International Journal of Engineering & Technology, 7 (4.11) (2018) 246-251

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Comparative Study of Document Clustering Algorithms

N. M. Ariff*, M. A. A. Bakar, M. I. Rahmad

School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
**Corresponding author E-mail: tqah@ukm.edu.my*

Abstract

Text clustering is a data mining technique that is becoming more important in present studies. Document clustering makes use of text clustering to divide documents according to the various topics. The choice of words in document clustering is important to ensure that the document can be classified correctly. Three different methods of clustering which are hierarchical clustering, k-means and k-medoids are used and compared in this study in order to identify the best method which produce the best result in document clustering. The three methods are applied on 60 sports articles involving four different types of sports. The k-medoids clustering produced the worst result while k-means clustering is found to be more sensitive towards general words. Therefore, the method of hierarchical clustering is deemed more stable to produce a meaningful result in document clustering analysis.

Keywords: *document clustering; text mining; hierarchical clustering; k-means; k-medoids.*

1. Introduction

Text mining covers various field of research including biology, biomedicine, social science and economy. For example, text mining had been used in biomedicine to investigate the relationship between genes and symptoms of breast cancer [1] as well as to retrieve information on the conditions of patients accurately and comprehensively through doctors' notes [2]. Meanwhile, text mining on financial statements had also been used to detect outliers in order to successfully identify fraud [3]. In social science, text mining was used on tweets from Twitter Application shared by ten libraries of renowned universities around the world to look at the similarities and differences of words used in the said social media [4].

Text clustering is a part of text mining and it is used for document clustering. Document clustering categorize documents into related topics based on the similarities found in texts from the various documents. For example, it had been used to differentiate documents into various fields such as business, sports, politics, technologies and entertainments [5, 6]. There are various types of clustering techniques, which had been used for document clustering such as hierarchical clustering, k-means and k-medoids [7, 8]. Each technique has its advantages and weaknesses in producing the most suitable clusters.

In this study, three clustering techniques which are hierarchical clustering, k-means and k-medoids are compared in order to find out which algorithm is the best in identifying clusters of documents. Although these three clustering algorithms are common, they are among the most popular clustering algorithms and are still widely used even in recent researches [9-14].

2. Text Mining

Before any text mining process can be done, texts in articles and documents have to first go through data cleaning process. Unimportant words which are also known as "stop words" have to be removed. These words are general words that contain little to no information at all, i.e. small value of entropy [15]. Hence, by removing these words, the results of text mining will be more accurate. Furthermore, punctuations are also considered not important and should be removed from texts in the documents. Next, words with affix such as '-s', '-ed' and '-ing' are treated so that they contain the same information as their base forms. This step is known as "word stemming" and it is one of the necessary step in text data cleaning.

After, the cleaning process is completed, one of the first step in text mining is to determine the frequency of words in the articles and documents used in the study. The remaining words are counted one by one to produce the frequency tables for all words in the documents. From these tables, the words with the highest frequencies and so on can be identified and word clouds can be obtained. Hence, patterns as well as the distributions of words in the documents can be explored and the texts in different documents can be compared and analyzed.

3. Document Clustering

Cluster analysis is an analysis that groups objects with the same characteristics into one group and objects with different characteristics into different groups [16]. The word frequencies found from the previous step are used to identify and cluster documents into various groups by incorporating cluster algorithms. Clusters produced by using all the words in the documents are often not significant. Hence, some words are selected carefully in order to cluster the documents accurately. Three methods of clustering are used in this study, which are hierarchical clustering, k-means and k-medoids. The clusters produced by these three methods are compared to find the most suitable algorithm for document clustering.

