Indonesian Journal of Electrical Engineering and Computer Science
Vol. 24, No. 2, November 2021, pp. 1054~1062
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v24.i2.pp1054-1062

New algorithm for clustering unlabeled big data

Marwan B. Mohammed, Wafaa AL-Hameed
College of Information Technology, University of Babylon, Baghdad, Iraq

Article Info

Article history:

Received Apr 18, 2021
Revised Sep 11, 2021
Accepted Sep 16, 2021

Keywords:

DBI
Hierarchical clustering
K-mean clustering
Lexical chain sentence
USE

ABSTRACT

The clustering analysis techniques play an important role in the area of data mining. Although from existence several clustering techniques. However, it still to their tries to improve the clustering process efficiently or propose new techniques seeks to allocate objects into clusters so that two objects in the same cluster are more similar than two objects in different clusters and careful not to duplicate the same objects in different groups with the ability to cover all data as much as possible. This paper presents two directions. The first is to propose a new algorithm that coined a name (MB Algorithm) to collect unlabeled data and put them into appropriate groups. The second is the creation of a lexical chain sentence (LCS) based on similar semantic sentences which are different from the traditional lexical word chain (LCW) based on words. The results showed that the performance of the MB algorithm has generally outperformed the two algorithms the hierarchical clustering algorithm and the K-mean algorithm.

Corresponding Author:

Marwan B. Mohammed
College of Information Technology
University of Babylon
Baghdad, Iraq
Email: marwan.bmr.phd@student.uobabylon.edu.iq

1. INTRODUCTION

In recent decades rapid growth in applications such as Internet search, digital imaging, etc and storage technology lead to the construction of a lot of high-volume, high-dimensional datasets. This data is stored digitally in electronic devices, consequently, improvement in different techniques such as classification, automatic data analysis, and retrieval techniques become required. Usually, this data tide is unstructured thus it is difficult to analyze them [1].

The clustering in general concept is an unsupervised aggregation technique that owns with a huge number of applications in several fields like medicine, business, imaging, marketing, image segmentation, chemistry, robotics, and climatology and usually, this technique is used to identify the identical class of elements based on their characteristics and it a subfield of data mining technique and it is very efficient to selecting out benefit information from the dataset [2], [3]. The Methods of cluster analysis are placed among statistics and informatics. One of the conditions of the clusters is that two objects from the same cluster are more similar than two objects from different clusters and the process of partitioning should achieve two important attributes. The first, homogeneity within the clusters (i.e data which belong to the same cluster should be similar). The second, heterogeneity between the clusters (i.e data which belong to two or more different clusters should be as dissimilar) [1], [4].

This work has two contributions. The first contribution is to create a lexical chain based on sentences lexical chain sentence (LCS) as a new idea instead of a traditional lexical chain common use based on words lexical word chain (LCW) to overcome two of the obstacles are: First, the word may have more than one sense

this named (ambiguous word) and thus the true sense must assign. Another challenge, a word may be associated with words in various chains .The second contribution is proposing a novel method coined named (MB algorithm) to collect numeric data unlabeled into clusters in general and the other side is contributing in forming LCS in particular. The MB algorithm differs from clustering algorithms in principle because it does not require identifying the number of clusters in the beginning, but it decides the number of clusters automatically based on the threshold value. Highlight the K − means algorithm and hierarchical algorithm to compared results with the proposed algorithm.

In the repetition cycle of hierarchical clustering (HC), either smaller clusters are merged into the larger clusters or larger clusters are divided into smaller clusters, the goal is to build a hierarchy of clusters which is called a dendrogram [1]. Hierarchical approaches have enjoyed substantial popularity in genomics and other fields for their ability to simultaneously uncover multiple layers of clustering structure [5] There are two kinds of approaches in HC. The first approach is named agglomerative clustering which runs on the principle of bottom-up, which is small clusters are combined into the larger ones. The second approach is called divisive clustering which depends on the principle of the top-down approach, which is the larger clusters are broken into smaller ones. Hierarchical clustering (HC) faces a fundamental problem lies in data analysis, where given data points and their pairwise similarities, in the form of a tree whose leaves correspond to data points and internal nodes, correspond to clusters. It is the suffering of slow, and the HC theory is considered underdeveloped Despite the abundance of HC algorithms, because of no "global" objective [6]-[8]. The k-means algorithm considers is the contrast of HC. Since it is one of the flat techniques [1] and treated as one of the most generally used clustering techniques for various applications [9], [10]. The idea of the k-means clustering includes the partitioning of a given number of data N into k clusters, where k is defined in prior, such that must be $k < N$ at the begging step in the algorithm requires initial assignment of objects into the selection of k cluster centroids so that the centroids have minimum similarity among themselves [4], [11]. The k-mean algorithm suffers several drawbacks. The first drawback, it that is unstable in selecting initial centroids for clusters, which densely affects the performance in terms of effectiveness [12]. The second drawback is the algorithm randomly chooses initial centroids by default. Finally, it does not supply any assurance of producing unique results after clustering. Therefore, to output effective results, must be initial cluster centroids are picking using a criterion based on standard deviation [13].

