

Max stable set problem to found the initial centroids in clustering problem

Awatif Karim¹, Chakir Loqman¹, Youssef Hami², Jaouad Boumhidi¹

¹LISAC Laboratory, Faculty of Science Dhar El Mehraz, University Sidi Mohamed Ben Abdellah, Fez, Morocco

²RTMCSA Research Team, School National of Applied Sciences, University Abd El MalekEssaadi, Tangier, Morocco

Article Info

Article history:

Received Feb 9, 2021

Revised Nov 5, 2021

Accepted Nov 26, 2021

Keywords:

Continuous hopfield network

Document clustering

Initial centroids

Maximum stable set problem

ABSTRACT

In this paper, we propose a new approach to solve the document-clustering using the K-Means algorithm. The latter is sensitive to the random selection of the k cluster centroids in the initialization phase. To evaluate the quality of K-Means clustering we propose to model the text document clustering problem as the max stable set problem (MSSP) and use continuous Hopfield network to solve the MSSP problem to have initial centroids. The idea is inspired by the fact that MSSP and clustering share the same principle, MSSP consists to find the largest set of nodes completely disconnected in a graph, and in clustering, all objects are divided into disjoint clusters. Simulation results demonstrate that the proposed K-Means improved by MSSP (KM_MSSP) is efficient of large data sets, is much optimized in terms of time, and provides better quality of clustering than other methods.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Awatif Karim

LISAC Laboratory, Faculty of Science Dhar El Mehraz, University Sidi Mohamed Ben Abdellah

Box 30003, Fez, Morocco

Email: awatif.karim@usmba.ac.ma

1. INTRODUCTION

Today, clustering is an important tool in data mining; it is an automatic learning process where the objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity without knowing the labels of the data. Many clustering strategies are available in the literature as [1], [2], among them two main categories known as hierarchical and partitioning clustering. A tree structure created in the process of hierarchical clustering and shows how objects are grouped (in an agglomerative method) or partitioned (in a divisive method). Dumond [3] have argued that the agglomerative hierarchical clustering algorithm starts with each object forming a separate group.

In partitioning methods, a data set is decomposed into k partitions. A cluster is a set of elements represented by the centroid (or prototype) of the cluster. This latter is formed in such a way that it is closely related (in terms of similarity function) to all objects in that cluster. The best known partitioning methods including, K-Means and its variants [4], fuzzy c-means, possibilistic c-means, hard c-means (HCM) and mean shift [5], K-medoids [6], partitioning around medoids (PAM), CLARA [7] and CLARANS [8]. Most of these algorithms select randomly the initial centers in advance, which considerably affect the quality of clustering results. In the same context, K-Means is an unsupervised learning method [9], which is widely used in text clustering. K-Means is an algorithm whose cluster number is given at the start and well constructed as well. But, the difficulty is that the results of clustering depends on the fixed number of classes and on the random choice of initial clusters, especially when the data set is large and we don't have assumptions about the data [10]. However, it can be stuck in a local minimum and cause an unstable result (if we reinitialize the algorithm with other values, it may converge to another local solution) [11]-[14].

To overcome the above deficiencies, most research in clustering analysis has been focused on the “automatic clustering algorithm”. The Elbow Method is one of the most popular methods to determine this optimal value of k . It consists of running K-Means clustering of the data set with a range of values of k , calculating the sum of squared errors for each k , and plotting them in a line chart. If the chart looks like an arm, the best value of k will be on the “elbow” [15].

With the help of metaheuristics algorithms, it was investigated as a new method in cluster analysis [16]. One of the major methods for this problem is a combination of swarm intelligence algorithms with cluster validity indices as an objective function [17], [18]. There are some criteria, derived from different approaches, for determining the optimal number of classes. We cite the criterion of Xie and Beni [19], which is based on a measure of separability and the compactness of classes. These two notions define criteria for evaluating a classification. Xie and Beni propose to choose the optimal k which minimizes the relationship between separability and compactness.

In [20], researchers have presented a method fuzzy silhouette index on dynamic data to find the optimal number of clusters. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k , the clustering result is unstable. In [21], the greedy algorithm was applied to obtain the number of centers and made the final clustering result features of higher accuracy rate and stability. Shen-Yi *et al.* [22] have initialized K-Means by the initial centers produced by hierarchical clustering algorithm, the clustering results found in terms of FM are not very convincing, and the dataset is from the Chinese text corpus. According to [23], the Fuzzy C-Means is generated to produce initial centers, and particle swarm optimization (PSO) is used to make optimum clusters.

Song *et al.* [24] improved Huang Min’s algorithm [25] to select initial clustering centers focusing on the distance between the samples that have the same maximum density parameters and compare it with the average distance of the dataset. Sherkat *et al.* [26] produced initial centers based on a method called deterministic seeding K-Means (DSKM). The key idea of the proposed method is to select k data points that are distant from each other, and at the same time have a high L1 norm. These data points are used to initialize the K-Means algorithm, the method has produced good results because it takes as parameter (k) the real number of classes, which means that it is sensitive to the number of cluster parameter.

