World Applied Sciences Journal 34 (8): 1026-1036, 2016
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2016.34.8.109

Document Clustering Approach to Detect Crime

Qusay Bsoul, Juhana Salim and Lailatul Qadri Zakaria

Knowledge Technology Research Group,
Universiti Kebangsaan Malaysia, 43600 Bangi Selangor, Malaysia

Abstract: Document Clustering aims at automatically detecting similar documents in one cluster using different types of extractions and cluster algorithms. This paper reviews Document Clustering method applications and their effectiveness in crime topics. The different methods applied in past research on Crime Document Clustering proved to be ineffective. Hence, there is the need to increase their performance and this is particularly true in cases where most important crime words focused on verbs and nouns that can be used to describe the event. Past research used cluster algorithm but failed to generate the number of clusters. More recent research used local clusters to evaluate extraction feature, where the local clusters has weakness when sign initial centroids, this problem effect on evaluate extraction methods. We prove this is new problem in our experiments in this paper. This work discusses two major sequential stages in document clustering which are "features extraction", "Clustering Algorithms", as well as the major challenges and the key issues in designing them. Finally, we propose two main process of feature extraction and suggest novel ways to generate number of clusters.

Key words: Intelligent Clustering · K-means · Affinity Propagation algorithm · Lemmatization Algorithm · Crime Detection

INTRODUCTION

Rapid technological advances in the application of computerized systems to trace and track crimes have motivated computer data analysts to take practical steps in assisting the law enforcement officers and detectives to enhance the process of solving crimes and increase its performance. In recent years, there has been an increasing interest on the part of researchers in detecting and tracking crime news stories based on clustering methods, where some of them used Text Clustering as well Data Clustering. Such an emerging interest is attributed to the social dilemma and epidemic disease represented and reflected by the occurrence of social crimes which pose tremendous threats to societies [1-2]. Since large amount of news concerning stories in general and others related to crime news in particular accumulated like flood over the web. Many challenges are encountered by the decision makers in the law enforcement departments in detecting, identifying and tracing or tracking crime events [1, 3]. Thus, tracking the social crimes or events according to their time line is becoming a tedious task. Such difficult challenges and complexities in organizing the news crime are generated from a huge dimensionality of crime data, which usually refers to the highly diverse embedded modalities such as criminal data and weapon data [4]. In other words, the law enforcement officers and detectives are provided by these modalities with justified explanations about the international or world view for crime patterns by carrying out identification of the relations between local patterns [4-5].

Document Clustering is considered as one of the most commonly used methods in detecting topics/events or types of crime documents [6]. It is a method of document clustering that involves three main processes [1, 7-8]. The first process is pre-processing of documents, whereby unimportant words and symbols are removed from the crime documents. The second process is the representation of the crime documents involving the extraction of the most important information from the document and showing the similarity among these documents. The last process of document clustering consists of applying the clustering algorithm to the groups of documents of topics/events or types of crime based on the similarities among the documents.

Corresponding Author: Qusay Bsoul, Knowledge Technology Research Group,
Universiti Kebangsaan Malaysia, 43600 Bangi Selangor, Malaysia.

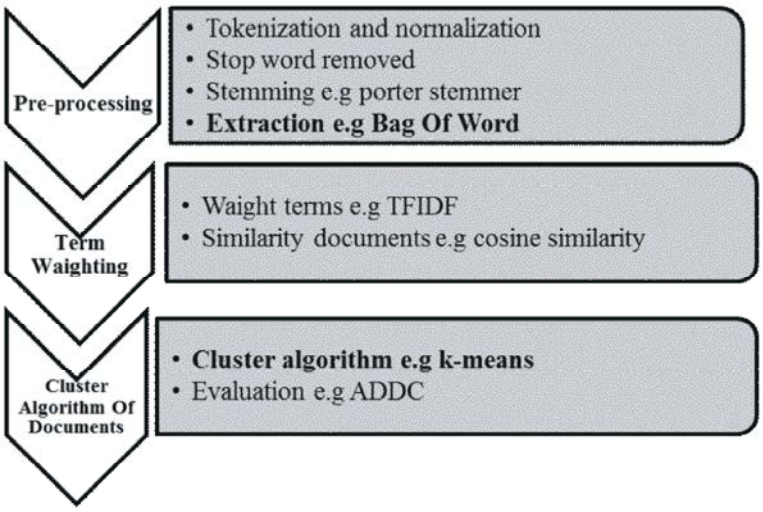

Fig. 1: Main process of Document Clustering

Past researchers provided the restraints on Document Clustering in two stages: the Extraction of Terms and Cluster Algorithms specifically in clustering documents relationship to crime domain. Section 2 presents a comprehensive review of the previous related work on the extraction of terms and cluster algorithms of Document Clustering. In section 3, we describe our suggested approach of Crime Document Clustering within two subsections. Lastly, section 4 presents the conclusion.

Process of Document Clustering: Document Clustering has three main processes namely; pre-processing, term weighting and cluster algorithm of document, where each process has sub-process as shown in Figure 1. This paper will focus on two processes, the first is the extraction of terms and the second is cluster algorithm.