Some of the studies previous whether related to hierarchical and k-means algorithms and also talks about the lexical chain. These studies start from 2015 into 2021. Wei *et al.* [14] attempted towards integrating WordNet with lexical chains to reduce from problems which still exist several challenges, like synonym and polysemy, high dimensionality, extracting core semantics from texts, and assigning appropriate description for the generated clusters. The authors proposed approach exploited ontology hierarchical structure and relations provide a more accurate assessment of the similarity between terms for word sense disambiguation. Also, they introduced lexical chains to extract a set of semantically related words from texts, which can represent the semantic content of the texts. Abualigah *et al.* [15] proposed a new algorithm that improved the performance of the text clustering technique, so that was combined two different measures (i.e. Euclidean distance and cosine similarity) as objective function jointly to make an accurate decision during the clustering process these became and they named this algorithm "multi-objective k-mean (MKM)". As the researchers showed caused in the combined multi-objective with k-means clustering is the multi-objective function in the text clustering domain is not popular, and it considers this essence issue that affects the performance of the text clustering technique. Therefore, the increased performance of the multi-objectives function was investigated by using the k-mean text clustering technique. Kimes and *et al.* [5] studied focus on the problem in cluster analysis is whether the identified clusters represent the important underlying structure or are artifacts of natural sampling variation. Since there few numbers from the approaches have been proposed which addressed this problem in the context of hierarchical clustering, this problem is further complicated by the natural tree structure of the partition, and the multiplicity of tests desired to parse the layers of nested clusters. Therefore, they solved this problem by proposing a Monte Carlo-based approach for testing statistical significance in hierarchical clustering which addressed these issues. This approach was implemented as a sequential testing procedure guaranteeing control of the family-wise error rate. Kalra *et al.* [16] proposed a framework for purpose analysis and data mining of heterogeneous data of the multiple heterogeneous data sources.it came to solve the challenging task of developing exploratory analytical techniques to explore clustering techniques on heterogeneous data consist of heterogeneous domains such as categorical, numerical, and binary or a combination of all these data through applied the k-Mean clustering algorithm in real life. The authors' succeed to achieve the goal of this work to retrieve the result individually from all the data sources into one format, analysis of all the heterogeneous sources including text corpus, social media, image, and homogeneous data, applying the clustering algorithm individually on each heterogeneous data source for extracting the hidden knowledge. But they showed that can occur loss information when converted data heterogeneous to homogeneous. Tiwari and Dembla [17] have proposed a novel algorithm for the automatic

text summarization system that utilized lexical chain calculation and it was implemented using eclipse Java development tool, enterprise edition for web developers. This method also involved the nouns and proper nouns in the computation of lexical chains. this algorithm addressed the most vital information and it is not longer than half of the source data and also, it is the best solution for the information overloading problem as do not have to scan through each line of long length documents and still receive the foremost important information. Therefore, in this approach, they have taken the concept of the significance and utility calculation for each chain so that the chains related to the documents are selected and used in the summary generation process. The advantage of this method is better output in terms of Execution time as compared to the existing algorithm, Improved match of words between the human-generated summary and proposed algorithm-generated summary, and better recall, which are commonly used criteria for summary evaluation. Chami *et al*. [6] proposed a new method called hyperbolic hierarchical clustering (HypHC) to displaying a direct correspondence from discrete trees to continuous representations through the hyperbolic embeddings of their leaf nodes and then back by a decoding algorithm that maps leaf embeddings to a dendrogram, which allows them to search the space of discrete binary trees with continuous optimization. They consider this method as the first continuous relaxation of Dasgupta's discrete optimization problem with provable quality guarantees so that they derived a continuous analog for the notion of the lowest common ancestor depend on analogies between trees and hyperbolic space.

This paper is organized as follows, section 2 explains the differences between the traditional lexical chain word and a new lexical chain sentence. Section 3 displays details of the proposed method, finally, section 4 illustrates the experiment's result which shows results and debates. The derived conclusion is shown in section 5.

