Proceedings of the New Trends in Information Technology (NTIT-2017).
The University of Jordan, Amman, Jordan. 25-27 April 2017.

β-Hill Climbing Technique for the Text Document Clustering

Laith Mohammad Abualigah*, Ahmad Mohammad Sawaie‡, Ahamad Tajudin Khader†,
Hasan Rashaideh§, Mohammed Azmi Al-Betar¶ and Mohammad Shehab‖
*†‖School of Computer Sciences, Universiti Sains Malaysia (USM), Pulau Pinang, Malaysia 11800.
*Email: lmqa15_com072@student.usm.my
†Email: tajudin@cs.usm.my
‖Email: moh.shehab12@gmail.com
‡§Department of Computer Science, Al-Huson University College, Al-Balqa Applied University, Salt, Jordan.
‡Email: a.sawaie86@gmail.com
§Email: eashaideh@bau.edu.jo
¶Department of Information Technology, Al-Huson University College, Al-Balqa Applied University, Irbid, Jordan.
¶Email: mohbetar@bau.edu.jo

Abstract

Regarding the increasing volume of document information (text) on Internet network pages, recent applications, and so on, the dealing with this knowledge has become incredibly complex because of the size. The text clustering is a proper technique used to arrange a tremendous amount of text information by classifying into a subset of clusters. In this paper, we present a novel local search method, namely, β-hill climbing technique to solve the text document clustering problem. The primary innovation of β-hill climbing technique is β. It has been introduced to perform a balance between local and global search. Local search (exploitation) methods are success applied to the text document clustering problem as the k-mean. Experiments conducted on five benchmark text datasets taken randomly from "Dmoz-Business" dataset with varying characteristics. The results prove that the proposed β-hill climbing achieved better results in comparison with the original hill climbing technique measured by F-measure, precision, recall, and accuracy. The results show that the proposed (β-hill climbing) obtained better results in comparison with the other original technique by tuning the tuning the parameter of the β-hill claiming.

Keyword: Text Document Clustering, β-Hill Climbing, Local Exploitation.

I. Introduction

Nowadays, the main important point on the level of the domain of the text analysis is how to represent a tremendous amount of text information in an easy form. As well, all web pages and advanced applications become hold an enormous amount of text information that the users need it to be tidy. Text clustering is one of the most efficient unsupervised methods used to solve the problem of partition many text documents into a subset of predetermined the clusters numbers. This method is much use in the area of text mining, data clustering, detection and disease clustering, open source clustering software, clustering the results of the search engine, time series clustering and wireless sensor clustering to perform comprehensive analysis for all the information [1]–[3].

Mostly, challenges that have a greater importance in all domains of text mining area and especially in the domain of text clustering are text document holds many uninformative features. These kinds of features affect achievement and performance of the text clustering process, where the uninformative features are unnecessary, unrelated, and noisy features [4], [5]. Hence, the clustering method needs a powerful technique to improve the clustering process through portion the smiler document together in the same clusters.

Recently, some researchers have proposed many text clustering methods to solve difficulties that face the text clustering process. Local search is one of the robust techniques easily used to generate a subset of document clusters. By using this technique, the information view became easier and the user time became less as well. The unsupervised text clustering technique performs its processes apart from knowing the given class label of documents [4], [6], [7].

In general, text clustering defined as an optimization problem in terms of maximizing or minimizing the performance of the clustering algorithm. In terms of minimizing, find the minimize distance value between the document with clusters centroids. However, In terms of maximizing, find the maximize similarity value between the document with clusters centroids. The text clustering has been success used in main domains include ontology-based text clustering, text mining, feature selection, automated clustering of newspapers, and text categorization [1], [6], [7].

Regarding the Vector Space Model (VSM), it is a popular pattern used in the area of the text mining especially in text clustering and text feature selection to facilitate the analysis process [8]. This pattern represents the component of each text document as a row (vector) of terms frequency; each term frequency represented as one position (dimension space). Therefore, the performance of the k-mean text clustering algorithm affected positively if the number of represented feature is small [2], [3].

