Journal of Computer Science 10 (7): 1197-1206, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.1197.1206 Published Online 10 (7) 2014 (http://www.thescipub.com/jcs.toc)

COMPARATIVE STUDY OF K-MEANS AND K-MEANS++ CLUSTERING ALGORITHMS ON CRIME DOMAIN

[1]Bashar Aubaidan, [2]Masnizah Mohd and [2]Mohammed Albared

[1]Drug Industry and Medical Appliances, Samarra-Iraq/SDI, Iraq
[2]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

Received 2013-11-21; Revised 2014-01-06; Accepted 2014-02-19

ABSTRACT

This study presents the results of an experimental study of two document clustering techniques which are k-means and k-means++. In particular, we compare the two main approaches in crime document clustering. The drawback of k-means is that the user needs to define the centroid point. This becomes more critical when dealing with document clustering because each center point represented by a word and the calculation of distance between words is not a trivial task. To overcome this problem, a k-means++ was introduced in order to find a good initial center point. Since k-means++ has not being applied before in crime document clustering, this study presented a comparative study between k-means and k-means++ to investigate whether the initialization process in k-means++ does help to get a better results than k-means. We proposes the k-means++ clustering algorithm, to identify best seed for initial cluster centers in clustering crime document. The aim of this study is to conduct a comparative study of two main clustering algorithms, namely k-means and k-means++. The method of this study includes a pre-processing phase, which in turn involves tokeniza-tion, stop-words removal and stemming. In addition, we evaluate the impact of the two similarity/distance measures (Cosine similarity and Jaccard coefficient) on the results of the two clustering algorithms. Exper-imental results on several settings of the crime data set showed that by identifying the best seed for initial cluster centers, k-mean++ can significantly (with the significance interval at 95%) work better than k-means. These results demonstrate the accuracy of k-mean++ clustering algorithm in clustering crime doc-uments.

Keywords: Crime Document Clustering, K-Means++, K-Means Algorithm, Similarity/Distance Measures

1. INTRODUCTION

This Clustering technique is a method that seeks to organize data into different classes that share identical characteristics. In this technique, intra-class similarities are maximized or minimized. It is useful for criminal investigators who want to sort crimes based on similarity or perpetrated through certain gang affiliations in order to effectively identify the criminals. In the clustering technique, it is not common for labels to be attached to data in advance. There are two important methods of this technique. Firstly, differences between crimes are determined through the process of partitioning that enables the identification of criminal trends patterns and changes that affect those trends and patterns. Secondly, data is classified into groups based on shared characteristics. This is useful when attempting to identify offenders. The clustering technique offers exciting possibilities in regards to crime detection due to the effective ways in which it can group, analyze and retrieve data. Furthermore, it carries the potential of predicting crimes based on an understanding of criminal trends and patterns derived from its sorting and arrangement of criminal data. The criminal data groups formed through the clustering technique presents the distribution of crimes in a color-coded geo-spatial manner. Criminal suspects are derived from the perpetrators of similar crimes and relationships are sought between past perpetrators and the sought after criminals. Specific characteristics of the perpetrator are

Corresponding Author: Bashar Aubaidan, Drug Industry and Medical Appliances, Samarra-Iraq/SDI, Iraq

collected from witnesses and fed into the database for both search and apprehension and archive purposes (Thongtae and Srisuk, 2008). The large volume of criminal information in its various forms presents a serious challenge in terms of how this data is stored and organized. This predicament is further exacerbated by the fact that data is not stored on a central server but is spread over a number of interfaces such as file servers, file storage facilities and personal workstations. This prevents the formation of a single comprehensive database that is structured according to a single uniform framework thus hindering the ability to effectively organize the information, identify trends and patterns and perform predictive tasks. Once collected in single storage mechanisms, the K-means algorithm can be applied to sort and process the data. It is among the most popular clustering algorithms used for large datasets over a variety of disciplines. When performing cluster analysis, this process however falls short due to its sensitivity to initial centroids or seeds (Agarwal *et al*., 2013; Bahmani *et al*., 2012; Arthur and Vassilvitskii, 2007). Its random selection of the first centroid for all documents is the source of its weakness resulting in poor clustering performance (Wu, 2012). To remedy this, this research used the k-means++ to avoid the problem of sensitivity to initial centroids. The k-means++ employs a mathematical formula to select the second initial centroid This is of particular importance for society as crime is as much a social dilemma and epidemic disease as it is a violation of the law (Alruily *et al*., 2010; Chandra *et al*., 2008). They are working in crime domain, used to build their own corpus by collecting data from multiple resources such as news portals and police databases. However, working within the crime data represents an interesting dilemma. This is because the diversity modalities of crimes and the difficulties of collecting the data due to the privacy issue. Alruily *et al*., (2010) presented a system that combines two text-mining techniques, namely information extraction and clustering. It functions by adopting a rule-based approach to extract information. This study is split into four main sections: In Section two, we discussed related works on crime processed document. Then in Section three, we described the How do we implementation out our review and Section four will be on the experimental findings and finally, Section five and will be conclusion our work.