2. Lexical chain sentence (LCS)

This section explains the difference between the lexical chain sentence (LCS) proposed and the lexical chain word (LCW). The lexical chain (i.e. LCW) is built by calculating the semantic distance between the words using WordNet. the lexical relationship exists between words, these lexical relations between words are extracted by using WordNet. At LCW Each word must belong to exactly one chain when lexical chains are computed. But there are two challenges are: First, there may be more than one sense for a word (ambiguous word) and thus the correct sense must be identified. Another challenge, a word may be related to words in different chains. The lexical chain aims to find the best way of grouping the words that will result in the longest and strongest chains [18], [19]. Consider lexical chaining as an example of "semantic approaches" or also known as "linguistic approaches" because word sense disambiguation tries to build relationships among words or sentences to lead to the partial comprehension of the document. Morris and Hirst were the first to implement an idea of lexical chaining in 1991. The lexical chain mainly deals with the problem of word sense disambiguation (WSD). It is created based on the same topic words of the document. Generally, lexical chains provide a better indication of discourse topic than does word frequency simply because different words may refer to the same topic. Even without sense disambiguation, these approaches can derive concepts [20].

While the idea of propose LCS is to make the LCS deal with the problem of sentence sense disambiguation (SSD) and how to make it in the correct chain. Lexical sentence chains are created based on similar sentences sense and another hand same based on topic sentences of the document. the proposed LCS is constructed by computing the semantic distance among sentences through using memetic between universal sentence encoder which proposed model [21] and cosine similarity distance coined universal sentence encoder cosine similarity (USECS) without using WordNet as LCW. Also, the LCS relationship that exists between sentences is extracted from USECS. Each sentence must belong to exactly one chain (cluster) when lexical chains are computed. The LCS overcome the challenge of LCW by taking sentences completely without tokenizing sentences into words like LCW. Each sentence has one sense difference about words which may be more than one sense. Hance, identify correct sentence sense becomes easy, also the LCS prevents reddened sentences in more than one chain (cluster). It collects sense sentence similarity in one chain sentence as much as possible. Thus, the LCS seeks to find the best method to collects sentences that will result in the longest and strongest chains. Figures 1(a) and (b) shown differences at work LCW and propose LCS respectively.

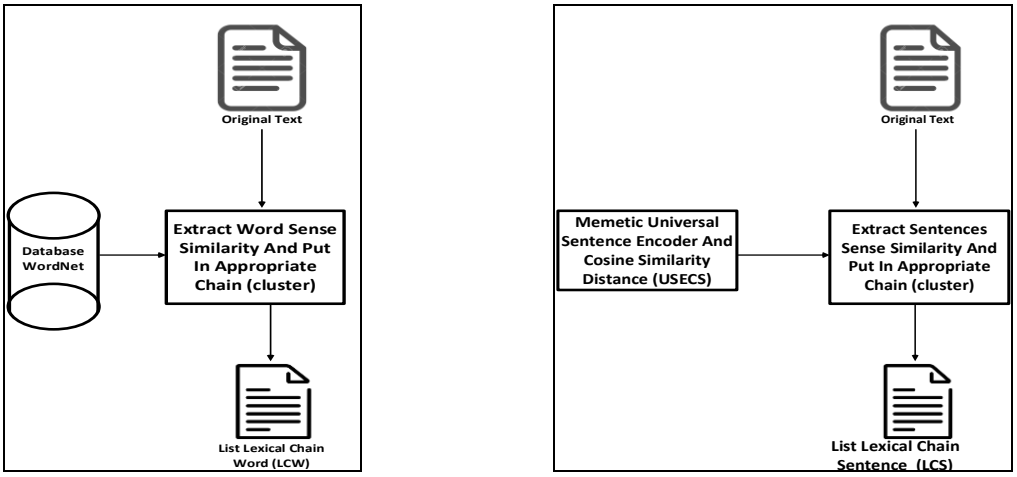

Figure 1. (a) Lexical chain word graph for extract word semantic similarity from text, and (b) Lexical sentences graph for extract sentences semantic similarity from the text

3. THE ALGORITHM PROPOSED

The proposed Algorithm has six-phases as showed in Figure 2. In this section, the main steps of the proposed algorithm are described listed as follows in detail:

