electronics

MDPI

Article

Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering

Laith Abualigah [1], Amir H. Gandomi [2,*], Mohamed Abd Elaziz [3], Husam Al Hamad [1], Mahmoud Omari [1], Mohammad Alshinwan [1] and Ahmad M. Khasawneh [1]

[1] Faculty of Computer Sciences and Informatics, Amman Arab University, Amman 11953, Jordan; Aligah.2020@gmail.com (L.A.); hhamad@aau.edu.jo (H.A.H.); omari@aau.edu.jo (M.O.); mohmdsh@aau.edu.jo (M.A.); a.khasawneh@aau.edu.jo (A.M.K.)
[2] Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia
[3] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt; abd_el_aziz_m@yahoo.com
* Correspondence: Gandomi@uts.edu.au

Abstract: This paper presents a comprehensive survey of the meta-heuristic optimization algorithms on the text clustering applications and highlights its main procedures. These Artificial Intelligence (AI) algorithms are recognized as promising swarm intelligence methods due to their successful ability to solve machine learning problems, especially text clustering problems. This paper reviews all of the relevant literature on meta-heuristic-based text clustering applications, including many variants, such as basic, modified, hybridized, and multi-objective methods. As well, the main procedures of text clustering and critical discussions are given. Hence, this review reports its advantages and disadvantages and recommends potential future research paths. The main keywords that have been considered in this paper are text, clustering, meta-heuristic, optimization, and algorithm.

Keywords: meta-heuristic; optimization algorithms; machine learning; optimization problems; big data; text clustering applications

Citation: Abualigah, L.; Gandomi, A.H.; Elaziz, M.A.; Hamad, H.A.; Omari, M.; Alshinwan, M.; Khasawneh, A.M. Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering. *Electronics* **2021**, *10*, 101. https://doi.org/10.3390/electronics10020101

Received: 29 November 2020
Accepted: 22 December 2020
Published: 6 January 2021

1. Introduction

Generally, clustering is a common text mining technique used to arrange a restricted set of clusters, often with a predefined number of clusters, to represent a dataset based on similarities between its objects [1]. The main clustering applications are classifying from market segmentation [2], text summarization [3], text classification [4], text document clustering [5], image processing [6], data clustering [7], text document categorization [8], wireless sensor networks [9], web mining [10], sentiment Analysis [11], Big data clustering [12], and others. One of the main application fields determined to be especially promising for clustering methods is bioinformatics [13]. Certainly, the effect of clustering gene expression data contained the help of micro-array and other related procedures, which have evolved fast and strongly during the last few years [14,15].

Clustering techniques are, in general, classified into three main classes: overlapping/nonexclusive, partitional, and hierarchical. The last two classes are linked in which hierarchical clustering is a nested classification of partitions clustering [16]. Therefore, they present poor performance when the separation of overlapping clusters is conducted. Meta-heuristic optimization algorithms are used in the partitioning clustering method. The partitional clustering partitions a dataset into a subset of groups based on a particular measure identified as a fitness function [17]. The fitness function straight influences the nature of the formation of groups. Once a suitable fitness function is chosen, the partitioning process is transformed into an optimization problem (i.e., partitioning based on minimization the distance measure or maximization the similarity measure between

patterns, otherwise optimizing their frequency in the N-dimensional space). These partitional methods are commonly used in different research areas due to their ability to cluster big datasets, such as in signal/image processing for image segmentation [18], analysis to classify the group of homogeneous users in economics [19], in relation to producing specific hidden equalizers [20], in robotics to effectively organize the humans according to their activities [21], in seismology to match the aftershocks from the general background situations [22], to achieve high dimensional data report [23], in computer science domain for web text mining and image pattern recognition [24], in control studies to manage the portfolio [25], in medical anthropology to classify diseases from a combination of patient records and genomic investigations [26], in wireless sensor network for distributing the sensors to improve lifetime and coverage area [27], and in library mathematics for grouping publications according to the content [28]. In these variant applications, the characteristics of patterns connected with the datasets are distinct from each other. Consequently, a single or basic partitional algorithm cannot wholly solve all clustering problems. Hence, given a problem inability, a user has to accurately examine the quality of the patterns associated with the given dataset and choose the suitable clustering algorithm [1,29,30].

From an optimization aspect, clustering can be officially presented as a kind of Non-deterministic (NP) -hard optimization problem [31]. This has encouraged the search for practical optimization algorithms, providing not only the performance of ad hoc learning for specific classes of problems but also the convenience of general-purpose optimization methods [32]. Unusually, meta-heuristic and evolutionary optimization algorithms are meta-heuristic techniques broadly considered to be useful in solving NP-hard problems, being capable of producing near-optimal solutions (optimal clusters) to the given clustering problems in a reasonable time. Under this supposition, many meta-heuristic algorithms for solving clustering problems, especially text clustering, have been introduced in the literature. These optimization algorithms are used to optimize the given objective function (i.e., fitness function) that controls the improvement search [1,33]. A simple example of the clustering process is presented in Figure 1. In this figure, unorganized documents mean that the given documents are not clustered; on the other hand, the documents from various topics are given together. Documents clusters indicate that the presented documents are clustered based on their contents; on the other hand, each similar documents are offered in a different cluster.

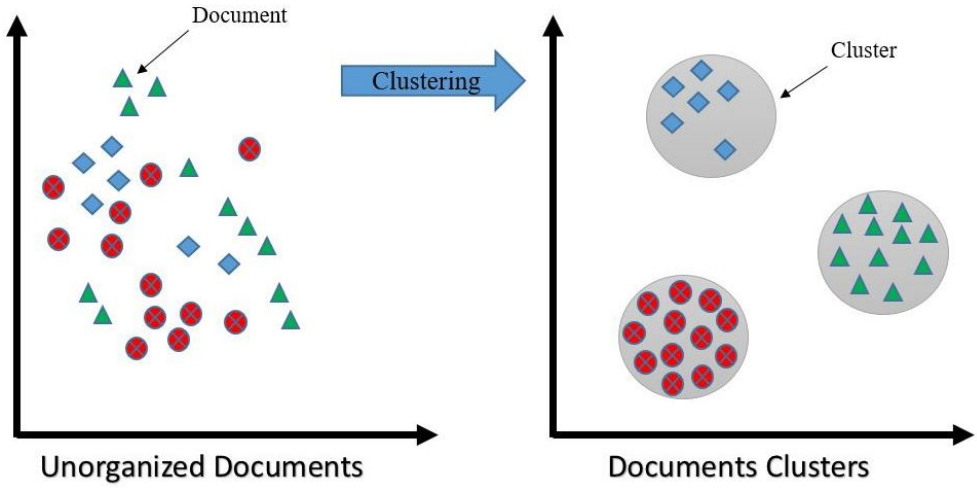

Figure 1. A simple example of the clustering process.

This paper aimed to provide readers with an accurate overview of the different meta-heuristic optimization algorithms available for big data and text clustering by comparing them analytically. However, this work studies the accessibility and application of an appropriate optimization algorithm for each class explicitly. From several large text datasets, it also provides experimental findings. When coping with large text clustering questions,

certain viewpoints need close consideration. Consequently, this study will assist researchers and clinicians in selecting methods and algorithms suitable for broad text clustering applications. Compared to conventional clustering approaches, the number of documents is the first and relatively significant factor to deal with when clustering big text and data. This includes significant developments in the design of clustering algorithms. Velocity is the other critical feature of big texts and results. This state refers to a high online computing demand. To trade with the data streams, processing velocity is required. The third aspect is the variety. From different references, such as sensors, computers, cell phones, vehicles, etc., various data representations, such as text, video, and image, are given. The key components of critical text and data that must be brought when selecting practical meta-heuristic clustering algorithms are these three aspects (i.e., Length, Velocity, and Variety).

Despite an enormous number of clustering algorithm survey papers prepared in the literature for different domains (such as machine learning, data mining, deep learning, data retrieval, pattern recognition, and semantic ontology) [34], it is difficult for users to select a priori which algorithm or methodology will be appropriate for dealing with big data and semantic ontology This is due to the restrictions that arise in the existing surveys: (1) The meta-heuristic algorithm properties are not well analyzed and studied. (2) Several new optimization algorithms have been provided by the domain, which has not been analyzed in these surveys. (3) To evaluate the superiority of one algorithm over another, no detailed functional analysis was performed. This paper aimed to research the field of meta-heuristic clustering algorithms and achieve the following objectives, as inspired by these analyses:

- To introduce a categorizing structure that groups the existing meta-heuristic clustering algorithms into classes and shows their advantages and shortcomings from a general point of view.
- To give a complete classification of the clustering evaluation criteria to be utilized for experimental research.
- To make a theoretical analysis for the most representative meta-heuristic optimization algorithms of each class.

The main parts of this paper are organized as follows. In Section 2, the main procedures of the text clustering are presented. Section 3 shows the variants of meta-heuristic algorithms that have been used in solving clustering problems. Evaluation criteria used in text clustering applications are discussed in Section 4. Discussion and theoretical analyses are given in Section 5. Finally, the conclusion of the survey and prospects for further investigation are presented in Section 6.

2. Main Procedures of the Text Clustering

Text clustering intends to produce optimal clusters that contain related documents (objects). Clustering is based on partitioning a collection of documents into a predefined number of associated groups, where each group includes a number of similar objects, but various groups have various objects.

2.1. Problem Descriptions and Formulations

In this section, the text clustering problem and its descriptions and formulations are given as follows.

- A collection of documents (D/objects) is grouped into a predefined number of clusters (K) [35].
- D can be demonstrated as a vector of objects $D = (d_1, d_2, d_3, ..., d_i, ..., d_n)$, d_2 presents the object number two, i is the number of the object and n presents the number of total objects given in D [36].
- Each group contains a cluster centroid, called c_k, which is represented as a vector of term weights of the words $c_k = (c_{k1}, c_{k2}, c_{k2}, ..., c_{kj}, ..., c_{kt})$.

- c_k presents the k_{th} cluster centroid, c_{k2} is the value of position two in the centroid of cluster number k, and t is the number of all unique centroid terms (features) in the given object.
- The similarity or distance measures is utilized to clustering each object to the closest cluster centroid [37–39].

2.2. Pre-Processing Steps

The chief purpose of the clustering technique is to produce groups according to the objects' intrinsic contents. Ere creating clusters, the text need model pre-processing steps, as follows: (i) tokenization, (ii) stop word removal, (iii) stemming, (iv) term weighting, and (v) document representation [40]. A brief demonstration of these pre-processing levels is presented, as follows.

2.2.1. Tokenization

Tokenization is the process of separating words into bits (words), called tokens, which are presumably missing individual letters simultaneously, such as punctuation. Usually, these tokens are connected to terms/words, but it is essential to distinguish between type/token. A token is an example in a text of a sequence of letters that is organized as a functional semantic unit. A sort is a set, including the same letter chain, of all tokens. A word is an example involved in the vocabulary of the search method [41].

2.2.2. Stop Words Removal

Standard and famous words, such as "which", "the", "our", "is in", "an", "that", "me", and some are stop-words, as well as other prominent phrases in the text that are exceptionally widely used and small useful words. These words should be omitted from the document given (text) because they are generally highly repetitive, thus diminishing the efficacy of the clustering techniques. The stop-word list (http:/www.unine.ch/Info/clef/) comprises more than 500 words in total [40].

2.2.3. Stemming

Stemming is the phase in which modern words are shortened to their root/stem. The stem method is not the same as the root morphological method; it is generally the same stem to outline words, even though it is not a real root in itself. Porter (Stemmer of Porter. The popular public stemming method used in text mining [41,42] is available at http:/tartarus.org/martin/PorterStemmer/) stemmer. Both pre-processing acts are extracted from Python NLTK Natural Language Processing Demonstrations (http://text-processing.com/demo/).

2.3. Document Representation

The Vector Space Model (VSM) is a powerful model used to define documents' content in an official format called [43]. It emerged at the beginning of the 1970s. Each paper is structured as a term weight vector to assist the calculation of similarities. To increase the efficiency of the clustering algorithm and lower the time cost [44], each term in the text communicates a dimension of the weighted performance. In different data mining fields, VSM is used, such as Reference [45] for knowledge extraction, Reference [46] for text labeling, and Reference [44] for text clustering.

Term weighting is given by the vector space model (VSM) to show the document in a standard form [47], as given in Equation (1). This model shows each document as a vector of words, as given in Equation (2) [37,44,48]. Equation (1) shows n documents and t terms in a standard format utilizing the VSM, as follows:

$$VSM = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,(t-1)} & w_{1,t} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{(n-1),1} & w_{(n-1),2} & \cdots & \cdots & w_{(n-1),t} \\ w_{n,1} & w_{n,2} & \cdots & w_{n,(t-1)} & w_{n,t} \end{bmatrix} \quad (1)$$

$$d_i = (w_{i,1}, w_{i,2}, w_{i,3}, ..., w_{i,j}, ..., w_{i,t}). \quad (2)$$

2.4. Solution Representation of Clustering Problem

The text clustering document is expressed as an optimization the problem that is applied based on optimization algorithms. Algorithms for optimization trade with numerous solutions to solve the given problem. The candidate solution to solve the clustering problem is defined by each solution (row). The solution is designed as a n dimension vector that defines each document's content in the D dataset that is given, and each location leads to a document. Figure 2 demonstrates the solution design. The solution's ith location contributes to the decision about the ith text. If the number of the clusters given is K, then the value in the range $(1, ..., K)$ is at each position of the solution. The set of K centroids [37] fits each component. The number of clusters is normally given in advance.