3.1. Hierarchical Clustering

Hierarchical clustering algorithm is regarded as one of the oldest and main streams clustering methods. It is widely used in many scientific applications [9]. This is because it is applicable to most types of data and it does not require any predefined parameter which makes it suitable for handling real-world data [10]. Two common types of hierarchical clustering are divisive and agglomerative approach. The divisive approach starts off with all the documents being in one cluster. Then, the cluster is divided by splitting documents that have big values of distances between each other into different clusters. These distances are based on the characteristics of variables under study. This splitting process is repeated until the targeted number of clusters have been reached or until each document is in its own cluster. Meanwhile, the agglomerative approach starts with each document representing a different cluster. These documents are then combined if the distances between documents are small, since small distances imply that the characteristics of variables in the documents are similar. This combining process is repeated until the targeted number of clusters are obtained or until all the documents are combined as one big cluster. This study focuses on the agglomerative approach since it is believed to obtain a more meaningful and reflective result. The steps for hierarchical clustering in this study can be simplified as follows:

1. The frequencies of selected words in each document are counted.
2. Distance between documents are calculated. The complete-linkage clustering using Euclidean distance is used in this study since the range of differences in word frequencies are quite big. The complete-linkage clustering method is defined as:

$$\mathrm{D}\left(X_i, Y_j\right) = \max_{X_i \in C_s,\, Y_j \in C_t} \mathrm{d}(X_i, Y_j) \tag{1}$$

where X_i and Y_j are documents i and j in cluster C_s and C_t respectively. The Euclidean distance measure $\mathrm{d}(X_i, Y_j)$ is obtained as follows:

$$\mathrm{d}(X_i Y_j) = \sqrt{\sum_{k=1}^{8} \left(x_{i,k} - y_{j,k}\right)^2} \tag{2}$$

with $x_{i,k}$ and $y_{j,k}$ are the frequencies of word k in X_i and Y_j respectively.

3. Clusters C_s and C_t which have minimum distance of $\mathrm{D}\left(X_i, Y_j\right)$ are then combined into a new cluster.
4. Step 2 and 3 are then repeated until the targeted number of clusters obtained.

3.2. k-means clustering

The k-means clustering algorithm is one of the common and popular algorithm due to its simplicity and ease of implementation [11]. Furthermore, it achieves faster convergence in finding an optimum local solution [12]. The method of k-means clustering is performed by getting k number of centroids and clustering nearby objects to the centroids [17]. The process of k-means algorithm can be simplified into five simple steps as follows:

1. The targeted number of clusters, k, is determined. Each cluster has a centroid. If k = 4, then four documents are selected randomly as centroids for each cluster respectively.
2. The Euclidean distance between each document and each of the centroid is calculated using in (2).
3. Each document is put into the group which has a centroid that provides the minimum distance value between the document and the group's centroid.
4. For each newly formed group, the mean value for the variables under study based on all the documents in the group is calculated and taken as the new centroid for the group.
5. Step 2 to 4 are repeated for the new centroids. This process is stopped once no changes to the members of each group and to the values of centroids can be seen.

3.3. k-medoids clustering

The method of k-medoids clustering is a clustering algorithm similar to k-means, where it uses medoids to replace centroids. Medoids are objects which are deemed as the most middle object for each cluster [18]. The function of medoids is similar to that of centroids, but the coordinates for a medoid is the real coordinate of the most middle object. The logic behind the algorithm is to minimize the total uniqueness between each object and its reference point which is the medoid [13]. Medoids can be looked at as the medians of the groups while centroids are the means of the groups. The algorithm is thus less sensitive to outliers than k-means [14]. In this study, the partitioning around medoids (PAM) algorithm is used and the steps in this algorithm can be simplified as follows:

1. The targeted number of clusters, k, is determined. Similar to k-means clustering, each cluster will have a medoid. Hence, if k = 4, then four documents are selected randomly as medoids for each cluster respectively.
2. The Euclidean distance between each document and each medoids is calculated using in (2).
3. Each document is assigned to the group, where it has the minimum distance between the group's medoid and the document.
4. From the newly formed groups, the sum for all the distances between all the documents and their respective medoids are obtained.
5. Exchange one of the medoid with another document which is not a medoid and repeat Step 2 to Step 4. If the sum of the distances in Step 4 increases, then the original medoid is used. While if the sum decreases, the new medoid is used.
6. Repeat Step 5 for all combinations of medoids.