- The population is taken from dataset document understanding conference (DUC) 2002 which contains a set of topics up the number to 59 topics. The Table 1 explaining in detail the content of this dataset. The sentences in documents are separated by the sentence tokenization process, after that using the universal sentence encoder (USE) model proposed [21] for creating an embedding sentence vector for each sentence with a fixed length. This model is interesting with sentences (i.e context-based representation) only so that transforms each sentence completely into an embedding sentence vector instead of learning vectors for individual words in the sentence, they compute a vector for sentences on the whole, by taking into account the order of words and the set of co-occurring words. Thus, this model is different from the word2vector model which deals with words based [22]. This model has overcome on sparse matrix problem which occurs in cosine similarity distance because it takes to consider sentences semantically, thus, this work depended on the principle memetic between USE and cosine similarity distance in this state.
- Calculate similarity between sentence embedding vectors by using cosin distance as shown in (1) [23]:

$$\text{Cosine Distance}(V, C) = \frac{\sum_{i=0}^{n} V_i \times C_i}{\sqrt{\sum_{i=0}^{n} V_i^2} \times \sqrt{\sum_{i=0}^{n} C_i^2}} \quad (1)$$

  where i is counter for vectors and centroid columns and V_i is represent vector sentence and C_i is represent centroid for each cluster. The results of this measure distance are put in a matrix named $SimiliratyMatrix$. The size of this matrix is $(m \times m)$ (i.e square matrix).
- In the proposed method the center selecting from the embedding sentences vector (ESV), this selecting is being sequential. This center attracts sentences that similar it semantically through similar value resultant from center and sentence which must be greater or equal to a threshold value. This method identifies the threshold value previously and does not identify a number of the clusters because it is deciding the number of clusters optimality based on the threshold value automatically. The sentences compatible with center according to threshold value on-base its index number to be placed in the cluster named $\text{Cluster ESV}_\text{i}$ where(ESV_i) is represent sentence vector index which becomes the center and cluster ESV_irepresent sub-matrix include sentence numbers that attracted. However, when selecting a new center (i.e $\text{ESV}_{\text{i+1}}$) to bring the rest of the sentences which not compatible with the center previous (i.e ESV_i) for purpose create another cluster. This state should be $\text{ESV}_{\text{i+1}}$ not mentioned in the content of other clusters that preceded it. Because selecting it as new centers maybe frequents same data in a new cluster and this leads to increase cluster numbers and weakens achieving optimality clusters which aim to cover all data without repeat. Thus, must ignore it and continue the loop to take a sentence as the new center is not mentioned pervious.
- After completing the collection process sentences in clusters and keep them in a list named Cluster ESV, the proposed method must ensure that clusters content free of redundant the same sentence in more than

one cluster. If such a situation exists, it compares similar values in the sentence in all clusters and survival of this sentence in the cluster which have a higher similarity value than other clusters, then the similarity values of this sentence are deleted in the rest of the clusterable.

- Check the number of sentences in the cluster created. this method is required that the cluster content must be greater than two to avoid that being content cluster less than two after duplicate removal.
- Now, maybe there exist sentences with similarity values with centers but not compatible based on a threshold value, thus became these sentences called outliers. To include these outliers through taking similarity values each one of them with centers only and compares among them and selecting a higher value and put it in the cluster that belongs to that center. This list is considered certified lexical chain sentences. In this step, the process ends create LCS proposed or set of clusters that coverage all sentences in the document.

