ICIC Express Letters
Volume **17**, Number **3**, March **2023**
ICIC International ©2023 ISSN 1881-803X
pp. **269–278**

THE DEVELOPMENT OF A NEW HYBRID K-MEANS AND ELBOW METHOD (C-ALGORITHM) FOR MULTIPLE DOMAIN CLUSTERING

Panithan Mekkamol and Chatklaw Jareanpon*

Polar Lab
Department of Computer Science
Faculty of Informatics
Mahasarakham University
Khamriang Sub-District, Kantharawichai, Mahasarakham 44150, Thailand
58011260501@msu.ac.th; *Corresponding author: chatklaw.j@msu.ac.th

Received June 2022; accepted August 2022

Abstract. *This research aims to develop a new clustering algorithm called C-Algorithm that the document can classify to the previous domain or create a new domain and solve the K-means problem. This problem comes from the distance measurement of similarity from the new document to the centroid of each group. The new document will classify the group so that the relationship between groups and the new document is analogous or divergent. This experiment observes the proper group numbers using the Elbow method before starting the process. After this process, the Threshold value will be calculated from the centroid of the document in the group and percentile. The new document will compare with the Threshold and decision to set to the group or create the new document. This research compares the performance of the weight between the TF-IDF and BM25. These results show that the best performance comes from the BM25, Euclidean distance, and 80-85 percentile. The result of this research is more accurate than the traditional K-means algorithm.*
Keywords: Document clustering, Distance similarity, K-means, C-Algorithm, Elbow method

1. **Introduction.** It is advantageous to utilize digital data in the digital era since it reduces paper consumption and communication time, but it produces enormous amounts of digital data from multiple devices. More than 80% of the data is a text that has consisted of unstructured and semi-structured material that has not been classified as valuable or ineffective information that can extract from the text-mining methods [1]. The K-means algorithm is often used for document clustering because of the good outcomes [2]. It is also commonly used to identify similarities between objects depending on distance vectors [3]. It is essential to provide the requisite number of clusters which might be difficult if the number of clusters or K-values is unknown. K-values have a significant role affected with algorithm efficiency [4]. K-means clustering does not achieve well with outlier datasets and noise [5]. After the clustering procedure is complete, the data in the cluster is no longer regarded as outliers. When submitting the new data, it should be assigned to the group with the similarity of the data using the K-means method. When data is closed to the center of a cluster, it is grouped in that cluster as same as the outlier data, which is possibly not related to the data in the cluster. This will make the mistaken result affected with affiancing of the clustering.

This research aims to develop an algorithm for clustering documents in multiple domains by determining the similarity of the documents in each domain, and the clustered data should belong to a group of domains or be separated into a new group. This research

DOI: 10.24507/icicel.17.03.269

evaluates the performance of weighting using the term frequency-inverse document frequency (TF-IDF) and Best Match 25 (BM25). By assessing the similarity of the new set of documents for testing, whether the dataset should be in a group or not by locating the percentile of a dataset in a cluster determining the Threshold, and determining whether the information should be in a group or out of a group, and evaluating the efficiency of clustering documents of the proposed algorithms.

2. Problem Statement and Preliminaries.

2.1. Document clustering.

Clustering documents is clustering similar documents to the same group or else to the separated groups to increase the recall/precision and reduce the search time. The clustering method is the need to find new structures that are never known before. Several research pointed out that techniques based on K-means used for clustering documents have significantly improved the performance, compared to other techniques [6-10]. Additionally, Kumar et al. [11] proposed the method for detecting cluster outliers in a multi-dimensional dataset. The efficiency was better than the existing COID algorithm.

2.2. The optimal number of clusters using the Elbow method.

An appropriate cluster is required to set the K-means method of document clustering (K-values, number of clusters). Selecting the number of segments affects the result of the integration. Incorporating the Elbow method will aid in identifying the optimal number of clusters effectively [16-18].

The idea of the Elbow method is to measure the distance between the data and the center of the cluster and show the trend of the sum of squared errors (SSE), and then they have grouped the dataset based on the specified range. For each group, the SSE value is calculated and expressed as a line in the graph. If the line bends, that is the best number of clusters. The Elbow method is shown in Equation (1).

$$W_k = \sum_{r=1}^{k} \frac{1}{n_r} Dr \tag{1}$$

where k is the number of clusters, n_r is a data point in the cluster r, and Dr is a sum of the distance between data points in the cluster.