β-hill climbing is an optimization technique introduced in 2016 by Mohammed Azmi Al-Betar at Al-Huson University

[9]. It can produce a search path in the available search space until moving to the local optimal solution. This technique has several extensions to overcome such problem such as Tabu Search and Simulated Annealing. One of the main characteristics of the β-hill climbing is that leads to escape stuck in local optima.

In this paper, the authors proposed a novel local search method, namely, β-hell climbing technique for the text document clustering problem (β-hill climbing). The primary idea in this method is applying the text clustering using β-hell climbing technique to find more related and coherent clusters. The proposed method seeks to make the performance of the text clustering higher in terms of clusters accuracy. Experiments results were conducted on five various text benchmark datasets taken randomly from "Dmoz-Business" dataset with varying characteristics to test the proposed method. The results prove that the proposed β-hill climbing for text clustering obtained better results in comparison with the original hill climbing (H-TC) measured by F-measure, precision, recall, and accuracy. As well, it improved the text clustering algorithm by dealing with a large number of clusters.

The remainder of this paper prepared as follows: Section 2 presents the related works in the domain of the text clustering. Section 3 reviews the proposed β-hill climbing technique for the text document clustering problem. Results and discussion are given in Section 4. Finally, Section 5 provided the conclusion.

II. Related Works

This section presents the most related works in the domain of text document clustering and β-hill climbing.

Recently, the text document clustering is a useful technique for partitioning an extensive amount of text information into associated clusters. Therefore, one of the fundamental problems that affect the clustering method is the appearance many uninformative and sparse features in the texts. Unsupervised feature selection (FS) is an essential technique for eliminating possible uninformative features to support the text clustering method. In this paper, the harmony search (HS) algorithm proposed, namely, feature selection based on harmony search algorithm for text clustering (FSHSTC), to solve the text feature selection problem [5]. At of the end, FSHSTC is done to enhance the text clustering by getting a new subset of informational text features. Experiments were carried out on four text benchmark datasets. The results prove that the proposed method is enhanced the effectiveness and performance of the text clustering algorithm (i.e., k-mean algorithm) in terms of F-measure and accuracy.

The metaheuristic optimization algorithms are actively employed to solve several complex optimization problems. In this paper, the genetic algorithm proposed, namely, feature selection based on genetic algorithm for text clustering (FSGATC) to solve the text feature selection problem [10]. At of the end, FSGATC is prepared to enhance the text clustering by creating a new subset of informational text features. Experiments were carried out on four text benchmark datasets and compared with another well-known algorithm in the same domain. The results prove that the proposed method (FSGATC) got better results according to the effectiveness and performance of the text clustering algorithm (i.e., k-mean algorithm) in comparison with k-mean and HS algorithms regarding F-measure and accuracy.

Due to the tremendous growth of web pages, and modern applications, text clustering has developed as a vital task to deal with many text documents. Unusual, web pages are simply browsed and tidily shown via applying the clustering method in order to distribution the documents into a subset of similar clusters. In this paper, the authors proposed two novel text document clustering methods based on krill herd algorithm to develop and enhance the text documents clustering [4]. In the first method, the basic krill herd algorithm utilizes all genetic operators. While in the second method, the basic krill herd algorithm utilizes without all genetic operators. Experiments conducted on four standard text datasets. The results revealed that the proposed krill herd algorithms overcome the k-mean text clustering algorithm in term of the clusters accuracy (i.e., purity and entropy).

One of the popular unsupervised text mining tools is text documents clustering. In text clustering algorithm, the correct decision for any document distribution is made using an objective function (i.e., similarity measurements or distance measurements). Text clustering algorithms work very poorly when the form of the objective function is not valid and complete. Hence, the authors proposed multi-objective function, namely, aggregation Euclidian distance and Cosine similarity measurements for adjusting the text clustering distribution [2]. The multi-objective function has been by combined two evaluating functions which rise as an efficient alternative in numerous clustering situations. In particular, the multi-objective function is not a popular in the domain of the text clustering. The authors said that the multi-objective function is a core problem that reduces the performance of the k-mean text clustering algorithm. Experiments conducted on seven standard benchmark text datasets, which is common in the domain of the text clustering. The results revealed that the proposed multi-objective function outperforms the other measure standalone in term of the achievement of the k-mean text clustering. The results evaluated by using two well-known clustering measures, namely, accuracy and F-measure.