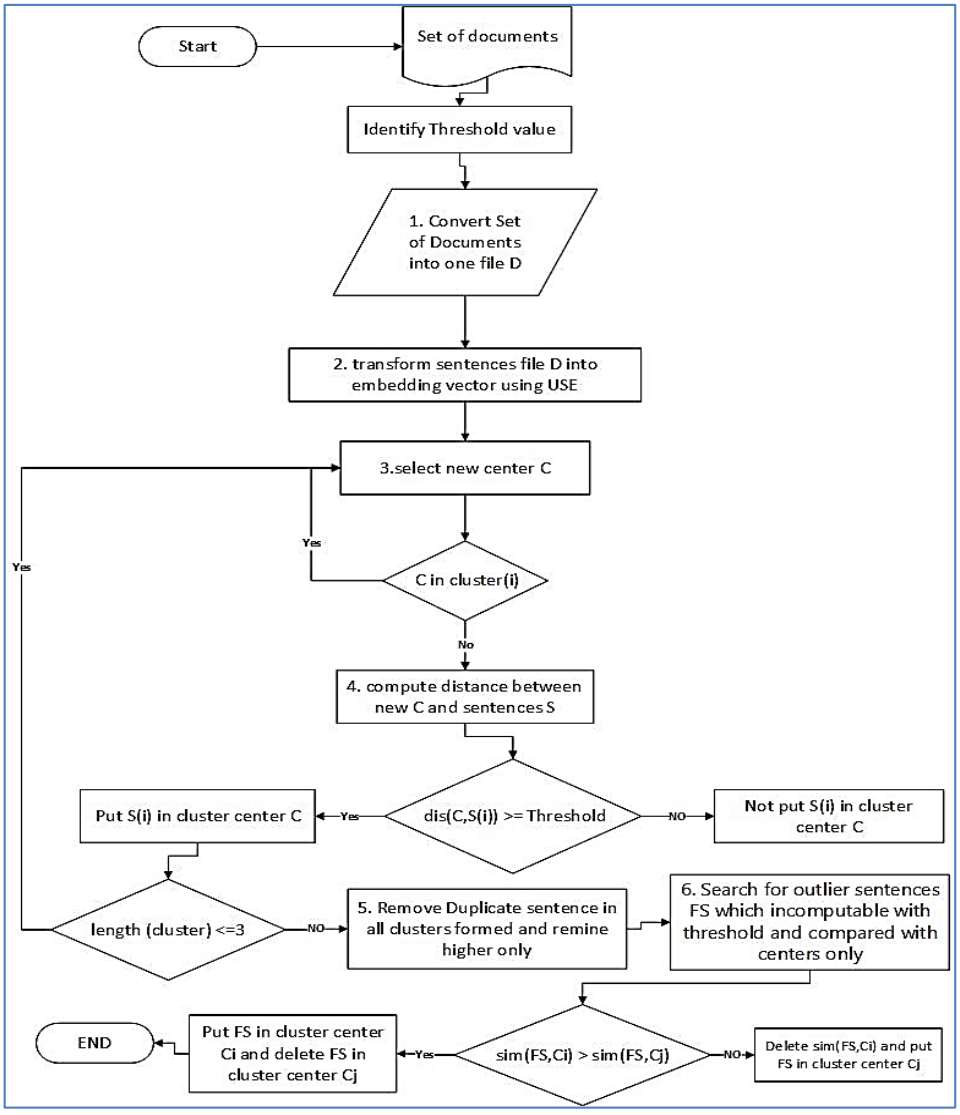

Figure 2. Flowchart explains steps MB algorithm proposed to collect sentences in the clusters according to the threshold value

Table 1. Description DUC 2002 dataset

Description	DUC 2002 dataset
Number of topics	59 (d061j through d120i)
Number of documents in each topic	~10
Total number of documents	567
Data source	TRES
Summary length	200 d 400 words

4. RESULTS AND DISCUSSION

This work deal with a text dataset named DUC 2002. The document understanding conference (DUC) is the most common benchmarking dataset used for text summarization [24]. The DUC 2002 contains a set of topics up the number to 59 topics, each topic includes a group of documents (articles) $D = \{d_1, d_2, \dots, d_n\}$ talking about that topic. each d_i contains a set of sentences $S = \{s_1, s_2, \dots, s_m\}$. All documents sentence special to a specific topic makes into one file $D^* = \{s_1, s_2, s_3, \dots, s_n\}$ for simplifying.

Using the Davies bouldin index (DBI) method to assess clusters and the strong relationship between them, another side to evaluate correlation content between them for each cluster. Finally, it gives a score. This score whenever a positive and low value is good and indicates that this method is strong and better. The DBI introduces a scattering measure SC_i to measure the scattering within the same cluster and maximizes the ratio of scattering measure to the cluster center isolation and to give the DBI for many clusters C. Can talk that the DBI considers the average case of each cluster by using the mean error of each cluster. Thus, the equation for DBI can be expressed as (4) [2], [25].

$$SC_{i,q} = \left(\frac{1}{|A_i|}\sum_{x_j \in A_i} \|x_j - v_i\|^q\right)^{\frac{1}{q}} \quad (2)$$

$$R_i = max_{j=C\&j\neq i} \frac{SC_{i,q} - SC_{j,q}}{\|v_i - v_j\|} \quad (3)$$

$$\therefore DBI = \frac{1}{C}\sum_{i=1}^{c} R_i \quad (4)$$

Where SC_i is scattering measure, A_i is the size of cluster I, x_j be an n-dimensional feature vector assigned to the cluster, v_i is the centroid of the cluster, R_i is a measure of how good the clustering scheme; C is a number cluster.