2. CRIME DOMAIN RELATED WORKS

This study presents a comparative study of two main clustering algorithms, namely k-means and k-means++. In this section, we will offer our review to the work of related to crime document compilation. Most of the work is based on reviewed in the machine learning approach unsupervised. Crime related data is often made private circulating predominantly among legal and law enforcement agencies. However, some data is made publically accessible. Public data is often in the form of news reports, which are often many and can greatly differing in their account of the crime. In order to utilize such data for crime prevention and containment, such data must be collected within a single framework and ordered according to a single comprehensive taxonomy. From here, crime patterns can be identified. Alruily *et al.* (2010) presented a system that combines two text-mining techniques, namely information extraction and clustering. It functions by adopting a rule-based approach to extract information. To achieve this, it searches for dependency relations between intransitive verbs and prepositions. This approach is ideal for identifying crime types and extracting them from a certain crime domain. This is followed by the clustering process that employs the Self Organizing Map (SOM) to cluster Arabic crime documents. The results are validated through experiments that indicate that these techniques are promising. 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. Bache and Crestani (2010) constructed the police dataset from solved cases, which they treated as unsolved. This was a strategic move as it tactfully maneuvers around red-tape and classified criminal documents. Such forms of privacy have proven to be formidable challenges in earlier attempts to develop a crime database. Aouf *et al*. (2008); (Ali *et al*., 2012) they compared the effectiveness of single pass clustering and k-means in detecting crime topics and aiding in the identification of events or crimes. They also experimented on enhanced k-means clustering in order to select the optimal initial centroid to be automatically compared with regular k-means in order to randomly choose the initial centroid. Jo (2009) finding revealed that using k-means generated the best results, not only at the level of internal measurement of the clustering index function, but also on real users' experimentation. Furthermore, when comparing k-means, single pass clustering and other approaches of clustering news topics revealed that k-means was better than single pass clustering.

3. MATERIALS AND METHODS

In this study framework for crime document clustering contains with the following Phases; (i) first phase– crime document preprocessing, (ii) second phase-build the Documents representation, (ii) third phase-documents are clustered based on the K-means and k-means++ apply, also similarity/distance measure for each algorithm, (iv) fourth phase-the comparative Analysis and evaluation of clustering is carried out by using overall purity and overall F-measure **Fig. 1** shows the framework of the crime document clustering.

3.1. Crimes Text Pre-Processing

The crime document clustering consists of three phases of crime document Pre-processing, which are; (1) Tokenizing (2) Stop Word Removal, (3) stemming. Detailed explanation was given in the following Subsections.

3.2. Tokenization

The first step of morphological analyses is the tokenization. The aim of the tokenization is the exploration of the words in a sentence. Textual data is only a block of characters at the beginning. All following processes in information retrieval require the words of the data set.Hence the requirement for a parser which processes the tokenization of the documents. This may sound trivial as the text is already stored in machine-readable formats. Nevertheless, some problems are still left, like the removal of punctuation marks. Other characters like brackets, hyphens require a processing as well. Furthermore, tokenized can cater for consistency in the documents. The main use of tokenization is identifying the meaningful keywords (Kumar and Chandrasekhar, 2012). According to (Bruce *et al*., 2009), conversion of the text of a document into data, which is suitable for analysing using with machine learning algorithm, usually requires that, the text should be broken into discrete units, separated by a space or other special marker, which is inserted among them, so that each unit corresponds to a word in the text:

- Goal: To separate text into individual words
- Example: “We have arrested the criminal.”→We_have_ arrested_the_criminal.