2.3. Document similarity measures.

Unsupervised machine learning with K-means relies on finding the distance between two points to predict the outcomes. Therefore, choosing the similarity measurement method for the data is very important because it affects the measuring document similarity. There were several techniques suggested in research to be used for discovering the accurate distance, for example, Cosine Similarity [5,14], Euclidean Distance [12,13], Manhattan [13], and Minkowski [12]. The equations of the four distance measurements are shown in (2)-(5), respectively.

Euclidean Distance

$$d(i,j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{in} - x_{jn}\right)^2} \tag{2}$$

Cosine Similarity

$$similarity(i,j) = \frac{i \cdot j}{\|i\| \times \|j\|} \tag{3}$$

Manhattan

$$d(i,j) = \left|\left(x_{i1} - x_{j1}\right)\right| + \left|\left(x_{i2} - x_{j2}\right)\right| + \cdots + \left|\left(x_{in} - x_{jn}\right)\right| \tag{4}$$

Minkowski

$$d(i,j) = \left(\left|\left(x_{i1} - x_{j1}\right)\right| p + \left|\left(x_{i2} - x_{j2}\right)\right| p + \cdots + \left|\left(x_{in} - x_{jn}\right)\right| p\right) 1/p \tag{5}$$

where x_{in} and x_{jn} are the values of the ith and jth variable, at points i and j, respectively.

2.4. **Feature selection by term weighting.** The features are selected by weighting the attribute using TF-IDF and BM25 methods for comparing the weighting efficacy of the results.

TF-IDF is a term weighting a document as a numerical statistic that shows how important the word is to represent a group of documents. TF-IDF values increase in proportion to the number of times words appear in a document. TF-IDF is formed by term frequency (TF) multiplied by IDF (inverse document frequency). TF-IDF methods are used to give weight to attributes representing sentences using (6).

$$tfidf_{td} = tf_{td} \times idf_t \tag{6}$$

where t is a term, d is the document, tf_{td} is term t frequency in the document d divided by total words in the document d, idf_t is $\log_2$ of total document divided by documents with the term t.

BM25 considered the influence of factors such as document length on word apparent frequency, which were not considered in TF-IDF weighting. In the BM25, documents were pre-proportioned and multiplied by the word frequency value. Therefore, genuinely unique word separations can be expected in long documents, as shown in (7).

$$score(d,q) = IDF(q) \times \frac{f(q,D) \times (k+1)}{f(q,D) + k \times \left(1 - b + b \times \frac{|D|}{avgdl}\right)} \tag{7}$$

where $f(q,D)$ is correlated to the term's frequency, defined as the number of times query term q appears in the document D, $|D|$ is the length of the document D in terms, $avgdl$ is the average document length over all the documents of the collection k, b are free parameters, and $IDF(q)$ is the inverse document frequency weight of the query term q.

Based on the relevant literature review, the researchers chose the K-means to group the document because of its higher performance and utilized the Elbow method to initialize the optimal number of groups. Moreover, the important parameters that are effected with the performance are 1) the similarity or distance measurement using several methods such as Euclidean Distance, and Cosine Similarity, and 2) the feature selection. These parameters were conducted and tested in this study. Additionally, this proposed method will compare the word weighting to test the most effective way for clustering.

3. **Proposed Method.** This research proposes an algorithm to determine whether the clustered data should belong to a cluster or be separated to find a new cluster.

Normally, when the new document feeds to the clustering system, the distance (C_{ij}) will calculate between the new document and the group's center, and it is set to the member of the closest group.

In addition to normal distance, the main proposed concept was adding the step of separating the group.

1) Calculate the distance ($C_{ij\,\max}$) between the furthest data of each group and the group's center.
2) If the (C_{ij}) is less than the $C_{ij\,\max}$, the new document will be set to the member of the closest group.
3) Else, the new document will separate and create a new group.

As shown in the following 1) find optimal K using the Elbow method; 2) calculate the number of source domains ($k = K - 1$) and splitting the dataset source domain and test document; 3) cluster the source domain (number of clusters = k); 4) calculate the center of each cluster for calculating the distances from all data in the group; 5) calculate the data distance to the center of each cluster with Euclidean Distance; 6) calculate the cluster Threshold by calculating the distance between every data within the cluster and the cluster center. When the distance of all data in each group is obtained, the distances will

arrange in ascending order. We use the percentile position to determine the appropriate Threshold. The Threshold of each cluster (called Local Threshold) is a value used to measure the distance of newly entered data to that group, each of which is different, and 7) calculate the distance between a test set and the center of an existing cluster using Euclidean Distance. If the document distance is close to any cluster and less than the Threshold of that cluster, the document will classify in that cluster. Else, the document will be in the new cluster.

