Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9768373, 1 page
https://doi.org/10.1155/2023/9768373

Retraction

Retracted: Application of *K*-Means Clustering Algorithm in Energy Data Analysis

Wireless Communications and Mobile Computing

Received 19 September 2023; Accepted 19 September 2023; Published 20 September 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Zhou, "Application of *K*-Means Clustering Algorithm in Energy Data Analysis," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 5914893, 8 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5914893, 8 pages
https://doi.org/10.1155/2022/5914893

WILEY | Hindawi

Research Article

Application of *K*-Means Clustering Algorithm in Energy Data Analysis

Ying Zhou

Lanzhou Resources & Environment Voc-Tech University, Lanzhou, Gansu 730022, China

Correspondence should be addressed to Ying Zhou; 11233234@stu.wxic.edu.cn

Received 31 March 2022; Revised 1 May 2022; Accepted 16 May 2022; Published 31 May 2022

Academic Editor: Aruna K K

In order to solve the problem of how to explore potential information in massive data and make effective use of it, this paper mainly studies news text clustering and proposes a news clustering algorithm based on improved *K*-Means. Then, the MapReduce programming model is used to parallelize the TIM-*K*-Means algorithm, so that it can run on the Hadoop platform. The accuracy and error are used as measurement indicators, and the collected datasets are used for experiments to verify the correctness and effectiveness of the TI value and TIM-*K*-Means algorithm. In addition, the Alibaba cloud server is used to build the Hadoop cluster, and the feasibility of parallelization transformation of TIM-*K*-Means algorithm is verified by accelerated comparison. The results show that the parallelized TIM-*K*-Means has a good acceleration ratio, can save about 30% of the time under the same conditions, and can meet the actual needs of processing massive data in the context of big data. In multidocument automatic summarization, news clustering algorithm can gather the news with the same topic and provide cleaner and accurate data for visual automatic summarization, which is of great significance in the fields of public opinion supervision, hot topic discovery, emergency real-time tracking, and so on.

1. Introduction

With the rapid development of network technology and cloud computing in recent years, the network has become an indispensable part of production and life, and the dissemination of information has become more and more rapid with the rise of the network [1]. People have more convenient and complete ways to obtain information [2]. The popularization of computer education has greatly reduced the technical threshold of software development. With the national attention to the information industry and the support of youth entrepreneurship, various media have sprung up rapidly. Online social media is loved by people of all ages and has a large number of users. With the huge number of users and high user activity, these network media not only speed up the dissemination of information but also produce a large amount of data. Traditional paper media have been impacted unprecedentedly, so they have invested a lot of resources to establish their own news portal or app to deal with the crisis. In these media, most of the information is transmitted in the form of text. However, there is a certain limit to the amount of information people obtain, which is far lower than the speed of information generation and dissemination, and this gap is expanding with the acceleration of network technology, resulting in the accumulation of massive information. How to mine valuable information from massive information and apply it to related fields has become a key research field [3].

The rapid development of data mining technology solves the problem of how to obtain potentially valuable information. Data mining uses relevant algorithms to analyze the data and get valuable rules or information hidden behind the information, so as to better find the potential value, optimize the production process, and provide useful information for scientific research. At present, data mining technology is quite active in the fields of social network, recommendation system, text analysis, and so on. Clustering is an important unsupervised learning method in data mining. The data is divided into several categories through the similarity between data. It is widely used in the fields of biological information, medical health, artificial intelligence, and so on [4]. As a classical clustering algorithm, the *K*-Means algorithm has the advantages of fast, simple, and easy to implement, but it also has some disadvantages, such as using random method to select the initial center point, resulting in local optimal solution and mistakenly selecting outliers as the center, resulting in reduced clustering accuracy and long running time. This paper improves the clustering effect by optimizing a step in the calculation process of the K-Means algorithm. In addition, researchers also integrate the K-Means algorithm with other models and algorithms and apply it to various fields such as finance, medicine, and image processing [5].