A new version of hill climbing technique method, namely, β-hill climbing, has been proposed [9]. The authors add a new stochastic operator in hill climbing called β-operator to establish a balance between the exploitation (i.e., intensification) and exploration (i.e., diversification) during the search. Experiments conducted on IEEE-CEC2005 global optimization functions. The results reveal that the proposed β-hill climbing obtained better results to the hill climbing providing reliable results when it compares with other comparative methods using the same IEEE-CEC2005 global optimization functions.

III. The proposed text document clustering method

A. Text document preprocessing

In this section, the proposed method starting with the original text dataset to convert it in the form of numerical matrx. This matrix needs to present each feature or term by its term frequency (TF) [3], [5], [10], [11]. Thus, to turn the dataset from text to digital (i.e., term frequency) from, three preprocessing steps were applied for the text document representation. These steps provided in the following subsection:

1) Tokenization: In the first stage, a process of cutting a continues letters of documents into token (i.e., words, phrases, symbols, and important elements). On the other hand, a process of removing an empty sequence by taking each token from the first letter to the last letter. This process saved the memory [6], [10].

2) Removal stops-words: In the second stage, a process removing all common words (i.e, "a", "against", "about", "am", "all", "above", "after", "and", "again", "any", "an" and so on)[1] [5]. Available list of stop-words contains 571 stop-words [5], [12].

3) Stemming: In the third stages, a process of dismantling some of the related words in terms of structure and meaning to be in the same form. That means every some of the related words will represent by one root (i.e., feature). Usually, This process was done by using the Porter Stemmer[2]. Porter Stemmer gets rid of some of the parties such as eliminating the prefixes and suffixes od each term (i.e., "ed", "ly", "ing", and so on). For example, "connection", "connective", "connective", "connections", "connecting", and "connected" all these words or terms have the common root connect. This root after these some of preprocessing will call feature [4], [6].

B. Calculating the term weighting (TF-IDF)

After finished the third preprocess step, we move to calculate the term weighting for each feature. In the area of text mining particularly in the domain of the text clustering, the term weighting scheme, namely, term frequency-inverse document frequency (TF-IDF), used to give a weighted score by Eq. (2) for each term (feature) [3], [5], [11]. In this paper, this scheme is used to distinguish between the document terms (classification) as an objective function.

TFIDF is a standard scheme used to find the weighting of document terms. This scheme based on the term frequency and inverse document frequency for representing each document. Each document d in the dataset D is represented as a row (vector) of terms weighting [2], [6], [12]. Eq. (1) shows the vector of document number i (d_i), document i represents as vector of the length t, t is number of all unique terms. $w_{i,2}$ denote to the weight value (score) of the feature 2 in the document number i.

$$d_i = (w_{i,1}, w_{i,2}, ..., w_{i,j}, ..., w_{i,t}) \tag{1}$$

[1]List of Stop-words. Website at http://www.unine.ch/Info/clef/

[2]Porter stemmer. Website at http://tartarus.org/martin/PorterStemmer/

$$w_{i,j} = tf(i,j) * idf(i,j) = tf(i,j) * log\left(\frac{n}{df_{(j)}}\right), \tag{2}$$

where $w_{i,j}$ is the weight value of the term number *j* in the document number *i*. *tf(i,j)* is the frequency of term number *j* in the document number *i*. *idf(i, j)* is a factor utilized to enhance the term based on the number of occurrence in each document. *n* is the number of documents in dataset, and *df(j)* is the number of documents that hold term *j*. The following matrix shows the documents using format of the vector space model [4], [5], [8].

$$VSM = \begin{bmatrix} w_{1,1} & \cdots & w_{1,(t-1)} & w_{1,t} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \vdots & \ddots & \vdots & \vdots \\ w_{(n-1),1} & \cdots & \cdots & w_{(n-1),t} \\ w_{n,1} & \cdots & w_{n,(t-1)} & w_{n,t} \end{bmatrix} \tag{3}$$

C. Text clustering based on β-hill climbing technique

In this section, we explain the proposed text clsutering method based on β-hill climbing.

