ICIC Express Letters
Volume **16**, Number **6**, June **2022**

 ISSN 1881-803X
pp. **583–593**

SIMILARITY THRESHOLD ESTIMATION FOR ENHANCED TEXT DOCUMENT CLUSTERING

Mohammad A. Hassan* and Yaser A. Al-Lahham

Computer Science Department
Zarqa University
P.O Box 132222, Zarqa 13132, Jordan
*Corresponding author: mohdzita@zu.edu.jo; yasirlhm@zu.edu.jo

Received September 2021; accepted December 2021

Abstract. *Traditional text document clustering techniques have a threshold value estimation challenge. In most cases, it is a human decision, and the value should be determined manually, which needs more implementation trials to get acceptable targeted results. In this paper, an effective threshold estimation method is proposed that can be generated automatically. The estimation of the threshold value depends on both the local connectivity between cluster elements, and the disconnection between elements of different clusters. Such threshold value will be determined during the clustering preparation processes, which are fully dependent on the content of the documents in the data set being clustered. Experimental evaluation over real data has been conducted to demonstrate the effectiveness of the proposed approach in estimating the threshold value.*
Keywords: Information retrieval, Threshold value, Document clustering, Hierarchical clustering, Cluster evaluation

1. **Introduction.** Clustering is generally an unsupervised learning technique, which means collecting similar objects (such as documents) into groups called clusters. Therefore, the documents within a cluster are very similar, but dissimilar to other clusters. The goal of a good document clustering is to minimize intra-cluster distances between documents while maximizing inter-cluster distances [1,2]. Some researchers noted that "clustering is especially useful in organizing documents to improve retrieval and support browsing" [3]. Some studies reported that although clustering is described as an unsupervised learning method, it can be used as an initial step that improves the accuracy of some supervised applications [4].

The quality of a clustering technique depends on determining similarity between documents. In most clustering techniques, cosine similarity between documents' vectors is used, which is based on the number of common words present in the documents, and the weight of each word. Two documents are defined to be related if their cosine similarity is above a certain value, called threshold. If the value of similarity is larger than the threshold, then the documents are assigned to the same cluster, so threshold plays as a discrimination criterion.

Choosing the proper threshold value is considered a major problem, since there are no specific rules to justify such choice [5], which makes it possible for false positive results to increase when searching for a cluster. Threshold value could be manually determined, where the domain expert is usually responsible for determining the threshold value that is most suitable for clustering. The manual estimation of the threshold value has a significant cost on the retrieval systems performance [6].

In some threshold manual selection approaches, the threshold is obtained simply by choosing any high percentile of the data. In such approaches, similarity is calculated

DOI: 10.24507/icicel.16.06.583

among several data items, and the highest similarity is selected as the threshold [7]. Selecting the highest value is not always the best threshold choice, as reported in [8]. Moreover, increasing the value of threshold is not always a good way to improve clustering quality [9].

In this research, we propose a new technique for threshold selection, which depends on the contents of the documents of the data set. The proposed technique depends on the distances between documents' vectors: internally in a cluster, and externally to other clusters. In the meanwhile, the local cosine similarity between document vectors in a cluster and its representative is also used in the evaluation process.

The rest of the paper is organized as follows. Section 2 presents a literature review related to clustering and threshold value estimation. Section 3 presents the proposed threshold estimation method. Section 4 introduces the implementation and evaluation over real-world data sets that showed efficiency, and Section 5 concludes the paper.

2. **Related Work.** Researchers in [10] proposed a soft clustering algorithm called SISC (Similarity-based Soft Clustering), based on a given similarity measure. Such algorithm uses a special suggested threshold value to determine similarity between two documents, and such threshold value is used in the "iterative step" of clustering. The iterative step examines each cluster and decides whether the centroids should change or not, meaning that a document is determined to be in a cluster centroid c, if it has a similarity $m(c, x)$ (the similarity of document x for cluster c) that is larger than the threshold. The algorithm terminates when no more changes are made.

Other researchers, as in [11] presented an incremental clustering algorithm based on maintaining high cluster cohesiveness, represented as a cluster similarity. They determined the quality of cluster cohesiveness by calculating the ratio of the count of similarities above a certain similarity threshold (ST) to the total count of similarities. That a higher ratio will yield more cohesive is the cluster.

