: 25-01-2008
Author: / BNHS KK ZS AO
Sharlinie-False
PETALING JAYA, Jan 25 (Bernama) -- Petaling Jaya OCPD ACP Arjunaidi Mohamed said today the rumour circulated via the short messaging service…..

***Fig. 1**: Describes The Structure Of The "Sharlinie" Topic Document*

3. CLUSTERING SYSTEM

3.1. Pre-Processing Of Topics/Events

The first phase of the proposed system in this study is concerned with conducting the processing of crime documents in accordance with the most common pre-processing methods. Based on the findings of studies by the previously mentioned researchers, it was found that the process of removing the stop words from the topics of crime, and stemming the words, led to the improving or enhancing of the clustering results by a factor of 5 – 15% [9] [13]. Part of the text pre-processing method applied is as follows:

i) Removal of stop words: stop word as a list of 571 stop words used in the smart system wasused [14]. This stop word list was obtained from [15].
ii) Stemming: this is concerned with stemming the words, and for the current research, Porter's Stemming Algorithm [16], which is the most commonly used algorithm for word stemming in English,was selected for this purpose.

3.2. Representation Terms

Bag-Of-Words (BOW) representation was seen as the most commonly used method of 'term type' representation. The advantage of using bag-of-words representation is its simplicity. In other words, the researcher only has to record the frequency of occurrence of a linguistic item (word) in the document; whereas, he/she can ignore or does not need all of the remaining things, such as the structure and order or organization of the words in the document. Therefore, in this current study, the bag-of-words method was proposed to be utilized for term extraction.

Concerning the common use of BOW, it is usually used in cases where a word is used as a term. In other words, each term t_n corresponds to a single word, and in this current study, we used all terms 'words' after removing the stop words and stemming different aspects from previously mentioned studies in our related work, where they used the top rank of highest terms. The reason why we chose all terms, is because they are clear (e.g., two types of news; first one about sport and the second one about economics), so that many similar words, such as 'said', 'want', 'write' (i.e., said player and said economic analyst), and can be common or shared by each group. Furthermore, it was expected that the frequency of this term (TF) would be very high. The document frequency would be high and include more than players and economics, as these two terms have two groups connected to each other. Each term of player and economics will be less than the frequency, and in taking the top terms, the index will more likely lose the most important terms.

Journal of Theoretical and Applied Information Technology
15th November 2012. Vol. 45 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A. Term Frequency x Inverse Document Frequency Weighting

As pointed out by previous researchers, the frequency of the term is not considered according to Boolean weighting and TF weighting throughout all the documents in the document corpus. Themost commonly used method is Term Frequency × Inverse Document Frequency (TF×IDF) weighting, because such a property is taken into consideration when using this method. Research using this approach can assign the weight of terms (i) in document (d) to the number of times that the item occurs in the document; such assigning is seen to be proportional. Moreover, it is in inverse proportion to the number of documents in the corpus in which the term appears.

$$w_i = tf_i \cdot \log\left(\frac{N}{n}\right) \quad (1)$$

Furthermore, in cases where the occurrence of a certain term is seen in most of the documents being investigated, it is pointed out that TF×IDF weighting approach gives weight to the frequency of a term in a document, with a factor discounting its importance. For instance, the applicability of this can be realized in such a case where the term is assumed to have little discriminating power.

3.3. Similarity Measure

One of the most popular and common similarity measures applied to text documents, such as in numerous information retrieval applications (as pointed out by Baeza-Yates et al., [17]) is Cosine similarity and clustering [18]. In measuring the given two documents $\overrightarrow{t_a}$ and$\overrightarrow{t_b}$, their cosine similarity is:

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) \ \frac{\overrightarrow{t_a}.\overrightarrow{t_b}}{|\overrightarrow{t_a}|\times|\overrightarrow{t_b}|}. \quad (2)$$

Where,$\overrightarrow{t_a}$and $\overrightarrow{t_b}$ are perceived as m-dimensional vectors over the term set T$\{t_1 \dots t_m\}$. Each term, with its weight in the document, is represented by a particular dimension, which is non-negative. Therefore, the cosine similarity is non-negative and bounded between [0, 1].