1) Mathematical model of the text clustering problem: Unsupervised text clustering problem formulated as an optimization problem in order to generate a new optimal subset of clusters. The proposed method deal with each document alone by generating 1000 iterations to produce a new subset of documents clusters [3], [4], [4], [6].

Definition 1 Let *D* a set of documents, where d_i presents the document number i as a vector of term weighting (see Eq. (4). Where, $w_{i,j}$ denote the weight value of term j in the document i, *t* is the number of all features in dataset, *i* is the document number and *j* is the feature number.

$$d_i = (w_{i,1}, w_{i,2}, ..., w_{i,j}, ..., w_{i,t}) \tag{4}$$

The generated subset of clusters by the proposed method (β-hill climbing) defined as the follows definition:

Definition 2 let D is a collection of text documents, D portion into K is the number of all clusters. Eq. (5) shows documents representation in the dataset. *n* is the number of all documents, where $d_i \in \{1, n\}$, $(i = 1, 2, ..., n)$. Eq. (6) shows clusters centroids representation, where each cluster has a one centroid such as c_k is the centroid of the cluster k.

$$D = (d_1, d_2, ..., d_i, ..., d_n) \tag{5}$$

$$C = (c_1, c_2, ..., c_k, ..., c_K) \tag{6}$$

2) Adapt β-hill climbing technique for the text clustering problem: In this section, adapted the β-hill climbing technique for the text clustering problem through a new method. This paper determined the solution characteristic of the β-hill climbing, then adjust the coding for the clustering problem, employed β to be improving operator, and finally, choose a suitable fitness function to evaluate the clusters solutions. As well, term weighting took by (TF-IDF), it is the objective function to evaluate each position (document). The primary motivation in this paper still obtaining optimal subset of documents clusters.

3) β-hill climbing technique search space: The optimization techniques are starting with a random initial solution or solutions and trying to increase the fitness function of them to reach the optimal global solution. Each position presents a document in the dataset ($d_i \in D$).

Definition 3 Let D is a collection of documents contain n=100 documents, where n is number of all documents. Hence, the search space of each solution equal K, where $d \in \{1, K\}$ [3], [4].

4) Solution representation of the feature selection problem: In the β-hill climbing technique for text clustering problem, the solution represents a subset of clusters by a number in range 1....K. The solution of the β-hill climbing represents as vector; each position represents one document to which cluster belong. The i_{th} position in the solution represents the situation of the i_{th} document [4].

Text clustering problem is applied based on β-hill climbing technique, which begins with the random initial solution and improves its solution by reaching a globally optimal solution. Each single document in the dataset reflects as a dimension of the search space. Fig. 1 shows the solution of the text clustering problem [4].

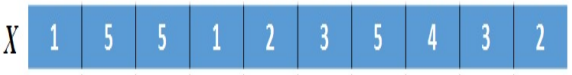

Fig. 1. The solution presentation of the text document clustering problem.

X represents the β-hill climbing solution for solving the text clustering problem. In case the value of position number i is equal *5*, deduce that the i_{th} document belongs to the cluster number 5. Fig. 1 shows ten documents distributed into five clusters. The first cluster hold two documents (i.e., 1 and 4). The second cluster hold two documents (i.e., 5 and 10). The third cluster hold two documents (i.e., 6 and 9). The fourth cluster hold one document (i.e., 8). Lastly, the fifth cluster hold three documents (i.e., 2, 3, and 7).