The results of the proposed method are compared with the results of two algorithms, $\mathrm{k-means}$, and hierarchical clustering algorithm to find efficiency and strength to collect sentences in correct clusters. Using the DBI method to evaluate the power of these algorithms. This work takes five topics (061-065) from DUC 2002 dataset to display how distributed sentences in the clusters based on threshold value as proposed method or based on the number of k as in $k-means$ algorithm or based on max distance because it deals with similar sentences as in hierarchical clustering.

A series of threshold values were experimented with within the proposed method for the five topics above to find out the number of clusters that the method decides based on the threshold value as shown in Table 2. The selected number of k clusters in the $\mathrm{K-means}$ algorithm is the same number of clusters that the method proposed generated. In this paper, cluster validity analysis was applied to ensure the validity of the number of clusters considered in each clustering algorithm, also, it offers numerical value for different groups' validity indices which indicate the number of clusters. One of the cluster validity indices used for cluster validation called davies bouldin index (DBI). The purpose is to evaluate the two algorithms in terms of efficiency and strength by using the DBI scale. As for the comparison of the proposed method with the hierarchical clustering, this will focus on the number of clusters generated in the two methods only as expensive or not, since the hierarchical clustering does not determine the k cluster in advance.

The Table 2 explaining that the number of clusters in the proposed method in all topics compared with hierarchical clustering is better. Because the number of clusters generated according to the threshold value is smaller than the number of clusters generated in the hierarchical clustering algorithm. Therefore, the proposed method considers less expensive than hierarchical clustering. Although threshold values are different. But the number of the clusters may be frequenting and this does not mean frequent same sentences or same centers. Due to it based on condition, mean maybe a set of sentences compatible with one center. Thus, become this a set within a content this center as like when T=0.59 the number cluster is 3 whereas when being T=0.55 the number of clusters is 4 for example in topic 062. Also, in this table exist the word 'null' which means the proposed method does not create clusters because the sentences are not maybe the agreement with the threshold value specially or with the condition generally. Table 3 explains the results evaluation k-mean algorithm and proposed method using DBI measure. Most experiments in evaluation number of clusters generated totally in the proposed method successfully in grouping clusters form more correlation and efficiency than the k-mean algorithm. Usually, DBI scores for the algorithms when being an algorithm lower score than another algorithm which means the algorithm is good. Since, as DBI score low this means good.

Table 2. Number of clusters generated from two algorithms

Topic name	Threshold	No.of. clusters in the proposed method	No. of. clusters in the hierarchical clustering
061	0.5	19	343
	0.55	15	
	0.56	12	
	0.57	10	
	0.58	9	
	0.59	9	
	0.6	7	
	0.61	5	
	0.62	4	
	0.63	4	
	0.64	3	
	0.65	3	
	0.66	2	
062	0.5	7	233
	0.55	4	
	0.56	3	
	0.57	3	
	0.58	4	
	0.59	3	
	0.6	3	
	0.61	3	
	0.62	4	
	0.63	2	
	0.64	2	
	0.65	Null	
	0.66	Null	
063	0.5	7	405
	0.55	7	
	0.56	6	
	0.57	3	
	0.58	2	
	0.59	2	
	0.6	2	
	0.61	2	
	0.62	2	
	0.63	2	
	0.64	2	
	0.65	Null	
	0.66	Null	
064	0.5	5	189
	0.55	Null	
	0.56	Null	
	0.57	Null	
	0.58	Null	
	0.59	Null	
	0.6	Null	
	0.61	Null	
	0.62	Null	
	0.63	Null	
	0.64	Null	
	0.65	Null	
	0.66	Null	
065	0.5	13	365
	0.55	7	
	0.56	4	
	0.57	2	
	0.58	2	
	0.59	2	
	0.6	Null	
	0.61	Null	
	0.62	Null	
	0.63	Null	
	0.64	Null	
	0.65	Null	
	0.66	Null	

The Table 3 is contain five topics coined (061,062,063,064,065). Each topic contains evaluations between the proposed algorithm and $K-$means algorithm by using DBI measure. The results Topic 061 display that the proposed method is not successful with k-mean when cluster number is 2 and 3, but in remain

clusters successful it. Topic 062 shown the proposed method better than the k-means method in all numbers of clusters except cluster number 2. Topic 063 the proposed method success only in cluster number 7 while in remain clusters to the same topic is failed. In topic 064 that the proposed method is advancing on the k-mean algorithm. In topic 065 the proposed method outperforms the $K-means$ algorithm in all clusters except cluster number 2.