In the era of big data, the data information star is growing exponentially, and the data to be processed each time can reach the level of GB, TB, or even Pb. Therefore, only relying on a single machine for data processing requires high-performance machines and takes a lot of time. If the operation process is unexpectedly interrupted due to machine problems, it needs to be rerun. The traditional parallel framework needs a lot of equipment. Although it can solve the problem of massive data processing to a certain extent, it has poor fault tolerance and scalability, and the cost is high. The emergence of Hadoop solves the time-consuming problem of massive data processing and uses its own fault tolerance to ensure the smooth operation of the program to a certain extent. Hadoop uses the distributed file system (HDFS) to store files, distributes data to multiple servers through MapReduce computing model for distributed computing, and schedules resources through yarn. The MapReduce computing model encapsulates the functions of data segmentation, task allocation, and fault-tolerant processing. Users only need to write task programs as required [6, 7]. Since its launch, Hadoop has been continuously improved and developed into a complete ecosystem with multiple components. At present, Hadoop has become the mainstream distributed platform, and major Internet companies are used as the basic platform for offline and streaming data processing. In the field of scientific research, the MapReduce programming model has become the first choice for researchers to parallelize algorithms [8]. In the era of mobile Internet, people get news information anytime and anywhere through mobile phones or computers. News has become one of the most important text information in daily contact. News information plays an extremely important role in spreading social positive energy, carrying forward traditional culture, setting a social example and guiding public opinion. News clustering is still a kind of text clustering, which gathers texts with similar or even the same topics through cluster analysis. In information retrieval, the direct use of keyword matching will lead to unsatisfactory retrieval results due to ambiguity and other factors. If we first cluster the text set and then search according to the keywords generated after clustering, we can retrieve the text categories that better match the user's goals. In multidocument automatic summarization, the news clustering algorithm can gather news with the same topic and provide cleaner and accurate data for visual automatic summarization, which is of great significance in the fields of public opinion supervision, hot topic discovery, emergency real-time tracking, and so on [9].

2. Literature Review

Text mining is the product of the combination of data mining and natural language processing. Through the analysis of text, we can get potentially valuable information. Text clustering is one of the important branches of text mining, which mainly includes two processes: feature extraction and clustering calculation [10]. In terms of feature extraction, the research mainly focuses on how to accurately express the text with words. The common main methods are frequent itemset mining after word segmentation, inverse indexing, and latent Dirichlet allocation (LDA) model or the combination of different methods. In clustering, the existing frequent itemset mining algorithm is used to form the feature vector of text documents, which not only reduces the vector dimension but also retains the commonality between documents for similarity calculation. In addition, researchers use the obtained frequent phrase sequence to represent the text and use the association rule miner to find the binomial set that meets the minimum support of the Apriori algorithm, which avoids the disadvantage that the traditional vector space model ignores the word sequence and improves the accuracy and accuracy of text clustering analysis. The text network can also be constructed based on the frequent itemsets according to the similarity between texts, and the text network can be divided by using the community division algorithm, so as to achieve the purpose of clustering [11]. After using frequent itemsets to extract feature vectors, the two similarity indexes are combined to produce a new similarity index. At the same time, fuzzy logic is used for clustering rules. Finally, the datasets are classified by support vector machine to verify the accuracy of the proposed algorithm. Using the labeled data to construct the strong category discrimination word set, the cosine similarity and the similarity based on the strong category discrimination word items are fused to form a new similarity calculation method, and a semisupervised short text clustering algorithm based on improved similarity and class center vector is formed. Then the harmony search algorithm is used for feature selection to obtain useful information or new subsets with features, so as to reduce the impact of information loss and sparsity on text clustering, so as to enhance the clustering effect. Four benchmark text datasets are used for experiments. The enhancement of unsupervised feature selection technology based on harmony search in the K-Means clustering algorithm is proved by measuring the F value and accuracy [12, 13]. The K-Means algorithm is a process of repeatedly moving the center point of the class based on similarity, in which the selection of the center point and the definition of similarity are particularly important. For the optimization of the center, it includes the maximum distance product algorithm, minimum variance optimization method, and maximum minimum similarity. In addition, it also combines LDA and other models to solve the problems of data space and semantic barriers [14, 15].

On the basis of weighted K-Means, the Minkows metric can be used to measure the distance, and the feature weight can be used as the feature scaling factor in the traditional K-Means criterion. At the same time, the anomaly clustering center is used to initialize the centroid and feature weight of weighted *K*-Means. Through experiments on the dataset of UCI machine learning library and the dataset of generated Gaussian clusters, it is proved that the Minkows metric plays an important role in the *K*-Means algorithm. The shortcomings of the *K*-Means initial point selection affecting the clustering effect are studied. The criteria are dynamically weighted according to the covariance integration of the dataset to avoid large differences in the cluster. The simulation shows that this method has a certain effect. In addition, genetic algorithm is used to optimize the selection of initial cluster center point in the *K*-Means algorithm, so as to improve the clustering accuracy. There is also the method of using the FP growth algorithm to find out the frequent itemsets and using the frequent itemsets to generate the initial clustering centroid and clustering *K* value. The improved *K*-Means algorithm not only improves the accuracy but also speeds up the convergence speed of clustering. There is a new point-to-point distance-*s* distance, and combined with labor heuristic, the *s*-*K*-Means algorithm is proposed. Compared with the traditional *K*-Means algorithm using Euclidean distance, the clustering effect of this algorithm is significantly enhanced, especially in the case of irregular category distribution. According to the theory that "the farthest sample points are most unlikely to be divided into the same cluster," the maximum distance method is proposed to select the initial center.