Another feature distinguishing the cosine similarity from others is that it does not depend on the length of the document. For instance, in a case where the cosine similarity between two identical copies of a document (d) are combined to get a new pseudodocument d_0 is measured, it is probable that the result will show the value of the cosine similarity between d and d_0 as 1. This is evidence of the similarity of these two documents, or, that these two documents are considered to be identical.

3.4. CLUSTERING ALGORITHM

In this current work, two clusters; namely k-means and single pass, will be compared to identify the similarities and differences between them.

A. K-means cluster

As stated by Hartigan [19], in using the k-means algorithm, the mean of the documents assigned to that cluster reflects or represents each of the k clusters; which is regarded as the main idea behind using the k-means algorithm. Another name for this is the centroid of that cluster. In discussing the k-means algorithm, Berkhin [20]identified two versions of it. The first version is known as the batch version, also called Forgy's algorithm, was attributed to Forgy [21]. The main steps followed in the K-means algorithm are as follows:

i) Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.
ii) Generate a new partition by assigning each pattern to its closest cluster's centre.
iii) Compute new cluster centres.

Therefore, enhancing the k-means proposed in this current study will be carried out by following these six steps in selecting an initial partition with the K clusters.

i) In the previous process of representing the terms, the number of each document after removingthe stop words ('size of documents after remove stop word') will be calculated.
ii) Splitting the data sets, based on the initial partition with the K cluster (e.g., our data set has247 documents) we select six cluster numbers of K, so that |247/6|=41 documents. For this, the whole date set will be split into six groups, and each group will contain about 41 documents. The reason for this step is to distribute the documents and the system will choose a centroid.
iii) Selecting the largest sized documents as the centroid; repeat steps 5 and 6 until the cluster membership stabilizes. This is based on the size of each document in each group.
iv) Assigning each pattern to its closest cluster'scentre.
v) Computing the new cluster centres.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

B. Single Pass Clustering

In clustering the documents using Single Pass Clustering, as created by Sylvester and Seth [22]; as its name suggests, a single, sequential pass over the set of documents,which is attempting to be clustered, is required by the researcher. The algorithm used is shown in Figure 2.

For each document (d) in the sequence loop
1. find a cluster c that maximises cos(c, d);
2. if cos(c, d)> t then include d in c;
3. else create a new cluster whose only document is d;

End loop.
Where, t is the similarity threshold value, which is usually derived experimentally.

Fig. 2: Single-Pass Algorithm

In using this algorithm, the next document is classified into the sequence according to a condition on the similarity of the function employed. At every stage, whether a newly seen document should become a member of an already defined cluster or the centre of a new one or not, is usually decided by the algorithm. In its most simple form, defining the similarity function mainly depends on the basis of some similarity measure between document-feature vectors.

4. EXPERIMENTAL METHODS

4.1. Data description

For this work, the corpora were collected from Bernama news, and the dataset being tested consisted of six categories of topics, including Canny Ong (which has five events), Mona Fandy, Noritta Samsudin, Nurin Jazlin (which has eight events), Sharlinie Mohd Nashar, and Sosilawati articles (as shown in Table 2). These topics constituted 247 documents that were to be used as the testing dataset [23] (as shown in Table 1). These events were used to test the dataset.