In the example given in Figure 2, nine documents and three clusters are presented. Each solution designs where the documents belong. In this case, documents 2, 5, and 7 are from the same group as label 1 (i.e., cluster number one). Meanwhile, documents 1, 3, 4, 8, and 9 belong to the same cluster as label 2 (i.e., cluster number two). Document number 6 belongs to cluster number 3.

Figure 2. Solution representation of the clustering problem.

2.5. Fitness Function

The fitness value is determined to assess each solution according to its positions. Each collection of documents belong to a collection of K centroids $C = (c_1, c_2, ..., c_k, ..., c_K)$, where c_k is the centroid of cluster k. The fitness function value for each candidate solution is calculated by the average similarity of documents to the cluster centroid (ASDC), as given in Equation (3) [44,49]. The similarity measure inside the fitness function in Equation (3) can be changed into another similarity or distance measures.

$$ASDC = \left[\frac{\sum_{j=1}^{K} \left(\frac{\sum_{i=1}^{n} Cos(d_i, c_j)}{m_i} \right)}{K} \right], \quad (3)$$

where K is the number of given clusters in the dataset, m_i is the number of documents that correctly belong to cluster i, and $Cos(d_i, c_i)$ is the similarity value between the centroid of cluster j and the document number i. Each solution is presented in binary matrix $a_{i,j}$ of size $n * K$ to calculate the clusters centroid, as given in Equation (4) [38].

$$a_{ij} = \begin{cases} 1, & \text{if } d_i \text{ is assigned to the } j\text{th cluster} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Equation (5) is utilized to calculate the k_{th} cluster centroid, which is given as a vector $c_k = (c_{k1}, c_{k2}, c_{k3}, ..., c_{kj}, ..., c_{kt})$ [35].

$$c_{kj} = \frac{\sum_{i=1}^{n} a_{ij}(d_{ij})}{\sum_{i=1}^{n} a_{ij}}, \quad (5)$$

where a_{ij} is a matrix contains the grouped data (see Equation (4)), d_{ij} is the jth feature weight of the document number i, and n is the number of all documents in the used dataset.

3. Meta-Heuristic Algorithms in Text Clustering Applications

In this section, the related works are given as follows.

3.1. Meta-Heuristic Algorithms

In this section, the most common meta-heuristic algorithms used to solve the text document clustering problems and have been published in the literature.

3.1.1. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is one of the powerful meta-heuristic optimization algorithms in the publications [50,51] used to solve clustering issues. He-Nian et al. recommended a procedure called OK-PSO [52] to cluster text based on k-means (KM) and a PSO algorithm. The KM is used to calculate the distance across each word and the centers of the clusters. To test the optimization of the clustering distance, the 2-D Otsu algorithm was used. The technique uses the PSO algorithm to look for the optimum threshold to speed up the threshold estimation. On datasets, the efficiency of the suggested approach was tested and compared with other clustering methods. Experimental results showed the utility of the proposed approach over the algorithms.

A hybrid approach is proposed to solve the linear text segmentation by improving the segmentation accuracy and computation complexity [53]. The hybrid algorithm is called TSHAC-DPSO based in Hierarchical Agglomerative Clustering that can efficiently generate a satisfactory solution without parameter setting or an auxiliary knowledge base. The algorithm adapts Discrete PSO to produce the optimal solution by refining the solution found by TSHAC. The algorithm was tested on several standard datasets and showed comparable performance with several known linear text segmentation algorithms.

Sunita et al. presented a comparative analysis of three clustering algorithms, namely KM, PSO, and hybrid PSO plus KM [54]. The performance of the aforementioned algorithms was tested on text in Nepali language. In the experiments, the texts are represented in terms of synsets corresponding to a word and synonyms from WordNet to group semantically related terms. The experiment evaluation was based on intra- and inter-cluster similarity. The results showed that the hybrid PSO plus KM outperforms both standalone PSO and KM algorithms.

Particle Swarm Optimization (PSO) algorithm is used as a feature selection technique for text document clustering in Reference [55]. The purpose of using PSO is to select the most informative features for representing each text document, thus improving the KM clustering algorithm performance. The proposed approach was tested on several standard datasets. The experimental results revealed an improved performance of KM due to the use of optimized feature set obtained by PSO algorithm, in addition to an improvement in computation time.

Jung Song et al. proposed an ensemble method for sentence clustering based on automatic population partitioning (APP) to improve document summarization [56]. The proposed method utilized the characteristic advantage of global search ability of GA and the local search ability of PSO algorithms. Experiments are conducted on standard dataset and used a normalized Google distance similarity measure to measure the similarity between the sentences. Results showed an improved summarization performance over other know sentence summarization methods.

The paper used spectral clustering, which is widely used in machine learning, augmented with PSO algorithm to improve the text clustering [57]. The proposed approach called SCPSO was tested on several standard datasets and compared with several well-

known algorithms, such as Spherical KM, Expectation Maximization Method, and the basic PSO algorithm. The experimental results showed that the proposed method outperforms the comparable algorithms is terms of clustering accuracy.

3.1.2. Gray Wolf Optimizer (GWO)

In Reference [58], a new algorithm called a GWO–GOA is proposed for enhancing GWO algorithm using GOA and the pragmatic approach. The algorithm pre-processed the document and extracted local optima feature to obtain best global optima using hybrid GWO–GOA and Fuzzy c-means (FCM) clustering to cluster the selected optima; the algorithm used eight datasets and showed better results in precision, specificity, sensitivity, F-measure, and recall compared to other algorithms

Text mining can be defined basically as the method by which high-quality information is extracted from the text. It is used extensively in applications, such as text clustering, categorization, and classification. The text clustering has currently become an interesting task employed to organize the text document. The accuracy of text clustering is reduced due to several unnecessary words and large dimensions. The semantic word processing and novel Particle Grey Wolf Optimizer (PGWO) for efficient text clustering are introduced in Reference [59]. First, the text documents are provided as input to the initial phase, which offers valuable keyword for clustering and feature extraction. The resulting keyword is then added to WordNet ontology to figure out every keyword's synonyms and hyponyms. Consequently, for every keyword employed to construct the text function library, the frequency is calculated. Since the larger dimension is contained in the text feature library, the entropy is used for identifying the most relevant feature. Finally, by merging the particle swarm optimization (PSO) with the grey wolf optimizer (GWO), the new Particle Grey Wolf Optimizer (PGWO) method is being designed. Therefore, the suggested algorithm is employed to add class labels for the generation of different text document clusters. To evaluate the efficiency of the proposed algorithm, the simulation is conducted and compared with existing methods. The proposed algorithm achieves 80.36% clustering accuracy for 20 Newsgroup datasets and 79.63% clustering accuracy for Reuter, which guarantees better automatic text clustering.

In several main areas, such as information retrieval, text mining, and natural language processing, text clustering problem (TCP) is a leading method. This poses the need for a robust algorithm for document clustering that could be employed efficiently to explore, analyze, and organize information to collect large amounts of data. In Reference [60], the authors suggested an extension, referred to as TCP-GWO, of the grey wolf optimizer (GWO) for TCP. Beyond what is necessary with meta-heuristic swarm-based methods, the TCP requires a degree of accuracy. Breaking text documents based on GWO into homogeneous clusters that are relatively reliable and efficient is the key problem to be tackled. Primarily, to continuously optimize the distance between the clusters of documents, TCP-GWO, or the document clustering method, employ the average document distance to the centroid cluster (ADDC) as the objective feature. The reliability of the suggested TCP-GWO has been illustrated based on a sufficiently large number of sample size documents selected at random from a sample of six publicly available data sets. In the evaluation process for evaluating the recall detection performance of the document clustering method, high complexity documents were also included. The experimental findings for a test collection of over a subset of 1300 documents revealed that, in approximately 15–20% of cases, inability to effectively cluster a document happened with a classification performance of more than 65% for an extremely complex data set. The high F-measure rate and ability to effectively cluster documents are significant advances resulting from this analysis. The suggested TCP-GWO approach was compared using randomly selected data sets to the other methods of text clustering. Incidentally, in terms of accuracy, recall, and F-measure rates, TCP-GWO outperforms comparable methods.

3.1.3. Cuckoo Search (CS)

Due to the hierarchy generated in document organization, Self-Organizing Map (SOM) is interesting. It is considered the best for those issues where clustering and visualization are required. In Reference [61], the main idea of this work is to find effective clustering and classification methods that could be significant in the document organization. Therefore, the authors proposed an associative cluster to classify the text data and employ a cuckoo search algorithm to find the optimal rank in the relevant class hierarchy. In contrast to the previous approaches, the authors applied the proposed methodology to 8 separate data for experimentation and obtained better results.

For data processing, clustering can be considered an essential technique. However, it requires more work to cluster robust data, such as papers to regain the intricacy of the relevant knowledge concealed in the space of multi-dimensionality. Based on meta-heuristics, document grouping algorithms have recently proven their efficacy in exploring the search field and achieving ideal global solutions instead of local solutions. However, a few of these algorithms are not reasonable and endure several disadvantages, as well as the need to identify the clusters in advance, are neither exponential nor customizable, and high-dimensional and indistinct matrix papers are classified.

A new hierarchical and incremental approach (cuckoo search (CS) latent semantic indexing (LSI)) for text clustering based on recent cuckoo search (CS) optimization and latent semantic indexing (LSI) was proposed by authors in Reference [62] to address the cited limitations. Four high-dimensional text datasets are experimented, demonstrating the usefulness of the LSI paradigm in minimizing dimensional space with greater precision and low processing time. The suggested CS LSI also calculates the number of clusters automatically by using a new proposed index based on the real measurement of the wavelength. In the gradual mode, this is often used later and retains a more coherent cluster to identify the outlier records. It shows the efficacy of CS LSI in achieving high clustering performance relative to standard algorithms for paper clustering.

3.1.4. Firefly Algorithm (FA)

Hierarchical text clustering plays an essential role in managing, organizing, and summarizing documents systematically. However, due to the use of KM as part of its procedure, the Bisect KM, this is a well-known hierarchical clustering method, can only produce local optimal solutions. In Reference [63], authors suggest replacing the KM with the firefly method, thus generating a hierarchical clustering Bisect FA. The firefly technique performs at each stage of the proposed Bisect FA to generate the optimal clusters. For validation, authors conducted experiments on 20 datasets which widely employed in the literature. Findings show that Bisect FA achieves more efficient and lightweight clustering than the Bisect KM, KM, and C-firefly methods. As a result, the proposed Bisect FA is considered an efficient algorithm for unsupervised learning.

Text clustering is organizing related documents into a cluster while assigning different documents to other clusters. In several disciplines, a well-known clustering tool, the KM method, is widely used. However, estimating the number of clusters using KM is a significant challenge. Therefore, authors in Reference [64] conducted a study to introduce a new clustering technique that takes the Firefly Algorithm for dynamic document clustering, called Gravity Firefly Clustering (GF-CLUST). GF-CLUST can classify the required number of clusters for a given test set, challenging in the clustering of texts. It identifies documents with strong force as centers and produces clusters based on the calculation of cosine similarity. This is supplemented by the selection of possible clusters and the combination of tiny clusters. To evaluate the proposed GF-CLUST, experiments are performed on different document datasets, such as 20 Newgroups, Reuters-21578, and TREC collection. The outcomes of GF-CLUST 's purity, F-measure, and entropy outperform those of current clustering techniques, such as KM, Particle Swarm Optimization (PSO), and Practical General Stochastic Clustering System (pGSCM). Besides, compared to pGSCM, the number of extracted clusters in GF-CLUST is like the real number of clusters.

Given that the ORC (Optical Character Recognition) system only recognizes the binarized image, text binarization is an essential step in text comprehension. The more precise the binary text is, the better the ORC method operates. A novel binarization technique is suggested in Reference [65] to binarize text from complex color images. Firstly, to combine similar color pixels into an image, the Fuzzy C-means method is employed. To eliminate noise components in the background, the flood filling method is then utilized. Eventually, to binarize the text image, the authors employed the Otsu global binarization technique. The experimental results indicate that the proposed technique outperforms the Otsu global binarization system.

3.1.5. Krill Herd Algorithm (KHA)

A hybrid krill herd algorithm with KMis proposed in Reference [66] for solving text clustering problem. The proposed strategy utilizes the local exploitation of KM to avoid local optimum and premature convergence. This hybridization resulted in a quick convergence to optimal solution. Several version of hybrid krill herd were developed and the best one is augmented with an objective function that combines two measures, cosine similarity and Euclidean distance, for improving the local search facility. The experimental results of several assessment measures showed that the proposed method outperforms all versions of krill herd algorithm beside several other text clustering methods found in the literature.

This paper introduced two text clustering algorithms based on krill herd algorithm for improving the clustering of web text documents [67]. One of the algorithms utilizes all the operators of krill herd algorithm while the other one neglects the genetic operator. The performance of the proposed method was compared with the KM algorithm in terms Purity and Entropy measures. The experimental results showed that the proposed algorithms outperform KM algorithm.

Abualigah et al. proposed a hybrid krill herd algorithm with harmony search algorithm for document and text clustering [68]. The purpose of the hybridization is to improve the global search ability by introducing the global search operator of the harmony search to the krill herd algorithm. This introduction of the operator improved the exploration search ability of krill herd by using the Distance factor as a new probability factor. The proposed algorithm was tested on several standard datasets. The results showed enhancement in clusters accuracy and high convergence rate.

Abualigah et al. proposed three new improved versions of krill herd algorithm for enhancing the results of the basic version [69]. The improvements involve different ordering of the crossover and mutation genetic operators, in which these operators are performed after the update of krills' position, thereby achieving an accurate global search. The experiment was conducted on several standard text datasets. The experimental results were compared with other published algorithms, such as GA, harmony search, and PSO, tested on the same datasets. The result showed that the proposed improved algorithms outperform the comparative algorithms on all benchmark datasets.