The K-means clustering technique is used to test the main proposed concept, that demonstrates the separation and creates the number of the group ability or the performance.

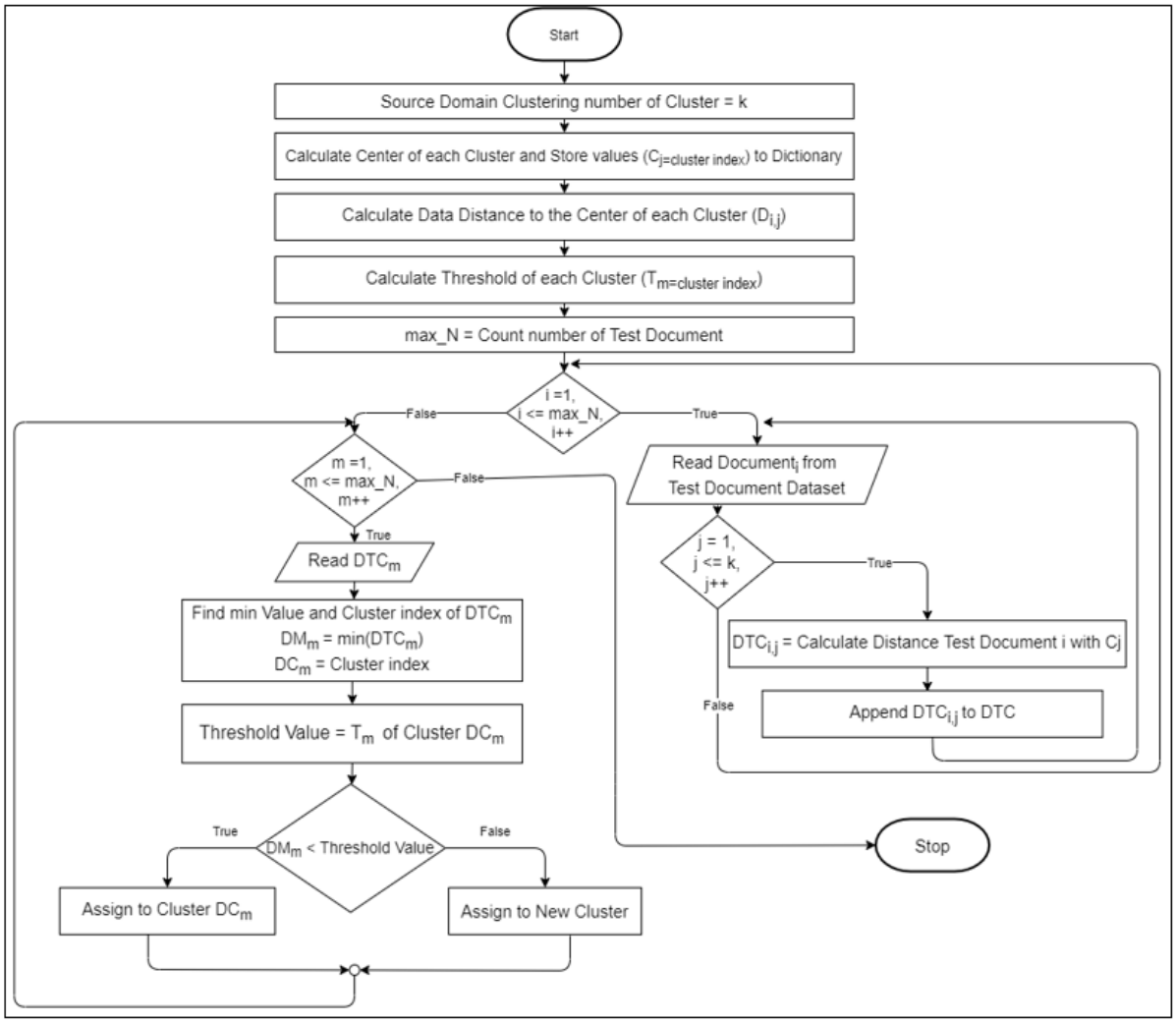

FIGURE 1. The proposed method diagram

4. **The Experimental Result.**

4.1. **Dataset.** In this research, we select to demonstrate the 2 datasets as shown in Table 1.