Table 1.Data Set Of Topics

Topics	No. of documents
Canny Ong	48
Mona Fandy	35
Noritta Samsudin	35
Nurin Jazlin	59
Sharlinie Mohd- nashar	35
Sosilawati	35

Table 2. Data Set Of Events Of Document

Topic	Event id	Event Description	No. of doc
Canny Ong	1	Investigation into Canny Ong case include medical report and trial	1 7
	2	Evidence/Suspect into Canny Ong case	1 3
	3	DNA test	6
	4	Family reacts into Canny Ong and negligence suit	3
	5	Court Sentence, plead guilty	9
Nurin Jazlin	1	Investigation into Nurin Jazlin case include trial	1 3
	2	Evidence/Suspect into Nurin Jazlin case	1 3
	3	DNA test	3
	4	Reward for the public	3
	5	Family react to Nurin Jazlin investigation	8
	6	Public reacts to Nurin Jazlin investigation	5
	7	Investigation into Jazimin suit	1 2
	8	Suit to the court	2

4.2. Evaluation

Two types of measuresare usually used for the purpose of evaluating the cluster quality [24], namely internal and external quality measures. It is stated that external knowledge, such as class label information for evaluating the produced clustering solution, is not used or utilized by the internal quality measure.

Both the precision and recall ideas from information retrieval are combined in the F-measure cluster evaluation metric. Each cluster is considered as being the results of a query and each class is perceived as being the desired set of documents for the query. In calculating the recall and precision for each cluster j and class i, the following is an illustration of this:

$$\text{Recall } (i,j) = \frac{n_{ij}}{n_i}. \quad (3)$$

$$\text{Precision } (i,j) = \frac{n_{ij}}{n_j}. \quad (4)$$

Journal of Theoretical and Applied Information Technology
15th November 2012. Vol. 45 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.3 Evaluation Users

Here, is used to represent the number of documents which have the class label i in cluster j, and reflects the number of documents which have the class label i. Finally, represents the number of documents in cluster j. Thus, the following presents the calculation of the F-measure of cluster j and class i:

$$F(i,j) = \frac{2\text{Recall}(i,j)\text{Precision}(i,j)}{\text{Recall}(i,j)+\text{Precision}(i,j)}. \quad (5)$$

In carrying out the calculation of the overall value for the F-measure, it is important to take the weighted average of all values for the F-measure into consideration, as follows:

$$F=\sum_i \frac{n_i}{N} \max F(i,j). \quad (6)$$

Thus, based on the previous calculation, it can be seen that the occurrence of the F-measure values is found at interval (0, 1) and the larger F-measure values correspond to the higher clustering quality.

5. EXPERIMENTS AND RESULTS

This study involves three sets of experiments. The first experiment was performed three times using various initial seed sets and thresholds, where the centroid for the k-means was selected randomly, but the number of clusters for the k-means static was selected based on the number of different types. In contrast, the threshold was selected three times, so that the first results would be shown as averaged values of all three times.

5.1. First Experiment

The first experiment was aimed at clustering the documents under different groups of topics and events, in order to examine the effect on clustering, so that there would be four groups of data sets, which included two topics (i.e., Canny Ong and Nurin Jazlin), six different topics (i.e., Canny Ong, mona fandy, noritta samsudin, nurin jazlin, sharlinie mohd nashar and sosilawati), five events of the topic canny ong, and eight events of the topic nurin jazlin. Based on the results for the four groups shown in Table 3, it was found that in the first group, the k-means achieved a better result than the single pass, where the F-measure for the k-means was 0.915 and the F-measure for single pass was 0.712. Therefore, the k-means was 1.2 times better than the single pass.

Based on the results for the second group, it was revealed that the k-means achieved a better result than the single pass, where the F-measure for k-means was 0.751 and the F-measure for the single pass was 0.615. Thus, the k-means was 1.22 times better than the single pass.

Concerning the results of the third group, the k-means was found to achieve a better result than the single pass, where the F-measure for the k-means was 0.796 and the F-measure for the single pass was 0.618. Therefore, the k-means was 1.2 times better than the single pass.

Based on the results for the fourth group, the results showed that the k-means achieved a better result than the single pass, where the F-measure for the k-means was 0.727 and the F-measure for the single pass was 0.607. Thus, the k-means was 1.19 times better than the single pass.