In [12], the researchers proposed a new technique for threshold automatic learning. This technique is mainly based on the usual performance evaluation measures (such as Receiver Operating Characteristic (ROC) and precision-recall curves). Such technique depends on presenting a system (S) for classifying data collocations into two classes: relevant (true positive) and irrelevant (true negative) collocations. The ultimate goal of threshold technique in this system is to maximize the rate of true positives and the rate of true negatives, while minimizing the rate of false positives and the rate of false negatives.

A comparative study that compares the performance of six feature threshold techniques has been provided in [13], namely, Averaging (Avg), Maximization (Max), Fixed Local (FLocal), Weighted Local (WLocal) in addition to their proposed techniques: Standard Deviation (STD), and Maximum Deviation (MD). The globalization techniques are evaluated using the original, weighted and normalized scores.

Depending on their results, researchers claimed that Maximum Deviation (MD) has limited improvement when using Documents Frequency (DF) as a scoring method, especially at low threshold values, while FLocal has the best performance compared to the remaining methods. This supports the claim that localization techniques are better than globalization methods, while DF scores are suitable for the threshold when they use MD in an even distributed data set.

Related works could be categorized into two groups. In the first group, researchers established threshold value method concerning their algorithm; the threshold method was extracted from algorithm inputs, as in [10,11]. Therefore, attempts to apply such method in other fields or algorithms will meet many challenges, and could be an impossible task. The other group depends on evaluation measurements in order to extract threshold value, for example [12,13], which is a complex and time-consuming approach before achieving the optimal threshold value. Algorithm implementation in such category would wait until

the evaluation results appear to test the threshold value, and this should continue until achieving the optimal results.

Consequently, most of the previous approaches cannot be generally applicable to all algorithms. So, there is a need to introduce an approach that could provide a threshold value method that is applicable to most algorithms.

This paper proposed a method for estimating threshold value in text documents clustering technique, depending on data set collection and distances similarity measure, which avoids additional complex mathematical calculations. It is identified in the early stages of the clustering operations.

3. The Proposed Method of Threshold Estimation.

The proposed threshold estimation method, which can be generated without human decision will be described in detail in following subsections. This threshold value will be determined during the clustering preparation processes, which are fully dependent on the content of the documents in the data set being clustered.

3.1. Threshold estimation.

Threshold estimation strategy follows multiple steps, starting with selecting a sample of documents out of the data set, applying the clustering process and finally determining the threshold value.

The sample is randomly selected and clustered using any of the hierarchical clustering methods. Documents should be indexed using the vector space model, where each document is represented as a vector of terms, and the weight of terms in these documents should be measured.

In the following step, a representative for each cluster will be selected. Representatives in our approach adopted the centroid selection method proposed in [14]. This method is suitable for hierarchical clustering. Such approach depends on selecting index terms of the parent documents in the hierarchy, combining these terms into a virtual document vector of entries that are composed of the average weight of each index term that is repeated more than once.

After preparing the representatives, the global and local distances should be calculated. The local distance is the distance between the representative of a cluster and the documents in that cluster, while the global distance is the distance between the resulted representatives of the clusters. The averages of the local and global distances should be calculated separately after that. Finally, the averages of local and global distances will be used for obtaining the proposed threshold value. Figure 1 illustrates the proposed strategy applied for selecting the threshold value.