Extraction Terms: Recently, electronic systems for reporting crimes have taken the place of the traditional manual reports written on paper in most police departments. Such more advanced or technology-based crime reports include different kinds of information being categorized into types of crime, date/time, location etc. [9]. The role of this categorization is to distinguish between reports of crimes such as 'theft, drugs, murder'. These categories in crime document are represented by a set of terms which are called features. Extracting the features from the indexed words is considered as a real challenge. This is because more than one feature or combinations of features are considered in this stage. Several researchers have focused on extracting information from the terms or 'features' indicating certain crime by employing *Name Entity* [10], *Bag Of Word* [10-11], *N-Gram* [12], *Frequent Word and Frequent Word Meaning* [13]. In order to make the extractions step better and more effective, researchers used *Concept Weighting Ontology* [14], *Word Net Lexical Ontology* [15] *Core Semantic Features* [16] *Semantic Relationships* [17], *Word-Net* [18], *Semantic Method* [19], *Semantic Preserving* [20] and *Verb-Centric Approach* [21]. However, the nature of text data often represents a challenging issue due to the high dimension and ambiguous/overlapping word senses. Previous studies proposed and developed various features selection methods to address high dimension such as BOW, N-gram and NER. Those methods are called syntactic extraction. Another approach is semantic extraction which involve overlapping word senses like semantic words by word-net and latent semantic index.

In this paper, we initially discussed the syntactic extraction and why researchers proceed to develop and use the semantic extractions. In syntactic extraction, Zhiwei *et al*. [8] compared between *bag of word* and *name entity* by using *Probabilistic Model Clustering*. Their findings revealed that the results gained through using the *name entity* approach were better and more effective than those results generated from the data using *bag of word*. Yanjunli *et al*. [13] used *Frequent Word and Frequent Word Meaning* to compare them with *bag of word*s using *Bisecting k-means*. Their result showed that *Frequent Word and Frequent Word Meaning* were better than *bag of word*.

In evaluating the graphical interface of Document Clustering, Masnizah *et al*. [22] distinguished between *Name Entity* and *bag of word* by using *single pass clustering*. Their findings revealed that the *name entity*

approach is better than *bag of word*. Using similar method of syntactic extraction on extract nouns, Fodeh *et al*.[16] extracted the nouns and compared with all terms extracted on diverse dataset and using *Spherical k-means* to cluster, they found the *noun* terms were better in all experiments than extracted all terms after pre-processing. In contrast, they did not compare with the baseline of *BOW*. Al-Shamari and Lin [23] extracted nouns and verbs by using a new method called *lemmatization algorithm*, which is an idea to extract information from sentences based on certain rules of the sentence. They generated a new technique involving *lemmatization algorithm* for Arabic Information Retrieval. It aims to extract nouns and verbs from Arabic documents based on the preposition words as well as some rules related to other linguistic elements such as the definite article "the". It was reported that this algorithm is better than BOW as it is effective at catching important words from two different lists of prepositions; one including proceeding *verbs* and one including proceeding *nouns*.

Sharma *et al*. [21] proposed an extraction algorithm for extracting verb-centric relationship *using Naïve Bayesian classifier*. By examining a sentence from biomedical text, their algorithm could identify if it was a relationship bearing sentence or not and then extracted the relationship depicting phrase from the sentence. The researchers were also able to extract the participating entities which involved around the relationship depicting phrase. Their algorithm is capable of handling missing, incomplete and conjoining entity issues involved in extraction of participating entities. The results showed that this algorithm gained a balanced precision from 0.86 to ~0.95 and a recall form 0.88 to ~0.92 based on their evaluation of three biomedical data sets. On the other hand, extracted terms as syntactic will hide the core meaning for example, most of the meaning related to the word will be hidden and that will lead to main problem related to syntactic extraction.

Though semantic can overcome the weakness of syntactic extraction [24] by extract meaning, semantic still has the following drawbacks:

- Most of words have synonymy and polysemy which will make impediments to extract the correct and suitable words instead of the original for document clustering. It is worth mentioning that we can overcome these hindrances owing to the Word Sense Disambiguation (WSD) process which enables the most appropriate ontology concept to substitute the original terms in a document. Yet, much work on the use of ontology has to be done. Unexpectedly, past research has failed to improve document clustering [16]. For instance, by substituting a word with its potential concepts, it is possible only to expand or augment the feature space without necessarily enhancing the clustering performance [16, 25].
- The main problem in text clustering is huge feature extracted from text, where the number of features may reach hundreds or thousands. Thus, when substituting a term (syntactic) with its potential concepts (semantic), the processing time may be augmented and the clustering performance may be reduced due to high dimension of feature space.
- Extracting core semantics from texts will reduce number of feature but this reduction was not always significant, where some of deleted terms will be important features as shown in Fodah *et al*. [16].