The author introduced a new technique for solving text document clustering by developing four different versions of krill herd algorithm KHA [70]. These versions are the basic KHA, hybrid KHA, and multi-objective hybrid KHA. Each one is considered as an incremental enhancement of the preceding one. The algorithms were tested on several benchmark datasets. The experimental results showed that the multi-objective hybrid version obtained the best results among the others and outperforms other comparative algorithms found in the literature.

The paper introduced a new feature selection method for improving text document clustering [71]. The proposed method enhances the performance of the hybrid krill herd algorithm by hybridizing it with swap mutation strategy. The hybridization is incorporated within a parallel membrane computing framework. The algorithm was tested on several standard text datasets and showed excellent convergence velocity and obtained superior results compared to popular feature selection algorithms.

3.1.6. Social Spider Optimization (SSO)

Scholars have commonly utilized evolutionary optimization algorithms to enhance accuracy and performance as the clustering issue can be mapped to the optimization method. Stochastic general-purpose approaches for addressing optimization issues are one of the evolutionary approaches. Swarm Intelligence is one of the major approaches that work with the aggregative behavior of swarms and their dynamic interactions without control. However, the Swarm intelligence model seems to be much more promising because of its reliability. Authors in Reference [72] suggested a swarm intelligence scheme named social spider optimization for text document clustering. This technique utilized social spiders' cooperative intelligent behavior. Each spider prefers to replicate a specialized behavior based on its gender. It also aims to substantially minimize premature convergence and local minimum issues in text documents clustering. Results have been compared with the KM clustering method and show good results.

It is reported that text document clustering is considered one of the primary data mining issues instigated by researchers. It helps classify text documents in such a way that there are similar text documents for each group. Several problems have been found when grouping text documents. The critical issues in text document clustering are reliability and performance. Scholars have commonly employed evolutionary optimization algorithms to enhance accuracy and performance as the clustering issue can be mapped to the optimization problem. Stochastic general-purpose approaches for addressing optimization issues is one of the evolutionary methods. Swarm Intelligence also is a specific method that works with the aggregative swarm's behavior and their dynamic interactions without any control. For textual document clustering, the authors proposed a new swarm intelligence method named Social Spider Optimization SSO in Reference [73]. Authors compared it to KM clustering and state-of-the-art clustering methods, including PSO, ant colony optimization (ACO), and Improved Bee Colony Optimization (IBCO), and found it to be more accurate. Authors then suggested two-hybrid clustering methods, called SSO + KM and KM + SSO, and discovered that SSO + KM clustering surpassed KM + SSO, KPSO (KM + PSO), KGA (KM + GA), KABC (KM + Artificial Bee colony) and IBCO clustering algorithms. To show the effectiveness of clustering techniques, the authors employed the Sum of intra-cluster distances, average cosine similarity, precision, and Inter cluster distance.

3.1.7. Gravitational Search Algorithm (GSA)

GSA is a research method proposed in Reference [74]; it is used to solve the optimization problems. This paper proposed a new method called GSA-KHM combined between GSA and K-harmonic means to improve dependency on the initialization [75]. The proposed method applied five datasets and showed better results than other methods.

3.1.8. Whale Optimization Algorithm (WOA)

WOA is a research method proposed by Mirjalili and Lewis in Reference [76], it is used to solve various optimization problems. Jagatheeshkumar and Selva in Reference [77] implemented clustering algorithm used Whale Optimization Algorithm (WOA) and Fuzzy C Means (FCM); the algorithm tested, over three datasets, precision, recall, F-measure, purity, and entropy, and the performance of the proposed algorithm showed better results compared to existing methods.

3.1.9. Ant Colony Optimization (ACO)

Ant colony optimization (ACO) was used to solve the clustering problems in many research fields [78]. In Reference [79], ACO is applied on multi-label text categorization with relevance clustering classification technique. The ant colony algorithm is used for feature optimization of text data. The proposed method is tested on several datasets, including WebKB of CMU test learning group, Yahoo web page dataset, and RCV1 (Reuters Corpus Volume 1). For evaluating the performance of the proposed method, other algorithms were

tested, including MLFRC and Rank SVM (RSVM). The experimental result show that the proposed method outperforms both MLFRC and Rank SVM.

In Reference [80], ACO algorithm is applied to solve the problem of fuzzy document clustering. The authors used a thesaurus for extracting documents features in order to have a language-independent feature vector for the purpose of measuring the similarities of documents written in different languages. The pheromone trails in the ACO algorithm are used to decide the membership values in clustering process. The performance of the algorithm is evaluated on a dataset of bilingual (English and Spanish) documents consisting of scientific research papers in various subject fields. The experimental results show that the proposed algorithm gives good performance in spite of the difficulties of the stated problem and the huge efforts in the pre-processing stage.

3.1.10. Genetic Algorithm (GA)

Mustafi et al. [81] developed the popular KM algorithm for the tasks of clustering through enhancing the mechanism of initializing the centroids and keeping the clusters number returning in each iteration. The proposed technique depended on using the genetic algorithm with the differential evolution algorithm, where the first is used for generating the original seeds, and the other for obtaining the clusters number. The authors applied the proposed algorithm for the text documents clustering, then compared their technique with the standard KM algorithm for solving the clustering problems.

Song et al. [82] developed the genetic algorithm based on the ontology science, where the algorithm can operate in a self organizing way for solving the problems of clustering texts. The authors used some concepts in the ontology, thesaurus-based and corpus-based, to solve the text clustering problems. Where they supposed two hybrid strategies using various similarity measures, thesaurus-based measure, transformed LSI-based measure which had the superiority than the traditional similarity measures. In addition, the proposed technique for clustering was more efficient than the standard GA and the standard KM.

Song et al. [82] tried to solve the text clustering problem through enhancing the genetic algorithm using a model known as latent semantic which is different from the popular vector space model, in which each part of the text or the vocabulary exemplifies one dimension. But, the authors' model implied by a query through the representation in a reduced space of the dimension where the model concerns the effects of synonymy and polysemy, which creates a semantic structure in textual data. The enhancement of the genetic algorithm depends on using variable string length.The proposed mechanism proved efficiency and high accuracy of clustering over the conventional GA.

Chun-hong et al. [83] proposed an efficient algorithm for the text clustering based on the latest semantic analysis, as well as with using the idea of optimization, the new model avoids the drawbacks of the vector space model and the KM algorithm. Then, the authors compared their algorithm with the vector space model which proved effeciency according to the precision and recall measures.

Shi et al. [84] proposed a patented algorithm for clustering texts by Genetic Algorithm Model (CGAM) which uses the genetic algorithm as the fitness function and the KM algorithm as a convergence criterion. For the Chinese texts clustering, the proposed algorithm constructs an innovative selection method of initial centers of GA and recommends the contribution of characteristics of various parts of speech. The results proved the superiority of CGAM than the KM and GA for clustering business intelligence system of Chinese and English texts.

Wahiba et al. [85] presented a developed evolutionary algorithm for texts clustering for the biomedical database of MEDLINE, where the developing depended on the genetic algorithm and the VSM and an agglomerative algorithm for generating the initial population. The authors tested their proposed mechanism on a sample composed of 500 abstracts, from which the algorithm proved efficiency. A document grouping algorithm based on the KM was presented by Garg et al. [86] with enhanced initial genetic algorithm-based

clusters. The analysis of previous different text databases demonstrated that clustering is much more precise compared to KM grouping by utilizing the proposed.

Wang et al. [87] introduced a text clustering algorithm based on the fuzzy concepts using the rough set and genetic algorithm as a hybrid. The weight parameters through the clustering are described by genetic algorithm; thus, it makes parameters more reasonable and operational and avoids some problems, like subjectivity and unreliability of describing weight parameters in other work.The results showed that the proposed algorithm is feasible.

Yu et al. [88] proposed a new algorithm for text categorization based on the fuzzy C-means algorithm, in which the idea is inspired from the fuzzy clustering, this with the genetic algorithm which is utilized for initializing the cluster centers. The proposed mechanism had high accuracy for classification and confirmed high efficiency for clustering.

Tohti et al. [89] proposed combining the KM method and the GAAC clustering method and using the two feature extraction for text representation and clustering of Uyghur. The proposed technique is composed of two principal stages. In the first one, the optimal initial cluster center can be gained from the small amount of text set by the GAAC method. In the second stage, the large amount of text set is fast clustered by the KM method. The experiments showed that the proposed algorithm has a significant raising on the accuracy of clustering and the time complexity.

Dong et al. [90] tried to solve the feature word weight expressing the text based on the genetic algorithm and the KM. The authors depended on the enhancement on developing According to the weight factor and feature vector through the combination and checking a behavior of pre-processing, which reflected the texts' diversities. The experimental results confirmed high accuracy of classifying and clustering of feature words.

Shao et al. [91] proposed a new method for solving the Text Clustering and Association Rules Mining problems have been introduced. The algorithm relied on the automatic generation of the hybrid conceptual framework, which takes a complete explanation of the characteristics of the correlation among concepts, uses text clustering technology to replace the relationship between artificial industrialized nations and assessment criteria, and hybridizes the classification techniques of item sets to generate the concept maps, and provides the coherence of response.

3.1.11. Harmony Search (HS)

Sailaja et al. [92] proposed clustering technique a Text Independent Speaker Identification with Finite Generalized Gaussian Mixture Model, which is also Multivariate with Hierarchical Clustering and used the EM algorithm for estimating the parameters. In addition, through Hierarchical clustering, the numbers of acoustic classes associated with each speech spectra are determined.

Zeng et al. [93] proposed an algorithm called HI-Rocchio, which depended on two stages; the first stage is represented by an incremental evolutionary method. The Rocchio algorithm is based on the Rocchio method, and enhancing the Hierarchical clustering approach is the final phase. The experimental results showed more advantages of the proposed technique, like that the proposed algorithm, which has multi-hierarchical relations for describing the text documents, which does not exist in the Rocchio algorithm, and, in the text cluster process, the HI-Rocchio algorithm velocity is more than Hierarchical clustering. Moreover, unlike the classical algorithms, the suggested methodology can create a new group to address the inconsistency between the relatively fixed characteristics of the training set and the concepts of the text draft have changed and evolved.

Lokhande et al. [94] presented an efficient technique for text summarization of web documents, which depended on then creating initialization of the natural language processing, tokenization, part of speech tagging, parsing, and chunking. Then, they implemented the Hierarchical clustering Algorithm and Expectation Maximization Clustering Algorithm to search the similarity in a sentence.

Rong et al. [95] tried to solve the text clustering problem based on a hybrid between the SC-KH (Staged text clustering) algorithm and the KM algorithm, where the proposed algorithm divides the process of clustering into two stages: splitting and merging. The KM algorithm is utilized in the the stage splitting, in which the initial values can be identified by utilizing the Canopy algorithm, and the other algorithm of the hybridization is utilized in the merging stage. The results proved that the proposed technique is superior to the standard KM and the standard hierarchical agglomeration algorithm.

Abualigah et al. [96] proposed a new technique for solving the feature selection problem based on the HS algorithm to search for the best subset of informative features. Then, the proposed method is utilized for clustering texts and is summarized as FSHSTC, FS technique using HS algorithm for the TC technique, where it can overcome the other methods drawbacks in improving the performance of the text clustering. The results showed that the performance text clustering is improved using the proposed method.

3.1.12. Other Meta-Heuristic Algorithms

Lipeng et al. [97] employed a new hybrid Differential Evolution (DE) and Invasive Weed Optimization (IWO), in addition to KM algorithms, called IWODE-KM algorithm, to optimize KM parameters for Chinese text clustering; the algorithm solved random initialization sensitivity and stuck at local optimum of KM algorithm, and the new proposed algorithm showed better results than its ancestors.

Jyotirmayee et al. [98] proposed a clustering method based on word sense disambiguation algorithm, the method used Word Sense Disambiguation (WSD) and Lesk algorithms to classify textual data by return the words identifiers in a Knowledge-Base and increase contextual overlap to increase accuracy of sense and context. The proposed algorithm tested on many datasets and obtained better results than KM and Vector Space Models (VSMs).

Shi et al. [99] improved clustering of incremental affinity propagation algorithm and semi-supervised learning by adjusting similarity matrix, and the results of the proposed algorithm performed very well compared to other clustering methods.

Nishant Agarwal [100] developed a real-time system using temporal clustering algorithm for detecting bursts in streaming data and for improving storage mechanism to perform evolutionary queries, the system analyses user behavioral to find related tasks, anomaly detection, and bursty patterns, and the used methods showed better hashtag precision results compared to others.

In Reference [101], the authors presented a Chinese text clustering method based on both the Self-Organizing Map (SOM) neural network and density. This algorithm is composed of two phases. Chinese text is converted into text vectors during the first phase, employed as SOM training data, and mapped through SOM training. An initial clustering result is obtained for text data, i.e., a set of virtual coordinates. Then, the virtual coordinates set is further clustered based on density during the second phase. This suggested technique is different from the current versions during the first phase. In addition, due to decreasing dimensions, it outperforms other methods in computation time in the second phase. Statistical experiments demonstrated the efficiency of the proposed method regarding clustering text data and high multi-dimensional data.