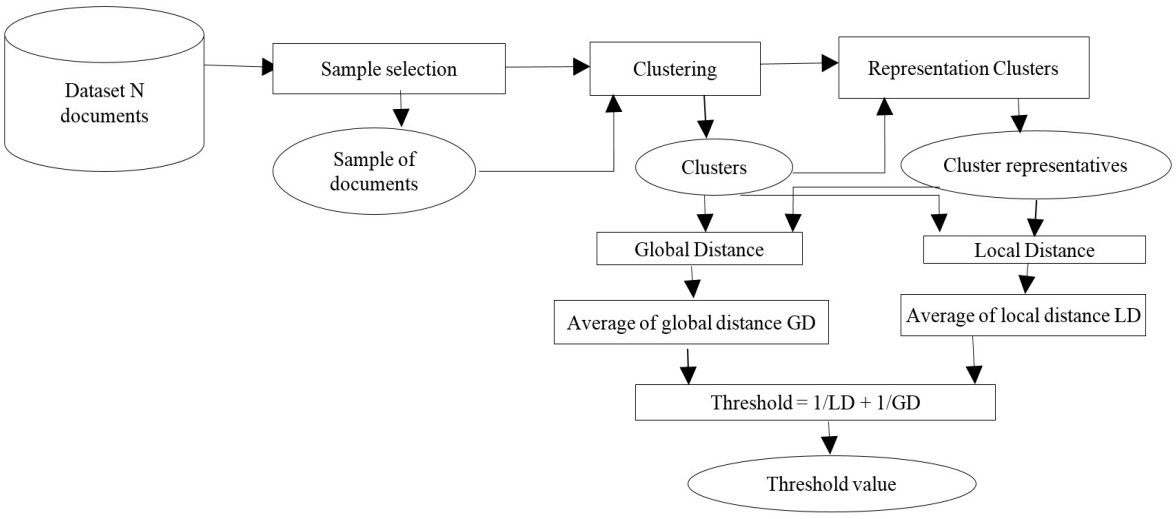

FIGURE 1. Threshold estimation procedure

The proposed approach views the threshold as a contribution of two parts: the internal properties of a cluster, and the relation between a cluster and other clusters of the collection. The internal properties are represented by the inverse average Euclidean distance between each document of a cluster, and a cluster representative. The relation between a cluster and other clusters is represented by the average distance between its representative, and the representatives of other clusters.

The Local Distance (LD) between any document in a cluster and its representative is expected to be small, and the overall average as well, while the relation between each cluster and others, which is referred to as Global Distance (GD), is expected to be high. In order to keep a high influence to internal properties, while reducing the influence of the external relationship to other clusters, the inverse of each value (of LD and DG) is used to express the proposed threshold, as seen in Equation (1):

$$T_{(threshold)} = \frac{1}{LD} + \frac{1}{GD} \tag{1}$$

3.2. **Clusters representatives.** Cluster representative, also known as "clustroid", is a vector of terms determined in different methods depending on the used clustering algorithm. This representative is selected to be close to all of the documents of the cluster, and as apart as possible to other representatives of other clusters.

The adopted centroid selection is the method proposed in [14], which is considered a suitable method for hierarchical clusters. Such approach depends on selecting index terms of the parent documents in the hierarchy, combining these terms into a virtual document vector of entries composed of the average weight of each index term that is repeated more than once. Average weight for a term t_k is given by the equation:

$$W(t_k, Cd) = \frac{\sum_{j=1} W(t_k, d_j)}{n} \tag{2}$$

where n is the number of term's; t_k is a term's value; d_j is a document that contains t_k, and $W(t_k, d_j)$ is the *tf-idf* weight of the term t_k in the document d_j.

The proposed method adopted a hierarchical clustering as it is shown to be an effective method of document clustering. It partitions a collection into high-level clusters that represent a broader topic, which are in turn partitioned into smaller clusters that represent tighter topics, and so on [15]. Hierarchical structure is usually classified into two methods depending on how the hierarchy is built, namely agglomerative and divisive. We used the agglomerative approach, which starts with an initial clustering of the term space, where all collection documents represent one cluster. The nearest clusters are merged using an inter-cluster similarity measure, and the process continues until only one cluster or a predefined number of clusters remains. Inter-cluster similarity is usually used in the Agglomerative method to classify documents. The most popular techniques are single-link, complete-link and group average. This proposed method applies the group average method, as it is not biased towards the border points of a cluster.

3.3. **Measuring distances.** Euclidean distance is a common dissimilarity measure in clustering techniques, in addition to the traditional cosine similarity. In our proposed approach, these two methods will be implemented and evaluated individually in order to achieve a better threshold value.