To enhance the clustering accuracy with a reduced number of terms, extracting a subset of the disambiguated terms with their relations (known as the core semantic features) that are ''cluster-aware'' is highly appreciated, Zheng *et al*. [17] combined detection of noun phrases with the use of WordNet as background knowledge to explore better ways of representing documents semantically for clustering using three divers clusters namely; *k-means, bisecting K-means and Hierarchical Agglomerative Clustering*. Based on noun phrases as well as single-term analysis, they exploited different document representation methods to analyze the effectiveness of *hypernymy, hyponymy, holonymy* and *meronymy* and they used Reuters-21578 as data-set. The results showed that the best method is *hypernymy*. Chen *et al*. [18] used *hypernyms* of WordNet as proposed to enhance document clustering using *Fuzzy-based Multi-label Document Clustering*. In their experiment, they used many datasets benchmark and their findings showed that the used *hypernyms* seemed better than without using the *hypernyms*. The researchers could increase the accuracy and effectiveness of text mining, but one weakness of their work is concerned with reducing the dimensionality of terms. On the other hand, Fodeh *et al*.[16] carried out an experiment using Spherical k-means on semantic features in which all the *polysemous* and *synonymous nouns* were extracted from the documents and a unique approach that was capable of permitting them to measure the *information gain* and disambiguate these nouns in an unsupervised learning setting. The purpose of developing this approach was to identify the *core subset of semantic features* representing a text corpus. Thus, based on this experiment, the results revealed that by employingo

Table 1: Summary of previous work on extraction

Authors	Dataset used	Baseline Extraction	Outperform Extraction	Cluster used
[8]	Financial news	BOW	NER	Probabilistic Model
[13]	Divers dataset	BOW	Frequent word, frequent word meaning	Bisecting k-means
[17]	Divers dataset	hyponymy, holonymy, meronymy	Hypernymy	k-means
[18]	Divers dataset	Without hypernyms	Hypernyms	Fuzzy Clustering
[21]	Medical " Medline database"	NP1-VP-NP2	Verb-centric Approach by VerbNet	Naïve Bayesian classifier
[16]	Divers dataset	Ontology and semantic	Core semantic features	Spherical k-means
[26]	News	BOW	Weighting ontology	k-means
[19]	Classic 3 and real crime	maximum number δ, minimum similarity τ	Nouns +verbs semantic 1.25 δ+ 0.175 τ	Single Pass
[22]	TDT	BOW	NER	Single Pass
[15]	News	BOW	Word-net ontology	Self-Organizing Map
[20]	Divers dataset	Core semantic features, All terms after stop word remove	semantic preserving vector space model	Spherical k-means

core semantic features for clustering, it is more possible to reduce the number of features by 90% or more. At the same time, it is possible to produce clusters that capture the major themes in a text corpus.

According to Hmway and Thi [14] and Gharib *et al*. [15], when they compared Concept Weighting ontology and Word Net Lexical ontology with *Bag Of Word* using k-means and Self Organizing Map Clusters, the performance of Concept Weighting ontology and Word Net Lexical ontology was much better than bag of word.

Table 1 showed the summary of research on extractions. The main gaps in this process is the evaluation using clustering search, where the clustering search deepened on the initial centroids of each clusters 'groups'. The researchers showed in their work on the times of independents runs of Document Cluster and then take the average. However, the average of independent runs is not enough to evaluate the extraction methods, to prove our stated. BOW was applied as extraction using K-means/ Spherical k-means. The dataset used were taken from TREC of TDT2 and TDT3 consisting of 1468 documents, which were split into 53 groups. To evaluate the two clusters, we used Average Distance of Documents to the cluster Centroid (ADDC) as evaluation measure. Figure 2 and 3 showed the 1000 times of runs for each one of K-means and Spherical k-means that has different result either, positive or negative. The best result was nearest to zero for all of them. The worst result in k-means was near to the best result in Spherical k-means. The next section, discuss other weakness related to Document Clustering.

Cluster Algorithms: The process related to the cluster approach [26] involves grouping objects of similar category. This approach indicates mainly two types of clustering: hierarchical and partitioning. Hierarchical clustering methods are often known to be more adequate. However, at the beginning of the clustering process, they

