Clustering Research Based on Particle Swarm Optimization (PSO) Algorithm and Large-Scale Short Text Processing in Hadoop Cluster Architecture

Wanle Chi

Department of Information Technology, Wenzhou Vocational & Technical College, Wenzhou, China

358455713@qq.com

Abstract

To achieve large-scale short text processing, the Hadoop cluster framework was introduced. In addition, based on Hadoop cluster framework, the distributed parallel clustering algorithm DPSOKmeans was adopted. The experiment showed that the clustering algorithm ran on the Hadoop framework and had a good convergence line. Meanwhile, the high clustering quality can handle massive data. Through the experiment and practice application for DPSOKmeans clustering algorithm, it was proved that the global optimization ability of the clustering algorithm DPSOKmeans was higher than the K-means clustering algorithm. It is concluded that DPSOKmeans clustering algorithm should be used in Hadoop platform in the face of massive short text data.

Keywords

Mass, short text, clustering, K-means, PSO, Hadoop.

1. Introduction

In recent years, due to the gradual growth of the mobile Internet and the popularization of computer technology, the society at the moment is generating data at an unimaginable speed. Some new communication platforms publish as many as 100,000 messages per minute. These platforms are represented by BBS, WeChat, Twitter and Weibo. These vast amounts of data affect our lives all the time, and it is an effective way for people to access information source. However, these data have the characteristics of massive data, diversification, dynamics and less word, so people also call it short text data [1]. At present, the most important thing is how to extract user's hidden information quickly, accurately and conveniently from these massive short text data, so as to understand users' needs and create business opportunities in time [2-4]. Therefore, in this case, the text mining technology is generated. Text mining can search out valuable information hidden by users from the massive text data, which is the purpose of our mining. However, the core of text mining technology is the algorithm. A good algorithm can improve the speed and quality of mining. Due to the mass and diversity of these short text data, it is suitable for short text processing based on the unsupervised learning feature [5-6].

2. Related theory

2.1 Text mining

Text mining involves many fields, including information technology, pattern recognition, database technology, text analysis, machine learning in artificial intelligence, statistics, information retrieval and so on. Therefore, from different research fields, researchers have given different definitions of text mining. From the perspective of data mining: Text mining uses the mining technology to automatically

discover hidden patterns from text. From the perspective of information processing, text mining is used to process the unstructured text, which is different from the structured data in traditional data mining. Therefore, it should be incorporated into text information processing based on information retrieval and natural language understanding.

This paper presents a distributed parallel text-mining: It deals with objects in the form of unstructured text. It's the process of obtaining hidden information, valuable knowledge, important patterns, and identifying the underlying relationships between different data [7]. The 21st century is the era of information. With the rapid development of information technology, the growth of text data is accelerating [8]. As described in chapter 1, the data generated daily reaches PB levels. In the face of such a large amount of text data, we need only a small number of relevant data. It is extremely difficult to simply look up and quickly obtain the information we want. Nowadays, the application of text mining is extensive, such as the safe city construction of government departments, the crack of DNA code of bioengineering, and the inquiry of hot events in information retrieval and so on. All these aspects have promoted our exploration of text mining technology [9]. An effective technique to deal with these large-scale unstructured text data is the focus of our future research [10].

2.2 Text mining process

The general process of text mining is shown in figure 1.