Table 3. *F-Measure Evaluation Clustering For Four Groups*

Groups	*K-means*	*Single pass*
2 topics	0.915	0.712
6 topics	0.751	0.615
5 events canny ong	0.796	0.618
8 events nurin jazlin	0.727	0.607

5.2. Second Experiment

The second experiment was performed to examine the effect on clustering of the six different groups of topics. The experiments were made to examine the effect of k-means when the selected number of clusters was less or more than the real number of clusters. Table 4 shows the negative results obtained when the selected number of clusters was less than the real number of clusters. However, when the chosen number of clusters was more than the real number of clusters, the results were satisfactory; when compared with the results of the exact number of clusters. In contrast, in comparing the k-means of the six types of topics, in cases where choosing the number of clusters was eight with the single pass clustering, it was found that the k-means was better than the single pass.

Table 4. *F-Measure Evaluation Of Six Categories With Different Numbers Of Clusters*

Clustering	K-means
K=4	0.614
K=5	0.653
K=6	**0.751**
K=7	0.739
K=8	0.718

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.3. Third Experiment

The third experiment was performed to examine the effect on clustering of the selected centroid, when the documents for each cluster had the highest number of terms, in comparison to when the medium or smallest size was selected. To do this, the number of terms 'words' in each document was calculatedafter the stop wordswere removed. Next, a calculation was carried out based on the number of clusters that the area of each centroid had. For instance, when there were six clusters, we calculated 247/6=41, so that each of the 41 documents would retrieve the highest number of terms from the largest, medium and smallest numberdocument. Table 5 shows the F-measure results for the k-means, where the best results were obtained from the large sized documents having the most terms to distinguish other classes of another type of topic.

However, the best single pass clustering results were obtained when a threshold of 2.5 was selected as the best threshold on two types of topics, and the threshold estimated at around 1.45 was chosen as the best threshold on six types of topics. For topicevents, the best results were obtained when choosing a threshold of 1.7 as the best threshold on five events of canny ong, and when a threshold of 1.35 was selected as the best threshold on eight events of Nurin Jazlin.

Table 5. F-Measure Evaluation Of K-Means On Big, Medium And Small Size Of Centroids

Clustering	Large size	Middle size	Small size
2 topics	0.96	0.94	0.83
5 events	0.90	0.76	0.72
8 events	0.82	0.75	0.60

The findings obtained from the third experiment showed that selecting the large sized documents to find the centroid,achieved better results than those obtained from medium and small sized documents. This was proved by comparing the centroid results of the k-means in the large, medium and small sized documents. Table 5 shows that the centroid on two types of topics was 0.96, but when it was randomly selected, it was 0.915 for five types of events. Meanwhile, eight types of events were 0.9 and 0.82, when large sized documents were selected, in comparison to the k-means. When selected randomly,it was 0.796 and 0.727, because when applying the k-means as a centroid randomly, it took longer until the cluster membership was stabilized. However, when choosing the centroid based on large sizes documents, the k-means took less time until the cluster membership was stabilized.

6. CONCLUSION AND FUTURE WORKS

The major findings of these experiments have evidently proved that using k-means was better than using single pass clustering. However, a weakness of k-means was related to choosing the centroid for each cluster and the number of clusters, so that the second and third experiments to solve those problems could be carried out successfully. The findings revealed that the number of clusters were more and better than those of the real number of clusters. Based on these findings, it is recommended that it is better to choose a larger number of clusters rather than a smaller number,due to the possible occurrence of problems in cases where the system splits one type of topic into two clusters. In other words, problems are more probable if two types of topics are merged into one cluster.

Based on the findings of this current study, several points are recommended for carrying out future research. First, when expanding this current study in the future, the same system cluster k-means with Name Entity Recognition (NER) and mixed terms back-of-word with name entity recognition can be applied to improve the clusteringsystem. Secondly, the generation of a new stemming method is suggested,which will be based on morphology, syntactic structure and semantics; because despite stemming and deleting the stop words from the documents (as indicated by the results), many words are not important in document clustering such as say, want, write and out (which are a just a few examples of words that belong to more than one cluster). Thirdly, it is recommended that future research carry out identification of a list of stop words for the crime domain and a measurement of their impact on clustering results.