5) Distance measure: Euclidean is a standard distance measure used in the domain of the text clustering to compute the dissimilarity (distance) score between each document with clusters centroids. This paper uses the Euclidean distance measure as the objective functions by Eq. (7). Normally, distance values are between (0, 1), although it is unlike the cosine similarity measure. Where, if the distance value close to 0, that means it is the best value. If the distance value close to 1, that means it is the worst value, on the other hand, the document not close to the current cluster centroid [2], [7], [12].

$$Dis(d_4, c_2) = \left(\sum_{j=1}^{t} |w_{d,j} - w_{c,j}|^2 \right)^{1/2}, \quad (7)$$

Eq. (7) presents the distance between the document number 4 and the cluster centroid number 2. Where, $w_{d,j}$ is the weight of term *j* in document number 4, and $w_{c,j}$ is the weight of term *j* in cluster centroid number 2.

6) Fitness function: The fitness function (F) is a class of the evaluation measure employed to evaluate the solution. Iteratively, the fitness function of each solution calculated. Finally, the solution, which has a greater fitness value is the optimal solution. The proposed method used the average distance of documents to the cluster centroid (ADDC) equation as the fitness function in β-hill climbing for the text clustering problem. Each position has a distance value as the objective function. Eq. (8) by the average distance of documents to the cluster centroid (ADDC) (i.e., fitness function) [4], [7], [12].

$$ADDC = \left[\frac{\sum_{j=1}^{K} (\frac{\sum_{i=1}^{n} Dis(d_j, c_i)}{r_j})}{K} \right], \quad (8)$$

.

where K is the number of clusters, r_j is the number of documents that belong to the cluster number j, and $Dis(d_i, c_i)$ is the distance measure between the document number i and the cluster centroid number j.

Definition 4 Let the solution has a set of K centroid $C = (c_1, c_2, ..., c_k, c_K)$, where c_k is the centroid of cluster number k, which represents as a vector of terms weighting $c_k = (c_{k,1}, c_{k,2}, ..., c_{k,j}, ..., c_{k,t})$ and is computed by Eq. (9).

$$c_j = \frac{1}{n_i} \sum_{d_i \in c_j} d_i \quad (9)$$

.

where d_i shows that document i belongs to the i_{th} centroid, and n_i represents documents that belong to cluster i [2]. The fitness function for the solution in theβ-hill climbing technique is determined by the average distance of documents to the cluster centroid (ADDC) as represented by Eq. (8).

7) β-hill climbing technique: The β-hill climbing (see Algorithm 1) begins with an absolute solution $X = (x_1, x_2, ..., x_i, ..., x_n)$. It iteratively produces a new solution $X^{'} = (x_1^{'}, x_2^{'}, ..., x_i^{'}, ..., x_n^{'})$ using two operators: β operator and neighborhood navigation. In the neighborhood navigation stage, the function improve the solution by using the acceptance rule where iteratively a random neighboring solution of the solution X is adopted as Eq. (10) [9].

$$x_i = 1 + rand \mod K, \quad (10)$$

In β operator stage, the positions of the new solution are selected values by one way of these two way: (i) according

Algorithm 1 Pseudo-code of the β-hill climbing technique for feature selection problem

Input: A collection of documents D.
Output: Generate a new subset of documents clusters K.
Termination criteria
$X^{'}$ = improve (X) by neighborhood navigation.
for $i = 1, ..., n$ **do** ▷ Note, n is the number of documents
  if $rand \leq \beta$ **then**
    $x^{'}$ = 1+ rand $\mod K$,
  end if
end for
if F($X^{'}$)≤F(X) **then** ▷ Note, F is the ADDC
  X=$X^{'}$
end if
Assign the final solution as a subset of documents clutters.
End

to the current values of the current solution (ii) randomly from possible search space (i.e., binary). This way based on a probability of β where $\beta \in$ [0, 1] as Eq. (11) [9]. β value is fixed (0.1).

$$x_i^{'} \leftarrow \begin{cases} x_p & \text{if } rand \leq \beta \\ x_i & \text{otherwise} \end{cases} \tag{11}$$

where $x_p \in X$ is the possible region for the decision variable x_i and $rand$ presents a random number either one or zero.