Dimension reduction contains two techniques: feature selection and feature extraction. The reduction of dimensions using the feature selection technique has a more significant effect on the cluster results than the feature extraction technique. To minimize dimensions, however, there would be a need for feature extraction techniques. For this purpose, the feature extraction approach requires an alternative method. The Self Organizing Map (SOM) is among the special artificial neural network models that can efficiently create spatial cognitive processes of input data or produce smaller data dimensions. This study proposed by Reference [102] examined the effects of SOM compared to Singular Value Decomposition (SVD) to reduce the data dimension of text documents prior to KM clustering. Findings

demonstrated that SVD remains better throughout the cluster efficiency index than SOM, yet SOM is faster in computation times than SVD.

3.2. Local Search Techniques

In this section are the most common local search techniques used to solve the text document clustering problems that have been published in the literature.

3.2.1. Heuristic Local Search

The paper developed a Local Search and KM (LSKM) text clustering algorithm by executing KM until converge and LSKM to calculate local extreme points [103]. The experiments showed better results than KM for clustering Big data datasets.

The paper implemented a new modified KM clustering algorithm using Max Term Contribution (MTC) to extract text character and dimension reduction [104]. The proposed algorithm calculates the contribution of each term in high dimension to extract the maximum contribution terms to find a low dimension using Simulated Annealing (SA) and modified KM clustering method, the algorithm showed better precision results than other algorithms.

The paper developed a new algorithm called LSKM stands to Cellular Automata Based Local Search and KM to identify regions of protein coding in non-overlapping and mixed exon-inton boundary DNA sequences [105]. The experimental showed better accuracy results compared to traditional algorithms. Reference [106] proposed a new algorithm called β-hill Feature Selection Text Clustering (β-FSTC) to get best informative features subset. The algorithm improved clustering using B-hill climbing algorithm, the experiment used four text dataset and showed better results compared to other algorithms.

The paper improved KM algorithm to solve fall KM into the local optimal using hierarchical agglomerative clustering algorithm and cosine similarity to measure the distance between the text and ensure the high quality of the center point [107]. The experiments showed good stability and better accuracy results than traditional algorithms.

The paper developed β-hill climbing technique to solve clustering problem by partitioning similar documents for placing them into the same cluster [108]. The β parameter performed a balance between global and local search methods in order to solve KM and k-medoid clustering problem; the experiments used eight standard benchmark datasets and achieved better results than other techniques.

3.2.2. K-Means (KM) Clustering Technique

In the text mining field, the Micro-blog hot topic discovery is one of the hot topic research areas. The low hot topic discovery happened because of the KM algorithm's distance function, which causes low clustering accuracy. In Reference [109], three different definitions are proposed to solve the hot topic discovery: distinguish between the body words and the title words, blend similarity-based distance, and place contribution-based weight. Further, several algorithms have been proposed to accomplish the text clustering. The biterm topic model (BTM) and global vectors for word representation (GloVe) similarity linear fusion are proposed to achieve the short text clustering. Jensen-Shannon divergence (JS) method was utilized to measure the text similarity based on the BTM algorithm. The Improved Word Mover's Distance (IWMD) is used to determine the GloVe word vector model text-similarity. Lastly, these two similarities approaches are linearly combined and applied as the distance function to achieve KM clustering. The result shows that the proposed method based on BTM and GloVe models significantly enhance the clustering accuracy comparing to traditional KM approaches.

In Reference [110], a text clustering approach is proposed based on a combination of the KM algorithm and Self Organizing Model (SOM). In the beginning, the text is preprocessed to process successfully. Then, the proposed approach used enhance the cluster center for the KM and determining the isolated point text. Guoping, Lin et al. propose an enhanced version of the Density-based spatial clustering of applications with noise

(DBSCAN) algorithm. The authors presented the enhancement to overcome the threading limitations and adopt the solution on a small part of data, which led to improving the text clustering. Moreover, the proposed work can enhance classification process accuracy. The result shows significant performance in text clustering [111].

This study used a modified approach based on the An Artificial Immune Network for Data Analysis (aiNet) algorithm to deal with high-dimensional data in text clustering [112]. This approach is based on cluster centers with a virtual practical correlative mapping method to neglect data vector dimensions. The data first cluster by KM to extract the High-dimensional text, then the text will be the input of the aiNet algorithm. In terms of increasing the selection of initial centers point, Wang, Yiyang, et al. present a combination between the KM algorithm and the agglomerative hierarchical clustering method. Furthermore, the text clustering iteration layered cohesion algorithm used to improve final cluster centers' efficiency. The proposed algorithm was verified based on the Sina micro-blogging samples and the micro-blogging theme analysis through the proposed approach, segmentation, and vector text design [113].

An approach is proposed in Reference [114] to enhance image retrieval by combining text and visual features. Image retrieval has two types based on the text-based retrieval, such as (keywords, descriptions, and caption) and content-based image retrieval (CBIR). CBIR method avoiding the textual image description, but image retrieves based on the content similarities (e.g., colors, shapes, and textures). In this paper, the authors proposed improving CBIR's performance using the KM algorithm by introducing a text-guided weighting system for visual features. KM algorithm uses to calculate the initial center to improve time elapsed and reduce the number of iterations.

In Reference [115], a novel approach is proposed to cluster the text documents. In the first stage, features are chosen by a genetic system. Then, the hybrid algorithm performs the clustering for the extracted keywords. The Must Link and Cannot Link algorithm (MLCL) used to identify the extracted keywords and create the initial clusters. Finally, the Gaussian parameters perform clusters. The Brown Corpus and Reuters-21578 datasets used to test the proposed method. The results show that the presented work improve the clustering performance better than other methods like fuzzy self-constructing feature.

A novel approach is proposed for text clustering based on three main features algorithms with dynamic dimension reductions and feature weight method to solve the text clustering. The text documents are divided into different related clusters regarding chosen informative features via an acceptable evaluation method. Feature selection approaches choose these informative features in every document. Algorithms, such as particle swarm optimization (PSO), harmony search (HS), and Genetic algorithm (GA), are used for feature selection. The introduced work is length feature weight (LFW) based on the term appearance and repetition of features in other documents. In terms of reducing the number of features, a unique dynamic dimension reduction (DDR) method has been proposed. The KM algorithm is used to cluster the text document based on features selected by DDR. This work shows an out-performance with a combination of KM and swarms algorithms; further, the experimental result applies to seven benchmarked datasets [116,117]. In Reference [118], the authors propose an approach to enhance the text clustering performance and obtain accurate evaluators. This method combines the similarity and distance measure based on the KM algorithm, called Multi-objective K-mean (MKM). MKM shows an improvement in text clustering because of obtaining the optimal initial clusters centers.

An enhanced KM technique is suggested for text clustering. The system firstly the text pre-processed and classified by inverted classify. Next, the classified (index) is using the Term Frequency–Inverse Document Frequency (TF-IDF), which is also used to build the term-document matrix. Then, extract the main feature from the matrix through the Latent Semantic Indexing algorithm. Further, the Pillar algorithm selects seeds. Finally, the KM algorithm works based on determining seeds. The result shows that the method shows out-performance comparing to the standard KM [119]. In Reference [120], the KM algorithm was modified to work on a heterogeneous dataset by using the Euclidean distance

algorithm. This approach proves that KM is efficient in analyzing the heterogeneous data clustering.

In Reference [121], a comparison study was presented to examine the clustering accuracy of similarity and dissimilarity measure based on eight clustering approaches. This study a significant performance in some similarity approaches, such as the Dice coefficient, Extended Jaccard, and cosine similarity. In Reference [122], we discuss that the KM algorithm measures the relationship among the data objects; however, the similarity or dissimilarity measure provides an accurate result. A comparison study was presented to examine the clustering accuracy of similarity and dissimilarity measures based on eight clustering approaches. The research shows significant performance in some similarity approaches, such as the Dice coefficient, Extended Jaccard, and cosine similarity.

In Reference [123], the authors analyze the KM algorithm performance and how the partitioned text document clustering is done. Further, the authors try to focus on choosing the KM algorithm K value (true value) as a disadvantage. The study concluded that the KM algorithm suffers from ambiguity and effect by noise and outliers, high dimensionality, etc. Yuan et al. introduce an enhanced KM to solve the arbitrary choosing initial clustering centroids issue, namely (DPMCSKM). The selection in the proposed method is based on the initial clustering center on density peaks. In addition, the MapReduce approach is used for the parallelization of the improved KM in terms of fulfilling large-scale calculations in the clustering process. The proposed solution shows an outperformance in clustering accuracy [124].

Liu et al. in Reference [125], study the text clustering for the Chinese language and present a method based on the KM algorithm and Evaluation approach, which replace the Cluster Head (CH) that calculated from the real text sample with the nearest text node in the entire text set. The Euclidean distance algorithm calculates the distance between a couples of text points. Finally, the Entropy method is used to evaluate the clustering effect. This approach shows a significant result in solving an empty cluster, finding the optimal CHs, and optimal clustering results.

Reference [126] introduced an improved Latent Dirichlet Allocation (LDA) text clustering algorithm based on the Short Text Clustering Algorithm (SKP) to semantic extraction and sentiment analysis for small Chinese micro-blog texts, called SKP-LDA. The sentiment word co-occurrence method was proposed to define the word bag. The short text is provided with emotional polarity. For clustering by the LDA method, the topic relation and the knowledge sets of unique topic words will be extracted and included in LDA. Next, Top30 unique words and hidden n topics are extracted from knowledge sets. Finally, the result of LDA clustering is clustered again by the KM algorithm. The SKP-LDA, compared with different approaches, such as Enriched LDA (ELDA), Latent Sentiment Model (LSM), Lifelong Topic Model (LTM), and the Joint Sentiment Topic (JST), proves to be a remarkable emotional topic clustering impact and semantic analysis capability.

3.2.3. C-Means Clustering Technique

C-means is one of the efficient clustering techniques used in the literature [127,128]. This paper proposed an attributed weighted fuzzy c-means algorithm to solve the problem of text clustering [129]. The method is based on identifying the weights of the attributed during the iterations of the c-means algorithm. The authors claim that this technique does not affect the overall algorithm performance. The algorithm is tested on a dataset of test documents and the results show that there is a good computation speedup and enhanced accuracy. The proposed algorithm can be reliability used of automatic documents abstracting.

This paper proposed a method based of Latent Semantic Analysis (LSA) and improved fuzzy c-means (FCM) algorithm to solve the problem of text clustering [130]. The proposed method used a new feature extraction technique that establishes word-text relation matrix and LSA for text sematic vectors. The method also employs genetic algorithm for optimizing clustering results. The experimental results of the proposed algorithm shown to have good precision and recall values.

3.3. Big Data Techniques

A pairwise text similarity method is used on massive data-sets with normalization of Term Frequency-Inverse Document Frequency (TF-IDF) method and Cosine Similarity metric [131]. It used MapReduce model processes parallel large data-sets and distribute algorithm on clusters, this enhanced scalability and speed of text processing compared with other traditional methods.

A SWCK-means algorithm is proposed to solve traditional text clustering algorithms, it explored KM clustering, Spark and Hadoop big data technique to solve efficiency of the high dimensional vectors [132]. The proposed algorithm reduced the text data dimensions by calculating the word vectors weights using Word2vec, and it identified the KMC initial cluster centers by clustering the weight data using Canopy algorithm in order to improve efficiency of Canopy and KMC parallel design; the proposed algorithm achieved better results, especially in dealing with a huge amount of data, than traditional algorithms.

The paper improved a modified KM algorithm using Hadoop platform using Max-Min-distance to find better initial centroids [133]. The improved algorithm showed faster result than the traditional algorithm. The paper used MapReduce model to design and implement the Minimum Spanning Tree (MST) clustering algorithm based on MST construction, graph construction, and feature extraction vector [134]. The proposed algorithm showed better scalability and accuracy results than MapReduce-based KM, but with less speed.

One paper implemented a new KM clustering algorithm using cosine similarity feature extraction [135]. The algorithm analyzed Hadoop platform for a large dataset in distributed system, and showed good results comparing others algorithms.

3.4. Hybrid Clustering Techniques

Document clustering is seen as an effective method for document organization and browsing in machine learning as it becomes an essential area of study. Fuzzy c-means (FCM), a highly exhaustive search method, has been widely used for categorization problems. As an optimization technique, however, it conveniently leads to local optimised clusters. The Particle Swarm (PSO) method is a heuristic algorithm optimization algorithm. In Reference [136], authors introduced a hybrid method based on fuzzy c-means and particle swarm optimization (PSO-FCM) to cluster text documents, which allow full utilization of the merits of both techniques. Not only would the PSO-FCM support the FCM clustering to escape from local optima, but it also overwhelms the limitations of the PSO algorithm's slow convergence speed. Experimental findings on two widely used data sets indicate that the introduced algorithm has better results than FCM and PSO methods.

There has been a rapid rise in the number of digitized text documents on the internet. It is crucial to group the documents into clusters for speedy retrieval of information regarding a high collection of web data. Document clustering is the set of documents in groups so that the documents in each group are identical to each other and not to documents belonging to those other groups. The performance of the results of the clustering depends significantly on the text classification and the clustering method. This work provided a comprehensive study of three techniques for clustering text documents utilizing WordNet, namely KM, Particle Swarm Optimization (PSO), and hybrid PSO + KM methods [54]. A bag of words is the standard way of describing a text document. The bag of terms is almost unsatisfactory because it does not take advantage of semantics. Texts are defined in the presented work in terms of synsets, referring to a word. WordNet synonyms also improve the bag of the terminology data representation of text. For Nepali language text clustering, KM, Particle Swarm Optimization (PSO), and hybrid PSO + KM methods are employed. Experimental work has been carried out using intra- and inter-cluster similarity.