In this paper, we are interested in evaluating the approach described in our previous work [27], which assumed good results in determining the number of clusters by using MSSP and community healthcare network (CHN), but it overlooked the selecting initial clustering centers. In order to evaluate this approach, we consider the nodes found by MSSP as the initial cluster centroids. Thus, we propose a method for the automatic detection of initial cluster centroids, which are the input parameters in several partitioning clustering methods. The proposed algorithm is executed before clustering, which means that it is independent of any clustering method that starts with k centers, it is efficient of large data sets and much optimized in terms of time. It consists to modelize the text document clustering problem as the MSSP, and use CHN to solve the MSSP problem to have document seeds. We have chosen K-Means as a clustering method to test our experimentation. This paper is structured: Section 2 describes the concept of text clustering. Section 3 discusses MSSP and CHN, which are the main components of the method and presents its steps. In section 4, we present the implementation and the experimental results of the KM_MSSP. Section 5 concludes this work.

2. FUNDAMENTAL CONCEPTS OF DOCUMENTS CLUSTERING

A document clustering is a difficult task in text mining, it consists of grouping similar documents. when the topics don’t know in advance. Its goal is to organize a collection of documents according to their topics and help users to have information access, its process consists of the following steps [28]

2.1. Text pre-processing

Each text is represented in the vector space model (VSM) which is an algebraic model for representing text documents as vectors of terms. Then, the stop words list and punctuation marks must be removed. Stemming which is the process of reducing words to their stem or root form is performed, and a pre-processing filter has already been applied to the data to eliminate terms that have a low frequency (count < 3), leading to significant dimensionality reduction without loss of clustering performance.

2.2. Term weighting

To represent quantitatively the term in a document, we use the term frequency-inverse document frequency algorithm (TF-IDF). It takes into account both the relative frequency of a stem in a document and the frequency of the stem within the corpus. A composite weight for each term t_i in each document d_j is w_{ij} , it is given by,

$$w_{ij} = TF_{ij} \times IDF_i$$

2.3. Similarity measure

There are several measures of similarity between documents in the literature. In particular, we find the Euclidean, Manhattan and Cosine distance D that we will be used in our experimentation, it is computed.

$$D(x_i, x_j) = 1 - \cos(x_i, x_j)$$

Where $\cos(x_i, x_j)$ is cosine similarity of two text vectors. Cosine value is 1 when the documents are similar and 0 when they are dissimilar.

2.4. Validity index

A clustering evaluation demands an independent and reliable measure for the assessment and comparison of clustering experiments and results. In this research, we focus on six validation measures: F-measure, purity, entropy, normalized mutual information, Xie-Beni Index and Fukuyama-Sugeno index [29]. The common basis of the indexes is that their computations are all based on a contingency table, which defines the association between two classifications on one same set of individuals E .

3. THE PROPOSED METHOD

In the following, we describe the maximum stable set problem and we introduce the continuous Hopfield network, which are the main components of our approach. Also, the CHN and MSSP problem are combined to find the initial centroids or documents seeds. We have chosen K-Means as a clustering method to test our approach.

3.1. Maximum stable set problem

We denote by $G = (V, E)$ an undirected graph. V is the set of nodes and E is the set of edges. A stable set of a graph G is the subset S of V with the property that each pair of S is not connected via an edge in graph G [30]. The MSSP consists to find the stable set in a graph of maximum cardinality.

The goal in this paper is to consider the data mining applications specifically the clustering of text documents the motivating application of the maximum stable set. The MSSP is a well-known NP-hard problem in combinatorial optimization, which can be formulated as a quadratic 0-1 programming. To solve this latter problem, we use the CHN.

3.2. Continuous hopfield network

At the beginning of 1980, Hopfield and Tank [31] introduce a Hopfield neural network. It has been extensively studied of neural networks and is trained efficiently to solve difficult problems [32] such as pattern recognition, model identification, and optimization. The major advantage of the continuous Hopfield network is in its structure which can be realized on an energy function approach by adding the objective and penalizing the constraints in order to solve some classification and optimization problems. Talavn and Yez [33] showed that when the function reaches a steady state of a differential equation system associated with the CHN, an approximate solution of several optimization problems is obtained. Their results encouraged a number of researchers to apply this network to different problems.

The CHN of size n is a fully connected neural network with n continuous-valued units (or neurons). Following the notation used in [33]: Let T_{ij} be the strength of the connection from neuron i and neuron j . The model assumes symmetrical weights ($T_{ij} = T_{ji}$), in most cases, zero self-coupling terms ($T_{ii} = 0$) and each neuron i has an offset bias i^b . Let u and x be the current state and the output of the neuron i , with $i \in \{1, \dots, n\}$. The system of the CHN is described by the differential equation,

$$\frac{du}{dt} = \frac{u}{\tau} + Tx + i^b$$

where τ is the value of the time constant of the amplifiers, and without loss of generality can be assigned a value of unity. The output function is a hyperbolic tangent of each neuron state. The expression of the energy function associated with the continuous Hopfield network is defined by,

$$E(x) = -\frac{1}{2}x^tTx - (i^b)^t x$$

typically, in the CHN, the energy function is initialized to the objective function of the optimization problem and for each constraint a penalty term is added.