For the application of *K*-Means algorithm in text clustering, researchers have also proposed a variety of improvement and optimization methods. Firstly, the particle swarm optimization algorithm is optimized, combined with the strong global search ability of particle swarm optimization algorithm and the strong local search ability of the *K*-Means algorithm to improve the effect of text clustering [16]. The transformation formula of cosine similarity and the Euclidean distance under standard vector is proposed. Based on this, a cosine clustering with close relationship and similar meaning with the Euclidean distance is defined. The selection method of initial center of the *K*-Means clustering is improved, so that the convergence speed is accelerated and the clustering accuracy is improved. Then, in the text preprocessing stage of text clustering, an alternative thesaurus is constructed according to the feature space of the document set, the text theme is obtained with the thesaurus, and the document times are replaced according to the theme and the corresponding domain dictionary. In the clustering stage, they proposed an improved *K*-Means algorithm based on *K*-value optimization [17]. Then use the cooccurrence word principle to calculate the text similarity, and divide it into $K + n$ class families according to the reading value. Then use the *K*-Means algorithm to cluster these class clusters, which solves the problem that *K*-Means is sensitive to the *K* value. Fair operation and clone operation are introduced to optimize the bee colony algorithm, and the *K*-Means algorithm is combined to improve the clustering quality [18].

In terms of application, the rapid development of social networks provides rich data for text clustering. Many researchers began to pay attention to the application of clustering algorithm in social networks. The *K*-Means algorithm is not only applied in the field of text mining but also applied in other aspects. The *K*-Means algorithm is integrated into the minimum spanning tree algorithm, and a fast minimum spanning tree algorithm based on the N-point complete graph is proposed, which reduces the theoretical time complexity from $O(N^2)$ to $O(N^{1.5})$ and overcomes the deficiency that the traditional minimum spanning tree algorithm cannot be applied to large datasets due to time complexity. *K*-Means clustering can also be used to develop image compression methods on low-power embedded devices, that is, using the similarity of pixel colors to group pixels and compress the original image, so as to reduce the power of wireless imaging sensor networks [19, 20].

3. Research Methods

3.1. K-Means Algorithm. In the *K*-Means algorithm, for the dataset, where *n* represents the number of data and *x* represents the dataset in the dataset, the similarity calculation adopts the Euclidean distance, and the calculation formula is as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)}, \tag{1}$$

where *n* is the number of attributes in each data. *x* and *y* represent the *i*-th attribute of data *x* and *y*, respectively. The algorithm randomly selects the initial cluster center, and in the iteration, the average value of all vectors in the last cluster is used as the new cluster center. The calculation formula of cluster center vector is as follows:

$$u_i = \frac{1}{h} \sum_{x \in C_i} x. \tag{2}$$

In the formula, *C* represents the *i*-th class family after clustering, and *h* represents the total number of data in this class cluster. When judging the conditions for the end of clustering, the criterion function adopts the square difference function, and the calculation formula is as follows:

$$E = \sum_{i=1}^{k} \sum_{x \in c_i} |x - u_i|^2. \tag{3}$$

The specific flow of the *K*-Means algorithm is shown in Figure 1.

3.2. Parallel Foundation. Hadoop is an open source distributed computing platform, which was separated from the project into a separate software in 2006. After a long period of development, Hadoop has formed an ecosystem covering various services such as computing model, data storage, workflow, and communication coordination between clusters [21]. Its core components are shown in Figure 2.

In the Hadoop ecosystem, the Hadoop distributed file system is the basic component, which distributes a large

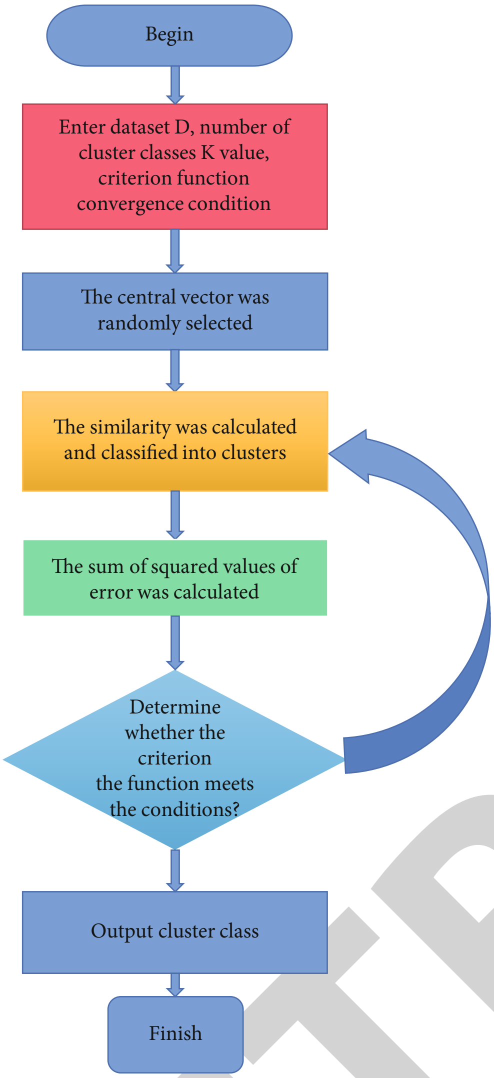
Begin
Enter dataset D, number of cluster classes K value, criterion function convergence condition
The central vector was randomly selected
The similarity was calculated and classified into clusters
The sum of squared values of error was calculated
Determine whether the criterion the function meets the conditions?
Output cluster class
Finish