This era could be stated as the era of zettabytes because of the fastest data growth. This leads to the design of an efficient method to organize data effectively. The need for this hour to handle all available data effectively is an important mechanism. Clustering is a method that helps in grouping together relevant documents. The authors conducted a

study in Reference [137] that considers the semantics and clusters the document with the hybrid of bisecting KM and the UPGMA method. Semantic analysis is made possible using a lexical database called WordNet. The results of the proposed methods are efficient, as the clusters are significant. Concerning accuracy, recall, F-measure, precision, and classification error, the efficiency of the presented technique is assessed. The experimental outcomes of the suggested study are satisfactory.

KM method can be considered as adaptive to the original points and easy to optimize in the local optimal. An enhanced GA-based CGHCM text clustering method is suggested in Reference [138] to prevent this issue. This algorithm has been evaluated to avoid falling into the local optimum, achieve efficient clustering performance, and obtain better results than literature.

Authors in Reference [139] presented a hybrid text clustering method based on dual particle swarm optimization and KM method; since the standard KM clustering method is sensitive to selecting initial cluster centers, the findings may correlate to the generic suboptimal solutions. It developed a self-adjusting weight vector technique that employed the optimal fitness shift rate to adjust the inertia weight automatically. In the evolutionary process, two populations utilized PSO depending on multiple inertia weight techniques. By sharing data between the two classes of offspring and offspring and parents to complete the evolution, two populations exchanged the best individual and removed the worst individual. This algorithm is called dual particle swarm optimization. The method merged global and local search capability to balance dual particle swarm optimization with effective KM. Every particle seems to have been a group of clustering centers, and fitness function was the reciprocal amount of scatter within the unit, then optimized with KM newborn particle. This was named a hybrid text clustering method based on dual particle swarm optimization and the KM method. Experimental results indicate that this method has high stability and efficient clustering compared to other text clustering methods, such as KM and PSO.

A text clustering method is an effective tool used to classify large amounts of text documents in groups. The size of documents influences text clustering by reducing its efficiency. Text documents consequently include sparse and uninformative features that decrease the efficiency of the underlying text clustering method and enhance computational time. Feature selection is a basic unsupervised learning method employed to select a better subset of appropriate text features to enhance text clustering efficiency and minimize computational time. To deal with the feature selection problem, authors in Reference [55] introduced a hybrid particle swarm optimization method with a genetic algorithm. KM is utilized to enhance the efficiency of the acquired features subsets. The experiments were carried out using eight datasets with different features. The findings demonstrate that by creating a new subset of more informative features, the introduced hybrid algorithm (H-FSPSOTC) enhanced the efficiency of the clustering method. The introduced method is compared with other techniques.

Krill herd (KH) is a new swarm-based optimization technique that when looking for food, mimics krill rationality [140]. To address the issue of text document clustering, an integration of objective features and the hybrid KH method, called MHKHA, is introduced [66]. The initial solutions of the KH algorithm are derived from the method of KM cluster analysis, and the decision about clustering is based on two mixed objective functions. Nine text datasets obtained from the Laboratory of Machine Learning were used to assess the efficiency of the proposed algorithm. Five measurement methods are used: precision, accuracy, recall, F-measure, and convergence behavior. The improved KH algorithm proposed is compared with various methods of clustering and thirteen algorithms. The improved KH algorithm proposed is compared with various methods of clustering and thirteen algorithms. The MHKHA performed best for all validation metrics and datasets used, compared to the other clustering methods evaluated.

The general classification of studied meta-heuristic optimization algorithms is given in Table 1. The number of published papers for the common meta-heuristic algorithms in the clustering domain is given in Figure 3.

Table 1. The general classification of meta-heuristic optimization algorithms.

Method Type	Single Objective	Multi-Objective
Evolutionary algorithms	Arithmatec Optimization Algorithm [141] Genetic Algorithm (GA) [69] Granular agent evolutionary algorithm [143] Evolutionary Strategy (ES) [145] Genetic Programming (GP) [84] Differential Evolution (DE) [97] Imperialist Competitive Algorithm (ICA) [147]	 NSGAII [142] SPEA, PESAII [144] Multi-objective ES [146] Multi-objective GP [82] Multi-objective DE [97]
Physical algorithms	Simulated Annealing (SA) [104] Memetic Algorithm (MA) [148] Harmony Search (HS) [96] Cultural Algorithm (CA) [150]	Multi-objective SA [104] Multi-objective MA [149] Multi-objective HS [96]
Swarm intelligence	Ant Colony Optimization (ACO) [80] Fish Swarm algorithm (FSA) [152] Artificial Bee Colony (ABC) [153] Particle Swarm Optimization (PSO) [136] Teaching Learning-based Optimization [154]	Multi-objective ACO [151] Multi-objective FSA [152] Multi-objective ABC [153] Multi-objective PSO [54]
Bio-inspired algorithms	Artificial Immune System (AIS) [155] Bacterial Foraging Optimization (BFO) [156] Krill Herd Algorithm (KHA) [70] Cuckoo Search (CS) Algorithm [62]	Multi-objective AIS [155] Multi-objective BFO [156]
Other meta-heuristic algorithms	Cat Swarm Opt. (CSO) [157] Invasive Weed Optimization Algorithm (IWO) [97] Cuckoo Search Algorithm [61] Gravitational Search Algorithm (GSA) [75] Firefly Algorithm (FA) [63] Bat Algorithm (BA) [14] Gray Wolf Optimizer (GWO) [60] Social Spider Optimization (SSO) [73]	Multi-objective CSO [157] Multi-objective IWO [97] Multi-objective Cuckoo [61] Multi-objective GSA[75] Multi-objective FA (MFA) [64] Multi-objective BA (MBA) [158]

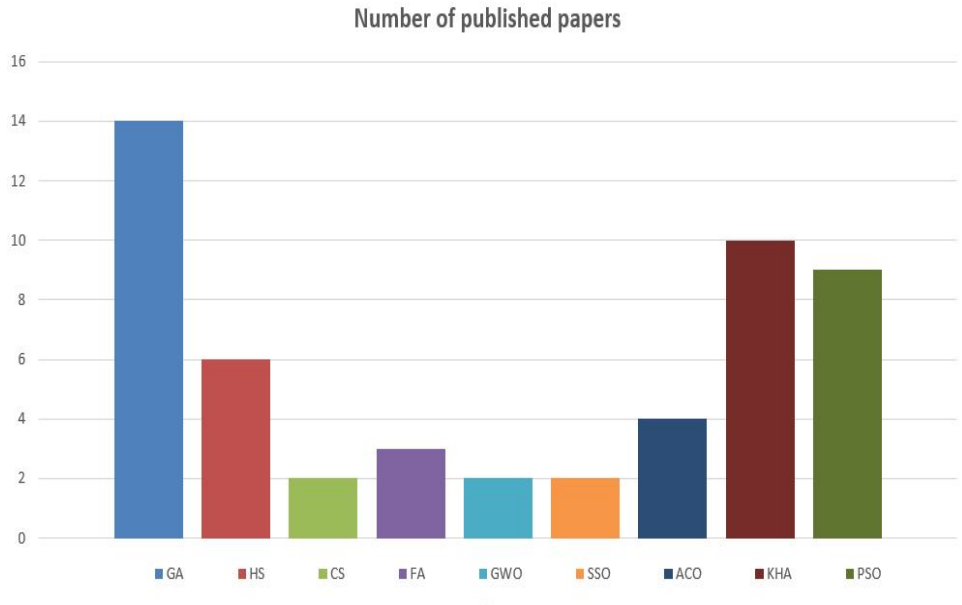

Figure 3. The number of published papers for the NI algorithms.

4. Evaluation Criteria Used in Text Clustering Applications

In the text clustering domain, there are several internal and external evaluation measures. These measures are given in the following subsections.

4.1. Internal Evaluation Criteria

Similarity and distance (dissimilarity) measures are the internal foundation for building the clustering optimization algorithms' procedures. As for quantitative results characteristics, the distance measure is favored to determine the correlation between data. Moreover, the similarity measure is favored when trading with qualitative data. Internal metrics are used without premonition of the text class mark to determine a collection of text documents to be given to their traditional cluster (i.e., cosine measure and Euclidean measure) [38,159].

4.1.1. Distance Measures

In this section, the commonly used distance measurements in the clustering applications are reviewed. Table 2 shows the common distance measurements.

Table 2. Distance measurements.

Name	Formula
Minkowski distance	$\left(\sum_{l=1}^{d}\left\|x_{li}-x_{lj}\right\|^{n}\right)^{1/n}$
Standardized Euclidean distance	$\left(\sum_{l=1}^{d}\left\|\frac{x_{li}-x_{lj}}{sl}\right\|^{2}\right)^{1/2}$
Cosine distance	$1-cos\ \alpha=\frac{x_i^T x_j}{\|x_i\|\|x_j\|}$
Pearson correlation distance	$1-\frac{cov(x_i,x_j)}{\sqrt{D(x_i)}\sqrt{D(x_j)}}$
Mahalanobis distance	$\sqrt{(x_i-x_j)^T\ S^{-1}(x_i-x_j)}$

4.1.2. Similarity Measures

In this section, the commonly used similarity measurements in the clustering applications are reviewed. Table 3 shows the common similarity measurements.

Table 3. Similarity measurements.

Name	Formula
Cosine similarity	$Cos(A,B)=\frac{A\cdot B}{\|A\|\times\|B\|}$
Jaccard similarity	$J(A,B)=\frac{\|A\cap B\|}{\|A\cup B\|}$
Sorensen similarity	$S(A,B)=\frac{2\times\|A\cap B\|}{\|A\|+\|B\|}$
Dice similarity	$Dice(A,B)=2\times\frac{\|A\cap B\|}{\|A\|+\|B\|}$

4.2. External Evaluation Criteria

The most popular evaluation measurements utilized in the text clustering domain are accuracy, purity, entropy, precision, recall, and F-measure [42,160,161]. The text clustering technique provides two sets of evaluation measures, namely internal and external measures [37]. External measurements are employed to assess the collected clusters' accuracy (correct) based on the given document's class labels in the dataset [47]. The following subsections describe the external evaluation measures used in assessing the output of the clustering algorithms.

4.2.1. Accuracy Measure

The accuracy test is employed to calculate the correct documents assigned to all groups in the given dataset [162–164]. This measure is determined using Equation (6).

$$AC = \frac{1}{n}\sum_{i=1}^{K} n_{i,i}, \tag{6}$$

where $n_{i,i}$ is the number of all correct candidiates of class i in cluster i, n is the number of all given documents, and K is the number of all given clusters in the dataset.

4.2.2. Purity Measure

The purity test is employed to calculate each cluster's percentage in a large class [165]. This test assigns each group to the most frequent class. An excellent value of purity is close to 1 because the percentage of large class sizes in each group, which is computed according to its size. Thus, the value of purity is in the interval $\left[\frac{1}{K^+}, 1\right]$. Equation (7) is utilized to determine the purity value of the cluster j:

$$P(c_j) = \frac{1}{n_j}\max_j n_{i,j}, \tag{7}$$

where $\max_j$ is the large class size in group j, $n_{i,j}$ is the number of all correct candidates of the class label i in cluster j, and n_j is the total number of members (documents) of cluster j. The purity test for all groups is determined using Equation (8):

$$P = \sum_{j=1}^{K} \frac{n_j}{n} P(c_j). \tag{8}$$

4.2.3. Entropy Measure

The entropy test examines the partitioning of class labels in each group [67,165,166]. This test centers on the containment of various cluster classes. A good sample has 0 entropy, indicating an excellent document clustering solution has a low entropy state. The entropy rate of cluster j according to the size of each group can be determined using Equation (9):

$$E(c_j) = -\sum_i p_{i,j} \log p_{i,j}, \tag{9}$$

where $p_{i,j}$ is the probability value of class' i members that belong to group j. The entropy test for all groups is determined using Equation (10):

$$E = -\sum_{j=1}^{K} \frac{n_j}{n} E(c_j). \tag{10}$$

4.2.4. Precision Measure

The precision (P) test for each cluster is calculated based on the given class label in the main datasets. The precision test is the ratio of important documents and the total number of documents in all groups [37,160]. Precision value for class i in cluster j is determined using Equation (11).

$$P(i,j) = \frac{n_{i,j}}{n_j}, \tag{11}$$

where $n_{i,j}$ is the number of correct candidates of the class labeled i in the group j, and n_j is the total number of objects in the group j.

4.2.5. Recall Measure

The recall (R) test for each cluster is calculated based on the assigned class label. The recall value is the ratio of important documents in all groups and the total number of relevant objects in the dataset [37,40]. Recall test for class i and cluster j is determined using Equation (12).

$$R(i,j) = \frac{n_{i,j}}{n_i}, \tag{12}$$

where $n_{i,j}$ is the number of correct candidates of the class i in group j, and n_i is the number of truly members of class i as the class labels given in the main dataset.

4.2.6. F-Measure

The F-measure aims to assess clusters of the examined partition clusters at the largest match of the class label partition clusters. This measure is a popular evaluation criterion in the clustering domain based on the aggregation of precision and recalls tests [37,39,160]. The F-measure value for group j is determined using Equation (13):

$$F(j) = \frac{2 \times P(i,j) \times R(i,j)}{P(i,j) + R(i,j)}, \tag{13}$$

where $P(i,j)$ is the precision of candidates of class i in group j, and $R(i,j)$ is the recall of candidates of class i in group j. Moreover, the F-measure for all clusters is determined using Equation (14).

$$F = \frac{\sum_{j=1}^{K} F_j}{K} \tag{14}$$

5. Theoretical Discussions

Clustering analysis is an essential mechanism that examines unlabeled data by constructing a hierarchical structure or developing a collection of clusters based on a predefined number of clusters. This process involves steps, arranging of pre-processing and algorithm progress, and solution efficacy and evaluation. Each one is tightly linked to each other and uses severe difficulties in scientific developments. The meta-heuristic optimization algorithms developed by different research communities intend to solve various clustering processes and have pros and cons. Nevertheless, we have already discussed many successful cluster analysis applications. There remain many open problems because of the presence of many inherent possible circumstances. These problems have already drawn and will proceed to draw intensive applications from broad disciplines.