IV. Evaluation measures of clusters

The comparative evaluations are done using four evaluation measures: accuracy (Ac), precision (P), recall (R) and f-measure (F). These measures are standard criteria utilized in this domain to evaluate the clusters precisely [2], [4], [5], [10].

F-measure (F) measurement is a standard measure used to gauge the percentage of the truly distributed document in each cluster. Two measures employed to find the F-measure value: precision (P) and recall (R) [6], [7].

$$P(i,j) = \frac{n_{i,j}}{n_j}, \tag{12}$$

$$R(i,j) = \frac{n_{i,j}}{n_i}, \tag{13}$$

where $n_{i,j}$ is the number of documents of class *i* in cluster number *j*, n_j is the number of documents of cluster *j* and n_i is the number of documents of class *i*.

$$F(i,j) = \frac{2 \times P(i,j) \times R(i,j)}{P(i,j) + R(i,j)}, \tag{14}$$

where *P(i,j)* is the precision of class *i* in cluster number *j*, *R(i,j)* is the recall of class *i* in cluster number *j* and F-measure value for all clusters calculated by find the average F-measure value of all clusters.

The accuracy (AC) measurement is one of the external measurements that used to compute the percentage of truly assigned documents to the clusters. Accuracy value calculated by Eq. (15) [2].

$$Ac = \frac{1}{n} \sum_{i=1}^{K} P(i,j) \tag{15}$$

Where, $P(i,j)$ is the precision value for class *i* in cluster *j*, *n* is the number of all documents in each cluster, *K* is the number of all clusters.

V. Experimental Results

We applied the proposed methods using MATLAB (R), version 8.3.0.532, 64-bit, and glnxa64. Text clustering algorithms (i.e., versions of hill climbing technique) run 20 times, each run 1000 iterations.

Table I displays five text datasets taken randomly from "Dmoz-Business" dataset, which applied to test and compared the performance of the versions of hill climbing technique. Datasets are available at http://http://sites.labic.icmc.usp.br/, by digital (numerical) form.

TABLE I
Characteristics of the text datasets

Datasets	Number of Documents (d)	Number of Terms (t)	Number of Clusters (K)
DS1	200	1012	3
DS2	240	1161	4
DS3	270	1256	5
DS4	400	1472	8
DS4	500	1779	10

A. Tuning parameter of the β-hill climbing

A set of experiments is given to evaluate the proposed algorithm or text clustering problem, where the beta parameter of -hill climbing with different settings are applied to examine the results of -hill climbing during the search process. The primary goal of these experiments is to obtain the optimal parameter value of -hill climbing for the text clustering problem that produces a balance between the exploration and exploitation technique.

In order to study the effectiveness of -hill climbing, 10 scenarios, each one with different (β) value as shown in table II. Each experimental is run ten runs using one dataset as a tester (i.e., DS1) with 1000 iteration.

The performance of the β values in terms of the evaluation measures is shown in table III and IV. It is clear that the proposed β-hill climbing obtained the better (best) results when the β value was (0.2). The experimental number 3 (i.e., No.3) obtained 3 best results in terms of the evaluation measures (F-measure, precision, and recall). While, the experimental number No.2 obtained one nest results among the evaluation measure (i.e., accuracy). We conclude that the optimal β is 0.2, which is archived by parameter No.2 (see table III). The next experiments will done by using the optimal β value (0.2) for the text clustering problem. This value obtained the bigger benefits from the extension version of hill climbing.