3.4. Silhouette Index

Silhouette index is used to compare the clusters obtained from the three methods under study. The index is defined as [19]

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{3}$$

with a_i is the intra-cluster distance which is the mean distance between document i and other documents in the same cluster. The mean distance between document i and documents from other clusters is also calculated and the cluster that gives the minimum mean distance is known as the neighbouring cluster for document i. The mean distance between document i and documents in the neighbouring cluster is taken as the value of b_i. The value of b_i is also known as the inter-cluster distance. The value for the silhouette index, s_i, is between the close interval [-1, 1]. If s_i is close to one, then document i is deemed to be in a suitable cluster for it; if s_i is close to -1, then document i should be in its neighbouring cluster instead; and if s_i has a value near zero, then document i is said to be at the border of its cluster and the neighbouring cluster [20]. The mean value of s_i for all documents provide a measure of suitability of the clusters formed. The mean of s_i is used to compare the clusters obtained and the clusters that give the mean value nearest to one are chosen as the best combination of clusters produced. A positive value shows that the combination of clusters is good since it implies that the intra-cluster distance is smaller than the inter-cluster distance. This means that on average, the differences between documents in the same cluster is smaller than the differences between documents in the cluster with documents in the neighbouring cluster.

4. Case Study

4.1. Data

Data for the case study is taken from articles from “The Star” website. The data set consists of online sports articles published by “The Star” from 8th October to 19th October 2016. Sixty articles from four different types of sports are chosen; badminton, football, tennis and motorsports. There are 15 articles for each type of sport. Each article contains 150 to 650 words and there are 16,526 words recorded all together.

4.2. Text Mining

After removing stop words and punctuations as well as stemming words with affix in the data cleaning process, only 11,570 words are left which consist of 3,587 different words. The frequencies of all 3,587 words are counted. The frequency table obtained is shown in Table 1. Based on Table 1, most of the words appear only once for all the documents. More than half of the words, which are 1914 words have a frequency of one. A drastic decrease can be seen on the number of words when the number of frequency increases.

Table 1: Frequency table for words in the data set

f	n	f	n	f	n	f	n	f	n
1	1914	12	17	23	4	39	1	58	1
2	668	13	18	24	8	41	2	62	1
3	278	14	11	25	3	42	2	63	1
4	173	15	14	26	7	43	2	68	2
5	109	16	8	27	3	44	2	69	1
6	71	17	6	28	3	46	1	72	1
7	59	18	9	29	3	47	1	86	1
8	38	19	8	30	1	48	1	116	1
9	48	20	3	33	2	49	1		
10	39	21	10	36	2	50	1		
11	20	22	5	37	1	56	1		

Note: f refers to frequency and n refers to the number of words.

Figure 1 shows the word cloud for 50 words with the highest frequencies in order to give a better representation of significant words in the data set. Words such as ‘said’, ‘will’, ‘team’ and ‘open’ can be clearly seen in Figure 1. These show that authors often quote statements from individuals in their articles. The word ‘will’ implies that the articles usually tell the plans or events that have not yet happen. Since the word ‘world’ is bigger than ‘Malaysia’, this shows that there are more articles on international sports than local ones. Words like ‘game’ ‘win’ and ‘match’ can also be seen in Figure 1, since these are common words in sports articles. This shows that there are correlations between texts in the different types of sports articles. Meanwhile, there are also frequent words that exist only in a particular type of sport such as ‘kyrgios’ which refers to a tennis player from Australia that was involved with a controversy and was banned from playing 8 weeks. The word ‘race’ is also only used by motorsports. Only once did the word ‘race’ appeared in an article about badminton, where the author is writing about the race among Malaysian athletes in order to qualify to the final badminton open. These exclusive words are useful in clustering the different types of sports articles.