As mentioned earlier, the goal of document clustering is to minimize intra-cluster distances (Local Distance (LD)) between documents, while maximizing inter-cluster distances (Global Distance (GD)) between clusters. In order to measure Local Distance (LD), which represents the distance between vectors of documents in clusters and their representatives, we apply Equation (3), with the following assumptions:

- There is a vector of terms $V(x_1, x_2, \ldots, x_t)$, where x represents the weight of a term, and t is the number of terms, for a document, located in a cluster C,

- The representative $R_c(y_1, y_2, \ldots, y_t)$ of the cluster C, where y represents the resultant of weights of terms (computed by Equation (2)) in each representative, and t is the number of terms.

$$LD(C) = \sum_{i=1}^{n} dist(R_c, d_i) \tag{3}$$

where n is the number of documents in a cluster C, and $dist(R_c, d_i)$ is the distance between the centroid of the cluster C and the document d_i.

In order to measure the Global Distance (GD) between all of the centroids, we apply the following Equation (4), with the following assumptions:

- There is a representative vector $R_{cj}(y_1, y_2, \ldots, y_t)$ for each cluster C_j (where y represents values of resultant terms weights in the representative, and t is the number of terms),
- Clusters $C(c_1, c_2, \ldots, c_m)$, (where m is the number of clusters),

$$GD = \sum_{i=1}^{m} \sum_{j=i+1}^{m-1} dist(R_{ci}, R_{cj}) \tag{4}$$

Then, the arithmetic mean is used to compute the average of internal distances results in clusters (LD), and external distances (DG). The average of internal cosine similarity between documents and its representative will be computed as well. The overall procedure of threshold estimation is summarized by Algorithm 1.

Algorithm 1: Threshold estimation

Begin
Input a collection of text documents N;
Randomly select a sample documents K of N;
$C = HAC(K)$; // Hierarchical Agglomerative Clustering, C: set of clusters
Foreach cluster C_i in C **do** R_{ci} = Representative (C_i);
Foreach cluster C_i in C **do**
Foreach documents d_j in C_i **do**
$D_d = D_d + dist(d_j, R_{ci})$;
$LD = LD + D_d/|C_i|$;
    For $i = 1$ to m **do**
        For $j = i + 1$ to $m - 1$ **do**
$D_c = D_c + dist(R_{ci}, R_{cj})$;
$GD = GD + D_c/|C|$
Threshold Value $T = 1/LD + 1/GD$;
End

4. **Implementation and Evaluation.** This section includes implementation details of the proposed threshold estimation method. The standard Arabic France Press (AFP) Newswire Corpus TREC-2001 is used for evaluation. The corpus includes 383,872 documents containing approximately 666,094 unique words. It is distributed as files, each containing news collected in one day.

The following evaluation procedure is employed, a subset of documents was randomly selected from the TREC-2001, and then the selected data was manually clustered in order to be used as a benchmark to evaluate automatically generated clusters. Thereafter, the estimated threshold is used to specify the similarity during automatic clustering over the selected subset of documents. Finally, the correctness of the resulted clusters using the

purity and F-measure will be presented. The selection criterion is applied in a way that forces the sample to have a diverse content, and represents the overall collection.

4.1. **Manual clustering of the selected sample.** The selected sample of documents was manually clustered. Clustering was performed according to features that separate the documents into groups based on the internal properties of the collection. Applying the hierarchical clustering on documents distributed them into four high-level clusters, which in turn were clustered into smaller fifteen clusters. Each cluster contains documents with similar news content, referred to as $C = \{c_1, c_2, \ldots, c_k\}$. The first four in the high level marked as (level 1) contain main news topics (political, economy, sport, and collections), while the second level (level 2) contains sub main topic.

4.2. **Threshold estimation applied to automatic clustering.** In this stage, the selected subset of documents is automatically clustered using a hierarchical clustering system. Clusters are generated twice; firstly, they are generated using local-global distance, and secondly, using the cosine similarity between documents and the representative of the cluster they belong to.

Prior to clustering, documents are pre-processed, and indexed to represent each document as a set of terms using the Vector Space Model (VSM), after removing stop words. Meanwhile, evaluation of the proposed method needs each cluster to be represented by one vector (centroid). The centroid is selected to represent the semantic of the desired cluster, as proposed in [14].