TABLE II
β VALUES FOR THE PARAMETER SITTING

Scenario No.	β Values
1	0.0
2	0.1
3	0.2
4	0.3
5	0.4
6	0.5
7	0.6
8	0.7
9	0.8
10	0.9

TABLE III
THE PERFORMANCE OF THE β VALUES IN TERMS OF THE EVALUATION MEASURES (PARAMETER NO.1 TO NO.5)

Measure	No.1	No.2	No.3	No.4	No.5
F-measure	0.4195	0.4351	**0.4714**	0.4651	0.4641
Precision	0.4333	0.4125	**0.5014**	0.4754	0.4765
Recall	0.4056	0.4356	**0.4448**	0.4449	0.4410
Accuracy	0.4550	**0.5264**	0.5250	0.5144	0.5140
Summation	0	1	3	0	0

B. Results and discussions

This section compares the proposed β-hill climbing with the original hill climbing for the text clustering problem using the best parameter value (β = 0.2). Table V shows that the proposed β-hill climbing has overcome the original version almost in the most given datasets according to the F-measure value (i.e., DS1, DS2, DS4, and DS5). According to precision measure, β-hill climbing obtained the best results of three of five datasets (i.e., DS1, DS2, and DS5). According to recall measure, β-hill climbing obtained the best results of four of five datasets (i.e., DS1, DS2, DS4, and DS5). Finally, according to accuracy measure, β-hill climbing obtained the best results of four of five datasets (i.e., DS1, DS3, DS4, and DS5).

It is evident in Table V that the β-hill climbing increased the performance of the clustering algorithm, which assessed by four measure on five datasets. The statistical analysis build according to the F-measure values in Table V display that the proposed β-hill climbing obtain the best results overall the experiments. Generally, β-hill climbing performed well in comparison with the original version.

For the average ADDC of twenty runs, β-hill climbing technique reaches global minima in all of the five datasets. Fig. 2 shows the convergence behavior of the β-hill climbing and hill climbing techniques for text clustering problem (i.e. DS1 and DS2). It is evident that the convergence of the β-hill climbing technique is smooth in comparison with original technique. As well, no shift carve during the converge. Finally, we observe that the β operator enhance the clustering technique because of the enhanced exploration and exploitation (i.e, local, and global search) capability.

TABLE IV
THE PERFORMANCE OF THE β VALUES IN TERMS OF THE EVALUATION MEASURES (PARAMETER NO.6 TO NO.10)

Measure	No.6	No.7	No.8	No.9	No.10
F-measure	0.4420	0.4305	0.4154	0.4215	0.4212
Precision	0.4515	0.4321	0.4014	0.4121	0.4199
Recall	0.4321	0.4299	0.4214	0.4321	0.4287
Accuracy	0.4845	0.4874	0.4654	0.4451	0.4501
Summation	0	0	0	0	0

TABLE V
ALGORITHM PERFORMANCE BASED ON CLUSTERS QUALITY

Dataset	Measure	Hill climbing	β-hill climbing
DS1	F-measure	0.4195	0.4714
	Precision	0.4333	0.5014
	Recall	0.4065	0.4448
	Accuracy	0.4550	0.5250
Rank		2	1
DS2	F-measure	0.3483	0.3866
	Precision	0.3869	0.4083
	Recall	0.3167	0.3672
	Accuracy	0.3625	0.4250
Rank		2	1
DS3	F-measure	0.2681	0.2228
	Precision	0.2752	0.2134
	Recall	0.2614	0.2325
	Accuracy	0.2777	0.3037
Rank		1	2
DS4	F-measure	0.1514	0.1874
	Precision	0.1501	0.1452
	Recall	0.1528	0.1802
	Accuracy	0.1800	0.2250
Rank		2	1
DS5	F-measure	0.1542	0.1744
	Precision	0.1509	0.1720
	Recall	0.1576	0.1768
	Accuracy	0.1640	0.1820
Rank		2	1
Mean rank		1.80	1.20
Final rank		2	1

VI. CONCLUSION

In this paper, a novel text clustering technique, namely, β-hill climbing for text clustering, is proposed to obtain an optimal subset of documents clusters. The β-hill climbing technique overwhelms the original hill climbing technique by developing text clustering results according to all evaluation measures that utilized in the experiments. Experiments were carried out on five portions text datasets taken randomly from "Dmoz-Business" dataset after the tuning parameter of the β-hill climbing has been done. The results explained that the performance of the clustering method is improved using the β-hill climbing technique in the most datasets. The performance of the text clustering is useful by adding the β operator to the hill climbing. For future work, the proposed β-hill climbing technique can be hybrid with another population-based method to enhance the global search in goal to find more accurate clusters.