3.3. MSSP and CHN to find initial centroids

We suppose that we have a dataset of n documents. So we want to divide a dataset into groups such that the members of each group are as similar as possible to one another. The overall flow of the algorithm is shown in Figure 1.

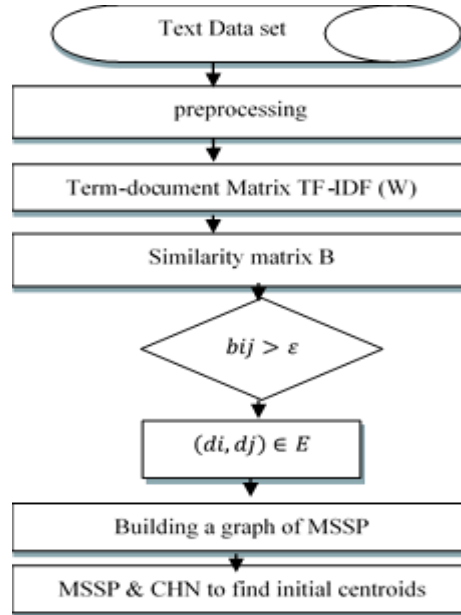


Figure 1. MSSP and CHN to find initial centroids

The specific process of the algorithm is:

- After the preprocessing step, we compute the term-document matrix W .
- Based on the cosine distance between documents we create a similarity matrix B . Therefore the nodes V and edges E of the graph $G(V, E)$ are building.
- Then we represent a text document clustering problem as the max stable set problem (MSSP).
- The documents are considered as nodes, To build the edges we calculate the similarity (cosine distance) between different nodes v_i . If similarity between two documents $b_{ij} > \epsilon$ we have an edge between them.
- After that we use quadratic integer programming formulation QP, to determine the stable set maximal S associated with this graph.

$$(QP) \begin{cases} \text{Min}F(x) = -\sum_{i=1}^n x_i \\ \text{Subject to} \\ x^t C x = 0 \\ x \in \{0,1\}^n \end{cases}$$

$$\text{such as } \forall i \in \{1, \dots, n\}, x_i = \begin{cases} 1 & \text{if } v_i \in S \\ 0 & \text{Otherwise} \end{cases}$$

$$C \text{ is an } n \times n \text{ symmetric matrix defined by: } c_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{Otherwise} \end{cases}$$

- The continuous Hopfield network is implemented to solve the QP problem.
- So the solution is denoted as vector that have values 0 and 1. If the value at position j is 1 that means j^{th} document is selected as centroid otherwise it is not selected.
- Finally, we obtain k documents fully disconnected (document seeds).

3.4. K-Means algorithm initialized by MSSP (KM_MSSP)

K-Means is a commonly used clustering method in text clustering, which serves centroids to represent clusters by minimizing the squared errors. The algorithm begins with a predefined set of centroids (which can be produced either randomly or by means of any other criterion, in this study by MSSP). It

achieves sequential repetitions of the rest of the sample according to the similarity (using cosine distance) with the centroids such as each document is assigned to the cluster with the most similar centroid. Then, the algorithm is iterative processing and adjust the center position, until no reassignment of patterns to new cluster centroids or minimal decrease in square error or the number of iterations has surpassed a threshold. So this produces a separation of the objects into groups from which the metric to be minimized can be calculated. The last step concerns improving the obtained solution found by MSSP and CHN as an input parameter in K-Means as shown in Figure 2, and evaluates the clustering results based on some validity criteria.

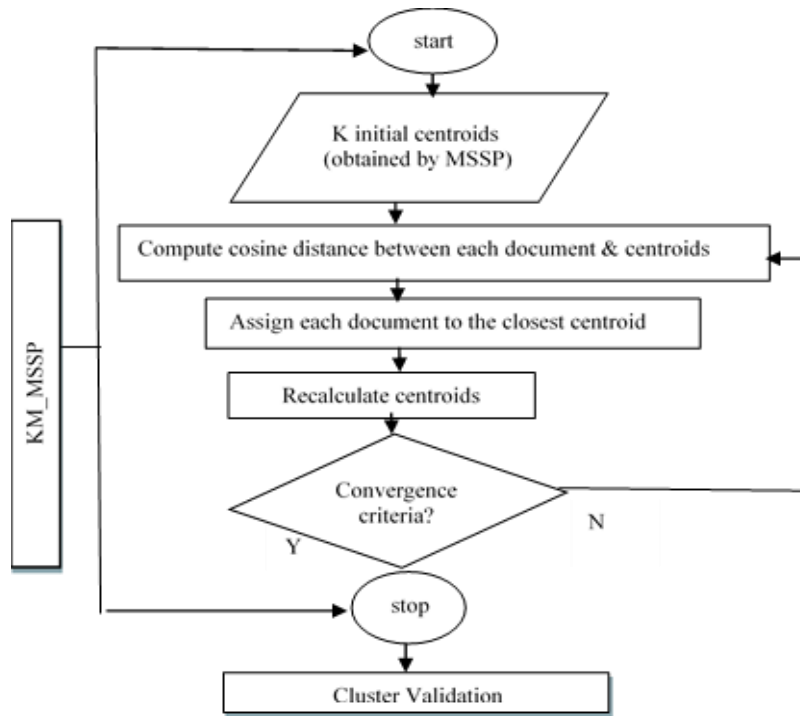


Figure 2. The proposed KM_MSSP algorithm

4. RESULTS AND DISCUSSION

In order to demonstrate the efficiency of our approach, we have affected a series of experiments of instances from the dataset british broadcasting corporation (BBC) and dataset 20NewsGroup. Most of these instances are created by varying the number of classes and the number of documents. In this section we present the dataset description, the results and the comparison of KM which uses random centroids with KM_MSSP which starts by k centroids found by MSSP.