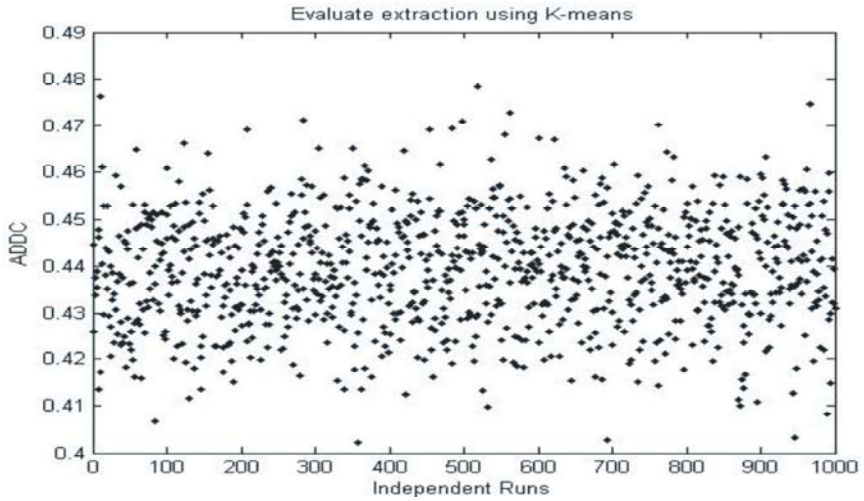

Fig. 2: BOW extraction using K-means for 1000 independent runs

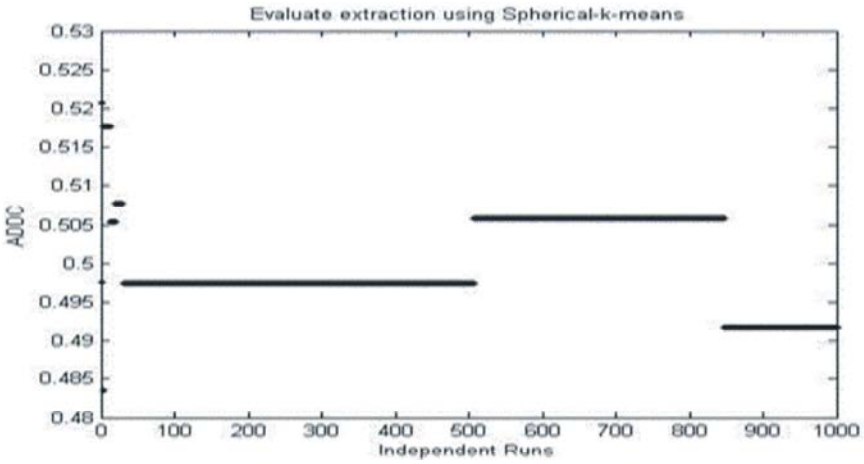

Fig. 3: BOW extraction using Spherical K-means for 1000 independent runs

cannot always ensure the reallocation of documents, which are likely to be misclassified [27]. The time complexity of hierarchical methods, in addition, is quadratic in the number of data objects [28]. When implemented, the partitioning clusters are more beneficial due to their relatively low computational complexity which allows them to include huge datasets [28-32]. K-means algorithm method plays a key function in partitioning clustering [33] and it has been mostly applied to partitional clustering with a linear time complexity [28]. Moreover, the main objective of k-means algorithm according to Hartigan [34] is that the mean of the documents assigned to that cluster is utilized to represent each of 'k' clusters and it is called the centroid of that cluster. The K-means algorithm however, lacks mainly the sensitivity to initialization and it requires the number of clusters from the start (i.e., the number of clusters must be known a priori). Besides, the initial centroids for each cluster have a key role in their performances which get stuck in local optima solutions [35] and the current experiments in Figure 2 and 3 showed the problem of local optima. Local optima means, the clusters cannot find good clusters during clustering process. Thus, this work aims at solving the number of cluster problem as shown in the previous work on cluster algorithms and some work focused on finding the number of clusters.

Many researchers have carried out empirical work for clustering algorithms on diverse data sets, where some of them tried to find the best cluster and on other hand tried to find the number of clusters. For example in finding the best cluster, Dai *et al*. [7] improved Agglomerative Hierarchical Clustering by taking into account the importance of the title of a story as there are cases when the occurrence of the term was found in the title and it was assigned as higher in weight. The findings showed that the proposed method is effective in clustering documents of financial news. Bação *et al*. [36] used self-organizing map (SOM) to cluster IRES as data clustering not text clustering and compared it with k-means. They found performance of SOM was better than k-means. Sheng-Tun *et al*. [37], used a fuzzy self-organizing map (FSOM) network to detect and analyze the patterns of crime trends from temporal crime activity data. Other researchers such as Fredrik and James [38], compared between scalable k-means, complete k-means and k-means for knowledge and data discovery (KDD) of data set. Their results proved that k-means is better than other algorithms. Bouras and Tsogkas [39] used the clustering methodologies including: single, maximum, linkage and centroid linkage hierarchical clustering, as well as regular k-means, k-medians and k-means++. Based on their findings, the k-means not only generated the best result at the level of internal measurement of clustering index function, but also produced best results on a real users' experimentation. Some researchers have focused on the clustering of topic or events so that they compared between k-means and single pass algorithms and other algorithms for the clustering of topics in news, Taeho Jo [40] revealed that k-means is better than single pass clustering.

In examining the hybrid algorithm of clustering, Dai, *et al*.[41] proposed a two-layer text clustering approach to detect retrospective news events by using the Affinity Propagation clustering (AP) as first layer. Researchers conducted a second feature selection on the generated clusters of (AP) cluster and they also adopted the usual agglomerative hierarchical clustering for the purpose of generating the ultimate news events. Finally, researchers

Table 2: Summary of previous work on cluster algorithms

Authors	Dataset used	Baseline algorithm	outperform clustering algorithm
[7]	Financial news	AHC	Enhance AHC
[36]	IRES	k-means	SOM
[37]	Data series of crimes	N/A	FSOM
[38]	KDD	Scalable, complete k-means	k-means
[39]	Web	Single, maximum, centroid AHC and k-medians, k-means++	k-means
[40]	News	Single pass	k-means
[8]	News	Probabilistic model cluster	Enhance Probabilistic model
[6]	Crimes	Single pass, k-means	Enhance k-means
[41]	News	k-means, AHC, AP	APAHC
[42]	Geographic map	Fuzzy c-means	k-means, k-medoids

chose the traditional Agglomerative Hierarchical Clustering (AHC) and the classic K-Means as comparative methods. The findings revealed that the proposed method achieved the highest precision measure followed by the Affinity Propagation (AP) clustering and the k-means and AHC clustering respectively in their ranks. As far as the recall measure is concerned, it was found that the proposed method and k-means obtained the highest result followed by the AHC clustering and AP clustering.