Figure 1: Flow chart of the K-Means algorithm.

amount of data to the computer cluster. The data is written once but can be read many times for analysis. MapReduce is a programming model for distributed parallel data processing. It is the main execution framework of Hadoop. It divides the whole job into two stages: map and reduce. HBase is a column-oriented NoSQL database, which can provide fast reading and writing of a large amount of data. Zookeeper and Oozie are mainly used for distributed coordination. Pig and Hive are abstract layers, which can analyze data by HQL statement and Latin statement, respectively. In addition, Hadoop also provides frameworks Sqoop and Flume for enterprise-level data integration. Among them, Sqoop is often used to transfer data between different types of databases, such as MySQL and HBase, while flume is used to efficiently collect, aggregate, and move a large amount of data from a single machine to HDFS [22].

3.3. File System HDFS. HDFS mainly includes four parts, NameNode, DataNode, Client, and SecondaryNameNode, and adopts the Master-Slaves mode. The SecondaryNameNode will back up the operation logs and image files in the NameNode at regular intervals. In the HDFS system, the NameNode stores metadata, including information such as directories, data block locations, and data size, and persists these information to the local disk. At the same time, the NameNode is responsible for managing the cluster and will judge the node and the data in the node according to the heartbeat signal sent by each node. The main function of the DataNode is to store data. The DataNode periodically verifies the stored data blocks and periodically sends a heartbeat signal to the NameNode. The heartbeat signal has two main functions. On the one hand, it indicates the storage information of the data block to the NameNode, and on the other hand, it indicates that the node is still working and not down. SecondaryNameNode receives fsimage and editlog for merging, then sends it to NameNode, and also saves the merged file locally to prevent data loss caused by NameNode crash. Client provides a file system interface for users to use, and Client accesses files in HDFS through NameNode and DataNode.

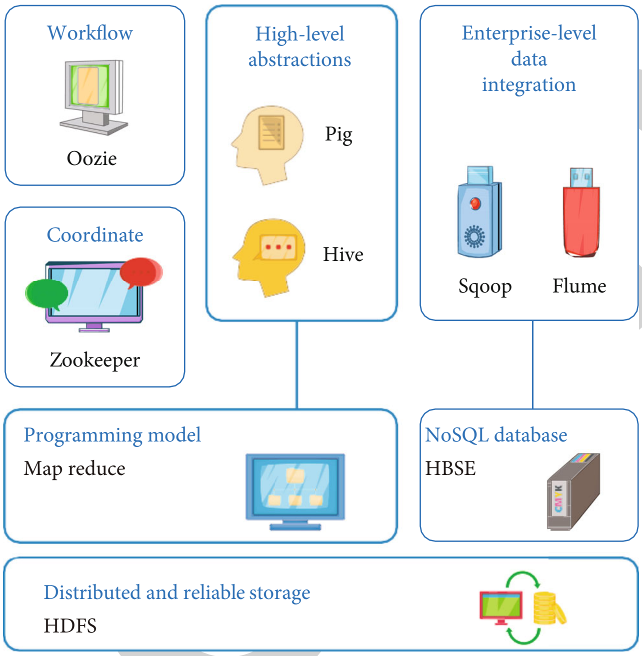

Figure 2: Core components of Hadoop ecosystem.

3.4. MapReduce Model. The MapReduce model is composed of multiple parts. When using it, users only need to write the program into map and reduce functions according to the format given by the model and then use the driver to configure the required components (including input and output format, combiner, and partition). Most components can be customized according to user requirements. For example, Inputformat and Outputformat define the input and input format, Recordreader defines the data reader, and Inputsplit controls slice size. At the same time, adjusting the parameter settings of these components can optimize the execution of MapReduce job, so as to improve the utilization of computing resources and reduce the consumption of task time.