4.1. Data set description

BBC Dataset consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005. The dataset is classified into five natural classes such as business, entertainment, politics, sport, and technology. It is available in [34]. A subset of the data sets is taken and separated into four categories, we use three subsets of the documents from the dataset BBC. The first subset contains 5 topics and 500 documents, the second contains 3 topics and 1328 documents, the third contains four topics and 1715 documents and each document has a single topic label, as shown in Table 1.

BBC sport dataset consists of 737 documents from the BBC Sport website corresponding to sports newsarticles in five topical areas from 2004-2005. The dataset is classified into five natural classes such as athletics, cricket, football, rugby, tennis. It is available in [34]. We use one subset of the documents from the dataset BBC, which contains 3 topics and 372 documents.

20NewsGroup: is a collection of approximately 20,000 newsgroup documents, partitioned in 20 groups. We have selected a subset of this dataset containing total 400 documents from over four categories. The 20 newsgroups collection is a popular data set for experiments in text classification and text clustering.

4.2. Experimental results

Simulative experiments are executed to show the advantages and drawing scientific remarks on the algorithm adaptability of the Improved KM_MSSP. The parameter of similarity ε is determined by several tests and is fixed as ($\varepsilon = 1$). As shown in Table 1, in dataset BBC_2225 we have 2225 documents organized into 5 different classes, in the fourth column (Number of clusters obtained by MSSP), our method gives 6 clusters for BBC_2225 and 5 clusters for BBC_1715 which are very close to the value of the third column (real number of classes). However, in BBC_500, BCC_1328, BBC_Sport1 and BBC_Sport2 the solution is equal to the real number of classes. Hence, we conclude that our method gives for each dataset a very close or an equal result to the real number of classes. In column 5, we observe that the initial centroids proposed by MSSP for each instance of the dataset are very dissimilar and belong to different groups. The execution time is very limited; it varies according to the size of dataset as shown in column 6.

To examine the quality of our approach, a statistical study was represented; this study is based on the calculation operator's performance,

$$Ratio = \frac{\text{number of classes obtained by MSSP}}{\text{real number of classes existing in the literature}}$$

in this context, the ratio minimum is always superior or equal to 1 then the KM_MSSP give the upper bound of real number of class. Moreover, if the ratio minimum is equals to 1, then the KM_MSSP has found the real number of classes existing in the literature.

Table 1. Initial centroids obtained by KM_MSSP out of 20 runs

	Dataset	Number of documents	Real number of classes	Number of clusters obtained	Ratio			Initial centroids obtained		CPU time(sec)
					mode	Mean	min	Clusters	Documents	
BBC_News	BBC_2225	2225	5	6	1.2	1.19	1.2	Sport politics sport entertainment tech business	199.txt 397.txt 324.txt 243.txt 334.txt 280.txt	4.2 s
	BBC_1328	1328	3	3	01	1.1	01	business politics tech	394.txt 397.txt 284.txt	1.2 s
	BBC_500	500	5	5	01	1.1	01	sport business entertainment politics tech	018.txt 080.txt 092.txt 074.txt 039.txt	0.18 s
	BBC_1715	1715	4	5	1.25	1.25	1.25	sport sport politics entertainment tech	191.txt 324.txt 397.txt 154.txt 284.txt	0.06s
BBC_Sport	BBC_Sport1	737	5	5	01	01	01	football athletics football cricket rugby	263.txt 026.txt 147.txt 077.txt 067.txt	0.01s
	BBC_Sport2	372	3	3	01	01	01	rugby cricket athletics	041.txt 104.txt 019.txt	0.002s
20Nes Group	20NG_400	400	4	4	01	01	01	misc.forsale comp.graphics rec.autos rec.motorcycles	76027.txt 38464.txt 103209.txt 104297.txt	0.16s

4.3. Comparison between classic K-Means (KM) and K-Means improved (KM_MSSP)

In order to demonstrate the efficiency of our approach, we compare KM which uses random centroids with KM_MSSP which starts by k centroids found by MSSP, using six validity measures: F-measure, purity, entropy, XieBeni index, Fukuyama Sugeno index and normalized mutual information (NMI). The comparison between KM and KM_MSSP has to be implemented on the same number of clusters

founded by MSSP and under the same environment by measuring the cosine distance between the given element representation and the centroid of the cluster. But the difference is that KM is based on the random choice of centers, whereas KM_MSSP starts with the centroids found by MSSP.