Velmurugan and Santhanan [42] compared three algorithms to cluster geographic map data set. Their result showed that k-means is good with small data, while, k-medoids is good with large data and fuzzy c-means is between k-means and k-medoids. Table 2 showed as a summary of research on clustering. As mentioned earlier in Selim and Ismail [35], they pointed out number of clusters is one of the common problems in clustering algorithms and hence some researchers have tried to solve this problem. Zhiwei Li *et al* [8], for instance, stated that the estimation of initial number of events is closely related to the article count-time distribution in their probabilistic model where the estimation of events number represents the initial (K) clusters. *Silhouette Width* (SW) was used by Sheng-Tun *et al*. [37] to determine the cluster numbers, whereby a higher value of SW justifies better discrimination among clusters. However the largest SW justifies the best clustering (number of cluster). In addition, numbers from the range of 2–10 are applied by Sheng-Tun *et al*. [30] to reach the optimal cluster number. When two clusters are identified, the best value of SW is approximately 0.1710. However, it is meaningless to uncover the trend of two crime categories as clusters are too few to be subsequently analyzed. Therefore, they used four clusters instead of two clusters. The splitting data is other way to split data as pairs.

Wei *et al*. [43] used Bisecting k-means as clustering, which this cluster can detect number of clusters in dataset. They compared the Bisecting k-means on diverse extraction methods, as they found the proposed extraction obtained the best f-measure for number of clusters. On the other hand, Affinity Propagation cluster which can generate a number of clusters automatically was proposed by Dueck and Frey 2007 [44]. AP is better than k-means in precision according to the shown result of Dai, *et al*.[41]. The f-measure assessment revealed that k-means is better than AP because k-means has proved to be better in recall.

Limitations of Feature Extraction and Cluster Algorithm: In general, the main limitations in the previous works summarized in Table 1 and 2 are presented as follows:

- The researchers assumed that the name entity is good for extracting information from crime documents without detecting the weakness of extracting the specific important information from crime documents such as extracting nouns and verbs, where events or the type of crime have nouns to describe locations, names, topics and dates. On the other hand, verbs can be used to describe events, types and reasons to commit crime and simultaneously avoiding extracting the unimportant features [45].
- Based on the comparison that was made between "the k-means with other algorithm without a hybrid", the researchers assumed that k-means has the best algorithm [46], without detecting the weakness of k-means for crime documents in order to find the number of clusters. Past researchers provided a solution for these weaknesses by integrating [41] with other algorithms such as SW [37], Bisecting k-means [43] and Affinity Propagation [44] to find the number of clusters. Their findings, showed that their approached was not effective in finding the correct number of clusters [37].
- In general, all the proposed works solved the weakness of clustering document in terms of enhancing part of the process, where the output of

each stage affected the accuracy of the next process [45]. This paper shows how the weakness of cluster effect on the evaluation of the extraction methods in Figure 2 and 3. Therefore, there is a need to detect the problems from bottom up, which means from algorithm for clustering which is used to extract the terms.

As mentioned above, there are three processes for clustering documents, from the bottom "cluster algorithm" to "extracting the terms from crime documents". Thus, the idea of bottom up guides us to describe or identify the problem of extracting information from crime documents when solving the weakness of k-means cluster algorithm. In addition, it is very important to extract words such as nouns and verbs related to topics/events of crimes, whereby events or type of crime have nouns to describe information such as locations, names, topics and dates. Furthermore, verbs can be used to describe information such as events, types and reasons to commit crime.

It can be synthesized that, all the existing methods for detecting and identifying document clustering have weakness, either on the method of extraction and clusters algorithms and either on the evaluations used by previous study. The problems related to cluster algorithm as already mentioned in sections 2.2 and in Figure 2 and 3, includes; the performance of k-means clustering highly depends on selecting the number of clusters and it is expected that the result of this method is often suboptimal [47] to the problem related to the extraction features [45] as explained in sections 2.1 and 2.3. In addition it has been elaborated in our experiment in figure 2 and 3. Each of these two problems have impact on other processes [45]. To overcome these obstacles, we provide alternatives to help detect and identify the groups of topics/events or types of crimes in document clustering with improved performance by using our proposed method as shown in the next section.