ACKNOWLEDGEMENTS

This project was supported by Arus Perdana Research Grant from Universiti Kebangsaan Malaysia (UKM-AP-ICT-21-2010).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES:

[1] ChandraB,Gupta M, Gupta M. "A multivariate time series clustering approach for crime trends prediction", in Proceedings of the International Conference on Systems, Man and Cybernetics, 2008. SMC, pp. 892-896, 2008.

[2] Meshrif Alruily, Aladdin Ayesh, Abdulsamad Al-Marghilani."Using Self Organizing Map to Cluster Arabic Crime Documents," Proceedings of the International Multi-conference on Computer Science and Information Technology, pp. 357–363, 2010.

[3] Chen H,Zeng D,Atabakhsh H,Wyzga W, Schroeder J. "COPLINK: managing law enforcement data and knowledge", Communications of the ACM, vol. 46,no. 1, pp. 28-34, 2003.

[4] Boo Y,Alahakoon D."Mining Multi-modal Crime Patterns at Different Levels of Granularity Using Hierarchical Clustering", CIMCA 2008, IAWTIC 2008, and ISE 2008, pp. 1268-1273, 2008.

[5] Bache R,Crestani F. "Estimating real-valued characteristics of criminals from their recorded crimes," pp. 1385-1386, 2008.

[6] 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.

[7] 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.

[8] Sheng-Tun Li, Shu-Ching Kuo, Fu-Ching Tsai."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.

[9] Bouras C, Tsogkas V. "Assigning Web News to Clusters," in Proceedings of Conference on Internet and Web Applications and Services, pp. 1-6, 2010.

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

[11] Aouf M, Lyanage L, Hansen S. "Review of data mining clustering techniques to analyse data 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.

[12] Almas Y, Kurshid A. "Lolo: a system based on terminologyfor multilingual extraction", in IEBeyondDoc 06: Proceeding of the Work shop on Information Extraction beyond the Document, pp.56–65, 2006.

[13] Varathan K D, Sembok T M T, Kadir R A. "Automatic lexicon generator", in Proceedings of International Conference on Information Retrieval and Knowledge Management, pp. 24-27, 2010.

[14] Salton G, Yang C, Wong A. "A Vector-Space Model for Automatic Indexing", Communications of the ACM, Vol. 18, no. 11, pp. 613–620, 1975.

[15] ftp://ftp.cs.cornell.edu/pub/smart/, 2004.

[16] Porter M F. "An Algorithm for Suffix Stripping," Program, Vol. 14, no. pp. 130–137, 1980.

[17] Baeza-Yates R, Ribeiro-Neto B."Modern information retrieval,"Addison-Wesley, New York, ACM press New York, 1999.

[18] Larsen B, Aone C. "Fast and effective text mining using linear-time document clustering," In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (KDD,99).pp.16-22, 1999.

[19] Hartigan J A. "Clustering algorithms," John Wiley & Sons, Inc., 1975.

[20] Berkhin, P. "Survey of clustering data mining techniques," Grouping Multidimensional Data: Recent Advances in Clustering, New York: Springer-Verlag, PP.25–71, 2006.

[21] Forgy E W. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications",Journal Biometrics,Vol. 21, PP. 768-769, 1965.

[22] Sylwester D, Seth S. "A trainable, single-pass algorithm for column segmentation," in Proceedings of the Third International Conference on Document Analysis and Recognition, pp. 615-618, 1995.

[23] http://blis.bernama.com/mainHome.do.

[24] Steinbach M, Karypis G."A comparison of document clustering techniques," KDD Workshop on Text Mining, vol. 34. p. 35, 2000.