1) Multi-Domain Sentiment dataset, collected by Blitzer et al. from Amazon.com, consisted of three different products: Books, DVDs, and electronics.
2) 20 News Groups text dataset, collected by Ken Lang, selected 3 domains by random selection method: alt.atheism, misc.for sale, and sci.electronics.

TABLE 1. The detail of the 2 datasets

Dataset	Number of domains	Domain	Number of document	Min term in document	Max term in document	Avg. term	Total
Multi-Domain Sentiment	3	Book	600	5	5,176	177	1,800
		DVDs	600	10	1,374	186	
		Electronics	600	8	1,345	97	
20 News Groups	3	alt.atheism	480	1	8,611	199	1,656
		misc.for sale	585	2	2,625	106	
		sci.electronics	591	1	11,765	134	

The data used in the experiment was the text in the document formatted in unstructured data. Therefore, the data must be transformed into structured data, select the feature and feed to the clustering process. This research selects to use the data preparation step as follows: 1) Word tokenization is the process of dividing a text into words, sentences, or symbols called Tokens; 2) Changing the words to the lowercase; 3) Selecting the words that are in English letters only because numbers and special characters are not necessary to group or classify the documents; 4) Removing stop words, which refers to the frequent words in a document that is hardly critical to clustering the document [15]. Moreover, it can speed up the process; 5) Remove punctuation; punctuation has no effect such as full stop (.), and comma (,); 6) Stemming, cuts off the various word to leave only the root of the word (basic form).

4.2. **Term weighting feature selection.** This research selects the features by two techniques, TF-IDF and BM25 weighting, and the attributes are used to represent the sentences. For the clustering algorithm to test the weighting efficacy, the defining number of groups of K-means is set to 3 groups according to the number of domains for both datasets. The results of the experiment are shown in Table 2.

TABLE 2. The comparison of the performance between TF-IDF and BM25

Method	Dataset	Avg. recall	Avg. precision	F1
TF-IDF	Multi-Domain Sentiments	0.88	0.90	0.88
	20 News Groups	0.81	0.87	0.82
BM25	Multi-Domain Sentiments	0.88	0.89	0.88
	20 News Groups	0.61	0.76	0.61

From Table 2, both methods average recall and F1 measure are the same in the multi-domain dataset but are different in 20 News Groups. Thus, this research will test the significance of the performance using a Paired-Sample T-test, and the hypotheses are as follows:

H0 = Term weighting with BM25 equal to TF-IDF

H1 = Term weighting with TF-IDF not equal to BM25

We test the clustering by using the K-means algorithm ten times on each consecutive dataset. The results of the experiment are shown in Table 3.

The significance (P value) is less than 0.05; therefore, it rejects the H0 and accepts the H1 hypothesis. It represents that TF-IDF term weighting has better clustering performance than BM25 term weighting. Thus, the TF-IDF is selected to use for the next step.

4.3. **Using Elbow to determine the optimal number of clusters.** To determine the optimal number of clusters for document clustering, the Elbow method will be grouped on the dataset by its range by submitting documents in all three domains of the two datasets. Using Elbow to determine the optimal number of clusters of the K-value is as shown in Figures 2 and 3.

TABLE 3. Term weighting Paired-Sample T-test on Multi-Domain Sentiment dataset and 20 News Groups dataset

Dataset		Paired differences					t	df	Sig (2-tailed)
		Mean	Std. deviation	Std. error mean	95% confidence interval of the difference				
					Lower	Upper			
Multi-Domain Sentiment	BM25	0.442	0.156	0.049	−0.5389	−0.304	8.34	9	0.0001
	TF-IDF	0.866	0.019	0.0062					
20 News Groups	BM25	0.396	0.112	0.035	−0.3608	−0.1425	5.21	9	0.0006
	TF-IDF	0.648	0.137	0.043					