3.3. Stop Word Removal

Generally, documents usually found to contain a lot of unnecessary words in English, such as, pronouns, prepositions, conjunctions and others, which are usually used by authors for the purpose of linguistically enhancing the structures and in particular, focusing on the syntactic or grammatical function of the language, rather than strengthening the semantic function or the meaning of the content. These words which are so frequently found in the texts and which do not provide more Valuable information about the text content are called stop words. Therefore, in this particular regard, the process of word removal is very common and of considerable importance to be involved in Document Clustering. This is because, by carrying out word removal. The dimensionality of the terms space will be drastically reduced. stop word as a list of 571 Stop words and are called so, these are generally regarded stop words because they tend to convey syntactic functions, rather than conveying more than they convey semantic functions, such as, carrying further meaning, which can enhance and strengthen the communicative or informational aspects of the document content. Thus, by carrying out the word removal process, conveying the meaning of the document or text content will be clearer and interpreting the meaning will be easier (Salton *et al*., 1975; Lazarinis, 2007). The Goal: To remove common words that is usually not useful for text classification:

Example: To remove words such as “a”, “the”, “I”, “he”, “she”, “is”, “are.

It is stated that, stop word removal has been carried out by many search engines, with the aim of supporting or providing users or text developers with queries, to gain better results by searching for meaning or information, rather than searching for functional words (Bruce *et al*., 2009).

3.4. Stemming

Word stemming is regarded as one of the most important factors of pre-processing tasks, which is expected to have effect on the effective impact the performance of Document Clustering systems. Stemming It is defined as the process of prefix removal (letters, which are added to the beginning of the word root) and suffix removal (letters, which are added at the end of the word root) (Larkey *et al*., 2002). In our study, we have used porter stemmer Goal: To normalize words derived from the same root.

Example: In applying the stemming process to the two variants of the same word “Arraignment”, “arraigned”, these variants need to be reduced or returned to their common representation “arraign”.

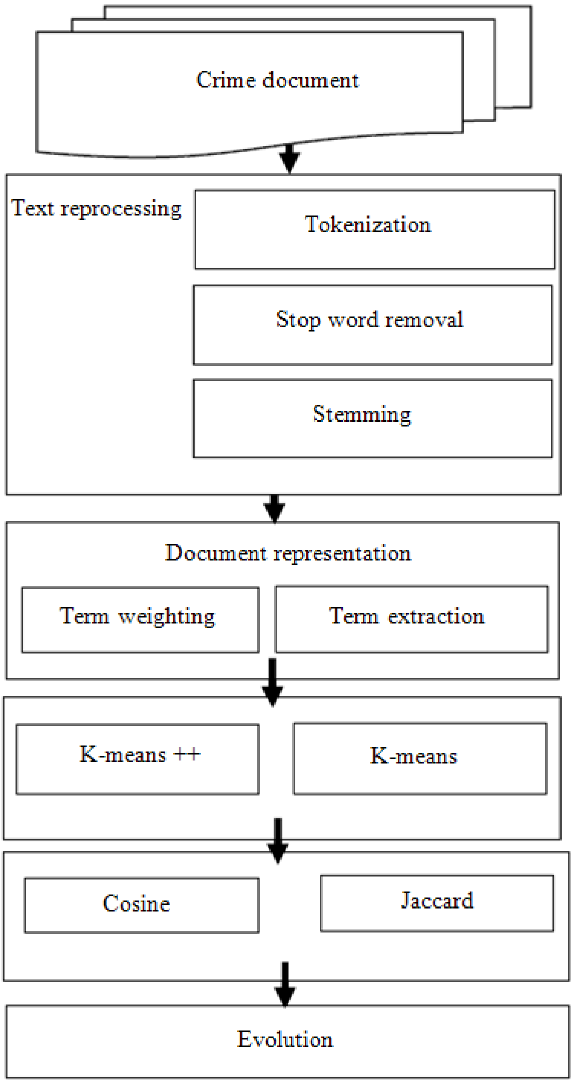

Fig. 1. The clustering methodology

3.5. Text Representation

Directly In applying most learning algorithms to text information, in a direct way without representing, it has been proved to be impossible, due to the complex nature of the text information. Therefore, before applying the text using to a machine learning method, it is essential to converting the content of a textual document to a compact representation is necessary. They are Document representation has been found to be efficiently used as a language-independent method, since they are it is independent of the meaning of the language and perform well in case of noisy text (Khreisat, 2006).

3.6. Term Extraction

In general, it is recognized that the indexing terms which represent documents. There are four kinds of term type representations namely; sub-word level (n-gram, which is used to re-print linguistic units, smaller than a word, such as, morphemes, syllables), word-level (which is used for a single token, representing a single word), multi-word level (phrases, sentences) and semantic or syntactic level. It is also stated by (Man and Lim, 2007) that, the bag-of-words representation is viewed as the most commonly used way among all these ways of for term type representation. It is most advantageous for being simple, because by using it, only the frequency of a word presented in the document has to be recorded, while all other things aspects such as, the structure and the ordering of the words are not needed or ignored. Therefore, in the current study, the Bag Of Words (BOW) was has been used as a term extraction.