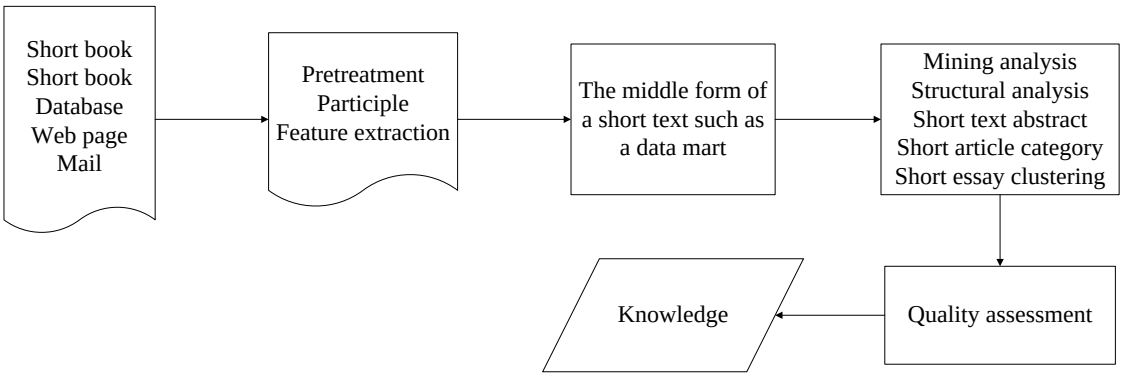

Figure. 1 General process of text mining processing

Text pre-processing: Text pre-processing is the basis of text mining. In this stage, text set is processed into feature vector set through the process of feature extraction, feature selection, segmentation, and stop and delete and so on, so as to facilitate text mining. The text feature is the most important factor that affects the performance of mining. Therefore, the efficient extraction of text features is of great significance for text mining. At present, the research of text feature mainly focuses on feature extraction and feature selection. The concrete steps are as follows: Text feature extraction: It is necessary to select the feature of the extracted text type, that is, to extract the items that can represent the text features [11]. As a result, the text is transformed into a structured form that can represent the content of the text. For example, information retrieval usually uses a vector space model (VSM). Text feature selection: The process of feature extraction is also the process of forming the original feature set. Because the original feature set is very large (it may reach a few million dimensions), the original feature sets need to be further screened before the text mining is made. The principle of screening is based on the purpose of mining. However, first, a feature item with powerful text information expression ability should be retained.

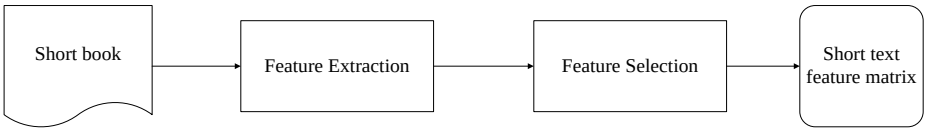

Figure. 2 The process of text preprocessing

The process of text mining: Once the text data is converted into a structured form, this form is the prerequisite for text mining. There are two steps in the excavation stage: First, All the support degree

generated in the text mining phase is not less than the minimum support threshold value. Second, the frequent item-sets produced in the text mining stage are used to construct the association rules satisfying the minimum reliability constraints.

Text mining quality assessment: The quality evaluation of text mining is an indispensable part of the mining stage, and it is the overall evaluation of the results of the mining. If the evaluation results meet the requirements, then the mining task is over. Otherwise, it will return to the last link to remake and adjust. Many cycles are carried out until the mining results meet the requirements, then the mining process is finished [12].

2.3 Text clustering

Text clustering is one of the most important methods in text mining. With the deep research of text mining technology, text clustering technology has attracted more and more scholars' attention. Nowadays, text clustering technology has been widely used, such as hot event extraction, spam filtering, web search and so on. The main task of text clustering: The text data to be processed is divided into several classes, and these text data are unruly. First of all, we find hidden information from these non-observable data. Through the hidden information, the texts can be classified into different categories according to the data connection. The clustering results show that the relationship between texts in the same category is most closely related, and the text data between different types of text have no connection. In text mining, the biggest difference between text classification and text clustering is whether there is predictive information. Text classification is the pre-classified information before classification, that is, class label, while text clustering is not. Its category label is automatically determined.