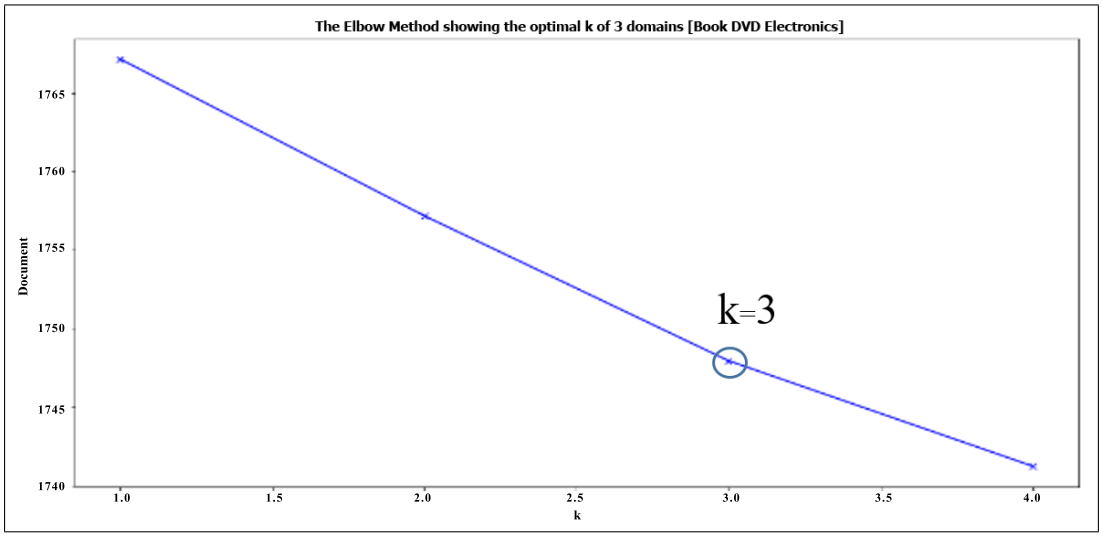

FIGURE 2. The Elbow method shows the optimal k of 3 domains from Multi-Domain dataset.

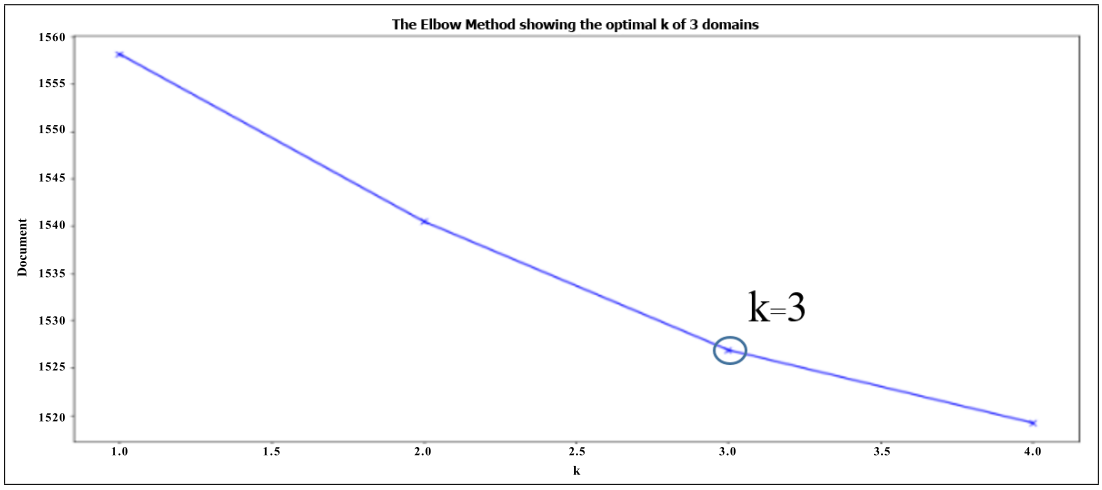

FIGURE 3. The Elbow method shows the optimal k of 3 domains from 20 News Groups dataset.

From Figures 2 and 3, the bend can see that the appropriate number of groups is three groups, so the value of k is set to 3, and then cut off the last group of documents for testing the performance of the proposed algorithm.

4.4. **The results of the tested experiment of the proposed method.** We extract the 476 documents from cluster 3rd of the Multi Domain Sentiment dataset and 20 News Groups dataset and submit them. To test whether it can be divided into the correct cluster correctly or not, this research varies the Threshold value used as the criterion for clustering the documents by calculating the percentages of the distance between the first and second groups of the documents at 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 and 100. Thus, the third group is used for testing the performance of the proposed method.