REFERENCES

[1] A. L. Bolaji, M. A. Al-Betar, M. A. Awadallah, A. T. Khader, and L. M. Abualigah, "A comprehensive review: Krill herd algorithm (kh) and its applications," *Applied Soft Computing*, vol. 49, pp. 437–446, 2016.

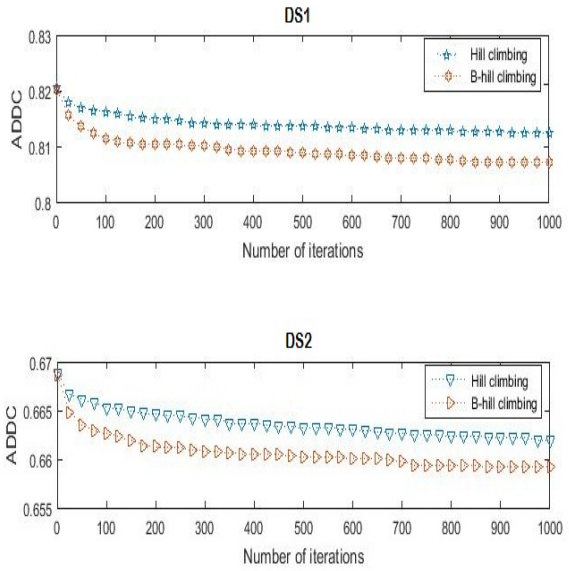

Fig. 2. The convergence behavior of the clustering techniques.

[2] L. M. Abualigah, A. T. Khader, and M. A. Al-Betar, "Multi-objectives-based text clustering technique using k-mean algorithm," in *Computer Science and Information Technology (CSIT), 2016 7th International Conference on*. IEEE, 2016, pp. 1–6.

[3] V. Tunali, T. Bilgin, and A. Camurcu, "An improved clustering algorithm for text mining: Multi-cluster spherical k-means," *International Arab Journal of Information Technology (IAJIT)*, vol. 13, no. 1, 2016.

[4] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, and M. A. Awadallah, "A krill herd algorithm for efficient text documents clustering," in *2016 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, May 2016, pp. 67–72.

[5] L. M. Abualigah, A. T. Khader, and M. A. Al-Betar, "Unsupervised feature selection technique based on harmony search algorithm for improving the text clustering," pp. 1–6, July 2016.

[6] K. K. Bharti and P. K. Singh, "Chaotic gradient artificial bee colony for text clustering," *Soft Computing*, vol. 20, no. 3, pp. 1113–1126, 2016.

[7] M. M. Zaw and E. E. Mon, "Web document clustering by using pso-based cuckoo search clustering algorithm," in *Recent Advances in Swarm Intelligence and Evolutionary Computation*. Springer, 2015, pp. 263–281.

[8] L. M. Q. Abualigah and E. S. Hanandeh, "Applying genetic algorithms to information retrieval using vector space model," *International Journal of Computer Science, Engineering and Applications*, vol. 5, no. 1, p. 19, 2015.

[9] M. A. Al-Betar, "\ beta-hill climbing: an exploratory local search," *Neural Computing and Applications*, pp. 1–16.

[10] L. M. Abualigah, A. T. Khader, and M. A. Al-Betar, "Unsupervised feature selection technique based on genetic algorithm for improving the text clustering," pp. 1–6, July 2016.

[11] L. M. Abualigah and A. T. Khader, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering," *The Journal of Supercomputing*, pp. 1–23, 2017.

[12] A. J. Mohammed, Y. Yusof, and H. Husni, "Document clustering based on firefly algorithm," *Journal of Computer Science*, vol. 11, no. 3, p. 453, 2015.