Therefore, can conclude that the evaluation DBI metric showed the proposed algorithm succeeded in evaluation impressively in many experiments, whether the number of clusters is 3 or more, or when the number of groups is small, regardless of the presence of some minor failures. Thus, it can be said that each algorithm has successes and failures. The proposed MB algorithm consider is the best compared with $K-means$ algorithm in terms of relationships and correlations between clusters, and with a hierarchical clustering algorithm from the numbers of clusters generated.

Table 3. Evaluation k-means and proposed methods by using DBI measure

No. cluster	DBI with k-means score	DBI with proposed method score
	Topic 061	
2	3.708	28.712
3	3.987	4.8568
4	3.717	1.6798
5	3.671	1.0024
7	3.201	1.6853
9	3.108	0.7809
10	2.937	0.5763
12	2.817	0.3108
15	2.646	0.0237
19	2.477	0.4966
	Topic 062	
2	3.810	9.4344
3	3.451	2.4584
4	3.435	0.8233
7	3.291	0.2568
	Topic 063	
2	4.270	13.7099
3	4.652	11.0973
6	3.971	9.6804
7	3.892	2.04537
	Topic 064	
5	3.538	0.3378
	Topic 065	
2	3.700	12.4589
4	4.708	1.2410
7	3.908	0.3986
13	3.050	0.0265

5. CONCLUSION

All sentences of documents are relevant to a specific topic gathered in one file for simplifying, then, similar semantic sentences will be collected from this file and put in an appropriate cluster coined-called chain sentence. After completing the assembly process, a lexical chain sentence (LCS) will be created. These proposed method characteristics are different from clustering algorithms in principle because it does not require identifying the number of clusters at the start, but it decides the number of clusters automatically based on the threshold value. While most cluster algorithms require identifying the number of clusters k in beginning like Kmeans algorithm. It is similar to the hierarchical clustering algorithm in principle not require the number of k clusters at the beginning. Whereas it differs from the hierarchical clustering algorithm because when wanting to merge an item in a cluster, the proposed method computes the distance between center and item only to decide merge or not without computing distance between content cluster and item. The merge condition whether min\max depends on the data type used in work also according to the threshold value. Thus, this method is less expensive. While in hierarchical clustering merge process item with cluster occurs by computing distance between content cluster with an item then select min\max distance this depends on the data type and not require to identify threshold value in order grouping points (sentences). Thus, hierarchical clustering considers more expensive. In general, this method is suitable for clustering any numerical data type. The collect numerical data are unlabeled in clusters. This is very important for easily dealing. The clustering algorithms help to place close data in a specific cluster. This work has taken two algorithms clustering are hierarchical clustering (HC), and K-clustering and compared them to the proposed method results. The output of this approach clarified that it is the best in most experiments conducted. Also, this paper success in creating a lexical chain based on sentence (LCS) thus becomes there flexible to deal with

sentences as a complete sentence in the chain based on semantic sentence similarity. The future work is to apply this proposed method to other datasets and compare it with other algorithms. Also, after success in creating a lexical chain based on sentence (LCS), will use to extract sentences to forming a summary in the next later.

REFERENCES

[1] S. V. Wazarkar and A. A. Manjrekar, "HFRECCA for clustering of text data from travel guide articles," in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Delhi, India, 2014, doi: 10.1109/ICACCI.2014.6968349.

[2] M. A. B. Siddique, R. B. Arif, M. M. Rahman Khan, and Z. Ashrafi, "Implementation of Fuzzy C-Means and Possibilistic C-Means Clustering Algorithms, Cluster Tendency Analysis and Cluster Validation," *ArXiv*, vol. 1, pp. 1-8, Nov 2018, doi: 10.20944/preprints201811.0581.v1.

[3] F. Kuwil, Ü. Atila, R. Abu-Issa, and F. Murtagh, "A novel data clustering algorithm based on gravity center methodology," *Expert Systems with Applications*, vol. 156, Oct 2020, doi: 10.1016/j.eswa.2020.113435.

[4] H. Rezanková and B. Everitt, "Cluster analysis and categorical data," *Statistika*, vol. 89, no. 3, pp. 216-232, 2009.

[5] P. K. Kimes, Y. Liu, D. N. Hayes, and J. S. Marron, "Statistical Significance for Hierarchical Clustering," *Biometrics*, vol. 73, no. 8, pp. 811-821, 2017, doi: 10.1111/biom.12647.

[6] I. Chami, A. Gu, V. Chatziafratis, and C. Ré, "From Trees to Continuous Embeddings and Back:Hyperbolic Hierarchical Clustering," *arXiv*, vol. 1, pp. 1-27, 1 Oct 2020, doi: 2010.00402v1.