Proposed Crime Document Clustering

Data Collection and Performance Measure: For this research, the datasets were collected from Bernama news [48] and the dataset tested six categories of topics to identify the events in each topic, including Canny Ong, Mona Fandy, Noritta Samsudin, Nurin Jazlin, Sharlinie Mohd Nashar and Sosilawati articles. To detect the types of crimes, the dataset tested ten types of crimes, including Traffic Violation, theft, sex crime, murder, kidnap, fraud, drugs, cybercrime and arson gang articles. These topics and types consist of 223, 2422 documents that were used as the testing dataset. We also used our experiments with Benchmark document datasets called TREC of TDT2 and TDT3 [22], which will allow other studies to make comparisons with our proposed crime document clustering, where all of mentioned datasets are available by this link "https:// www.dropbox.com/ sh/ htt omsm9pus5jga/ AACPF1jMgfJskyVIZQeCdQ8ha?dl=0". This study employed an overall purity measure and an overall F-measure in order to measure the external quality. These two measures are popular document clustering measures [26, 28]. As the existence of higher overall purity and F-measures gives the best cluster, we suggest to employ *Average Distance of Documents to the cluster Centroid* (ADDC) measure [49] in order to determine the locus of the cluster centroids so that the intra-cluster similarity can be maximized, in other word, minimizing the intra-cluster distance, while at the same time, the inter-cluster similarity can be minimized or maximizing the distance between clusters.

Proposed Extraction Features: As mentioned in section 2.3 regarding the weakness of extraction, Name Entity was suggested to be used [50] to extract information based on three questions namely; who, where and when. These questions play an important role to be extracted, because most of these features are usually used to describe crime in general. We also would like to extract nouns and verbs by using Word-Net to provide the most important features that could be used to describe specific types or events. This can be done by detecting whether a feature possibly could be a noun or verb via examination of the stemmed feature that exists in the Word-Net noun or verb database (as show in Fig.4). Events and types of crimes could be detected and identified by looking at nouns which could be used to identify location, names, topics, date, etc., as well as verbs to identify the events and types of event. In addition redundant features extracted from NE with Verbs and Nouns "(NER ∪ Nouns ∪ Verbs)" will be removed.

Proposed Affinity Propagation Algorithm: In order to overcome the problem of k-means clustering as mentioned earlier, we used Affinity propagation algorithm [46] to generate the number of clusters as illustrated in Fig. 5.

As shown in Figure. 5 and 6, then K-means clustering will be used, where there is a description of the k-means Algorithm techniques as follows:

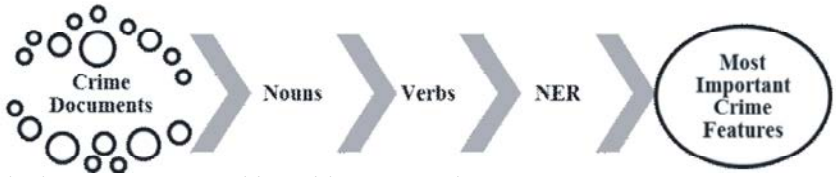

Fig. 4: proposed method to extract NE combine with nouns, verbs.

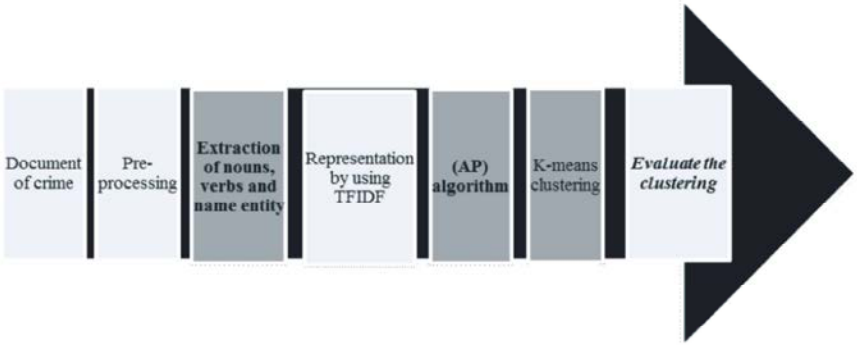

Fig. 5: Our proposed crime document clustering of enhances clustering and extraction.
** The gray coloris proposed in this work.

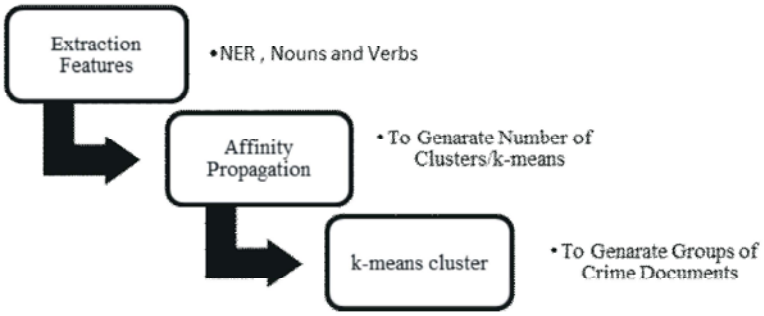

Fig. 6: Our proposed of order techniques.