4.3.1. F-measure comparison of KM and the proposed KM_MSSP algorithm

As can be seen from Figure 3, the average F-metric of the clustering algorithm of KM_MSSP is 94% on dataset BBC_1328, 92% on dataset BBC_2225, 86% on BBC_500, 81% on BBC_1715, 68% on BBC_Sport1 and 96% on BBC_Sport2. Which is higher than the average F-metric of the KM clustering of 81% on BBC_1328, 75% on dataset BBC_2225 and 73% on BBC_500. So, we notice that there is an increase of F-Measure values in KM_MSSP compared with those of KM in all available datasets.

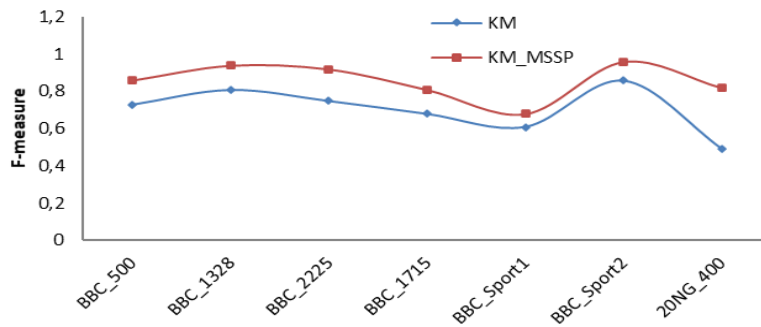


Figure 3. F-measure comparison of KM and the proposed KM_MSSP algorithm

4.3.2. Purity comparison of KM and the proposed KM_MSSP algorithm

A comparison of purity between KM and KM_MSSP is depicted in Figure 4. It is pictured that KM_MSSP outperforms KM in terms of purity in all available datasets. The purity of KM_MSSP is 94% on dataset BBC_1328, 92% on dataset BBC_2225 and 86% on BBC_500, which is higher than the purity of the KM of 81% on BBC_1328, 72% on dataset BBC_2225 and 66% on BBC_500.

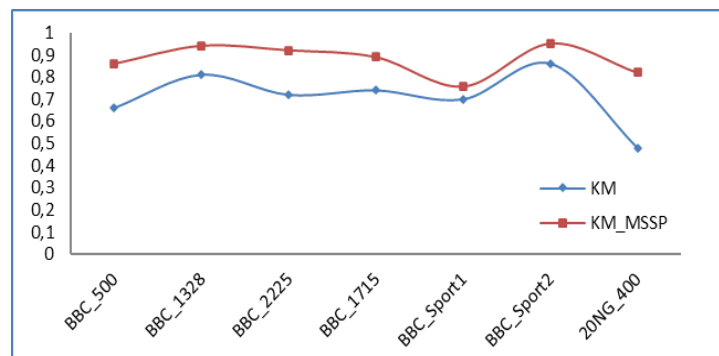


Figure 4. Purity comparison of KM and the proposed KM_MSSP algorithm

4.3.3. Entropy comparison of KM and the proposed KM_MSSP algorithm

Also, we observe that, there are a decline of entropy as shown in Figure 5, values in KM_MSSP compared with those of KM. The entropy of the KM_MSSP is 32%, which is lower than the entropy of the KM clustering of 61% on the BBC_2225 dataset. Thus KM_MSSP outperforms KM in terms of entropy in all available datasets.

4.3.4. NMI comparison of KM and the proposed KM_MSSP algorithm

A comparison of NMI between KM and KM_MSSP is depicted in Figure 6. It is pictured that KM_MSSP outperforms KM in terms of NMI in all available datasets. The NMI of KM_MSSP is 78% on dataset BBC_1328 and dataset BBC_2225, 63% on BBC_500, 85% on BBC_Sport2, which is higher than the NMI of K-Means of 51% on BBC_500 and dataset BBC_1328, 77% on BBC_2225, 64% on BBC_Sport2.

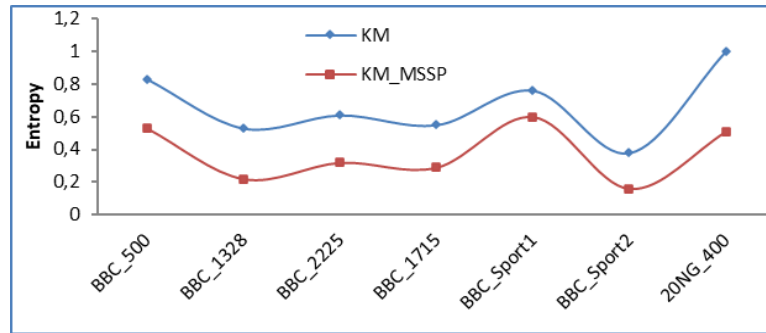


Figure 5. Entropy comparison of KM and the proposed KM_MSSP algorithm

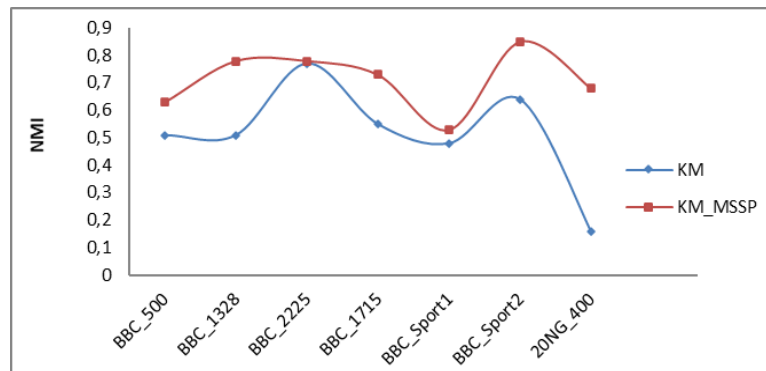


Figure 6. NMI comparison of KM and the proposed KM_MSSP algorithm

4.3.5. Time comparison of KM and the proposed KM_MSSP algorithm

The comparisons of the CPU time on all datasets are shown in Figure 7. According to the graph, KM_MSSP spends a very little time on all datasets than KM. So, our approach improves KM in terms of CPU time.