Text clustering is much more superior to classification for today's text data. It can be adapted to unclassified data. This can greatly reduce the time spent on class marking. The text data to be processed is pre-processed to obtain the combination of feature vectors, which is very important for text clustering. If the selected feature data is not appropriate, the clustering result will be very bad. The choice of ideal feature data is very similar to the characteristic data in the same cluster, and the difference is very large in different clusters. It is the core of text clustering to select the appropriate clustering algorithm to deal with the text data. A good clustering algorithm can make the text clustering work easy, and get a very good pedigree chart. Using the pedigree step, an appropriate and suitable threshold is selected after considering various factors [13]. As long as the selected value is proper enough, the ideal clustering results can be obtained from the results of previous step [14].

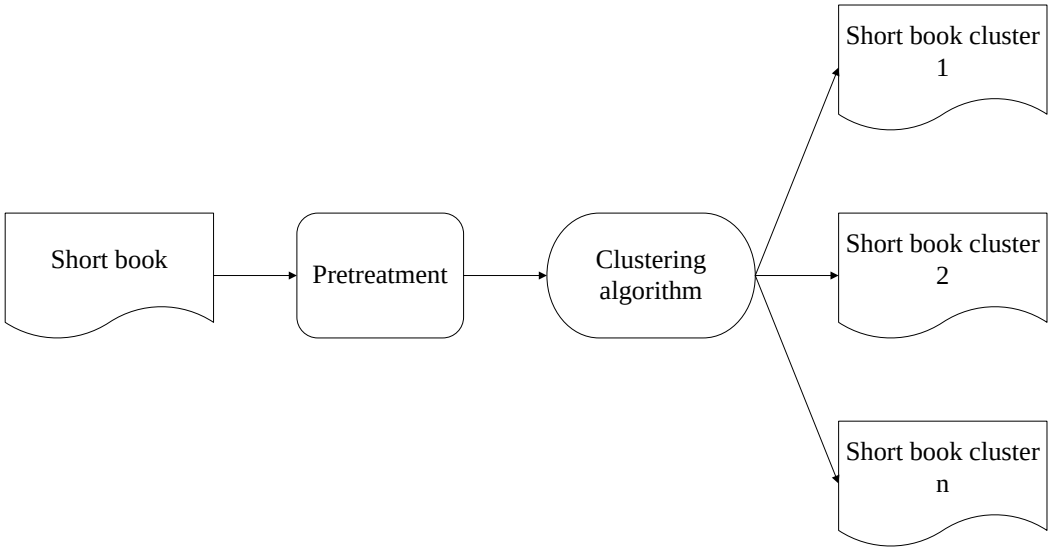

Figure. 3 Text clustering process

2.4 Large scale data processing architecture Hadoop

To improve processing performance, the design of the Hadoop system architecture is to: In the massive data environment, the localized computation is achieved as far as possible. The design rule is that there will be a Master node in the cluster to manage and control the normal operation of the whole system. At the same time, each Slave node in the management system is coordinated to complete the data calculation and storage. At this point, each Slaver node plays two roles: the data calculation and the data storage. The Master node uses the heartbeat mechanism to detect and check whether a Slave node is invalid in the system. The failure judgement is that Slave cannot respond to Heartbeat information in time [15].

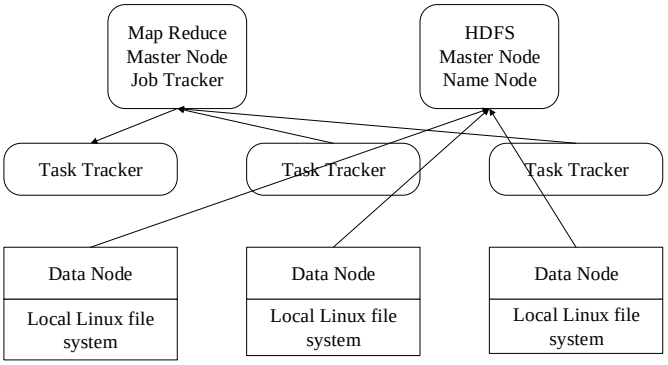

Figure. 4 Hadoop system architecture

3. Method

3.1 PSO algorithm

Particle swarm optimization (PSO) is an evolutionary algorithm (EA) developed to simulate the bird predator behaviour. This algorithm is used to solve the optimal solution. The algorithm begins with a random solution and searches the optimal solution by iteration. At the same time, the quality of the solution is evaluated by the fitness. The algorithm can search the global optimal solution by the present optimal solution. At the same time, the algorithm has the advantages of easy realization, high accuracy and fast convergence speed. Therefore, it has been widely applied in practical problems, and the algorithm is also a parallel algorithm.