Table 4. The comparisons of the recall value of similarity function

Percentile	On Multi-Domain Sentiment dataset				On 20 News Groups dataset			
	Recall				Recall			
	Euclidean	Manhattan	Minkowski	Cosine	Euclidean	Manhattan	Minkowski	Cosine
45	**0.9243697**	**0.842437**	**0.331933**	**0.09**	**0.95**	**0.59448**	**0.7326**	0
50	0.894958	0.831933	0.289916	0.07	0.95	0.54807	0.68508	0
55	0.8613445	0.794118	0.262605	0.06	0.95	0.5105	0.6453	0
60	0.8193277	0.756303	0.222689	0.05	0.95	0.47624	0.59669	0
65	0.7836134	0.731092	0.189076	0.04	0.94	0.43204	0.55359	0
70	0.7436975	0.691176	0.170168	0.02	0.94	0.38453	0.49503	0
75	0.6617647	0.638655	0.138655	0.02	0.93	0.34807	0.43204	0
80	0.5987395	0.581933	0.113445	0.01	0.91	0.27403	0.37348	0
85	0.4789916	0.512605	0.092437	0	0.89	0.22431	0.30608	0
90	0.2983193	0.392857	0.060924	0	0.85	0.1547	0.22983	0
95	0.1470588	0.283613	0.042017	0	0.79	0.10829	0.16354	0
100	0.0042017	0.079832	0	0	0.42	0.01657	0.02099	0

From Table 4, it was found that the 3rd group of documents submitted for testing with the designed algorithm can be separated from the 1st and 2nd groups with the most accuracy at the 45th percentile. The third group of documents (Tested Documents) submits to test the designed algorithm, which can be separated from the existing clusters. The most accuracies are came from percentages between 45-60 with the Euclidean, Manhattan, and Minkowski methods.

4.5. **Experimental results with all documents of all clusters.** To verify the proposed algorithm, we submit the documents of both datasets for clustering the first, second, and third, respectively, assigning into the original cluster or separating. The results are shown in Table 5.

Table 5. The average recall and F1 score values for all groups of documents by means of similarity measurements on the Multi-Domain Sentiment dataset

Percentile	Recall				F1			
	Euclidean	Manhattan	Minkowski	Cosine	Euclidean	Manhattan	Minkowski	Cosine
45	0.61	0.43	0.41	0.05	0.58	0.33	0.43	0.07
50	0.63	0.44	0.43	0.04	0.62	0.35	0.45	0.06
55	0.64	0.43	0.43	0.04	0.62	0.34	0.45	0.05
60	0.64	0.42	0.44	0.04	0.63	0.33	0.46	0.05
65	0.69	0.46	0.49	0.03	0.69	0.37	0.50	0.04
70	0.71	0.46	0.52	0.03	0.71	0.37	0.52	0.03
75	0.72	0.46	0.54	0.02	0.72	0.38	0.53	0.03
80	**0.73**	0.46	0.56	0.02	**0.74**	0.38	0.54	0.02
85	**0.73**	0.45	0.58	0.02	0.73	0.37	0.55	0.02
90	0.70	0.43	0.61	0.02	0.69	0.35	0.55	0.02
95	0.68	0.41	0.63	0.02	0.64	0.33	0.56	0.00
100	0.67	0.36	0.65	0.02	0.58	0.25	0.55	0.00

From Table 5, the results showed that the recall value of all documents for the three groups of Euclidean Distance similarity measurements was the best, with 73% at 80th and 85th percentiles. For the F1 measurement, Euclidean Distance is the best at 74% at the 80th percentile.

TABLE 6. The average recall and F1 score values for all groups of documents by means of similarity measurements of the 20 News Groups dataset

Percentile	Recall				F1			
	Euclidean	**Manhattan**	**Minkowski**	**Cosine**	**Euclidean**	**Manhattan**	**Minkowski**	**Cosine**
45	0.62	0.36	0.55	0	0.66	0.34	0.56	0
50	0.65	0.38	0.57	0	0.70	0.35	0.58	0
55	0.67	0.37	0.57	0	0.71	0.35	0.57	0
60	0.68	0.37	0.57	0	0.73	0.36	0.56	0
65	0.75	0.42	0.62	0	0.79	0.39	0.60	0
70	0.78	0.43	0.63	0	0.81	0.40	0.59	0
75	0.81	0.44	0.65	0	0.83	0.41	0.59	0
80	0.84	0.45	0.66	0	0.85	0.40	0.58	0
85	0.86	0.46	0.67	0	**0.87**	0.41	0.56	0
90	0.88	0.47	0.68	0	**0.87**	0.40	0.54	0
95	**0.89**	0.49	0.69	0	0.86	0.40	0.52	0
100	0.81	0.49	0.67	0	0.69	0.36	0.43	0

From Table 6, the results showed that the recall of all documents of the three groups of Euclidean distance similarity measures was the best 89% at the 95th percentile, and the F1 values were best 87% at 85th and 90th percentiles.