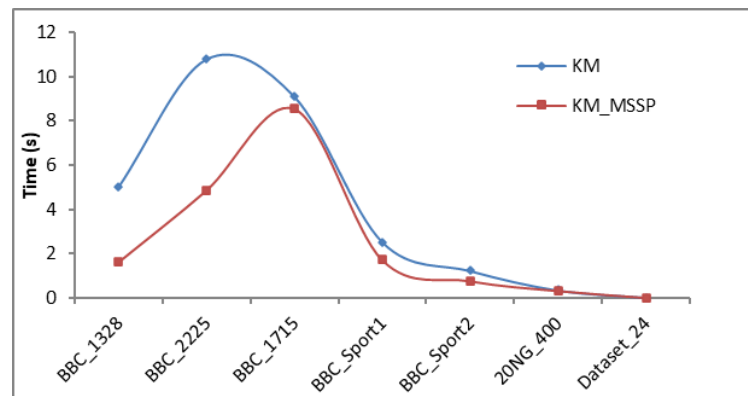


Figure 7. Time comparison of KM and the proposed KM_MSSP algorithm

4.3.6. Comparing Xie-Beni Index (XB) of KM and the proposed KM_MSSP algorithm

A comparison of the Xie-Beni index (XB) between KM and KM_MSSP is depicted in Table 2. It is pictured that KM_MSSP outperforms KM in terms of the XB index in all available datasets. The XB of KM_MSSP is 3% on dataset BBC_1328, 7% on dataset BBC_2225 and 8% on BBC_500, which is lower than the XB of the KM of 45% on BBC_1328, 47% on dataset BBC_2225 and 52% on BBC_500.

Table 2. Comparing Xie-Beni index (XB) of KM and the proposed KM_MSSP algorithm

Dataset	KM	KM_MSSP
BBC_500	0.52	0.08
BBC_1328	0.47	0.03
BBC_2225	0.45	0.07
BBC_1715	0.50	0.07
BBC_Sport1	0.53	0.09
BBC_Sport2	0.52	0.06
20NG_400	1.08	0.17

4.3.7. Comparing Fukuyama-Sugeno index of KM and the proposed KM_MSSP algorithm

From Table 3 we can see that there is a decline of Fukuyama-Sugeno index (FS) values in KM_MSSP compared with those of KM. The FS index of the KM_MSSP is 11%, which is lower than the FS of the KM clustering of 17% on the dataset_24. Thus KM_MSSP outperforms KM in terms of FS index in all available datasets.

Table 3. Comparing Fukuyama-Sugeno index of KM and the proposed KM_MSSP algorithm

Dataset	KM	KM_MSSP
BBC_500	0.87	0.42
BBC_1328	0.55	0.13
BBC_2225	0.75	0.4
BBC_1715	0.38	0.21
BBC_Sport1	0.20	0.16
BBC_Sport2	0.37	0.24
20NG_400	0.53	0.48

It is clearly observed that the results of KM_MSSP are far better for all datasets in terms of entropy, purity, F-measure, NMI, and CPU time. KM_MSSP begins with a number of clusters (number of clusters obtained by MSSP) very close or equal to the real number of classes and starts with initial cluster centroids selected by MSSP (which guarantees an independent set of documents). But KM obtains the initial centers by using a random method which gives unstable clustering results and too many iterations, it affects the quality of clustering and costs a great amount of time during the process of clustering.

4.4. Comparison between KM MSSP and other deterministic method

In order to demonstrate the effectiveness of our approach, we compare our approach with DSKM [26] in terms of Normalized Mutual Information NMI on dataset BBC_2225. From Table 4, the comparison between KM_MSSP and DSKM was done on the same number of clusters founded by MSSP (6 for Dataset BBC_2225 (see Table 2)). The NMI of Ké2& is greater than the DSKM algorithm.

Table 4. Comparing clustering NMI score between KM_MSSP and DSKM on dataset BBC_2225 and number of clusters equal to number founded by MSSP (k=6)

Methods	NMI
DSKM	0.681
KM_MSSP	0.78

5. CONCLUSION

This research aims to develop a method for determining automatically both, number of clusters and initial cluster centroids which are the background knowledge in classic K-Means or other methods of clustering. To achieve this goal and obtain the highest quality of clustering, the MSSP is generalized to apply to the clustering of text documents using CHN. The method is independent of any clustering method that starts with k centroids. Experimental results show that our method can effectively find the optimal number of clusters, find a correct set of centroids, obtain better clustering results in a short time, and also a large number of documents can be easily handled. To demonstrate the efficiency of our approach, we compare classic K-Means which uses random centroids with KM_MSSP which starts by K centroids found by MSSP, using six validity measures: F-measure, purity, entropy, XB, FS and NMI.