3.5. TIM-K-Means Algorithm. Text mining is a tool and method that takes documents as data to find potential valuable information targets of documents. It uses the relevant knowledge of natural language processing to map the document into data and processes the data corresponding to the document through the relevant algorithms in data mining or machine learning, so as to find the hidden law or knowledge. It can be seen that text mining is an extension of data mining, and data mining is the basis and essence of text mining. Text clustering is an important branch of text mining. Clustering analysis of news by the *K*-Means algorithm is essentially to mine the news content by using the clustering algorithm, gather similar news information together, and find the valuable information hidden behind the news content. When extracting features, the TF-IDF value is often used to represent the weight of a word. The main idea of the TF-IDF value is that if a word appears frequently in an article and rarely in its text, it is considered that the word can represent the article to a certain extent and can be regarded as an important feature to distinguish it from other texts [23].

Term frequency (TF) refers to the frequency of words in the file. The higher the TF value, the more the word appears in the text, which means that the word is more important in the text. The TF value is calculated by the following:

$$\mathrm{TF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}. \tag{4}$$

In the formula, $n_{i,j}$ represents the number of occurrences of document word t_i in document d_j, and the denominator is the sum of the number of occurrences of all words in document *dj*.

Inverse document frequency (IDF) measures the universality of a word. The larger IDF value of a word means that the word is widely used in the text. It cannot distinguish the text from other texts by virtue of the word and cannot be used as a feature to distinguish the text. The calculation formula is shown in the following:

$$\mathrm{IDF}_i = \log \frac{|D|}{1+\sum D \supseteq t_i}. \tag{5}$$

In the formula, $|D|$ represents the total number of all documents in the corpus, and $\sum D \supseteq t_i$ represents the number of documents containing the word t_i.

The TF-IDF calculation formula of this word is shown in the following:

$$\mathrm{TF}-\mathrm{IDF} = \mathrm{TF}_{i,j} \times \mathrm{IDF}_i. \tag{6}$$

The flow chart of feature extraction is shown in Figure 3.

Aiming at the disadvantages of the traditional *K*-Means clustering algorithm, such as the clustering effect, *K*-value sensitivity, randomness of initial clustering center selection, and possible local optimal solution, researchers propose a new method to calculate the similarity of the initial classification points or new methods to improve the effect of clustering. At the same time, in addition to improving and optimizing the algorithm itself, this paper also continues to try to integrate the *K*-Means algorithm with other algorithms, such as the ant colony algorithm, frequent itemset mining algorithm, and genetic algorithm, and combine the advantages of the two algorithms to work together on text clustering. This paper puts forward the concept of the TI value for the structural characteristics of news information and unifies the news title, introduction, and text, so as to make the feature words more representative. Then the news clustering algorithm which combines the TI value with the improved maximum distance algorithm is called the TIM-*K*-Means algorithm. The TIM-*K*-Means algorithm improves the *K*-Means algorithm in terms of text feature vector composition and initial center point selection and does not change its structure and process. Therefore, its parallelization process is similar to that of *K*-Means. The map function of the parallel TIM-*K*-Means algorithm puts the distance center generated by the optimized maximum distance algorithm into the central file to calculate the distance between the data sample points and all central points. Then add the data sample point to the cluster class represented by the cluster center point with the smallest distance, and pass it to the reduce function in the mode of the < key, value > key value pair. Key is the flag of the cluster center point, and value represents the sample point [24].

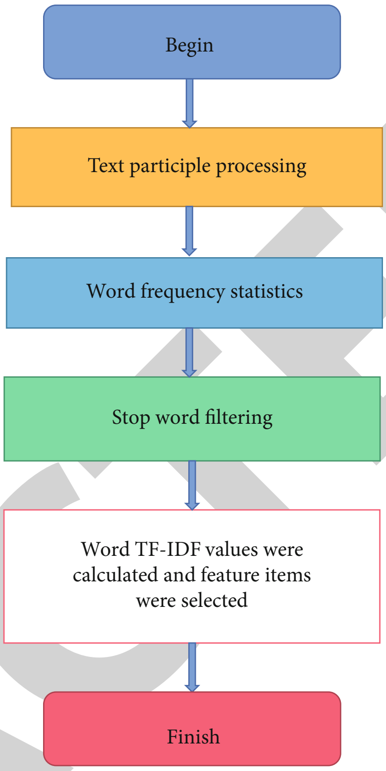
Begin
Text participle processing
Word frequency statistics
Stop word filtering
Word TF-IDF values were calculated and feature items were selected
Finish

Figure 3: Flow chart of text feature vector extraction.