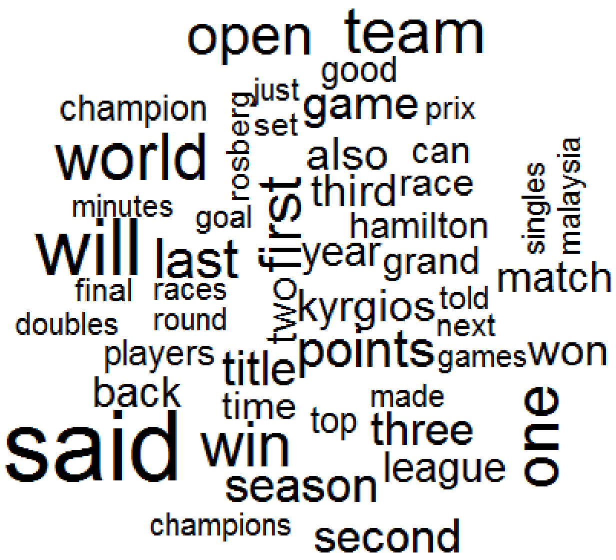


Fig. 1:. Word cloud for 50 words with the highest frequencies

4.3. Document Clustering

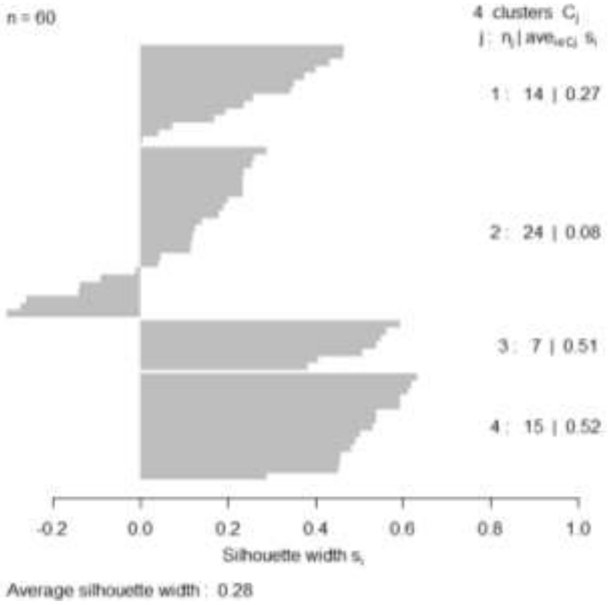
The sixty documents will be grouped into four clusters using hierarchical, k-means and k-medoids clustering algorithms. It is believed that a good combination of clusters will be able to represent the four different types of sports and the articles of each sports are correctly grouped into their respective clusters. Selected words are required to correctly cluster the documents. The chosen words are those that are bias to the type of sports they represent. For each type of sports, two of the most significant words are taken as variables to use in the clustering process. These words are ‘badminton’, ‘superseries’, ‘goal’, ‘league’, ‘race’, ‘championship’, ‘tennis’ and ‘match’. Before clustering analysis is done, the frequencies of words are normalized by finding the ratio of word frequencies to the total words in an article in order to reduce errors in the study.

Table 2 shows the results of cluster analysis from all three methods and the efficiency rate for all the clusters produced. The table shows that hierarchical clustering is able to produce accurate clusters with 100% efficiency for badminton, football and motorsports since all the documents within the three clusters are relevant documents. However, there are nine documents; 1 badminton article and 8 football articles, are wrongfully grouped into tennis which causes the efficiency rate of the tennis cluster drop to 65%. Meanwhile, k-means clustering method is able to correctly cluster 54 out of 60 documents. The k-means algorithm also produces a more balance combination of clusters since each cluster contains at least 13 out of 15 relevant articles for their respective sports. Lastly, 55 out of 60 documents are correctly grouped by the k-medoids algorithm. This is much better than the hierarchical and k-means clustering.