4.3. **Local and global distances experiment.** The average distance between documents' vectors in each manually-generated cluster and the representative of its cluster is calculated as the Local Distance (LD), and the distance between centroids of all clusters, which represents the Global Distance (GD), or the inter-cluster average distance, is used to determine the appropriate threshold value, as described previously in Section 3. The resulted average local distances between documents and the representative of cluster in the first level, second level and the third level of the hierarchy are presented in Figures 2-4, respectively. It could be noticed that the average distance in some lower-level clusters is high, which means that the documents are talking about the same topic using different aspects, and so different terminology.

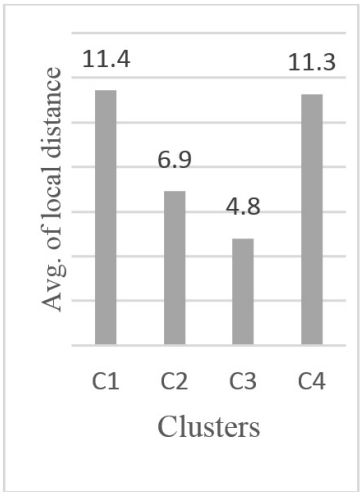

FIGURE 2. Average of local distances between documents and representative in level 1

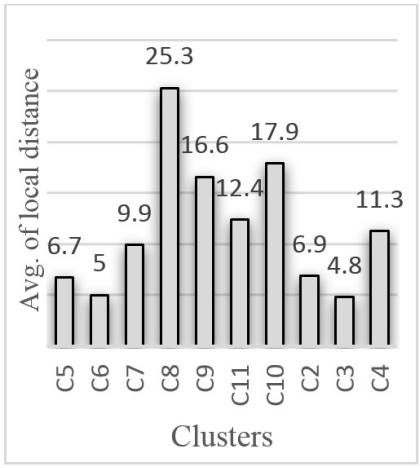

FIGURE 3. Average of local distances between documents and representative in level 2

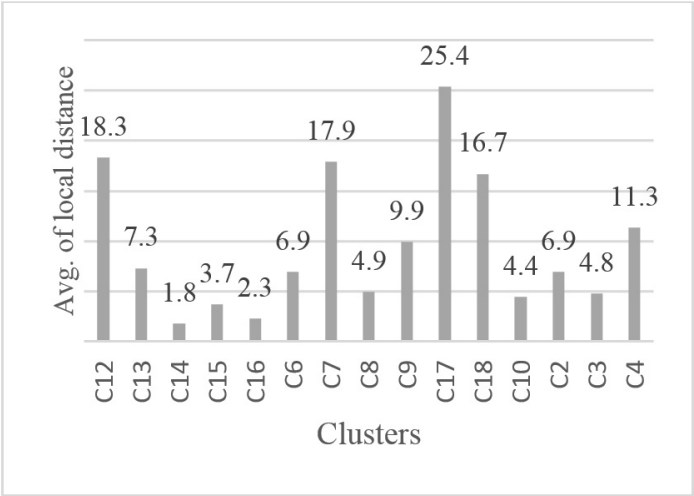

FIGURE 4. Average of local distances between documents and representative in level 3

According to the above results, the estimated threshold value is calculated based on the local and the global average distances in level-one of the hierarchical clustering, and presented in Table 1.

TABLE 1. Suggested threshold value (T)

	AvgofLD	**1/LD**	**AvgofGD**	**1/GD**	**T**
Level 1	8.6	0.12	11.5	0.087	**0.20**

4.4. **Cosine similarity experiment.** Threshold is estimated again as the average cosine similarity of each document and the representative of its cluster. The selected sample of documents is re-clustered using the resulted threshold. Figures 5-7 display results of the average of local cosine similarity between documents in clusters and their representatives for the three levels 1, 2, and 3 of the hierarchical clustering.

We exclude using the global cosine similarity and exclude calculating the combination of both local and global similarity, since the resulted threshold in these two cases will be too high and will in turn produce too small sized clusters. Therefore, the average of local cosine inter-clusters similarity will be used for threshold value estimation. The

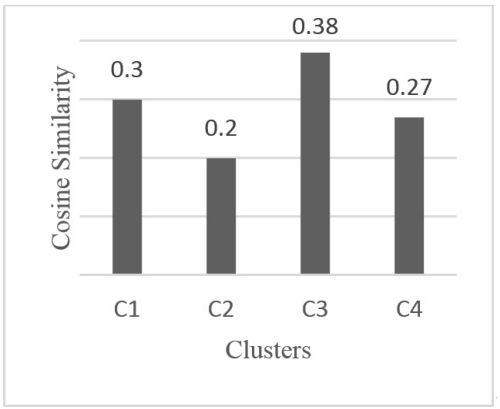

FIGURE 5. Average of local cosine similarity between documents and their representative in level 1