4. Result Analysis

4.1. Experimental Environment. In this paper, the TIM-*K*-Means algorithm is parallelized. In order to make the

Table 1: *K*-Means clustering accuracy under different coefficients.

	Lead weight *n*				
Title weight *m*	0	0.5	1.0	1.5	2.0
0	55.46%	55.73%	55.24%	55.34%	54.82%
0.5	55.36%	54.14%	54.92%	57.64%	55.47%
1.0	55.48%	54.73%	55.78%	55.41%	54.86%
1.5	55.74%	54.85%	55.74%	55.34%	54.68%
2.0	55.69%	55.26%	55.49%	56.49%	54.86%

Table 2: Algorithm running time.

Number of news Algorithm	5 W (s)	10 W (s)	25 W (s)	50 W (s)
K-Means	33.56	63.64	179.68	558.23
TIM-*K*-Means	32.46	48.35	168.98	540.16
Parallel TIM-*K*-Means	40.35	47.12	112.46	303.74

experimental results more convincing, experiments are carried out in a single machine and distributed environments, respectively.

(1) Single machine experiment environment setting: Intel® Core™ i5-6500, 8 G memory, 930 G memory, Win7 64-bit operating system. Java JDK1.8.0_ 161, python 2.7.13

(2) Distributed environment settings: the Hadoop platform is built by five Alibaba cloud servers

4.2. Verification and Analysis of TI Value. In the feature extraction of news text, this paper puts forward the concept of TI value based on the TF-IDF value, that is, the words in news title and introduction are integrated into the calculation of feature value, so that the obtained feature vector is more representative and accurate. For the convenience of later description, this paper calls the *K*-Means algorithm that only combines the news headline factors in feature extraction as the T_ *K*-Means algorithm, the algorithm that only combines the news lead factors as I *K*-Means, and the algorithm that combines the TI value as the TI *K*-Means algorithm.

In order to verify the TI value, this paper uses a dataset of 2000 news, including military, NBA, and science and technology. After word segmentation of the news title, introduction, and text, the distribution of words in these three structures is irregular. In order to determine the corresponding weight of the title and the introduction, this paper uses the progressive method for the experiment, with a step size of 0.5. The accuracy of each method is the average value after 10 times of operation, and the accuracy is expressed in percentage. *m* and *n*, respectively, represent the weights of news headlines and leads when calculating TI values. The experimental results are shown in Table 1.

News headlines and leads have a certain impact on news clustering. When giving appropriate weights to news headlines and leads, the feature vector constructed by the TI value is conducive to improve the clustering accuracy and

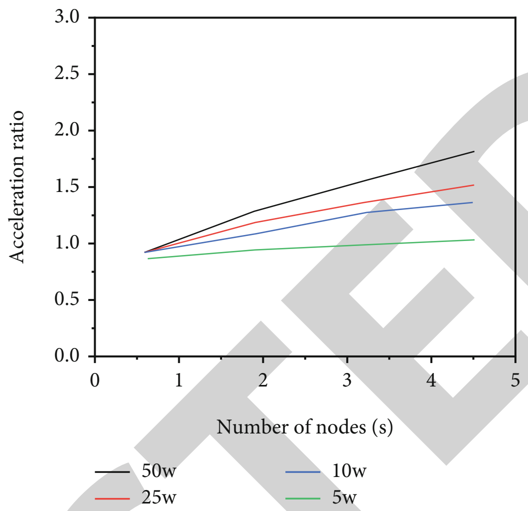

Figure 4: Parallel TIM-*K*-Means speedup.

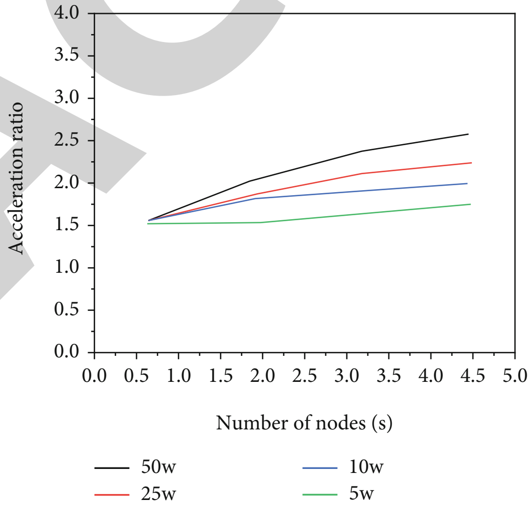

Figure 5: Parallel *K*-Means speedup ratio.

prove the correctness of the TI value. And in the TI value, the weight of title and introduction should be 0.5 and 1.5, respectively, that is, $m = 0.5$ and $n = 1.5$. This weight is in line with the objective fact that the news lead contains more information than the title.