In a set, any element in the space is moved at a certain speed v. In the set C, any element $C(\mathrm{j}_1, j_2, ..., j_n)$ is searching for the best solution of the search individual and the best of the whole. The whole set will change according to the best solution that the element is looking for. This element is composed of three parts, and its specific representation is as follows:

Current location: $\mathrm{x}_\mathrm{j} = (x_{j1}, x_{j2}, ..., x_{jc}),\ \mathrm{j} = 1,2,..., \mathrm{n}$

The flight speed of the j-th particle: $\mathrm{v}_j = (v_{j1}, v_{j2}, ..., v_{jc})$

The historical optimal solution of the j-th particle: $\mathrm{p}_\mathrm{j} = (p_{j1}, p_{j2}, ..., p_{jc})$;

Among them, the best solution that the whole set C can find is: $\mathrm{p}_\mathrm{g} = (p_{g1}, p_{g2}, ..., p_{gc})$. The current position is described as a set of coordinates of the space point. Each time the algorithm is updated, the current position will be solved as a problem. If it is better than the historical position, the coordinates of the target position will have second pj. The evolution equation of the particle swarm can be expressed as:

$$\mathrm{v}_{\mathrm{jc}}(m+1) = v_{jc}(m) + b_1.ramdom.(p_{jc}(m) - x_{jc}(m)) + b_2.ramdom().(p_{gc}(m) - x_{gc}(m)) \quad (1)$$

$$\mathrm{x}_{\mathrm{jc}}(m+1) = x_{ic}(m) + v_{ic}(m+1) \quad (2)$$

In the formula, the subscript c and j denote the particle c and the particle j, respectively. t represents the number of particle updates. b1 and b2 are the acceleration constants, and the positive values are generally within 2. random () generally takes a random number of [0~1]. Vmax is the maximum value of limted speed, and this constant has a user definition. The velocity of the particle is [-Vmax~ Vmax]. After the speed formula is updated, if:

$$\text{if } v_{jc} < -v_{max} \text{ then } v_{jc} = -v_{max} \tag{3}$$

$$\text{if } v_{jc} > v_{max} \text{ then } v_{jc} = v_{max} \tag{4}$$

PSO algorithm steps:

The first step is to start the initialization: The position xj and velocity vj of the particles are randomly generated in the solution of the space C dimension;

The second step is to evaluate particle j: The applicable value of the C dimensional optimization function is evaluated;

The third step is to update the optimal: If the applicable value of the particle is superior to the individual optimal solution pBest, the pBest position is the position of the current particle j. If the applicable value of the particle is better than the optimal solution gBest of the group, the gBest position is the position of the current particle j.

The fourth step is updating the particle: The position xj and the velocity vj of the particle are changed according to the formula (1) and the formula (2);

The fifth step is to end the condition: It is circulated to the step (2) until the stop condition is satisfied (applicable value and maximum number of iterations).

3.2 Particle swarm optimization (K-means) clustering algorithm

Because the PSO algorithm can overcome the shortcomings of the K-means clustering algorithm, a K-means clustering algorithm based on particle swarm is proposed.