3.7. Term Frequency Weighting

Term Frequency (TF) weighting is also recognized as a simple method for term weighting:

$$W_i = tf_i \, .\log\left(\frac{N}{n_i}\right)$$

According to this method, there is an equality of the weight of a term in a document and the number of times of appearance of this term in the document, i.e., to the raw frequency of the term in the document.

3.8. Term Frequency×Inverse Document Frequency Weighting

It is pointed that Boolean weighting and TF weighting do not take the frequency of the term into consideration throughout all the documents in the document corpus. Term Frequency × Inverse Document Frequency (TF×IDF) weighting is seen as the most popular method used for term weighting, since it considers this property. By using this approach, assigning the weight of term i in document d to the number of times the term appears in the document is proportional and it is in inverse proportion to the number of documents in the corpus, in which the term appears.

3.9. Similarity Measures

Document clustering is the process in which similar documents are grouped to form a coherent cluster. However, complications arise in how to determine if a pair of documents is similar or different. This is not always a straightforward process. In view of the variety of scales, distance measurements (or metrics) between clusters need to be carefully selected. The difference between two patterns is commonly calculated by means of the distance between clusters. The accuracy of clustering depends on a precise definition of the closeness between a pair of objects, in terms of either the pair wise similarity or distance. This research will focus on well-known measures of distance between patterns. In this regard, this study focuses on the cosine similarity and Jaccard coefficient as similarity or dis-tance measures (Steinbach *et al*., 2000).

3.10. Cosine

Cosine similarity is one of the most well-known similarity measures which are applied to text docu-ments such as in numerous information retrieval appli-cations (Baeza-Yates and Ribeiro-Neto, 1999) and clustering (Larsen and Aone, 1999). In measuring the given two documents $\vec{t_a}$ and $\vec{t_b}$, their cosine similarity is:

$$\mathrm{SIM}_c\left(\vec{t}_a, \vec{t}_b\right).\frac{\vec{t}_a.\vec{t}_b}{\left|\vec{t}_a\right| * \left|\vec{t}_b\right|}$$

3.11. Jaccard

The Jaccard coefficient, which is another similarity measure, also known as the Tanimoto coefficient, is used to measure the similarity in the intersection divided by the union of the objects. For text document, the use of Jaccard coefficient is to make a comparison of the sum weight of shared terms and the sum weight of terms presented in either of the two documents but in condition that they are not the shared terms. The formal definition is as follows:

$$\mathrm{SIM}_j\left(\vec{t}_a, \vec{t}_b\right).\frac{\vec{t}_a.\vec{t}_b}{\left|\vec{t}_a\right|^2 * \left|\vec{t}_b\right|^2 - \vec{t}_a.\vec{t}_b}$$

3.12. Clustering Techniques

The clustering technique has advantage of the classification technique in that it has better predictive capabilities as it draws from both solved and unsolved criminal cases as opposed to the classification technique, which solely draws from information collected from solved crimes.

3.13. K-Means Clustering

K-means is the most well-known clustering method due to its easy implementation and rapid convergence (Macqueen, 1967). However, this method is limited in that it is significantly influenced by the choice of initial solution. Its time complexity when t iterations are per-formed on a sample size of m items, each characterized by n attributes:

- t: Iteration
- n: Attributes
- k: Number of clusters
- m: Item

K denotes the number of clusters. While it has a linear time complexity both in the number of instances and attributes, it would be very slow when t and k are fixed but the number of instances or the dimension of data increases. Many works have been proposed to improve the efficiency of the original k-means (Kanungo *et al*., 2002; Jain and Dubes, 1988) defines that clus-ter analysis is the process of classifying objects into groups of similar objects based on a similarity/distance measure. Although k-means has been applied in:

- A wide number of different fields including text mining, information retrieval and ma-chine learning
- K-means remains the most commonly used technique due to its simplicity

As such, the k-means approach is selected instead of its fast variants for comparative purposes throughout the experiments conducted in this research. Moreover, as k-means is adopted in the algorithm proposed in this research, fast variants of k-means can be used to improve its speed K-means is a method that has been widely used for partitional clustering with a linear time complexity (Steinbach *et al*., 2000). As stated by (Hartigan, 1975), the k-means algorithm argues that the mean of the documents assigned to that cluster represents each of the k-clusters and as a result the k-means technique is largely regarded as the centroid of that cluster. The benefit of k-means clustering are simple and flexible easy to understand and can be easily to implemented However the disadvantages of k-means clustering are user need to identify the number of cluster in advance (Vora and Oza, 2013). According to (Berkhin, 2006), there are two versions of k-means algorithm. Following two major steps:

- Reassigning all the documents to their nearest centroids
- Recomposing centroids of newly assembled groups

3.14. K-Means++ Clustering

K-means++ is a simple probabilistic means of initializing for k-means clustering that not only has the best known theoretical guarantees on expected out-come quality, but works very well in practice. In this regard, the essential component required is the preser-vation of the diversity of seeds while ensuring that the outliers remain robust. The primary concern of the k-means problem is to identify cluster centers that mini-mize intra-class variance by reducing the distances from each clustered data point. This can be achieved through an effective and well-designed cluster-initialization technique. In applied statistics, k-means++, as in (Arthur and Vassilvitskii, 2007), is an algorithm for choosing the initial values (or "seeds") for the k-means clustering algorithm. It was proposed by (Arthur and Vassilvitskii, 2007). A way of avoiding the sometimes poor clustering found by the standard k-means algorithm. The k-means algo-rithm begins with an arbitrary set of cluster centres. We have proposed a specific way of choosing these cen-tres. At any given time, let D(x) denote the shortest distance from a data point x to the closest centre we have already chosen. Then, we define the following algorithm, which we call k-means++, as in Arthur and Vassilvitskii (2007). The algorithm is similar to k-means:

- Choose an initial center c1 uniformly at random from X
- Hoose the next center c_i, selecting ci x' ϵX with probability $D(\dot{X})^2 / \sum_{x \in X} D(x)^2$
- Repeat Step 2 until we have chosen a total of k centers
- Repeat Step 2-4 with the standard k-means algorithm

We call the weighting used in Step 2 simply "D2 weighting". This seeding method gives out considera-ble improvements in the final error of k-means. Alt-hough the initial selection in the algorithm takes extra time, the k-means part itself converges very fast after this seeding and thus the algorithm actually lowers the computation time too. It is noteworthy that, the authors as in (Arthur and Vassilvitskii, 2006) have tested their method with real and synthetic datasets and obtained typically 2-fold improvements in speed and for certain datasets close to 1000-fold improvements in error. Additionally, the authors as in (Arthur and Vassilvitskii, 2006) they have also calculated an approximation ratio for their algorithm. The k-means++ algorithm guarantees an approximation ratio O (log k), where k is the number of clusters used. This is in contrast to k-means, which can generate clustering arbitrarily worse than the optimum, as in (Charikar *et al*., 2004). It is hoped that this work will contribute significantly to the area of document clustering criminal news. This major contribution which can be advocated by the current study to this area is represented in its comparison between k-means and k-means++ to investigate whether the ini-tialization process in k-means++ leads to results better than those produced by k-means. In this context, this study proposes the k-means++ clustering algorithm to identify the best seed for initial cluster centres for clus-tering crime documents. This study presents a compar-ative study of two main clustering algorithms, namely the k-means and k-means++.

4. EXPERIMENTAL RESULTS OF DOCUMENT CLUSTERING

The first experiment was aimed at clustering the documents under different groups of topics and events, in order to examine the effect on clustering, in this research uses four experiments based on the number of topics and events used. Two topics of Canny Ong and Mona Fandy were used in the first experiment. The second experiment was performed to examine the ef-fect on clustering of the four different groups of topics. The experiments were made to examine the effect the topics are Canny Ong, Mona Fandy, Noritta Samsudin and Nurin Jazlin were used in the second experiment. Six topics were used in the third experiment. The fourth experiment used 168 events for all data set. The crime data set and testing data in this study were pro-cessed by tokenization and stemming to avoid prepro-cessing text.

4.1. Data Description

The crime dataset used in this study includes 247 documents collected from the website of Bernama news (http://www.blis.bernama.com). This dataset iscomposed of six topics, which includes articles of Canny Ong, Mona Fandy, Noritta Samsudin, Nurin Jazlin, , Sharlinie Mohd Nashar and Sosilawati articles which consist of 168 events Shown in **Table 1**.

4.2. Evaluation Metrics

In evaluating cluster quality, two kinds of measures namely; internal quality measure and external quality measure (Steinbach *et al*., 2000) are used for this purpose. The internal quality measure does not make a use of the external knowledge, such as, class label information, for evaluating the produced clustering solution.