4.3. Algorithm Parallelization Verification Analysis. In this paper, *K*-Means, TIM-*K*-Means, and parallel TIM-*K*-Means algorithms are experimented with datasets with 50000, 100000, 250000, and 500000 news, and the time taken to complete clustering is recorded [25]. Its running time is shown in Table 2.

In this paper, we use different amounts of news data to experiment with parallel *K*-Means and TIM-*K*-Means algorithms in a cluster environment with 1, 2, 3, and 4 data nodes, respectively. The speedup ratios of the two algorithms are recorded, respectively, as shown in Figures 4 and 5.

The parallel *K*-Means algorithm and parallel TIM-*K*-Means have similar speedup. It shows that the TIM-*K*-

Means algorithm can still run stably after parallelization transformation without destroying the original characteristics of the algorithm. And with the increasing number of datasets and data nodes, the acceleration ratio growth trend of TIM-*K*-Means is more obvious than that of *K*-Means. Therefore, from the aspect of acceleration ratio, the parallelization transformation of TIM-*K*-Means algorithm is feasible, which can accelerate the implementation of the algorithm to a certain extent and solve the problem of time-consuming clustering of massive news information.

5. Conclusion

Based on the in-depth understanding of clustering analysis and *K*-Means algorithm, this paper studies and improves news clustering. Firstly, this paper introduces the research background and significance of news clustering analysis and introduces the research status of these two aspects. Secondly, the basic knowledge of clustering analysis and algorithm parallelization technology is introduced. Thirdly, according to the organizational structure of news text, the concept of the TI value is defined and optimized. Combining the two, the TIM-*K*-Means algorithm is proposed, and the TIM-*K*-Means algorithm is parallelized by using the MapReduce programming model, so that it can adapt to the massive data environment. Finally, this paper verifies the above concepts and algorithms in stand-alone and distributed environments, respectively. The main research work of this paper is as follows:

(1) Combined with news headlines and leads, the concept of the TI value is defined, and the weight of headlines and leads in TI value is determined. When extracting the features of news text information, this paper gives different weights to the words in the news title and news lead and adds them to the TF-IDF value of the text feature word to obtain the TI value. Compared with the original feature extraction, the TI value fully considers the organizational structure of news, making feature words more representative. Through Tencent News data, it is proved that when the weight values of news title and lead are 0.5 and 1.5, respectively, the TI value is the most representative, which improves the accuracy of clustering to a certain extent

(2) The TIM-*K*-Means algorithm is parallelized. According to the calculation process of the TIM-*K*-Means algorithm, this paper deduces the error calculation formula and obtains the calculation method of clustering error in distributed environment. The parallel transformation of TIM-*K*-Means is carried out by using the MapReduce programming model. Experiments show that the parallelized TIM-*K*-Means has a good speedup ratio and can meet the actual needs of processing massive data in the context of big data

The TIM-*K*-Means news clustering algorithm proposed in this paper fully combines the organizational structure information of news and improves the selection method of the initial clustering center, which improves the clustering accuracy and stability to a certain extent and reduces the clustering error, but there are still deficiencies in some aspects, and further investigation and research are needed. The main research directions in the future are as follows:

How to accurately find the K value? The determination of the K value in news clustering requires certain prior knowledge, and in the absence of any prior knowledge, it can only be determined manually by the operator's work experience. How to automatically discover the K value more accurately before news clustering needs further research.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author has no conflict of interest to declare.

References

[1] Y. Fan, Y. Liu, H. Qi, F. Liu, and X. Ji, "Anti-interference technology of surface acoustic wave sensor based on K-Means clustering algorithm," *IEEE Sensors Journal*, vol. 21, no. 7, pp. 8998–9007, 2021.

[2] D. Zheng, X. Sun, S. K. Damarla, A. Shah, J. Amalraj, and B. Huang, "Valve stiction detection and quantification using a K-Means clustering based moving window approach," *Industrial & Engineering Chemistry Research*, vol. 60, no. 6, pp. 2563–2577, 2021.

[3] B. S. Aski, A. T. Haghighat, and M. Mohsenzadeh, "Evaluating single web service trust employing a three-level neuro-fuzzy system considering K-Means clustering," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 1, pp. 1–15, 2021.

[4] Z. Chen, "Using big data fuzzy K-Means clustering and information fusion algorithm in English teaching ability evaluation," *Complexity*, vol. 2021, no. 5, Article ID 5554444, 9 pages, 2021.

[5] D. Lou, M. Yang, D. Shi, G. Wang, and Y. Chen, "K-Means and c4.5 decision tree based prediction of long-term precipitation variability in the Poyang lake basin, China," *Atmosphere*, vol. 12, no. 7, p. 834, 2021.