We review and arrange the survey by listing some critical issues and potential research trends for optimization cluster algorithms.

- No universally clustering algorithm can be utilized to solve all clustering problems. Typically, algorithms are produced with specific theories and support some biases. For this reason, it is not reasonable to say "best" in the circumstances of clustering algorithms, although some observations are conceivable. These examples are often based on particular applications under specific requirements, and the results may match quite differently if the conditions vary.
- New mechanisms have produced more complicated and challenging tasks, needing more robust clustering algorithms. The following features are essential to the performance and efficiency of a novel clustering algorithm. Create random patterns of clusters rather than be restricted to some selective pattern; manage a big amount of data, as well as high-dimensional characteristics, with satisfactory storage and time complexities; identify and eliminate potential noise and outliers; reduce the confidence of clustering algorithms on users adjusting parameters; have the ability of trading with anew happening data without relearning from scratch; be protected from the impacts of order of input clusters; give some prudence for the number of possible groups

without prior information; give dependable data visualization and give users results that can clarify the analysis; and the ability to manage both statistical and simple data or be quickly flexible to some other data representation. Of course, some more particular specifications for unique applications will influence these characteristics.

- Feature selection, extraction, and cluster validation are crucial as the clustering algorithms at the post-processing and pre-processing stages. Choosing relevant and essential features can significantly reduce the difficulty of subsequent schemes, and result evaluations indicate the level of trust to which we can depend on the produced clusters. Unfortunately, both methods lack universal leadership. Finally, the tradeoff among various criteria and processes are still reliant on the applications themselves.

6. Conclusions and Future Works

More than 100 research papers have been reviewed in this survey paper to discover the robustness and weaknesses of meta-heuristic optimization algorithms. This paper summarizes the entire literature until the end of the year 2020 exhaustively. Most of the gathered papers describe the optimization methods that have been used in text clustering applications. Several variants of algorithms, including standard, basic, modified, hybrid methods, and others, are studied.

Meta-heuristic optimization algorithms proved its performance in solving various kinds of text clustering problems. However, local optima can be trapped because of its focus on exploration (i.e., global search) instead of exploitation (i.e., local search). This issue may be improved over time, as to how well the sets of rules governing various search algorithms work are better understood. There are two main problems in the text clustering application: the initial cluster centroids and the number of clusters. Parameter tuning will also play a critical role in future studies since the parameters' values and settings govern the algorithm's overall performance. From this discussion, we see that meta-heuristic optimization algorithms are robustly feasible for continuing use in machine learning domains. This survey helps the researcher work in this domain by describing how these algorithms have been employed, pointing out its weaknesses and strengths, and proving its effectiveness.

Finally, we suggest new future research directions on text clustering-based meta-heuristic optimization algorithms. The most vital features of these algorithms (i.e., GA, GWO, KHA, PSO, and HSA) might be blended for better overall performance in solving the text clustering problems. New hybrid and modified algorithms can be proposed to solve the text clustering problems. Moreover, several new meta-heuristic optimization algorithms have been proposed recently, which can be utilized to solve the clustering problems. New hybrid and modified algorithms can be proposed to solve the text clustering problems. Moreover, several new meta-heuristic optimization algorithms have been proposed recently, which can be utilized to solve the clustering problems. These algorithms are Slime Mold Algorithm, Marine Predators Algorithm, Equilibrium Optimizer, Sine Cosine Algorithm, Salp Swarm Algorithm, Harris Hawks Optimization, Henry Gas Solubility Optimization, Arithmetic Optimization Algorithm (AOA), and others.

Author Contributions: L.A.: Conceptualization, supervision, methodology, formal analysis, resources, data curation, writing—original draft preparation. A.H.G.: Conceptualization, supervision, writing—review and editing, project administration, funding acquisition. M.A.E.: Conceptualization, writing–review and editing, supervision. H.A.H.: Conceptualization, writing—review and editing, supervision. M.O.: Conceptualization, writing—review and editing. M.A.: Conceptualization, writing—review and editing. A.M.K.: Conceptualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hruschka, E.R.; Campello, R.J.; Freitas, A.A. A survey of evolutionary algorithms for clustering. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2009**, *39*, 133–155. [CrossRef]
2. Shi, T.T.; Liu, X.R.; Li, J.J. Market segmentation by travel motivations under a transforming economy: Evidence from the Monte Carlo of the Orient. *Sustainability* **2018**, *10*, 3395. [CrossRef]
3. Abualigah, L.; Bashabsheh, M.Q.; Alabool, H.; Shehab, M. Text Summarization: A Brief Review. In *Recent Advances in NLP: The Case of Arabic Language*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–15.
4. Jang, B.; Kim, M.; Harerimana, G.; Kang, S.; Kim, J.W. Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism. *Appl. Sci.* **2020**, *10*, 5841. [CrossRef]
5. Abualigah, L.; Diabat, A.; Geem, Z.W. A Comprehensive Survey of the Harmony Search Algorithm in Clustering Applications. *Appl. Sci.* **2020**, *10*, 3827. [CrossRef]
6. Hoeppner, F. Fuzzy shell clustering algorithms in image processing: fuzzy c-rectangular and 2-rectangular shells. *IEEE Trans. Fuzzy Syst.* **1997**, *5*, 599–613. [CrossRef]
7. José-García, A.; Gómez-Flores, W. Automatic clustering using nature-inspired metaheuristics: A survey. *Appl. Soft Comput.* **2016**, *41*, 192–213. [CrossRef]
8. Lee, L.H.; Wan, C.H.; Rajkumar, R.; Isa, D. An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization. *Appl. Intell.* **2012**, *37*, 80–99. [CrossRef]
9. Khasawneh, A.M.; Kaiwartya, O.; Abualigah, L.M.; Lloret, J. Green computing in underwater wireless sensor networks pressure centric energy modeling. *IEEE Syst. J.* **2020**, *14*, 4735–4745. [CrossRef]
10. Krishnapuram, R.; Joshi, A.; Nasraoui, O.; Yi, L. Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Trans. Fuzzy Syst.* **2001**, *9*, 595–607. [CrossRef]
11. Abualigah, L.; Alfar, H.E.; Shehab, M.; Hussein, A.M.A. Sentiment Analysis in Healthcare: A Brief Review. In *Recent Advances in NLP: The Case of Arabic Language*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 129–141.
12. Abd Elaziz, M.; Li, L.; Jayasena, K.N.; Xiong, S. Multiobjective big data optimization based on a hybrid salp swarm algorithm and differential evolution. *Appl. Math. Model.* **2020**, *80*, 929–943. [CrossRef]
13. Higham, D.J.; Kalna, G.; Kibble, M. Spectral clustering and its use in bioinformatics. *J. Comput. Appl. Math.* **2007**, *204*, 25–37. [CrossRef]
14. Alomari, O.A.; Khader, A.T.; Al-Betar, M.A.; Abualigah, L.M. MRMR BA: A hybrid gene selection algorithm for cancer classification. *J. Theor. Appl. Inf. Technol.* **2017**, *95*, 2610–2618.
15. Alomari, O.A.; Khader, A.T.; Al-Betar, M.A.; Abualigah, L.M. Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm. *Int. J. Data Min. Bioinform.* **2017**, *19*, 32–51. [CrossRef]
16. Manuel, A.J.; Deverajan, G.G.; Patan, R.; Gandomi, A.H. Optimization of Routing-Based Clustering Approaches in Wireless Sensor Network: Review and Open Research Issues. *Electronics* **2020**, *9*, 1630. [CrossRef]
17. Nanda, S.J.; Panda, G. A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm Evol. Comput.* **2014**, *16*, 1–18. [CrossRef]
18. Mahata, N.; Kahali, S.; Adhikari, S.K.; Sing, J.K. Local contextual information and Gaussian function induced fuzzy clustering algorithm for brain MR image segmentation and intensity inhomogeneity estimation. *Appl. Soft Comput.* **2018**, *68*, 586–596. [CrossRef]
19. Harrigan, K.R. An application of clustering for strategic group analysis. *Strateg. Manag. J.* **1985**, *6*, 55–73. [CrossRef]
20. Chen, S.; Mulgrew, B.; Grant, P.M. A clustering technique for digital communications channel equalization using radial basis function networks. *IEEE Trans. Neural Netw.* **1993**, *4*, 570–590. [CrossRef]
21. Bien, Z.Z.; Lee, H.E. Effective learning system techniques for human–robot interaction in service environment. *Knowl.-Based Syst.* **2007**, *20*, 439–456. [CrossRef]
22. Sornette, D.; Werner, M.J. Apparent clustering and apparent background earthquakes biased by undetected seismicity. *J. Geophys. Res. Solid Earth* **2005**, *110*. [CrossRef]
23. Özyer, T.; Alhajj, R. Parallel clustering of high dimensional data by integrating multi-objective genetic algorithm with divide and conquer. *Appl. Intell.* **2009**, *31*, 318. [CrossRef]
24. Srivastava, A.N.; Sahami, M. *Text Mining: Classification, Clustering, and Applications*; CRC Press: Boca Raton, FL, USA, 2009.
25. Nanda, S.; Mahanty, B.; Tiwari, M. Clustering Indian stock market data for portfolio management. *Expert Syst. Appl.* **2010**, *37*, 8793–8798. [CrossRef]
26. Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **1999**, *6*, 281–297. [CrossRef] [PubMed]
27. Safaldin, M.; Otair, M.; Abualigah, L. Improved binary gray wolf optimizer and SVM for intrusion detection system in wireless sensor networks. *J. Ambient. Intell. Humaniz. Comput.* **2020**, 1–18. doi:10.1007/s12652-020-02228-z. [CrossRef]
28. Brulles, D.; Peters, S.J.; Saunders, R. Schoolwide mathematics achievement within the gifted cluster grouping model. *J. Adv. Acad.* **2012**, *23*, 200–216. [CrossRef]
29. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv. (CSUR)* **1999**, *31*, 264–323. [CrossRef]
30. Alshaer, H.N.; Otair, M.A.; Abualigah, L.; Alshinwan, M.; Khasawneh, A.M. Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application. *Multimed. Tools Appl.* **2020**, 1–18, doi:10.1007/s11042-020-10074-6. [CrossRef]