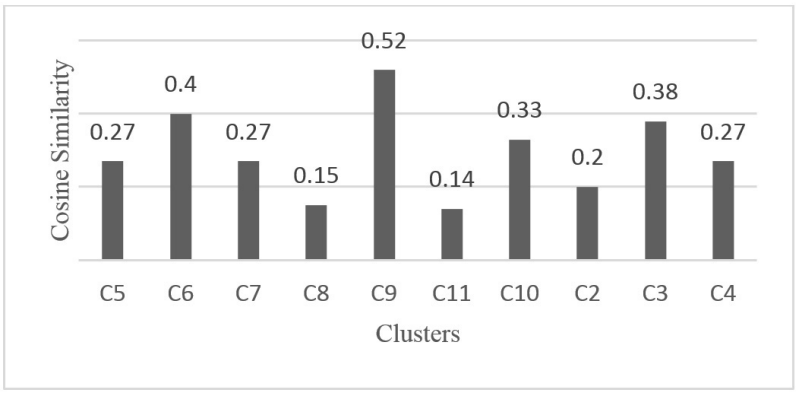

FIGURE 6. Average of local cosine similarity between documents and their representative in level 2

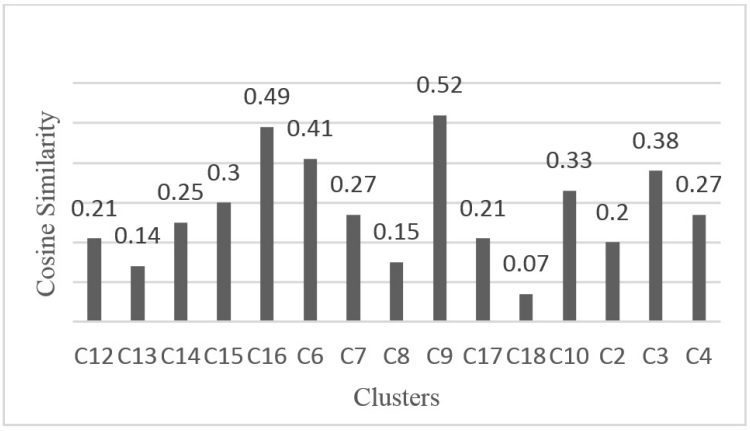

FIGURE 7. Average of local cosine similarity between documents and their representative in level 3

local cosine similarity will be compared to the proposed distance threshold estimation. The computed cosine similarity values are presented in Table 2, and the value estimated threshold is 0.2875. The estimated threshold values, either computed according to the local cosine similarity and to the local-global distance, are shown for the three levels in Figure 8. It could be found that the local-global distance shows lower threshold values in three levels.

TABLE 2. Average of local cosine similarity in level 1

Clusters	Avg.LocalCosinSim.
C1	0.3
C2	0.2
C3	0.38
C4	0.27
LAvg	**0.2875**

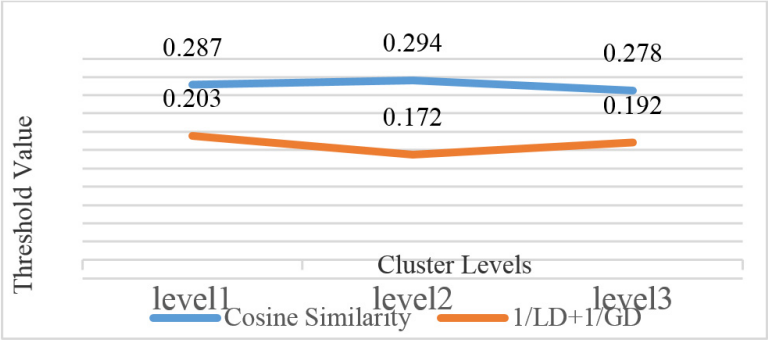

FIGURE 8. Suggested threshold value by cosine similarity and local-global distance in the three levels

4.5. **Evaluation – Purity measure.** Purity is a simple and transparent evaluation measure in clustering. To compute purity, each cluster is assigned to a class that is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and divided by N, as in Equation (5):

$$purity(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_{j} |w_k \cap c_j|, \tag{5}$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ is the set of clusters and $C = \{c_1, c_2, \ldots, c_j\}$ is the set of classes.