PSOKmeans algorithm thought: PSO not only has the ability of local optimization, but also has the ability of global optimization. Therefore, the combination of PSO algorithm and K-means algorithm to improve K-means is to make full use of the advantage of PSO. The PSOKmeans algorithm first uses the K-means clustering algorithm to get the center of a group of clusters. Then, when the particle swarm is initialized, it is given to a particle. The remaining particles are randomly initialized, and then the PSO algorithm is used to complete the clustering. Because the K-means algorithm has strong local optimization ability, the convergence speed of the PSOKmeans algorithm can be greatly improved. The mathematical description of the PSOKmeans algorithm is as follows:

Assuming there is a particle swarm D={Di,i=1,2,...,m} and Di is a pattern vector of X dimension. The clustering problem is to assign particles in D to K clusters C={C1,C2,...,CK} ($1 \le i$， $j \le k$ and $i \ne j$， $Ci \subset D$， $Ci \not\subset \varnothing$ and $Ci \cap Cj = \varnothing$).

The difference between particle $p \subset Cj$ and cluster Cj's centroid is measured by $dis(p \subset Cj)$ (The Euclidean distance between the particle and the centroid of the cluster). It is the sum of the square E of the error between all the particles in the cluster and the centroid of the cluster. That is,

$$E = \sum_{j=1}^{k} \sum_{p \subset C_j} dis(p, C_j)^2 \tag{5}$$

The fitness of the particles is used as follows:

$$f(A_j) = \frac{i}{E_j} \tag{6}$$

In the formula, i is constant and depends on the situation. Ej is the sum of the total inters class dispersion. If the fitness of the particle and the sum of dispersion are positive correlation, then the smaller the Ej is, the larger the f(Aj) is.

PSOKmeans algorithm steps:

Input: A number of categories K and data set $C(j_1, j_2,..., j_n)$ are generated;

Step 1: In initialization process, each particle in the particle swarm is randomly clustered, the centroid of each cluster is calculated, and the initial value of the particle position is given. According to the formula (6), the fitness of the particle is calculated and the initial value of the particle velocity is given. Because the particle group contains M particles, it is necessary to repeat M.

Step 2: The value of the fitness is represented by the element in the set $C(j_1, j_2,..., j_n)$. If it is less than the value of the current fitness, then the pBest is updated;

Step 3: According to the value of the fitness of the whole element in the set $C(j_1, j_2,..., j_n)$, if it is less than the current global optimal solution, then the gBest is updated;

Step 4: The formula (1) and the formula (2) are calculated to adjust the v and position of the element;

Step 5: K-means value optimization of particles: The clustering division of new individuals is used to calculate the new cluster centroid, update the particle fitness value and replace the original coding value according to the recent principles and the cluster center coding of particles;

Step 6: If the condition is satisfied, the algorithm is terminated; otherwise it turns to step 2, which satisfies the given value or the maximum cycle number.

4. Implementation of DPSOKmeans clustering algorithm

PSOKmeans clustering algorithm will also fall into local optimal when it meets large scale data. At the same time, in the era of large data, the input data will continue to increase, and the operation of the machine will gradually increase the consumption of resources. This will lead to a continuous decline in machine performance, and even the phenomenon of insufficient memory. Based on the above reasons, a distributed data mining (DDM) algorithm is proposed. DDM distributes data on different sites and uses distributed technology to implement data mining. According to this theory, the massive data is distributed on different machines in parallel to achieve data mining. This method improves the speed of data mining and overcomes the shortcomings of the clustering algorithm.

4.1 Design and implementation of PSOKmeans algorithm based on MapReduce

Short text data can be decomposed into many small data sets, and these small data sets can be processed in parallel. The DPSOKmeans clustering algorithm has no restrictions on the scope of the clustering, while the massive short text data is independent. Therefore, MapReduce can be used to implement parallel computing. According to the description of the distributed parallel mode MapReduce workflow, the distributed parallelization process of the DPSOKmeans clustering algorithm is shown in figure 5.

There are three steps to implement the DPSOKmeans clustering algorithm in MapReduce.

First, all points in the original data set are scanned, and the K points are randomly selected as the centroid of the initial cluster. The initial location is given to each element.