31. Falkenauer, E. *Genetic Algorithms and Grouping Problems*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1998.
32. Rayward-Smith, V.J. Metaheuristics for clustering in KDD. In Proceedings of the 2005 IEEE Congress on Evolutionary Computation, Edinburgh, UK, 2–5 September 2005; Volume 3, pp. 2380–2387.
33. Ghiasi, S.; Srivastava, A.; Yang, X.; Sarrafzadeh, M. Optimal energy aware clustering in sensor networks. *Sensors* **2002**, *2*, 258–269. [CrossRef]
34. Fahad, A.; Alshatri, N.; Tari, Z.; Alamri, A.; Khalil, I.; Zomaya, A.Y.; Foufou, S.; Bouras, A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Trans. Emerg. Top. Comput.* **2014**, *2*, 267–279. [CrossRef]
35. Bharti, K.K.; Singh, P.K. Chaotic gradient artificial bee colony for text clustering. *Soft Comput.* **2015**, *20*, 1113–1126.
36. Prabha, K.A.; Visalakshi, N.K. Improved Particle Swarm Optimization Based K-Means Clustering. In Proceedings of the IEEE 2014 International Conference Intelligent Computing Applications (ICICA), Coimbatore, India, 6–7 March 2014; pp. 59–63.
37. Forsati, R.; Mahdavi, M.; Shamsfard, M.; Meybodi, M.R. Efficient stochastic algorithms for document clustering. *Inf. Sci.* **2013**, *220*, 269–291. [CrossRef]
38. Forsati, R.; Keikha, A.; Shamsfard, M. An improved bee colony optimization algorithm with an application to document clustering. *Neurocomputing* **2015**, *159*, 9–26. [CrossRef]
39. Basu, T.; Murthy, C. A similarity assessment technique for effective grouping of documents. *Inf. Sci.* **2015**, *311*, 149–162. [CrossRef]
40. Bharti, K.K.; Singh, P.K. Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering. *Appl. Soft Comput.* **2016**, *43*, 20–34. [CrossRef]
41. Zhong, N.; Li, Y.; Wu, S.T. Effective pattern discovery for text mining. *Knowl. Data Eng. IEEE Trans.* **2012**, *24*, 30–44. [CrossRef]
42. Bharti, K.K.; Singh, P.K. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Syst. Appl.* **2015**, *42*, 3105–3114. [CrossRef]
43. Salton, G.; Wong, A.; Yang, C.S. A vector space model for automatic indexing. *Commun. ACM* **1975**, *18*, 613–620. [CrossRef]
44. De Vries, C.M. Document Clustering Algorithms, Representations and Evaluation for Information Retrieval. Ph.D. Thesis, Queensland University of Technology, Brisbane City, QLD, Australia, 2014.
45. Abualigah, L.M.Q.; Hanandeh, E.S. Applying genetic algorithms to information retrieval using vector space model. *Int. J. Comput. Sci. Eng. Appl.* **2015**, *5*, 19.
46. Hong, S.S.; Lee, W.; Han, M.M. The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification. *Int. J. Adv. Soft Comput. Appl.* **2015**, *7*, 2074–8523.
47. Mahdavi, M.; Abolhassani, H. Harmony K-means algorithm for document clustering. *Data Min. Knowl. Discov.* **2009**, *18*, 370–391. [CrossRef]
48. Ghanem, O.; Alhanjouri, M. Evaluating the Effect of Preprocessing in Arabic Documents Clustering. Ph.D. Thesis, Computer Engineering Department, Islamic University of Gaza, Gaza, Palestine, 2014.
49. Forsati, R.; Mahdavi, M. Web text mining using harmony search. In *Recent Advances in Harmony Search Algorithm*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 51–64.
50. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *J. Comput. Sci.* **2018**, *25*, 456–466. [CrossRef]
51. Baloochian, H.; Nazari, M. Clustering-Based Text Improvement and Summarization Based on Collective Intelligence Algorithm. *Spec. J. Electron. Comput. Sci.* **2018**, *4*, 7–15.
52. Chen, H.N.; He, B.; Yan, L.; Li, J.; Ji, W. A text clustering method based on two-dimensional OTSU and PSO algorithm. In Proceedings of the 2009 IEEE International Symposium on Computer Network and Multimedia Technology, Wuhan, China, 18–20 January 2009; pp. 1–4.
53. Wu, J.W.; Tseng, J.C.; Tsai, W.N. A hybrid linear text segmentation algorithm using hierarchical agglomerative clustering and discrete particle swarm optimization. *Integr. Comput.-Aided Eng.* **2014**, *21*, 35–46. [CrossRef]
54. Sarkar, S.; Roy, A.; Purkayastha, B. A comparative analysis of particle swarm optimization and K-means algorithm for text clustering using Nepali Wordnet. *Int. J. Nat. Lang. Comput. (IJNLC)* **2014**, *3*, doi:10.5121/ijnlc.2014.3308. [CrossRef]
55. Abualigah, L.M.; Khader, A.T. Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. *J. Supercomput.* **2017**, *73*, 4773–4795. [CrossRef]
56. Lee, J.S.; Hah, H.; Park, S.C. Less-redundant text summarization using ensemble clustering algorithm based on GA and PSO. *Wseas Trans. Comput.* **2017**, *16*, 10–17.
57. Janani, R.; Vijayarani, S. Text document clustering using spectral clustering algorithm with particle swarm optimization. *Expert Syst. Appl.* **2019**, *134*, 192–200. [CrossRef]
58. Purushothaman, R.; Rajagopalan, S.; Dhandapani, G. Hybridizing Gray Wolf Optimization (GWO) with Grasshopper Optimization Algorithm (GOA) for text feature selection and clustering. *Appl. Soft Comput.* **2020**, *96*, 106651. [CrossRef]
59. Vidyadhari, C.; Sandhya, N.; Premchand, P. Particle Grey Wolf Optimizer (PGWO) Algorithm and Semantic Word Processing for Automatic Text Clustering. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2019**, *27*, 201–223. [CrossRef]
60. Rashaideh, H.; Sawaie, A.; Al-Betar, M.A.; Abualigah, L.M.; Al-Laham, M.M.; Ra'ed, M.; Braik, M. A grey wolf optimizer for text document clustering. *J. Intell. Syst.* **2018**, *29*, 814–830. [CrossRef]
61. Jain, V.; Shrivastava, N.; PCST, B. Class Based Clustering with Cuckoo Search Rank Optimization for Text Data Categorization. *Int. J. Master Eng. Res. Technol.* **2015**, *2*, 82–87.

62. Ishak Boushaki, S.; Kamel, N.; Bendjeghaba, O. High-dimensional text datasets clustering algorithm based on cuckoo search and latent semantic indexing. *J. Inf. Knowl. Manag.* **2018**, *17*, 1850033. [CrossRef]
63. Mohammed, A.J.; Yusof, Y.; Husni, H. Integrated bisect K-means and firefly algorithm for hierarchical text clustering. *J. Eng. Appl. Sci.* **2016**, *11*, 522–527.
64. Mohammed, A.J.; Yusof, Y.; Husni, H. GF-CLUST: A nature-inspired algorithm for automatic text clustering. *J. Inf. Commun. Technol. (JICT)* **2016**, *15*, 57–81. [CrossRef]
65. Le, H.P.; Nguyen, T.D.; Park, J.; Lee, G. Combining Fuzzy C-means Clustering and Flood Filling Algorithm for Enhancing Text Binarization. *J. Korean Multimed. Soc.* **2009**, 333–336. Available online: https://www.semanticscholar.org/paper/Combining-Fuzzy-C-means-Clustering-and-Flood-for-Le-Nguy%C3%AAn/26691a4cb30b68b0e3435dacc07556481062b326 (accessed on 31 December 2020).
66. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. *Eng. Appl. Artif. Intell.* **2018**, *73*, 111–125. [CrossRef]
67. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A.; Awadallah, M.A. A krill herd algorithm for efficient text documents clustering. In Proceedings of the 2016 IEEE symposium on computer applications & industrial electronics (ISCAIE), Batu Feringghi, Malaysia, 30–31 May 2016; pp. 67–72.
68. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S.; Gandomi, A.H. A novel hybridization strategy for krill herd algorithm applied to clustering techniques. *Appl. Soft Comput.* **2017**, *60*, 423–435. [CrossRef]
69. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. Hybrid clustering analysis using improved krill herd algorithm. *Appl. Intell.* **2018**, *48*, 4047–4071. [CrossRef]
70. Abualigah, L.M.Q. *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*; Springer: Berlin/Heidelberg, Germany, 2019.
71. Abualigah, L.; Alsalibi, B.; Shehab, M.; Alshinwan, M.; Khasawneh, A.M.; Alabool, H. A parallel hybrid krill herd algorithm for feature selection. *Int. J. Mach. Learn. Cybern.* **2020**, 1–24. [CrossRef]
72. Chandran, T.R.; Reddy, A.; Janet, B. A social spider optimization approach for clustering text documents. In Proceedings of the 2016 IEEE 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, India, 27–28 February 2016; pp. 22–26.
73. Chandran, T.R.; Reddy, A.; Janet, B. Text clustering quality improvement using a hybrid social spider optimization. *Int. J. Appl. Eng. Res.* **2017**, *12*, 995–1008.
74. Rashedi, E.; Nezamabadi-Pour, H.; Saryazdi, S. GSA: A gravitational search algorithm. *Inf. Sci.* **2009**, *179*, 2232–2248. [CrossRef]
75. Mirhosseini, M. A clustering approach using a combination of gravitational search algorithm and k-harmonic means and its application in text document clustering. *Turk. J. Electr. Eng. Comput. Sci.* **2017**, *25*, 1251–1262. [CrossRef]
76. Mirjalili, S.; Lewis, A. The whale optimization algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [CrossRef]
77. Gopal, J.; Brunda, S. Text Clustering Algorithm Using Fuzzy Whale Optimization Algorithm. *Int. J. Intell. Eng. Syst.* **2019**, *12*.. [CrossRef]
78. Ma, S.X.; Liu, D.; Jia, S.J. Text Clustering Algorithm Based on Ant Colony Algorithm. *Comput. Eng.* **2010**, *8*. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-JSJC201008074.htm (accessed on 30 December 2020).
79. Nema, P.; Sharma, V. Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique. In Proceedings of the 2015 IEEE International Conference on Computers, Communications, and Systems (ICCCS), Kanyakumari, India, 2–3 November 2015; pp. 1–5.
80. Cobo, A.; Rocha, R. Document management with ant colony optimization metaheuristic: A fuzzy text clustering approach using pheromone trails. In *Soft Computing in Industrial Applications*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 261–270.
81. Mustafi, D.; Sahoo, G. A hybrid approach using genetic algorithm and the differential evolution heuristic for enhanced initialization of the k-means algorithm with applications in text clustering. *Soft Comput.* **2019**, *23*, 6361–6378. [CrossRef]
82. Song, W.; Li, C.H.; Park, S.C. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. *Expert Syst. Appl.* **2009**, *36*, 9095–9104. [CrossRef]
83. Chun-hong, W.; Li-Li, N.; Yao-Peng, R. Research on the text clustering algorithm based on latent semantic analysis and optimization. In Proceedings of the 2011 IEEE International Conference on Computer Science and Automation Engineering, Shanghai, China, 10–12 June 2011; Volume 4, pp. 470–473.
84. Shi, K.; Li, L. High performance genetic algorithm based text clustering using parts of speech and outlier elimination. *Appl. Intell.* **2013**, *38*, 511–519. [CrossRef]
85. Karaa, W.B.A.; Ashour, A.S.; Sassi, D.B.; Roy, P.; Kausar, N.; Dey, N. Medline text mining: An enhancement genetic algorithm based approach for document clustering. In *Applications of Intelligent Optimization in Biology and Medicine*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 267–287.
86. Garg, N.; Gupta, R. Performance Evaluation of New Text Mining Method Based on GA and K-Means Clustering Algorithm. In *Advanced Computing and Communication Technologies*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 23–30.
87. Wang, M.-C.; Wang Z.-O. Text Fuzzy Clustering Algorithm Based on Rough Set and Genetic Algorithm. *J. Electron. Inf. Technol.* **2005**, *4*. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-DZYX200504011.htm (accessed on 30 December 2020).
88. Yu, S.Y.; Ding, H.F.; Fu, Z.C. Study on text categorization based on genetic algorithm and fuzzy clustering. *Comput. Technol. Dev.* **2009**, *4*. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-WJFZ200904037.htm (accessed on 30 December 2020).

89. Tohti, T.; Ablat, A.; Aniwar, M.; Hamdulla, A. Combined algorithm of GAAC and K-means for Uyghur text clustering. *Comput. Eng. Sci.* **2013**, *7*, 30.
90. Dong, Y.H.; Guo, S.C. Text clustering algorithm with improved weighting factor and feature vector. *Comput. Eng. Des.* **2015**, *4*, 42.
91. Shao, Z.; Li, Y.; Wang, X.; Zhao, X.; Guo, Y. Research on a New Automatic Generation Algorithm of Concept Map Based on Text Clustering and Association Rules Mining. In *International Conference on Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 479–490.
92. Sailaja, V.; Srinivasa Rao, K.; Reddy, K. Text independent speaker identification with finite multivariate generalized gaussian mixture model and hierarchical clustering algorithm. *Int. J. Comput. Appl.* **2010**, *11*, 975–8887. [CrossRef]
93. Zeng, A.; Huang, Y. A text classification algorithm based on rocchio and hierarchical clustering. In *International Conference on Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 432–439.
94. Lokhande, M.M.P.; Gawande, M.N.; Koprade, M.S.; Bewoor, M.M. Text summarization using hierarchical clustering algorithm and expectation maximization clustering algorithm. *Int. J. Comput. Eng. Technol. (IJCET)* **2015**, *6*, 58–65.
95. Rong, Y. Staged text clustering algorithm based on K-means and hierarchical agglomeration clustering. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; pp. 124–127.
96. Abualigah, L.M.; Khader, A.T.; AlBetar, M.A.; Hanandeh, E.S. Unsupervised text feature selection technique based on particle swarm optimization algorithm for improving the text clustering. In Proceedings of the 1st EAI International Conference on Computer Science and Engineering. European Alliance for Innovation (EAI), Penang, Malaysia, 11–12 November 2016; p. 169.
97. Lipeng, Y.; Fuzhang, W.; Chunmei, F. A Text Clustering Algorithm based on Weeds and Differential Optimization. *Int. J. Database Theory Appl.* **2016**, *9*, 121–130.
98. Choudhury, J.; Kimtani, D.K.; Chakrabarty, A. Text clustering using a WordNet-based knowledge-base and the Lesk Algorithm. *Int. J. Comput. Appl.* **2012**, *48*, 20–24.
99. Shi, X.; Guan, R.; Wang, L.; Pei, Z.; Liang, Y. An incremental affinity propagation algorithm and its applications for text clustering. In Proceedings of the 2009 IEEE International Joint Conference on Neural Networks, Atlanta, GA, USA, 14–19 June 2009; pp. 2914–2919.
100. Agarwal, N. A Real-time Temporal Clustering Algorithm for Short Text, and Its Applications. Ph.D. Thesis, University of California San Diego, La Jolla, CA, USA, 2017.
101. Meng, Z.; Zhu, H.; Zhu, Y.; Zhou, G. A clustering algorithm for Chinese text based on SOM neural network and density. In *International Symposium on Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 251–256.
102. Jambak, M.I.; Jambak, A.I.I. Comparison of dimensional reduction using the Singular Value Decomposition Algorithm and the Self Organizing Map Algorithm in clustering result of text documents. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2019; Volume 551, p. 12046.
103. Liu, X. An Improved K-Means Text Clustering Algorithm Based on Local Search. In Proceedings of the 2008 IEEE 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, China, 12–14 October 2008; pp. 1–4.
104. Guoli, L.; Xiaohua, W.; Rongbo, W. Text Clustering Research on the Max Term Contribution Dimension Reduction and Simulated Annealing Algorithm. *Data Anal. Knowl. Discov.* **2008**, *24*, 43–47.
105. Sree, P.K.; Raju, G.; Raju, S.V.; Devi, N.U. NTCA: A Novel Text Clustering Algorithm Build on Cellular automata Based local search and K-Means Algorithm For Identifying the Protein Coding Regions in Genomic DNA. In Proceedings of the International Congress for Global Science and Technology, 2008; p. 39. Available online: https://www.researchgate.net/profile/Ashraf_Aboshosha/publication/283713969_AIML-Volume8-issue1-P1121546431/links/564449a608ae54697fb6b751.pdf#page=43 (accessed on 30 December 2020).
106. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A.; Alyasseri, Z.A.A.; Alomari, O.A.; Hanandeh, E.S. Feature selection with β-hill climbing search for text clustering application. In Proceedings of the 2017 IEEE Palestinian International Conference on Information and Communication Technology (PICICT), Gaza City, Palestine, 8–9 May 2017; pp. 22–27.
107. Qian, S.Y.; Liu, H.H.; Li, D.Y. Research and Application of Improved K-means Algorithm in Text Clustering. *DEStech Trans. Comput. Sci. Eng.* **2018**. [CrossRef]
108. Abualigah, L.M.; Hanandeh, E.S.; Khader, A.T.; Otair, M.A.; Shandilya, S.K. An Improved B-hill Climbing Optimization Technique for Solving the Text Documents Clustering Problem. *Curr. Med Imaging* **2020**, *16*, 296–306. [CrossRef]
109. Wu, D.; Zhang, M.; Shen, C.; Huang, Z.; Gu, M. BTM and GloVe Similarity Linear Fusion-Based Short Text Clustering Algorithm for Microblog Hot Topic Discovery. *IEEE Access* **2020**, *8*, 32215–32225. [CrossRef]
110. Xinwu, L. Research on Text Clustering Algorithm Based on K_means and SOM. In Proceedings of the 2008 IEEE International Symposium on Intelligent Information Technology Application Workshops, Shanghai, China, 21–22 December 2008; pp. 341–344.
111. Guoping, L. Algorithm of Web Text Classification Based on Hierarchical and Density Clustering. *J. Taiyuan Norm. Univ. (Nat. Sci. Ed.)* **2008**, *3*, 16.
112. Yan-Ting, X.Y.S.L.; Jia-Ning, X. The Two-stage Text Clustering Algorithm Based on K-mesans and aiNet. *Microcomput. Inf.* **2009**, *2009*, 81.
113. Wang, Y.; Wang, L.; Qi, J.; Qian, Z.; Xu, B.; Lei, C.; Yang, Y.; Cai, H. Improved text clustering algorithm and application in microblogging public opinion analysis. In Proceedings of the 2013 IEEE Fourth World Congress on Software Engineering, Hong Kong, 3–4 December 2013; pp. 27–31.