High purity is easy to achieve when the number of clusters is large – in particular, purity is 1 if each document gets its own cluster. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters [16]. Results of applying purity to our proposed approach are shown in Figure 9. It is observed that purity ranges from 0.77 to 0.9 for the distance method, and from 0.6 to 0.86 for the cosine similarity method. The distance method achieved higher purity than cosine method. This result

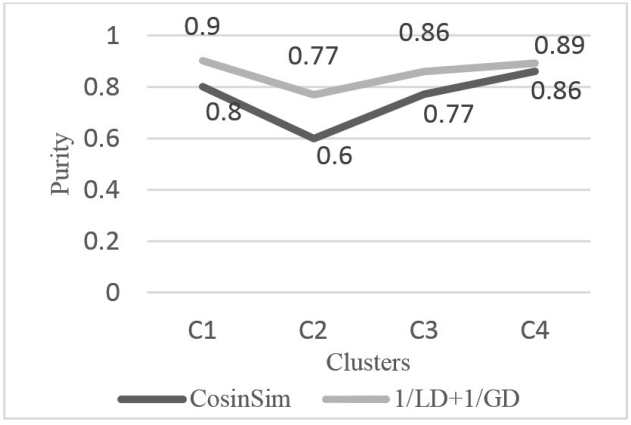

FIGURE 9. Comparison of purity between cosine similarity and distances

could be explained since the threshold value computed by the average cosine similarity is higher, so it is expected that resulted clusters will be of smaller size, and consequently the chance of some similar documents, that could use different terms to refer to the same topic, to belong to a cluster that contains that topic becomes lower, which results in a lower purity value.

4.6. **Evaluation – F-measure.** F-measure is a commonly used measure in clustering validation that combines the well-known precision and recall concepts. To evaluate a cluster (C_i), the F-measure equation will be as Equation (6).

$$F(c_i) = \max \frac{2P_jR_j}{P_j + R_j}, \quad F(c_i) = \max_{j=1,\ldots,m} \frac{2P_jR_j}{P_j + R_j},$$
$$\text{where } P_j = \frac{|c_i \cap k_j|}{|k_j|},\ R_j = \frac{|c_i \cap k_j|}{|c_j|} \tag{6}$$

where P is the precision, and R is the recall; the final F-measure for the entire set is given as in Equation (7).

$$F = \sum_{i=1}^{m} F(i) \frac{|c_i|}{N} \tag{7}$$

where N is the total number of documents.

In hierarchical clustering, the F-measure is the maximum value that occurs at any cluster in the tree, and an overall value for the F-measure is computed by taking the weighted average for each class, as given by Equation (8):

$$F = \sum_{i} \frac{n_i}{N} \max\{F(i,j)\} \tag{8}$$

where n_i is the number of documents in the ith cluster, the max is taken over all clusters and N is the number of documents; higher F-value indicates higher clustering quality [17]. Evaluation results of the proposed method using F-measure showed acceptable values compared to the results achieved by other researchers, such as [18]. Figure 10 shows the results of the F-values in level 1, and the max F-value is **0.61**.