Thus, KM_MSSP outperforms KM in all available datasets, for example, in dataset BBC_1328 the purity and the F-measure are 94%, NMI is 78%, entropy is 22% and time is 1.63s which is higher than K-Means of the purity and the F-measure are 81%, NMI is 51%, entropy is 53% and time is 5.01s. The field of

MSSP is still open to many challenges that provide future scope for improvement in the document-clustering problem. The future work includes: (i) apply our approach to other areas, (ii) Improve CHN by combining it with a metaheuristic to obtain a global minimum, and (iii) Use a deep learning concept to improve the pre-processing step.




REFERENCES

- [1] J. Han, J. Pei, and M. Kamber, "Data Mining Concepts and Techniques," Morgan Kaufman Publishers, 3rd ed, USA, 2012.
- [2] D. Larose, "Discovering knowledge in data: an introduction to data mining," John Wiley & Sons, Inc., Hoboken, New Jersey, 2014.
- [3] M. Dumont, P.-A. Reninger, A. Pryet, G. Martelet, B. Aunay, and J.-L. Join, "Agglomerative hierarchical clustering of airborne electromagnetic data for multi-scale geological studies," *Journal of Applied Geophysics*, vol. 157, pp. 1-9, Oct. 2018, doi: 10.1016/j.jappgeo.2018.06.020.
- [4] S. Bandyopadhyay and K. R. Varadarajan, "On variants of k-means clustering," *arXiv preprint arXiv:1512.02985*, 2015.
- [5] Md. A. B. Siddique, R. B. Arif, M. M. R. Khan, and Z. Ashrafi, "Implementation of Fuzzy C-Means and Possibilistic C-Means Clustering Algorithms, Cluster Tendency Analysis and Cluster Validation," *Preprints* 2018, doi: 10.20944/preprints201811.0581.v1.
- [6] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert SystAppl*, vol. 36, no. 2, pp. 3336-3341, 2009, doi: 10.1016/j.eswa.2008.01.039.
- [7] L. Kaufman and P. J. Rousseeuw, "Finding groups in data: An introduction to cluster analysis," Wiley, Hoboken, vol. 344, 2008, doi: 10.1002/9780470316801.
- [8] R. T. Ng and J. Han, "Clarans: a method for clustering objects for spatial data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 5, pp. 1003-1016, Sep. 2002, doi: 10.1109/TKDE.2002.1033770.
- [9] J. Wu, "Advances in K-Means Clustering: A Data Mining Thinking," *Springer Theses*, Springer Berlin Heidelberg, 2012.
- [10] W. Ashour and C. Fyfe, "Improving Bregman K-Means," *International Journal of Data Mining, Modelling and Management*, vol. 6, no. 1, pp. 65-82, Jan. 2014, doi: 10.1504/IJDM.2014.059981.
- [11] L. I. Kuncheva and J. C. Bezdek, "Selection of cluster prototypes from data by a genetic algorithm," in: *Proc. 5th European Congress on Intelligent Techniques and Soft Computing (EUFIT)*, Aachen, Germany, vol. 18, 1997, pp. 1683-1688.
- [12] J. C. Bezdek and R. J. Hathaway, "Optimization of fuzzy clustering criteria using genetic algorithms," in *Proc. First IEEE Conf. Evolutionary Computation, Piscataway, NJ: IEEE Press*, 1994, pp. 589-594, doi: 10.1109/ICEC.1994.349993.
- [13] B. Kövesi, J.-M. Boucher, and S. Saoudi, "Stochastic K-means algorithm for vector quantization," *Pattern Recognit. Lett.*, vol. 22, no. 6-7, pp. 603-610, 2001, doi: 10.1016/S0167-8655(01)00021-6.
- [14] M. Sarkar, B. Yegnanarayana, and D. Khemani, "A clustering algorithm using an evolutionary-based approach," *Pattern Recognition Letters*, vol. 18, no. 10, pp. 975-986, 1997, doi: 10.1016/S0167-8655(97)00122-0.
- [15] D. J. Ketchen and C. I. Shook, "The application of cluster analysis in strategic management research: an analysis and critique," *Strategic Management Journal*, vol. 17, no. 6, pp. 441-458, Jun 1996, doi: 10.1002/(sici)1097-0266(199606)17:6<441::aid-smj819>3.0.co;2-g.
- [16] A. J.-Garcia and W. G.-Flores, "Automatic clustering using nature-inspired metaheuristics: A survey," *Applied Soft Computing*, vol. 41, pp. 192-213, Apr. 2016, doi: 10.1016/j.asoc.2015.12.001.
- [17] S. Das, A. Abraham, and A. Konar, "Spatial information-based image segmentation using a modified particle swarm optimization algorithm," *ISDA '06: Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications*, Oct. vol. 2, 2006, pp. 438-444, doi: 10.1109/ISDA.2006.253877.
- [18] R. Kuo, Y. Syu, Z.-Y. Chen, and F.-C. Tien, "Integration of particle swarm optimization and genetic algorithm for dynamic clustering," *Information Sciences*, vol. 195, pp. 124-140, Jul. 2012, doi: 10.1016/j.ins.2012.01.021.
- [19] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841-847, Aug. 1991, doi: 10.1109/34.85677.
- [20] C. Subbalakshmi, G. Krishna, K. Rao, and P. Rao, "A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set," *Procedia Computer Science*, 2015, vol. 46, pp. 346-353, doi: 10.1016/j.procs.2015.02.030.
- [21] X. Tong, F. Meng, and Z. Wang, "Optimization to K-Means initial cluster centers," *Computer Engineering and Design*, vol. 32, no. 8, pp. 2721-2723, 2011.
- [22] Q. S.-Yi, L. H.-Hui, and L. D.-Yi, "Research and Application of Improved K-Means Algorithm in Text Clustering," *DEStech Transactions on Computer Science and Engineering (pcmm)*, Jun. 2018, doi: 10.12783/dtce/pcmm2018/23653.
- [23] S. Karol and V. Mangat, "Evaluation of text document clustering approach based on particle swarm optimization," *Proceedings of Central European Journal of Computer Science*, vol. 3, no. 2, Jun. 2013, pp. 69-90, doi: 10.2478/s13537-013-0104-2.
- [24] J. Song, X. Li, and Y. Liu, "An Optimized K-Means Algorithm for Selecting Initial Clustering Centers," *International Journal of Security and Its Applications*, vol. 9, no. 10, pp. 177-186, 2015, doi: 10.14257/ijasia.2015.9.10.16.
- [25] M. Huang, Z. S. He, X. L. Xing, and Y. Chen, "New K-Means clustering center select algorithm," *Computer Engineering and Applications*, vol. 47, no. 35, pp. 132-134, 2011.
- [26] E. Sherkat, J. Velcin, and E. E. Miliotis, "Fast and Simple Deterministic Seeding of K-Means for Text Document Clustering," *9th Conference and Labs of the Evaluation Forum (CLEF), Proceedings*, Sep. 2018, pp. 76-88, doi: 10.1007/978-3-319-98932-7_7.
- [27] A. Karim, C. Loqman, and J. Boumhidi, "Determining the Number of Clusters using Neural Network and Max Stable Set Problem," *The 1st. Int. Conf. On Intelligent Computing in Data Sciences. Procedia Computer Science*, 2018, vol. 127, pp. 16-25, doi: 10.1016/j.procs.2018.01.093.
- [28] M. R. Aliguliyev, "Clustering of document collection, A weighting approach," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7904-7916, 2009, doi: 10.1016/j.eswa.2008.11.017.
- [29] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," in: *Proc. 5th Fuzzy Syst. Symp.*, Jan. 1989, pp. 247-250.
- [30] Y. Chung and M. Demange, "The 0-1 inverse maximum stable set problem," *Discrete Applied Mathematics*, vol. 156, no. 13, pp. 2501-2516, 2008, doi: 10.1016/j.dam.2008.03.015.
- [31] J. Hopfield and D. Tank, "Neural computation of decisions in optimization problems," *Biological Cybernetics*, vol. 52, no. 3, pp. 1-25, 1985, doi: 10.1007/BF00339943.