1 **Input**: a collection of training crime documents D= {d_1, d_2… d_m}, number of clusters K derived from output of AP algorithm
2: **Output**: an assignment matrix A of crime documents to set of K clusters
3: **Select** K crime documents as the initial cluster centers generated from AP cluster
4: **repeat**
5: **Choose** C= number of initial centroids randomly
6: **Initialize** A as zero
7: for all *di* in *D* **do**
8: let j = $\mathrm{argmin}_{K\epsilon\{1,2,\ldots K\}} D(\boldsymbol{d_i}, \boldsymbol{c_k})$
9: assign *di* to the cluster *j*, i.e. A[i][j] = 1
10: **end do**
11: **Update** the cluster means as $\boldsymbol{c_k} = \frac{\sum_{i=1}^{m}(\sum_{k=1}^{k} A[i][k]\boldsymbol{d_i})}{\sum_{i=1}^{n}(\sum_{k=1}^{k} A[i][k])}$ for k=1, 2… K
12: **until** meeting a given criterion function or convergence criterion

CONCLUSION

This paper detects and identifies some limitations in Crime Document Clustering. Firstly, it addresses the fault detection and identification in the k-means algorithm and Spherical k-means. Secondly, it examines the weakness of extracting terms from documents as stated above, where we have to identify the gap in k-means which is the number of clusters without k-means and the weakness of extracting the specific important information such as Nouns and Verbs, where we show the important Verbs in section 2.3, refers to the gap with the NER extraction. Therefore, this study aims at enhancing the reliability of Document Clustering of crime report by efficient k-means as well as the extraction features of crime document.

The convergence criterion of k-means cluster will be affected by the AP cluster which will obtain the number of clusters. K-means is used for crime document clustering,

since its' results are the best testimony for its efficiency. This is due to the fact that it aims at enhancing the k-means algorithm for Document Clustering as well as the extraction of information including which group topics/events of crimes can outperform the original Document Clustering and other Document Clustering based on two criteria of time and complexity of k-means. Another purpose of using extraction is that the extraction of nouns, verbs and NER will be helpful to identify the most important terms needed to be extracted from the crime documents.

REFERENCES

1. Chandra, A., B. Gupta and C. Gupta, 2008. A multivariate time series clustering approach for crime trends prediction, in Proseeding of International Conference on Systems, Man and Cybernetics, SMC, pp: 892-896.
2. Al-Marghilani, A.A., 2010. Using Self Organizing Map to Cluster Arabic Crime Documents, Proceedings of the International Multiconference on Computer Science and Information Technology, ppp: 357-363.
3. Chen, A., B. Zeng, C. Atabakhsh, D. Wyzga and E. Schroeder, 2003. COPLINK: managing law enforcement data and knowledge, Communications of the ACM, 46(1): 28-34.
4. Boo, A. and B. Alahakoon, 2008. Mining Multi-modal Crime Patterns at Different Levels of Granularity Using Hierarchical Clustering, CIMCA 2008, IAWTIC 2008 and ISE 2008, 1268-1273.
5. Bache, A. and B. Crestani, 2008. Estimating real-valued characteristics of criminals from their recorded crimes, pp: 1385-1386.
6. Mohd Bsoul, Ali Noah, Saad Omar and Ab, Aziz, 2012. Optimal Initialcentroid In K-Means For Crime Topic, Journal of Theoretical and Applied Information Technology, 45(1): 19-26.
7. Dai, X., Q. Chen, X. Wang and J. Xu, 2010. 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.
8. Wang, Z., M. Li and Wei-Ying, 2005. 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.
9. Dutelle, A., 2011. An Introduction to Crime Scene Investigation, United State of America.
10. Kumaran, G. and J. Allan, 2004. Text classification and named entities for new event detection, in Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval.
11. Can, F., 2010. New event detection and topic tracking in Turkish, Journal of the American Society for Information Science and Technology, 61: 802-819.
12. Mustafa, H. and Q. Al-Radaideh, 2004. Using N-grams for Arabic text searching, J. Am. Soc. Inf. Sci. Technol., 55(11): 1002-1007.
13. Li, Yanjun, Chung, M. Soon, Holt and D. John, 2006. Text document clustering based on frequent word meaning sequences, 8th International Conference on Enterprise Information Systems (ICEIS' 2006), pp: 381-404.
14. Tar and Nyunt, 2011. Enhancing Traditional Text Documents Clustering based on Ontology, International Journal of Computer Applications, 33(10): 38-42.
15. Gharib Fouad and Mashat Bidawi, 2012. Self Organizing Map -based Document Clustering Using WordNet Ontologies, International Journal of Computer Science Issues, 9(1)2: 88-95.
16. Punch, F. and N. Tan, 2011. On ontology-driven document clustering using core semantic features, Journal of Knowl Inf Syst, Springer-Verlag London.
17. Zheng, H.T., *et al*. 2009. Exploiting noun phrases and semantic relationships for text document clustering. Information Sciences, 179(13): 2249-2262.
18. Chen, C.L., *et al*., 2010. An integration of WordNet and fuzzy association rule mining for multi-label document clustering. Data & Knowledge Engineering 69(11): 1208-1226.
19. Kabi G. Daghir, 2011. Semantic Document Clustering For Crime Investigation', Thesis Of Master of Applied Science In Information Systems Security, Concordia University, Montréal, Québec, Canada.
20. Michael Howard, semantic preserving text representation and its applications in text clustering', master of computer science, missouri university of science and technology, USA, 2012.
21. Abhishek Sharma, Rajesh Swaminathan and Hui Yang, 2010. A Verb-centric Approach for Relationship Extraction in Biomedical Text, IEEE Fourth International Conference on Semantic Computing (ICSC), pp: 377-385.