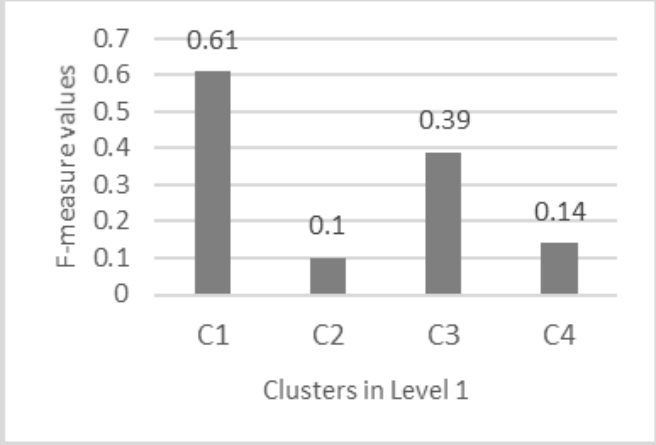

FIGURE 10. F-measure in level 1

5. **Conclusion.** In consistency with the clustering characteristics, and in order to make intra-cluster distances as close as possible, and inter-cluster distances as far as possible, we proposed an equation that depends on the local and global distances to estimate the similarity threshold value. Following the standard estimation of the local cosine similarity as a threshold value, against our proposed method, we concluded that our proposed

method yielded the best results in algorithm evaluation. Results of applying purity and F-measure showed the feasibility of the proposed method as compared to the cosine similarity or with previous studies and research works.

As a future work, the proposed method could be applied for further testing on different clustering algorithms, and on different data sets.

REFERENCES

[1] P. N. Tan, M. Steinbach, A. Karpatne and V. Kumar, *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co., Boston, 2005.

[2] K. Li, X. Cao, X. Ge, F. Wang, X. Lu, M. Shi, R. Yin, Z. Mi and S. Chang, Meta-heuristic optimization-based two-stage residential load pattern clustering approach considering intra-cluster compactness and inter-cluster separation, *IEEE Trans. Industry Applications*, vol.56, no.4, pp.3375-3384, 2020.

[3] P. Anick and S. Vaithyanathan, Exploiting clustering and phrases for context-based information retrieval, *Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, pp.314-323, 1997.

[4] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Inc., NJ, 1988.

[5] J. Hou and B. Zhang, Cluster merging based on a decision threshold, *Neural Computing and Applications*, vol.30, no.1, pp.99-110, 2018.

[6] F. Fkih and M. Omri, Estimation of a priori decision threshold for collocations extraction: An empirical study, *IJIRR*, vol.8, no.3, pp.34-49, 2013.

[7] W. H. DuMouchel, Estimating the stable index in order to measure tail thickness: A critique, *The Annals of Statistics*, vol.11, no.4, pp.1019-1031, 1983.

[8] L. Bednarik and L. kovács, Efficiency analysis of quality threshold clustering algorithms, *Production Systems and Information Engineering*, vol.6, pp.15-26, 2013.

[9] T. Boongoen and N. Iam-On, Cluster ensembles: A survey of approaches with recent extensions and applications, *Computer Science Review*, vol.28, pp.1-25, 2018.

[10] K. I. Lin and R. Kondadadi, A similarity-based soft clustering algorithm for documents, *Proc. of the 7th International Conference on Database Systems for Advanced Applications*, Hong Kong, pp.40-47, 2001.

[11] K. M. Hammouda and M. S. Kamel, Efficient phrase-based document indexing for web document clustering, *IEEE Trans. Knowledge and Data Engineering*, vol.16, no.10, pp.1279-1296, 2004.

[12] F. Fkih and M. N. Omri, Information retrieval from unstructured web text based on automatic learning of the threshold, *IJIRR*, vol.2, no.4, pp.12-30, 2012.

[13] N. M. Wanas, D. A. Said, N. H. Hegazy and N. M. Darwish, A study of local and global thresholding techniques in text categorization, *Proc. of the 5th Australasian Conference on Data Mining and Analystics*, Sydney, pp.91-101, 2006.

[14] M. A. Hassan and Y. A. M. Hasan, Efficient approach for building hierarchical cluster representative, *IJCSNS*, vol.11, no.1, pp.178-184, 2011.

[15] P. Willett, Recent trends in hierarchic document clustering: A critical review, *Information Processing and Management*, vol.24, no.5, pp.577-597, 1988.

[16] V. Rijsbergen, *Information Retrieval*, Butterworth-Heinemann Newton, MA, 1979.

[17] H. Xiong, J. Wu and J. Chen, K-means clustering versus validation measures: A data-distribution perspective, *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol.39, no.2, pp.318-331, 2008.

[18] B. C. M. Fung, K. Wang and M. Ester, Hierarchical document clustering using frequent item sets, *Proc. of SIAM International Conference on Data Mining*, San Francisco, pp.59-70, 2003.