- [32] Y. Hami and C. Loqman, "Quadratic Convex Reformulation for Solving Task Assignment Problem with Continuous Hopfield Network," *IJCA*, vol. 20, no. 5, Dec. 2021, doi: S1469026821500243.
- [33] P. Talavn and J. Yez, "The generalized quadratic knapsack problem. A neuronal network approach," *Neural Networks*, vol. 19, no. 14, pp. 416-428, 2006, doi: 10.1016/j.neunet.2005.10.008.
- [34] D. Greene and P. Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," *Proc. ICML 2006*, pp. 377-384, doi: 10.1145/1143844.1143892.

BIOGRAPHIES OF AUTHORS






Awatif Karim    received the diploma of engineer in networks and computer systems in Cadi Ayyad University, Morocco, 2010. Now she is a phd.D. student in Sidi Mohamed Ben Abdellah University, Maroc, 2014. Her research interests include artificial intelligence and data mining. She can be contacted at email: awatif.karim@usmba.ac.ma.






Chakir Loqman    is a Professor of Computer Science at the Faculty of Sciences, Fez Morocco. He received the PhD in Computer Science from The University of Sidi Mohamed ben Abdellah. His research interests include image processing, artificial intelligence and optimization. He can be contacted at email: loqman.chakir@usmba.ac.ma.



Youssef Hami    is a Professor in National School of Applied Sciences, Tangier Morocco. He received Ph.D. degrees in operational research and computer science from the University of Sidi Mohamed ben Abdellah in 2014. His research interests include operational research, computer science and optimization. He can be contacted at email: yhami@uae.ac.ma.



Jaouad Boumhidi    is a Professor of Computer Science at the Faculty of Sciences, Fez Morocco. He received the PhD in Computer Science from The University of Sidi Mohamed ben Abdellah in 2005. His research interests are in machine learning, deep learning and intelligent transportation systems. He can be contacted at email: jaouad.boumhidi@usmba.ac.ma.