4.6. **The comparison of the document clustering between the proposed algorithm and the traditional K-means algorithm.** In this experiment, we use the Threshold value at the 80th percentile and the Euclidean Distance measurement method and submit the third group of 476 documents on the Multi-Domain Sentiment Dataset to test the clustering with the original two groups. The results are shown in Table 7.

TABLE 7. The comparison of the re-clustering of documents between the traditional K-means algorithm and the proposed algorithm

Re-clustering	**Traditional K-means algorithm**		**Proposed algorithm**	
	Documents	**%**	**Documents**	**%**
Organized into 1st cluster	366	76.89	129	27.10
Organized into 2nd cluster	110	23.11	62	13.03
Organized into a new cluster	**0**	**0**	**285**	**59.87**

From Table 7, the proposed method clusters and sends the re-clustering document to a new group of 285 documents (59.87%) separated from the two existing groups, and 221 documents (40.13%) will be categorized into both groups. It shows that the documents clustering with the proposed algorithm can separate documents; it solves the problem of clustering with K-means. The proposed method considers the most minor similar data from each cluster to be combined in the same cluster to determine the data in the new group.

The proposed method was successful in distinguishing a set of documents for a new domain from another domain. When the data is in the center of the cluster, it can solve the clustering problem of the K-means algorithm, which assesses the similarity of the data. The data will be organized in that category in the same way that normal data contains.

We used the idea of outlier detection to determine if a document should be a group in an old cluster or a new cluster. As a result, a new cluster of documents or an existing cluster

of documents relocated to the new cluster is not deemed outliers unless the document is a non-similarity in the group.

5. **Conclusion and Discussion.** As a result, the proposed method can solve the problem of connecting information TF-IDF term weighting performed significantly better than BM25 on both datasets, agreed with the experiments of Afrizal et al. [19] and Kadhim [20].

We propose the approach revealing the best performing method that integrated document similarity measurement with the Euclidean Distance determining whether new documents will be in a cluster or excluded from a previous cluster by calculating the distance of new documents and the center of all clusters. If the distance of the new document was closest to the center of any cluster, then the document should be in that cluster but must compare with the Threshold of each cluster. If the distance is greater than the Threshold of that cluster, the new document will leave the cluster.

In the Threshold values configuration, the best way to separate the new documents from a cluster is the median of the distance of the documents toward the center. This is consistent with the experimental results of Barai and Dey [21], but if the new documents are the same cluster, the results of merging the documents into the same cluster are not good. Therefore, the best performance of the proposed Threshold configuration method of the dataset in the group is between 80th-85th. This will perform the best overall performance for all documents in all clusters. The efficiency of the proposed method can solve the problem of the clustering method with K-means.

A new cluster of documents is less similar to the documents in the existing cluster. This indicates that the key terms appearing in the new clusters of a document are not found or hardly in the original cluster, which is not enough to fit into the current cluster, the words are crucial for recognizing new documents that are relevant to selecting the features for classifying the original documents. The effectiveness of clustering is determined by the used attributes. In this study, we tested by excluding terms that occurred and appeared in fewer than two documents in the Multi-Domain Sentiment dataset; fewer than eight documents in the 20 News Groups dataset showed the greatest performance of both datasets. Before applying the proposed method, it is recommended to explore the suitable values for the dataset or pick the relevant attributes. The newly extracted cluster does not mean that they are similar documents within the cluster. Instead, it refers to a cluster of documents that are not similar to an existing cluster. Consequently, the similarity within the new cluster should be considered by repeating the proposed algorithm that can separate the new sub-cluster.

REFERENCES

[1] A. Kaushik and S. Naithani, A comprehensive study of text mining approach, *International Journal of Computer Science and Network Security (IJCSNS)*, vol.16, no.2, p.69, 2016.