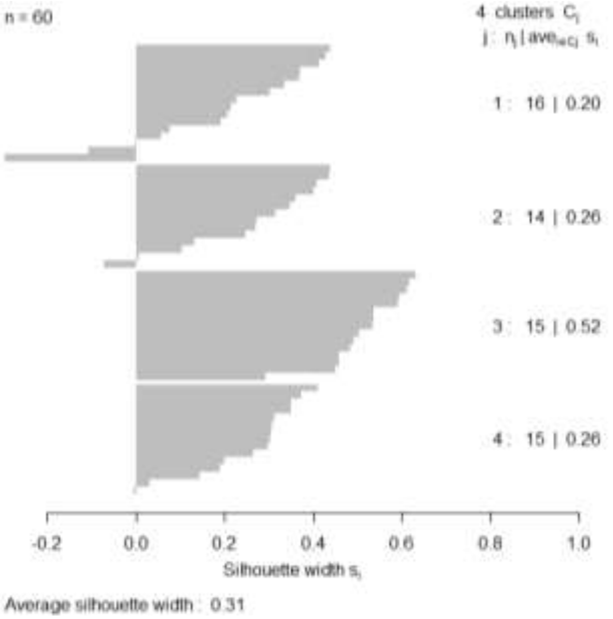
Figure 2 shows the silhouette plots while Table 3 gives the mean value of s_i for the clusters obtained from the three clustering methods. Figure 2 and Table 3 show that the mean values of s_i are positive for all three methods. This means that all three methods produce good clusters of documents. However, based on Figure 2, there are a few documents that have negative values of s_i for each clustering method. These negative values greatly affect the mean values of s_i for the second cluster of the hierarchical and k-medoids clustering results. Both the mean values are nearing zero with 0.08 and 0.02 respectively. Based on the silhouette index, k-means clustering method produce the best clusters, followed by the hierarchical clustering and k-medoids clustering method. The k-medoids clustering shows the lowest value of s_i although it gives the highest efficiency rate in Table 2. Overall, there are not much difference between the mean values of s_i for all three methods.

Table 2: Efficiency rate for clusters produced

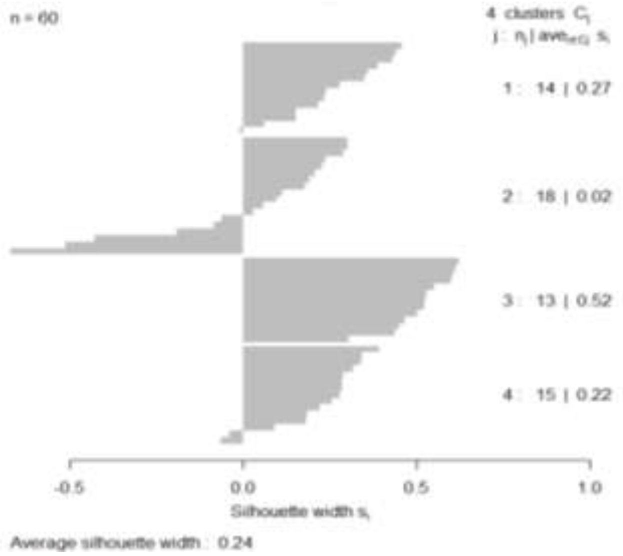
Clustering Algorithm	Cluster	Type of Sports	Total of Documents	Relevant Documents	Efficiency Rate
Hierarchical	1	Badminton	14	14	100%
	2	Football	7	7	100%
	3	Tennis	24	15	62.5%
	4	Motorsports	15	15	100%
		Total	60	51	85%
k-means	1	Badminton	16	13	81.25%
	2	Football	15	13	86.67%
	3	Tennis	14	13	92.86%
	4	Motorsports	15	15	100%
		Total	60	54	90%
k-medoids	1	Badminton	14	13	92.86%
	2	Football	13	12	92.31%
	3	Tennis	18	15	83.33%
	4	Motorsports	15	15	100%
		Total	60	55	91.67%