Table 1. The used dataset statistics

Topic	Event	Number of crime document
1.Canny ong	30	48
2.Mona fandy	30	35
3.Noritta samsudin	27	35
4.Nurin jazlin	28	59
5.Sharlinie Mohd- nashar	24	35
6.Sosilawati	29	35
Total	168	247

In contrary, the external quality measure mainly depends on the labeled test of the document corpora. Its methodology is to make a comparison between the resulting cluster and labelled classes and to measure the extent, to which documents from the same class or category are assigned to the same cluster. In the current study, purity is used as an external quality measure and another anther external quality measures known as F-measure, which is the most commonly used measures in text mining:

4.3. PURITY

Purity measures the degree of occurrence of documents from primarily one class in each cluster. For a specific cluster j of size n_j, purity of this cluster is defined as:

$$P_j = 1/n_j \max n_{ij}$$

where, n_{ij} is used to indicate a number of documents of class i being assigned to cluster j. So p_j, is defined as the fraction of the overall cluster size that is the largest class of documents which are assigned to that cluster which constitutes. The overall purity of the clustering solution is gained by the total weighted sum of indi-vidual cluster purities:

$$P = \sum_j \frac{n_j}{n} p_J$$

Whereas N is used to refer to a total number of documents in the document collection. In general, when the values of purity are larger, the clustering solution is found to be better.

4.4. F- MEASURE

The F-measure cluster evaluation metric has a combination of the precision and recall ideas from information retrieval. Each cluster is regarded as if it was the results of a query and each class is perceived as if it were the desired set of documents for the query. The calculation of the recall and precision for each cluster j and class i am presented as follows:

$$\text{Recall} = \frac{a}{a+c}$$

$$\text{Precision} = \frac{a}{a+b}$$

- $n_{i,j} = a$
- $n_i = a+c$
- $n_j = a+b$

Here n_{ij}, represents the number of documents having the class label i in cluster j and n_i refers to the number of documents having the class label i. Finally, n_j is the number of documents in cluster j. The calculation of the F-measure of cluster j and class i is presented as follows:

4.5. F(i,j) =2 Recall(i,j) Precision(i,j)/Recall(i,j)+Precision(i,j)

To calculate the overall value for the F-measure, the weighted average of all values for the F-measure is taken as follows:

$$F = \sum_i \frac{N_i}{N} \max F(i,j)$$

Thus, it is noticed that the F-measure values occur at the interval (0, 1) and the larger F-measure values are correspondent to the higher clustering quality.

4.6. Experiments and Result

These experiments were measured using the overall F-measure and overall purity, on this section discusses the four experiments based on the number of topics and events used as shown in **Table 2**.

The crime data set and testing data in this study were processed by tokenization and stop word removal and stemming, with two similarity distance measure cosine and jaccard were then used on k-means and k-means++ based on overall f-measure and over all purity. These experiments are to evaluate the difference in the results, when the number of topic is increased. A system was established to cluster the algorithms. The results were reported using the standard measurement evaluation performance in **Table 3-6** which show the results of the overall purity and overall f-measure evaluation of the experimental method. Purity and more effectiveness on the purity and more effec-tiveness on the crime document Clustering of.

Table 2. Illustrated the four experiments setting

Experiment	Topic
The first	2 topics (Canny ong, T1 and mona fandy, T2)
The second	4 topics (Canny ong, T1; mona Fandy, T2; noritta samsudin, T3; and nurin jazlin, T4)
The third	6 topics: Canny ong, T1; mona Fandy, T2; noritta samsudin, T3; and nurin jazlin, T4; sharlinie mohd Nashar, T5; and sosilawati, T6
The fourth	168events

Table 3. First experiment: Overall f-measure evaluation and overall purity on two topics (canny ong and nurin jazlin)

Evaluation	K-MEANS++		K-MEANS	
T1&T2	Cosine	Jaccard	Cosine	Jaccard
Overall F-measure	0.910	0.834	0.802	0.681
Overall purity	0.916	0.868	0.838	0.756

Table 4. Second experiment: Overall f-measure and overall purity evaluation on 4 topics (canny ong, mona fandy, noritta samsudin, nurin jazlin)

Evaluation	K-MEANS++		K-MEANS	
T1&T2&T3&T4	Cosine	Jaccard	Cosine	Jaccard
Overall F-measure	0.651	0.637	0.532	0.582
Overall purity	0.646	0.661	0.601	0.592

Table 5. Overall f-measure and overall purity evaluation on 6 topics topics (canny ong, mona fandy, noritta samsudin, nurin jazlin, sharlinie mohd nashar and sosilawati articles)