[6] F. Tao, R. Suresh, J. Votion, and Y. Cao, "Graph based multilayer K-Means++ (g-mlkm) for sensory pattern analysis in constrained spaces," *Sensors*, vol. 21, no. 6, p. 2069, 2021.

[7] P. Arjun and K. G. Manoj, "Improved hybrid bag-boost ensemble with K-Means-smote–enn technique for handling noisy class imbalanced data," *The Computer Journal*, vol. 65, p. 1, 2021.

[8] Z. Zhu and N. Liu, "Early warning of financial risk based on K-Means clustering algorithm," *Complexity*, vol. 2021, no. 24, Article ID 5571683, 12 pages, 2021.

[9] C. Y. Peng, U. Raihany, S. W. Kuo, and Y. Z. Chen, "Sound detection monitoring tool in cnc milling sounds by K-Means clustering algorithm," *Sensors*, vol. 21, no. 13, p. 4288, 2021.

[10] M. Zhao, H. Gao, Q. Han, J. Ge, W. Wang, and J. Qu, "Development of a driving cycle for Fuzhou using K-Means and ampso," *Journal of Advanced Transportation*, vol. 2021, no. 2, Article ID 5430137, 15 pages, 2021.

[11] V. Utomo and J.-S. Leu, "Automatic news-roundup generation using clustering, extraction, and presentation," *Multimedia Systems*, vol. 26, no. 2, pp. 201–221, 2020.

[12] B. Liang, N. Li, Z. He, Z. Wang, and T. Lu, "News video summarization combining surf and color histogram features," *Entropy*, vol. 23, no. 8, p. 982, 2021.

[13] L. Wang, S. Li, W. Wang, W. Yang, and H. Wang, "A bank liquidity multilayer network based on media emotion," *The European Physical Journal B*, vol. 94, no. 2, pp. 1–23, 2021.

[14] D. Fuentealba, M. Lopez, and H. Ponce, "Effects on time and quality of short text clustering during real-time presentations," *IEEE Latin America Transactions*, vol. 19, no. 8, pp. 1391–1399, 2021.

[15] Z. Gou, Y. Li, and Z. Huo, "A method for constructing supervised time topic model based on variational auto encoder," *Scientific Programming*, vol. 2021, no. 12, Article ID 6623689, 11 pages, 2021.

[16] H. Li and D. Han, "A novel time-aware hybrid recommendation scheme combining user feedback and collaborative filtering," *Mobile Information Systems*, vol. 15, no. 4, 16 pages, 2021.

[17] C. Hu, Z. Pan, and T. Zhong, "Leaf and wood separation of poplar seedlings combining locally convex connected patches and K-Means++ clustering from terrestrial laser scanning data," *Journal of Applied Remote Sensing*, vol. 14, no. 1, p. 1, 2020.

[18] I. H. Hannah, A. T. Azar, and G. Jothi, "Leukemia image segmentation using a hybrid histogram-based soft covering rough K-Means clustering algorithm," *Electronics*, vol. 9, no. 1, p. 188, 2020.

[19] F. Deng, W. Gu, W. Zeng, Z. Zhang, and F. Wang, "Hazardous chemical accident prevention based on K-Means clustering analysis of incident information," *IEEE Access*, vol. 8, pp. 180171–180183, 2020.

[20] J. Wu, L. Shi, W. P. Lin, S. B. Tsai, and G. Xu, "An empirical study on customer segmentation by purchase behaviors using a rfm model and K-Means algorithm," *Mathematical Problems in Engineering*, vol. 2020, no. 6, Article ID 8884227, 7 pages, 2020.

[21] Z. Chen and W. Liu, "An efficient parameter adaptive support vector regression using K-Means clustering and chaotic slime mould algorithm," *Access*, vol. 8, pp. 156851–156862, 2020.

[22] M. Bradha, N. Balakrishnan, A. Suvitha et al., "Experimental, computational analysis of Butein and Lanceoletin for natural dye-sensitized solar cells and stabilizing efficiency by IoT," *Environment, Development and Sustainability*, vol. 24, no. 6, pp. 8807–8822, 2021.

[23] A. Sharma and R. Kumar, "A framework for pre-computated multi- constrained quickest QoS path algorithm," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, 2017.

[24] R. Huang, S. Zhang, W. Zhang, and X. Yang, "Progress of zinc oxide-based nanocomposites in the textile industry," *IET Collaborative Intelligent Manufacturing*, vol. 3, no. 3, pp. 281–289, 2021.

[25] L. Xin, M. Chengyu, and Y. Chongyang, "Power station flue gas desulfurization system based on automatic online monitoring platform," *Journal of Digital Information Management*, vol. 13, no. 6, pp. 480–488, 2015.