(a) Hierarchical clustering



(b) k-means clustering



(c) k-medoids clustering

Fig. 2: Silhouette plots for clusters produced

Table 3: Mean values of s_i for clusters produced

Clustering Algorithm	Mean s_i
Hierarchical	0.28
k-means	0.31
k-medoids	0.24

4.4. Sensitivity Analysis

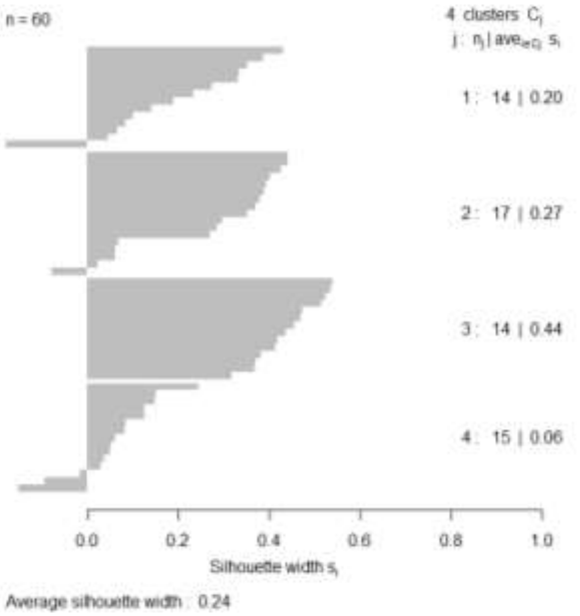
By wrongly choosing a general word as one of the selected words for document clustering is believed to disturb the accuracy of the clusters produced. A general word refers to a common word for all four types of sports under study. The cluster analysis is repeated by including two general words; ‘Malaysia’ and ‘world’, to check the sensitivity for all three clustering methods under study.

Table 4 shows the clusters produced and their efficiency rates for all three methods with the additional common words. Based on Table 4, the hierarchical clustering method is able to correctly grouped 57 documents with a mean efficiency rate of 95%. However, only 49 documents are correctly assign into their relevant clusters compared to the previously 55 documents when using the k-means clustering method. The k-medoids clustering method also shows a decrease in efficiency rate by correctly assigning only 50 documents.

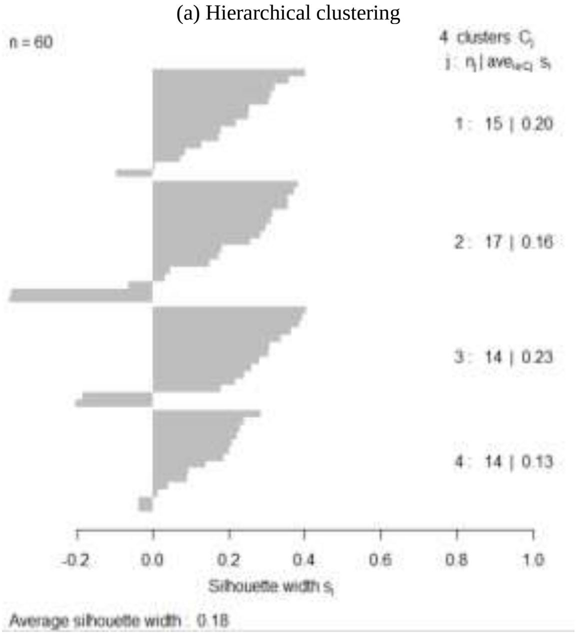
The silhouette plots for clusters produced with the addition of common words using all three clustering methods are given in Figure 3 while their mean values of s_i are given in Table 5. The mean values of s_i for all three methods are still positive even with the addition of general words, which means that all three clustering methods are suitable to be used for document clustering. However, the mean values are all smaller than the mean values obtained without the addition of general words. This shows that the selection of words is important in obtaining the clusters. Table 5 also shows that both the k-means and k-medoids clustering methods are more sensitive towards the addition of general words compared to the hierarchical clustering method. This can also be seen in Figure 3. The silhouette plot for k-means clustering shows that none of the clusters produced have a mean value of s_i as high as 0.5 as was seen in Figure 2.