Evaluation T1&T2&T3	K-MEANS++		K-MEANS\	
T4&T5&T6	Cosine	Jaccard	Cosine	Jaccard
Overall F-measure	0.73	0.68	0.620	0.61
Overall purity	0.74	0.68	0.642	0.62

Table 6. Overall F-measure and overall purity evaluation on (168 event)

Evaluation	K-MEANS++		K-MEANS	
168 event	Cosine	Jaccard	Cosine	Jaccard
Overall F-measure	0.890	0.810	0.819	0.622
Overall Purity	0.912	0.827	0.781	0.688

This has been evidenced by k-means++ better than k-means and the best value of k-means++with overall purity esti-mated based on cosine similarity is (0.916) and the overall F-measure (0.910) as compared to clustering with methods based on k-means (overall purity-0.838 overall F-masure-0.0802) purity of document clustering evaluation which has not previously been applied in this area and particularly in crime document clustering crime document Clustering of. This has been evidenced by k-means++ better than k-means and the best value of k-means++with overall purity. Estimated based on cosine similarity is (0.916) and the overall F-measure (0.910) as compared to clustering with methods based on k-means (overall purity-0.838 overall F-masure-0.0802) purity of document clustering evaluation which has not previously been applied in this area and partic-ularly in crime document clustering.

Base on the results of the four Experiments, K-means ++ has been proved to be better and more accurse than the k-means clustering regardless of the two similarity measures used: Cosine and Jaccard.

5. CONCLUSION

This study was aimed to investigate the best similarity in k-means and k-means++ for crime document and to evaluate and compare the performance of k-means and k-means++ in clustering. In this study, we had used crime Dataset collected from Bernama news and have tested six categories of topics. Based on the results in section 4, the K-means++ algorithm has the best results with Cosine similarity compared to Jaccard similarity. Experimental method, based on K-means ++, has been proved to be better and more accurse than the k-means clustering, in crime document clustering The results show that the k-means++ outperforms the k-means and that cosine similarity performs better than the Jaccard coefficient. The reason for this is due to the fact that the k-means identifies the first initial centroid randomly, while the k-means++ algorithm selects the second initial centroid mathematically through proba-bility proportional to the square of the distance over summation of the square distance for the current point As for the performance of the cosine similarity, it outperformed the Jaccard coefficient because it is independent of document length and the data set consisted of documents with different lengths. Based on these findings, it is recommended that it is better to choose a smaller number of topic rather than a larger number, due to the possible occurrence of problems in the rate of similarity between few topics are easy to detect while difficulty of detecting when the rate similarity between the many topics in the other word there are problem when the thousands topic. Based on the find-ings of this study, some points are future work for pur-suer future research:

- A combination of different similarity/distance measure is planned as to make a representation of the docu-ments so that it can be more enriching with terms weighting
- It is also recommended that the extension of this work in the future can be done by applying the (k-means++) algorithm for other languages such as Malay 3. A combination of different data set (art sport eco-nomic) or large document dataset can be experiments on several real-world datasets

6. ACKNOWLEDGMENT

The study was supported by the University Arus Perdana Research Grant (UKM-AP-ICT-21-2010).

7. REFERENCES

Agarwal, M., R. Jaiswal and A. Pal, 2013. k-means++ under Approximation Stability. In: Theory and Applications of Models of Computation, T.H., L.C. Lau and L. Trevisan (Eds.)., Springer Berlin Heidelberg, ISBN-10: 3642382355, pp: 84-95.

Ali, N.M., M. Mohd, N.F. Yacob, S. Saad and N. Omar *et al*., 2012. Investigating user perception in a develop-ment of crime news retrieval system. Int. J. Digital Content Technol. Applic.

Alruily, M., A. Ayesh and A. Al-Marghilani, 2010. Using self organizing map to cluster arabic crime documents. Proceedings of the International Multiconference on Computer Science and Information Technology, Oct. 18-20, IEEE Xplore Press, Wisla, pp: 357-363.

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. Proceedings of the International Conference on Service Systems and Service Management, Jun. 30-Jul. 2, IEEE Xplore Press, Melbourne, VIC., pp: 1-5. DOI: 10.1109/ICSSSM.2008.4598505

Arthur, D. and S. Vassilvitskii, 2006. Worst-case and smoothed analysis of the icp algorithm, with an application to the k-means method. Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, (CS ‘06), DC, USA, pp: 153-164. DOI: 10.1109/FOCS.2006.79

Arthur, D. and S. Vassilvitskii, 2007. K-Means++: The advantages of careful seeding. Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms, (DA ‘07), PA, USA, pp: 1027-1035.

Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval Addison-Wesley. 1st Edn., New York, ACM press New York, pp: 463.

Bahmani, B., B. Moseley, A. Vattani, R. Kumar and S. Vassilvitskii, 2012. Scalable K-Means++. Proc. VLDB Endowment 5: 622-633.

Berkhin, P., 2006. A Survey of Clustering Data Mining Techniques. In: Grouping Multidimensional Data, Jacob Kogan, C. Nicholas and M. Teboulle (Eds.)., Springer, New York, ISBN-10: 3540283498, pp: 25-71.

Bruce, W., C. Metzler and T. Strohman, 2009. Search Engines: Information Retrieval in Practice. 1st Edn., Pearson Education, Upper Saddle River, ISBN-10: 0131364898, pp: 524.

Bache, R. and F. Crestani, 2010. An approach to indexing and clustering news stories using continuous language models. Proceedings of the Natural Language Processing and Information Systems, Jun. 23-25, Springer Berlin Heidelberg, Cardiff, pp: 109-116. DOI: 10.1007/978-3-642-13881-2_11

Chandra, B., M. Gupta and M. Gupta, 2008. A multi-variate time series clustering approach for crime trends prediction. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Oct. 12-15, IEEE Xplore Press, Singapore, pp: 892-896. DOI: 10.1109/ICSMC.2008.4811393

Charikar, M., C. Chekuri, T. Feder and R. Motwani, 2004. Incremental clustering and dynamic information retrieval. SIAM J. Comput., 33: 1417-1440. DOI: 10.1137/S0097539702418498

Hartigan, J.A., 1975. Clustering Algorithms. 1st Edn., John Wiley and Sons, Inc.

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. 1st Edn., Prentice-Hall, ISBN-10: 013022278X, pp: 320.

Jo, T. 2009. Clustering news groups using inverted index based ntso. Proceedings of the 1st International Conference on Networked Digital Technologies, Jul. 28-31, IEEE Xplore Press, Ostrava, pp: 1-7. DOI: 10.1109/NDT.2009.5272194

Kanungo, T., D.M. Mount, N.S. Netanyahu, C.D. Piatko and R. Silverman *et al*., 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Patt. Anal. Mach. Intell.., 24: 881-892. DOI: 10.1109/TPAMI.2002.1017616

Khreisat, L., 2006. Arabic text classification using N-gram frequency statistics a comparative study. Proceedigns of the Conference on Data Mining (IN ‘06), pp: 78-82.

Kumar, A.A. and S. Chandrasekhar, 2012. Text data pre-processing and dimensionality reduction techniques for document clustering. Int. J. Eng. Res. Technol.

Larkey, S., L. Ballesteros and E. Connell, 2002. Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 11-15, ACM New York, NY, USA., pp: 275-282. DOI: 10.1145/564376.564425

Larsen, B. and C. Aone, 1999. Fast and effective text mining using linear-time document clustering. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 15-18, ACM New York, NY, USA., pp: 16-22. DOI: 10.1145/312129.312186

Lazarinis, F., 2007. Engineering and utilizing a stopword list in greek web retrieval. J. Am. Soc. Inform. Sci. Technol., 58: 1645-1652. DOI: 10.1002/asi.20648

Macqueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, (SP’ 67), University of California Press, pp: 14-14.

Man, L. and C.T. Lim, 2007. Text representations for text categorization: A case study in biomedical domain. Proceedings of the International Joint Conference on Neu-ral Networks, Aug. 12-17, IEEE Xplore Press, Orlando, FL., pp: 2557-2562. DOI: 10.1109/IJCNN.2007.4371361

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620. DOI: 10.1145/361219.361220

Steinbach, M., G. Karypis and V. Kumar, 2000. A comparison of document clustering techniques. Proceedings of the KDD Workshop on Text Mining, (TM’ 00), pp: 525-526.

Thongtae, P. and S. Srisuk, 2008. An analysis of data mining applications in crime domain. proceedings of the IEEE 8th International Conference on Computer and Information Technology Workshops, Jul. 8-11, IEEE Xplore Press, Sydney, QLD, pp: 122-126. DOI: 10.1109/CIT.2008.Workshops.80

Vora, P. and B. Oza, 2013. A survey on k-mean clustering and particle swarm optimization.

Wu, J., 2012. Cluster Analysis and K-Means Cluster-ing: An Introduction. Advances in K-Means Clustering.