[7] J. Leskovec, R. Anand, and D. U. Jeffrey, "Mining of massive data sets," Cambridge university press, 2020, pp. 566, doi: 10.1017/cbo9781139924801.001.

[8] M. Charikar, V. Chatziafratis, and R. Niaza, "Hierarchical Clustering better than Average-Linkage," in *Proceedings of the 2019 Annual ACM-SIAM Symposium on Discrete Algorithms*, 2019, doi: 10.1137/1.9781611975482.139.

[9] M. A. Rajab and L. E. George, "Stamps extraction using local adaptive k- means and ISODATA algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 137-145, Jan 2021, doi: 10.11591/ijeecs.v21.i1.pp137-145.

[10] S. Yu, S. Chu, C. Wang, Y. K. Chan, and T. C. Chang, "Two improved k-means algorithms," *Applied Soft Computing*, vol. 68, pp. 747-755, Jul 2018, doi: 10.1016/j.asoc.2017.08.032.

[11] S. Guha and N. Mishra, "Clustering data streams. In Data stream management," *Data-Centric Systems and Applications*, pp. 169-187, 2016, doi: 10.1007/978-3-540-28608-0_8.

[12] S. Zahra, M. A. Ghazanfar, A. Khalid, M. A. Azam, U. Naeem, and A. P. Bennett, "Novel Centroid Selection Approaches for KMeans-Clustering Based Recommender," *Information Sciences*, vol. 320, pp. 156-189, 2015, doi: 10.1016/j.ins.2015.03.062.

[13] A. Mustaqeem, S. M. Anwar, and M. Majid, "A Modular Cluster Based Collaborative Recommender System for Cardiac Patients," *Artificial Intelligence in Medicine*, vol. 102, p. 101761, Jan 2020, doi: 10.1016/j.artmed.2019.101761.

[14] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2264-2275, Mar 2015, doi: 10.1016/j.eswa.2014.10.023.

[15] L. M. Abualigah, A. T. Khader, and M. A. Al-Betar, "Multi-objectives-based text clustering technique using K-mean algorithm," in *7th international Conference on Computer Science and Information Technology (CSIT)*, Amman, Jordan, 2016, doi: 10.1109/CSIT.2016.7549464.

[16] M. Kalra, N. Lal, and S. Qamar, "K- Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data," *Information and Communication Technology for Sustainable Development*, pp. 61-70, 2018, doi: 10.1007/978-981-10-3920-1_7.

[17] A. Tiwari and D. Dembla, "A Novel Algorithm for Automatic Text Summarization System Using Lexical Chain," *Advances in Intelligent Systems and Computing*, vol. 904, pp. 103-112, 2019, doi: 10.1007/978-981-13-5934-7_10.

[18] M. Berker and T. Güngör, "Using Genetic Algorithms with Lexical Chains for Automatic Text Summarization," *Proceedings of the 4th International Conference on Agents and Artificial Intelligence*, 2012, doi: 10.5220/0003882405950600.

[19] S. Saxena and A. Saxena, "An Efficient Method based on Lexical Chains for Automatic Text Summarization," *International Journal of Computer Applications*, vol. 144, no. 1, pp. 47-52, 2016, doi: 10.5120/ijca2016910104.

[20] C. Mallick, M. Dutta, A. K. Das, A. Sarkar, and A. K. Das, "Extractive Summarization of a Document Using Lexical Chains," *Soft Computing in Data Analytics*, pp. 825-836, 2019, doi: 10.1007/978-981-13-0514-6_78.

[21] D. Cer *et al*., "Universal Sentence Encoder," *arXiv*, vol. 2, 2018, doi: arXiv:1803.11175v2.

[22] A. Joshi, S. Karimi, R. Sparks, C. Paris, and C. R. MacIntyre, "A Comparison of Word-based and Context-based Representations for Classification Problems in Health Informatics," in *18th BioNLP Workshop and Shared Task*, Florence, Italy, 2019, doi: 10.18653/v1/w19-5015.

[23] R. Ahuja and W. Anand, "Multi-document Text Summarization Using Sentence Extraction," *Springer*, vol. 517, pp. 235-242, 2017, doi: 10.1007/978-981-10-3174-8_21.

[24] N. Moratanch and S. Chitrakala, "A Survey on Abstractive Text Summarization," in *International Conference on Circuit, Power and Computing Technologies [ICCPCT]*, Nagercoil, India, 2016, doi: 10.1109/ICCPCT.2016.7530193.

[25] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224-227, 1979, doi: 10.1109/tpami.1979.4766909.