Table 4: Efficiency rate for clusters produced with the addition of general words

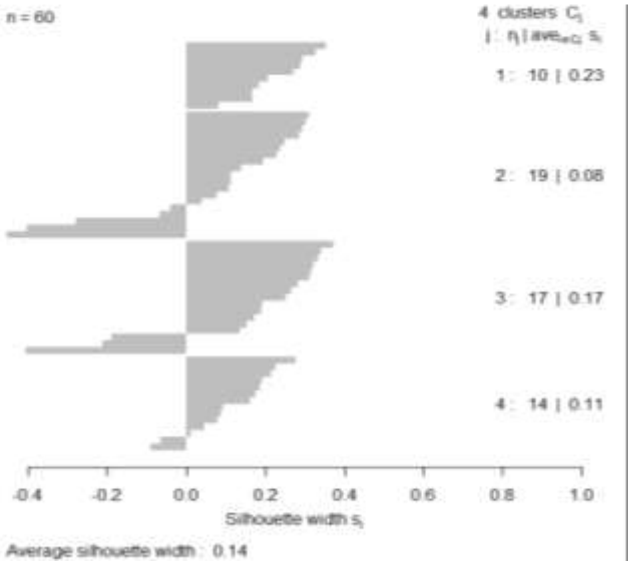
Clustering Algorithm	Cluster	Type of Sports	Total of Documents	Relevant Documents	Efficiency Rate
Hierarchical	1	Badminton	14	14	100%
	2	Football	17	15	88.24%
	3	Tennis	15	14	93.33%
	4	Motorsports	14	14	100%
		Total	60	57	95%
k-means	1	Badminton	14	9	64.29%
	2	Football	17	15	88.24%
	3	Tennis	15	11	73.33%
	4	Motorsports	14	14	100%
		Total	60	49	81.67%
k-medoids	1	Badminton	10	9	90%
	2	Football	14	13	92.86%
	3	Tennis	17	14	82.35%
	4	Motorsports	19	14	73.68%
		Total	60	50	83.33%



(a) Hierarchical clustering



(b) k-means clustering



(c) k-medoids clustering

Fig. 3: Silhouette plots for clusters produced with the addition of general words

Table 5: Mean values of s_i for clusters produced with the addition of general words

Clustering Algorithm	Mean s_i
Hierarchical	0.24
k-means	0.18
k-medoids	0.14

5. Conclusion

This study aims to investigate and compare document clustering algorithms. Three clustering methods are used which are the hierarchical, k-means and k-medoids clustering method. The comparison between the three methods are assessed by looking at their efficiency rates and silhouette index, s_i. Sixty sports articles are used as case study for all three methods.

The k-means clustering method shows the most significant result with the highest mean value of s_i. However, it is very sensitive towards the addition of general words. Hence, if the words chosen for cluster analysis are suitable, k-means algorithm would give the best clusters for the documents but if the wrong words are used, the results could be greatly affected. The hierarchical clustering method provides a significant result and is less sensitive towards the addition of common words. Meanwhile, k-medoids clustering gives the smallest mean value of s_i compared to the other two methods and it is also sensitive towards the addition of general words. The method of k-means is said to be better than k-medoids clustering method since big size documents give an advantage to k-means to reduce the risk of errors in clusters [10]. Thus, k-medoids is believed to be more suitable for smaller size documents. Overall, the hierarchical clustering method is deemed the most suitable method for document clustering since it provides significant clusters with high efficiency rates and is more stable compared to k-means and k-medoids clustering in terms of sensitivity towards general words.

These clustering methods may be improved in future researches by integrating dimension reduction methods before performing cluster analysis. Furthermore, other methods such as Latent Dirichlet Allocation (LDA) could also be looked at and compare with methods used in this study. LDA differs from methods used in this study in that it assigns a document to a combination of topics and hence characterized each document by one or more topics.