[2] S. V. Gaikwad, A. Chaugule and P. Patil, Text mining methods and techniques, *International Journal of Computer Applications*, vol.85, no.17, 2014.

[3] C. Sreedhar, N. Kasiviswanath and P. C. Reddy, Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop, *Journal of Big Data*, vol.4, no.27, DOI: 10.1186/s40537-017-0087-2, 2017.

[4] P. Bholowalia and A. Kumar, EBK-means: A clustering technique based on elbow method and k-means in WSN, *International Journal of Computer Applications*, vol.105, no.9, pp.17-24, 2014.

[5] W. Usino, A. Prabuwono, K. Hamed, S. Allehaibi, A. Bramantoro, A. Hasniaty et al., Document similarity detection using k-means and cosine distance, *Intl. J. on Advanced Computer Science and Applications*, vol.10, no.2, pp.165-170, 2019.

[6] R. C. Balabantaray, C. Sarma and M. Jha, Document clustering using k-means and k-medoids, *arXiv Preprint*, arXiv: 150207938, 2015.

[7] M. O. Adebiyi, E. B. Adigun, R. O. Ogundokun, B. Adeniyi, P. Ayegba and O. O. Oladipupo, Semantics-based clustering approach for similar research area detection, *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol.18, no.4, pp.1874-1883, 2020.

[8] S. Al-Anazi, H. AlMahmoud and I. Al-Turaiki, Finding similar documents using different clustering techniques, *Procedia Computer Science*, pp.8228-8234, 2016.

[9] C. Fry and S. Manna, Can we group similar Amazon reviews: A case study with different clustering algorithms, *2016 IEEE 10th International Conference on Semantic Computing (ICSC)*, pp.374-377, 2016.

[10] T. Bezdan, C. Stoean, A. A. Naamany, N. Bacanin, T. A. Rashid, M. Zivkovic et al., Hybrid fruit-fly optimization algorithm with k-means for text document clustering, *Mathematics*, vol.9, no.16, 1929, DOI: 10.3390/math9161929, 2021.

[11] A. Kumar, A. Kumar, A. K. Bashir, M. Rashid, V. A. Kumar and R. Kharel, Distance based pattern driven mining for outlier detection in high dimensional big dataset, *ACM Trans. Management Information System (TMIS)*, vol.13, no.1, pp.1-17, 2021.

[12] M. K. Arzoo and K. Rathod, K-means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8.2, *Int. Res. J. Eng. Technol.*, vol.4, no.4, pp.2363-2368, 2017.

[13] S. A. Salihu, I. P. Onyekwere, M. A. Mabayoje and H. A. Mojeed, Performance evaluation of Manhattan and Euclidean distance measures for clustering based automatic text summarization, *FUOYE Journal of Engineering and Technology*, vol.4, no.1, 2019.

[14] R. M. Ravindran and A. S. Thanamani, K-means document clustering using vector space model, *Bonfring International Journal of Data Mining*, vol.5, no.2, pp.10-14, 2015.

[15] K. V. Ghag and K. Shah, Comparative analysis of effect of stop words removal on sentiment classification, *2015 International Conference on Computer, Communication and Control (IC4)*, pp.1-6, 2015.

[16] M. Syakur, B. Khotimah, E. Rochman and B. D. Satoto, Integration k-means clustering method and elbow method for identification of the best customer profile cluster, *IOP Conference Series: Materials Science and Engineering*, vol.336, no.1, pp.12-17, 2017.

[17] R. Nainggolan, R. Perangin-Angin, E. Simarmata and A. F. Tarigan, Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the elbow method, *Journal of Physics: Conference Series*, vol.1361, no.1, pp.12-15, 2015.

[18] C. Yuan and H. Yang, Research on K-value selection method of K-means clustering algorithm, *J – Multidisciplinary Scientific Journal*, vol.2, no.2, pp.226-235, 2019.

[19] A. D. Afrizal, N. A. Rakhmawati and A. Tjahyanto, New filtering scheme based on term weighting to improve object based opinion mining on tourism product reviews, *Procedia Computer Science*, vol.161, pp.805-812, 2019.

[20] A. I. Kadhim, Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF, *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pp.124-128, 2019.

[21] A. Barai and L. Dey, Outlier detection and removal algorithm in k-means and hierarchical clustering, *World Journal of Computer Application and Technology*, vol.5, no.2, pp.24-29, 2017.