22. Masnizah Mohd, Fabio Crestani and Ian Ruthven, 2012. Evaluation of an interactive topic detection and tracking interface' Journal of Information Science, 38(4): 383-398.
23. Al-Shammari, patent application publication of "lemmatizing, stemming and query expansion method and system", Pub.no.:US 2010/0082333 A1, Pub.Date: Apr.1, http:// www.google.com/ patents/ US20100082333?printsec=abstract&source=gbs_overview_r&cad=0#v=onepage&q&f=false, 2010.
24. Wei, T., *et al*., 2015. A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications, 42(4): 2264-2275.
25. Hotho, A., S. Staab and G. Stumme, 2003. WordNet improves text document clustering. In Paper presented at the in SIGIR international conference on semantic Web Workshop.
26. Jain, K. and C. Dubes, 1988. Algorithms for clustering data, Ed.
27. Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: a review, ACM Computing Surveys (CSUR), 31(3): 264-323.
28. Steinbach, M. and G. Karypis, 2000. A comparison of document clustering techniques, KDD Workshop on Text Mining, 34: 35.
29. Chiang, M.C., C.W. Tsai and C.S. Yang, 2011. A time-efficient pattern reduction algorithm for K-means clustering, Information Sciences, 181(4): 716-731.
30. Kennedy, J., R.C. Eberhart and Y. Shi, 2001. Swarm Intelligence, Morgan Kaufmann, New York.
31. Mahdavi, M. and H. Abolhassani, 2009. Harmony K-means Algorithm for Document Clustering, 18(3): Springer, pp: 370-391.
32. Zhao, Y. and G. Karypis, 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering, Machine Learning 55(3): 311-331.
33. MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability, University of California Press, Berkley, CA, pp: 281-297.
34 Hartigan, J.A., 1975. Clustering Algorithms. John Wiley and Sons.
35. Selim, S.Z. and M.A. Ismail, 1984. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality, pattern analysis and machine intelligence, IEEE Transactions on PAMI, 6: 81-87.
36. Bação, Lobo, Painho, 2005. Self-organizing Maps as Substitutes for K-Means Clustering, Computational Science – ICCS Lecture Notes in Computer Science, 3516: 476-483.
37. Li, Kuo Tsai, 2010. An intelligent decision-support model using FSOM and rule extraction for crime prevention, Expert Systems with Applications, Elsevier, 37(10): 7108-7119.
38. Farnstrom and Lewis, 2007. Fast, single-pass K-means algorithms, www.citeulike.org/ user/ zador/ article/ 1772993, 2007.
39. Bouras, C. and V. Tsogkas, 2010. Assigning Web News to Clusters, in Proceedings of Conference on Internet and Web Applications and Services, pp: 1-6.
40. Taeho, J., 2009. Clustering News Groups using Inverted Index based NTSO, NDT, First International Conference on Networked Digital Technologies, pp: 1-7.
41. Dai He Sun, 2010. A Two-layer Text Clustering Approach for Retrospective News Event Detection, International Conference on Artificial Intelligence and Computational Intelligence, IEEE Computer Security, pp: 364-368.
42. Velmurugan, T. and T. Santhanan, 2011. A survey of partition based clustering algorithms in data mining: An experimental approach, Information Technology Journal, 10(3): 487-484.
43. Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, Xianyu Bao, 2015. A semantic approach for text clustering using WordNet and lexical chains, Journal of Expert Systems with Applications, (42).
44. Dueck and Frey, 2007. Non-metric af?nity propagation for unsupervised image categorization, IEEE 11th International Conference of Computer Vision, ICCV, pp: 1-8.
45. Ramadan and Mohd, 2011. A Review of Retrospective News Event Detection, International Conference on Semantic Technology and Information Retrieval, Putrajaya, Malaysia, pp: 209-214.
46. Jain, 2009. Data clustering: 50 years beyond K-means, Journal of Pattern Recognition Letters, pp: 136-713.
47. Aouf, M., L. Lyanage and S. Hansen, 2008. Review of data mining clustering techniques to analyze data with high dimensionality as applied in gene expression data, Proceeding of International Conference on Service Systems and Service Management, pp: 1-5.

49. Rana Mehrdad and Mehrnoush, Mohammad, 2013. Efficient stochastic algorithms for document clustering, Journal of Information Sciences, (220).

50. Chau, M., J. Xu and H. Chen, 2002. Extracting Meaningful Entities from Police Narrative Reports, 02 Proceedings of the 2002 Annual National Conference on Digital Government Research, pp: 1-5.