Acknowledgement

The authors would like to express great gratitude to Universiti Kebangsaan Malaysia (UKM) for the allocation of research grant, KRA-2017-019, and the infrastructures used for this study.

References

[1] Cohen K.B., Hunter L. Getting started in text mining. PLoS Computational Biology, 2008, 4: 1-3.

[2] Neustein A., Imambi S. S., Rodrigues M., Teixeira A., Ferreira L. Application of text mining to biomedical knowledge extraction: Analyzing clinical narratives and medical literature. Text Mining of Web-Based Medical Content, 2014, pp. -31.

[3] Kamaruddin S.S., Hamdan A.R., Abu Bakar A., Mat Nor F. Outlier detection in financial statements: A text mining method. WIT Transactions on Information and Communication Technologies, 2009, 42: 71-82.

[4] Al-Daihani S.M., Abrahams A. A text mining analysis of academic libraries' Tweets. Journal of Academic Librarianship, 2015, 1: 1-9.

[5] Kadhim A.I., Cheah Y.N., Ahamed N.H. Text document preprocessing and dimension reduction techniques for text document clustering. Proceedings of the 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, 2014, pp. 69-73.

[6] Mythily R., Banu A., Raghunathan S. Clustering Models for Data Stream Mining. Procedia Computer Science, 2015, 46: 619-626.

[7] Steinbach M., Karypis G., Kumar V. A comparison of document clustering techniques. Proceedings of the KDD Workshop on Text Mining, 2000, pp. 1-20.

[8] Balabantaray R. C., Sarma C., Jha M. Document clustering using K-Means and K-Medoids. International Journal of Knowledge Based Computer System, 2013, 1: 7-13.

[9] Balcan M.-F., Laiang Y., Gupta P. Robust Hierarchical Clustering. Journal of Machine Learning Research, 2014, 15: 4011-4051

[10] Bouguettaya A., Yu Q., Liu X., Zhou X., Song A. Efficient agglomerative hierarchical clustering. Expert Systems with Applications, 2015, 42(5): 2785-2797.

[11] Slamet C., Rahman A., Ramdhani M.A., Darmalaksana W. Clustering the verses of the Holy Qur'an using K-Means algorithm. Asian Journal of Information Technology, 2016, 15(24): 5159-5162.

[12] Younus Z.S., Mohamad D., Saba T., Alkawaz M.H., Rehman A., Al-Rodhaan M., Al-Dhelaan A. Content-based image retrieval using PSO and k-means clustering algorithm. Arabian Journal of Geosciences, 2015, 8(8): 6211-6224.

[13] Peker M. A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM. Journal of Medical Systems, 2016, 40: 116.

[14] Al-Anazi S., AlMahmoud H., Al-Turaiki I. Finding similar documents using different clustering techniques. Procedia Computer Science, 2016, 82: 28-34.

[15] Feinerer I., Hornik K., Meyer D. Text Mining Infrastructure in R. Journal of Statistical Software, 2008, 25(5): 1-54.

[16] Swathi B.V., Govardhan A. Find-k: A new algorithm for finding the k in partitioning clustering algorithms. International Journal of Computing Science and Communication Technologies, 2009, 2(1): 268-272.

[17] Park H.S., Jun C.H. A simple and fast algorithm for K-Medoids clustering. Expert System with Applications, 2009, 36: 3336-3341

[18] Arora P., Deepali D., Varshney S. Analysis of K-Means and K-Medoids algorithm for big data. Procedia Computer Science, 2015, 78: 507-512.

[19] Hennig C., Liao T.F. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C (Applied Statistics), 2013, 62(3): 309-369

[20] Rao A.R., Srinivas V.V. Regionalization of watersheds by hybrid-cluster analysis. Journal of Hydrology, 2006, 318(1): 37-56.