114. Nisha, S.N.; Ban, M.K.M.; Student, P.; Svcet, P. An Enhanced Image Retrieval Using K-Mean Clustering Algorithm in Integrating Text and Visual Features. Available online: http://www.ijiset.com/v1s1/IJISET_V1_I1_03.pdf (accessed on 30 December 2020).
115. Rose, J.D.; Dev, D.D.; Robin, C.R. A novel approach for text clustering using must link and cannot link algorithm. *J. Theor. Appl. Inf. Technol.* **2014**, *60*. Available online: http://www.jatit.org/volumes/Vol60No1/10Vol60No1.pdf (accessed on 30 December 2020).
116. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A.; Alomari, O.A. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. *Expert Syst. Appl.* **2017**, *84*, 24–36. [CrossRef]
117. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. A novel weighting scheme applied to improve the text document clustering techniques. In *Innovative Computing, Optimization and its Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 305–320.
118. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A. Multi-objectives-based text clustering technique using K-mean algorithm. In Proceedings of the 2016 IEEE 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–14 July 2016; pp. 1–6.
119. Adinugroho, S.; Sari, Y.A.; Fauzi, M.A.; Adikara, P.P. Optimizing K-means text document clustering using latent semantic indexing and pillar algorithm. In Proceedings of the 2017 IEEE 5th International Symposium on Computational and Business Intelligence (ISCBI), Dubai, UAE, 11–14 August 2017; pp. 81–85.
120. Jain, H.; Grover, R.; LIET, A. Clustering Analysis with Purity Calculation of Text and SQL Data using K-means Clustering Algorithm. *IJAPRR* **2017**, *4*, 47–58.
121. Jia, Y.; Kwong, S.; Hou, J.; Wu, W. Semi-supervised non-negative matrix factorization with dissimilarity and similarity regularization. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*. [CrossRef]
122. Afzali, M.; Kumar, S. An Extensive Study of Similarity and Dissimilarity Measures Used for Text Document Clustering using K-means Algorithm. *I.J. Inf. Technol. Comput. Sci.* **2018**, *9*, 64–73. [CrossRef]
123. Naeem, S.; Wumaier, A. Study and implementing K-mean clustering algorithm on English text and techniques to find the optimal value of K. *Int. J. Comput. Appl.* **2018**, *182*, 7–14. [CrossRef]
124. YUAN, Y.; LIU, H.; LI, H. An Improved K-Means Text Clustering Algorithm Based on Density Peaks and Its Parallelization. *J. Wuhan Univ. (Nat. Sci. Ed.)* **2019**, *5*, 6.
125. Liu, W.; Liu, M.; Huang, M. Study on Chinese Text Clustering Algorithm Based on K-mean and Evaluation Method on Effect of Clustering for Software-intensive System. In Proceedings of the 2020 IEEE International Conference on Computer Engineering and Application (ICCEA), Guangzhou, China, 18–20 March 2020; pp. 513–519.
126. Wu, D.; Yang, R.; Shen, C. Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm. *J. Intell. Inf. Syst.* **2020**, 1–23. [CrossRef]
127. Zhang, Q.; Yang, L.T.; Chen, Z.; Li, P. PPHOPCM: Privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing. *IEEE Trans. Big Data* **2017**. [CrossRef]
128. Zhang, Q.; Yang, L.T.; Chen, Z.; Li, P. High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT. *Inf. Fusion* **2018**, *39*, 72–80. [CrossRef]
129. Tan, Y.J.; Li, C.X. Study and Simulation of Text Clustering Using Attribute Weighted Fuzzy C-means Algorithm. *Comput. Simul.* **2011**, *5*. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-JSJZ201105056.htm (accessed on 30 December 2020).
130. Wen-xia, W. The Text Clustering Algorithm Based on LSA and FCM. *J. Shanxi Datong Univ. (Nat. Sci. Ed.)* **2016**, 3. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-YBSF201601003.htm (accessed on 30 December 2020).
131. Victor, G.S.; Antonia, P.; Spyros, S. Csmr: A scalable algorithm for text clustering with cosine similarity and mapreduce. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 211–220.
132. Wang, H.; Zhou, C.; Li, L. Design and Application of a Text Clustering Algorithm Based on Parallelized K-Means Clustering. *Rev. D'Intell. Artif.* **2019**, *33*, 453–460. [CrossRef]
133. Zhao, Q.; Shi, Y.; Qing, Z. Research on Hadoop-based massive short text clustering algorithm. In Proceedings of the Fourth International Workshop on Pattern Recognition. International Society for Optics and Photonics, Nanjing, China, 31 July 2019; Volume 11198, p. 111980A. [CrossRef]
134. Yang, K.; He, G.; He, G. Research and application of MapReduce-based MST text clustering algorithm. In Proceedings of the 2012 IEEE International Conference on Information Science and Technology, Wuhan, China, 23–25 March 2012; pp. 753–757.
135. Dangol, S.; Pokhrel, S. Analysis of Document Clustering Using K-means Algorithm with Cosine Similarity for Large Scale Text Documents with and without Hadoop. Available online: https://www.semanticscholar.org/paper/Analysis-of-Document-Clustering-Using-K-means-with-Dangol-Pokhrel/3904fcc4bc8d8b53ff3fca6821b614df1ab22d3f (accessed on 30 December 2020).
136. Kang, J.; Zhang, W. Combination of fuzzy C-means and particle swarm optimization for text document clustering. In *Advances in Electrical Engineering and Automation*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 247–252.
137. Loshma, G.; Hedge, D.N.P. Semantic analysis based text clustering by the fusion of bisecting k-means and UPGMA algorithm. *ARPN J. Eng. Appl. Sci.* **2006**, *11*, 3.
138. Shi, K.; Li, L.; He, J.; Zhang, N.; Liu, H.; Song, W. Improved GA-based text clustering algorithm. In Proceedings of the 2011 4th IEEE International Conference on Broadband Network and Multimedia Technology, Shenzhen, China, 28–30 October 2011; pp. 675–679.

139. Wang, Y.G.; Lin, L.; Liu, X.G. Hybrid text clustering algorithm based on dual particle swarm optimization and K-means algorithm. *Appl. Res. Comput.* **2014**, 12. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-JSYJ201402012.htm (accessed on 30 December 2020).
140. Gandomi, A.H.; Alavi, A.H. Krill herd: A new bio-inspired optimization algorithm. *Commun. Nonlinear Sci. Numer. Simul.* **2012**, *17*, 4831–4845. [CrossRef]
141. Abualigah, L.; Diabat, A.; Mirjalili, S.; Abd Elaziz, M.; Gandomi, A.H. The Arithmetic Optimization Algorithm. *Comput. Methods Appl. Mech. Eng.* **2021**.
142. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
143. Pan, X.; Jiao, L. A granular agent evolutionary algorithm for classification. *Appl. Soft Comput.* **2011**, *11*, 3093–3105. [CrossRef]
144. Corne, D.W.; Jerram, N.R.; Knowles, J.D.; Oates, M.J. PESA-II: Region-based selection in evolutionary multiobjective optimization. In *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; pp. 283–290. [CrossRef]
145. Babu, G.P.; Murty, M.N. Clustering with evolution strategies. *Pattern Recognit.* **1994**, *27*, 321–329. [CrossRef]
146. Xia, H.; Zhuang, J.; Yu, D. Novel soft subspace clustering with multi-objective evolutionary approach for high-dimensional data. *Pattern Recognit.* **2013**, *46*, 2562–2575. [CrossRef]
147. Aliniya, Z.; Mirroshandel, S.A. A novel combinatorial merge-split approach for automatic clustering using imperialist competitive algorithm. *Expert Syst. Appl.* **2019**, *117*, 243–266. [CrossRef]
148. Sheng, W.; Liu, X.; Fairhurst, M. A niching memetic algorithm for simultaneous clustering and feature selection. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 868–879. [CrossRef]
149. Zhang, Y.; Liu, J.; Zhou, M.; Jiang, Z. A multi-objective memetic algorithm based on decomposition for big optimization problems. *Memetic Comput.* **2016**, *8*, 45–61. [CrossRef]
150. Alami, J.; El Imrani, A.; Bouroumi, A. A multipopulation cultural algorithm using fuzzy clustering. *Appl. Soft Comput.* **2007**, *7*, 506–519. [CrossRef]
151. İnkaya, T.; Kayalıgil, S.; Özdemirel, N.E. Ant colony optimization based clustering methodology. *Appl. Soft Comput.* **2015**, *28*, 301–311. [CrossRef]
152. Cheng, Y.; Jiang, M.; Yuan, D. Novel clustering algorithms based on improved artificial fish swarm algorithm. In Proceedings of the 2009 IEEE Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China, 14–16 August 2009; Volume 3; pp. 141–145.
153. Karaboga, D.; Ozturk, C. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Appl. Soft Comput.* **2011**, *11*, 652–657. [CrossRef]
154. Satapathy, S.C.; Naik, A. Data clustering based on teaching-learning-based optimization. In *International Conference on Swarm, Evolutionary, and Memetic Computing*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 148–156.
155. Timmis, J.; Neal, M. A resource limited artificial immune system for data analysis. In *Research and Development in Intelligent Systems XVII*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 19–32.
156. Wan, M.; Li, L.; Xiao, J.; Wang, C.; Yang, Y. Data clustering using bacterial foraging optimization. *J. Intell. Inf. Syst.* **2012**, *38*, 321–341. [CrossRef]
157. Kulkarni, P.K.H.; Jesudason, P.M. Multipath data transmission in WSN using exponential cat swarm and fuzzy optimisation. *IET Commun.* **2019**, *13*, 1685–1695. [CrossRef]
158. Alsalibi, B.; Abualigah, L.; Khader, A.T. A novel bat algorithm with dynamic membrane structure for optimization problems. *Appl. Intell.* **2020**, 1–26. [CrossRef]
159. Zhong, S.; Ghosh, J. Generative model-based document clustering: A comparative study. *Knowl. Inf. Syst.* **2005**, *8*, 374–384. [CrossRef]
160. Kaur, S.P.; Madan, N. Document Clustering Using Firefly Algorithm. *Artif. Intell. Syst. Mach. Learn.* **2016**, *8*, 182–185.
161. Kumar, L.; Bharti, K.K. A novel hybrid BPSO–SCA approach for feature selection. *Nat. Comput.* **2019**, 1–23. [CrossRef]
162. Del Buono, N.; Pio, G. Non-negative Matrix Tri-Factorization for co-clustering: An analysis of the block matrix. *Inf. Sci.* **2015**, *301*, 13–26. [CrossRef]
163. Inbarani, H.H.; Bagyamathi, M.; Azar, A.T. A novel hybrid feature selection method based on rough set and improved harmony search. *Neural Comput. Appl.* **2015**, *26*, 1859–1880. [CrossRef]
164. Bharti, K.K.; Singh, P.K. A three-stage unsupervised dimension reduction method for text clustering. *J. Comput. Sci.* **2014**, *5*, 156–169. [CrossRef]
165. Chen, L.; Liu, M.; Wu, C.; Xu, A. A Novel Clustering Algorithm and Its Incremental Version for Large-Scale Text Collection. *Inf. Technol. Control.* **2016**, *45*, 136–147. [CrossRef]
166. Singh, V.K.; Tiwari, N.; Garg, S. Document clustering using k-means, heuristic k-means and fuzzy c-means. In Proceedings of the IEEE International Conference on Computational Intelligence and Communication Networks (CICN), Gwalior, India, 7–9 October 2011; pp. 297–301.