Second, every Map node reads data from HDFS, generates clustering set by DPSOKmeans clustering algorithm. Meanwhile, it generates a new central point at Reduce function. This stage is repeated until the end condition is satisfied.

Finally, all data are classified according to the cluster core generated by the final generation.

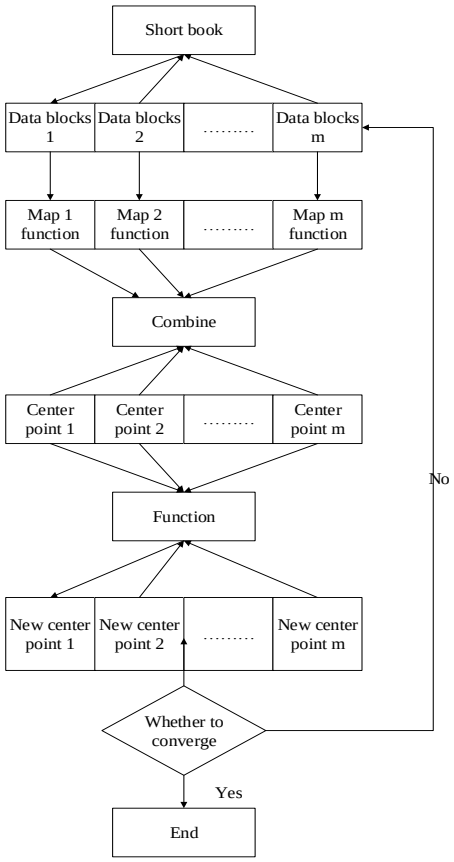

Figure. 5 MapReduce parallelization process of DPSOKmeans clustering algorithm

4.2 Implementation of super large scale short text clustering based on Hadoop

The implementation of the distributed parallel clustering algorithm DPSOKmeans includes the following classes:

DPSOKmeans clustering function class: Two functions of Mapper and Reducer of the DPSOKmeans clustering function are realized, and the clustering process of short text is realized in these two functions. The global optimal cluster is output.

TFIDF class: The number of different feature words in all short texts is counted and used in the Mapper function technique.

Cosine class: Taking the short text as the basic unit, all the characteristic words are extracted, and the cosine distance of the short text is calculated, which is output in the form of the cosine distance matrix.

ClusterGenrator class: The main function of this class is to check the results of this cluster with the conditions of adaptive value, pBest, gBest and so on. Compared with the result of last clustering, if the condition value is less than the last condition value, the center of this cluster should be added to the set, or else to continue to iterate. The operation is finished until the condition is satisfied.

SingleDPSOKmeans class: The main purpose is to implement DPSOKmeans clustering algorithm in single machine mode to deal with large scale short text data, so as to compare with massive data in Hadoop system architecture.

ClusterNum class: This class records the clustering effect of the clustering algorithm proposed in this paper, and compares it with the evaluation criteria of text clustering results.

Distributed and parallelized system architecture: The system architecture used in this article is shown in figure 6. In this system, NameNode and JobTasker are used. Two of them share one computer (and also can be multiple computers) as the main node Master. The main task of Master is to cut the of the short text data, while it is responsible for the maintenance of NameSpace, the distribution of JobTasker and the distribution of the data block. TaskTracker and DataNode also use this machine as the slave

node, which are mainly responsible for the storage of data block. Meanwhile, the state and storage position of DataNode is returned within the specified time. At the same time, it is responsible for the execution of MapReduce task. The functions of components in the system architecture are as follows:

Master control node (JobTracker): Each Hadoop cluster system has only one primary node. The function of this node is to assign an execution plan for the MapReduce task. It assigns the execution nodes to the Map and the Reduce, and monitors the execution of the entire task. Slave node (TaskTracker) is divided into two classes. One is Map and the other is Reduce. As the name implies, the slave node of the Map is responsible for the Map task assigned by the master node, which can have more than one node. However, the slave node of the Reduce is responsible for the Reduce task assigned by the master node, and the node can be more than one.

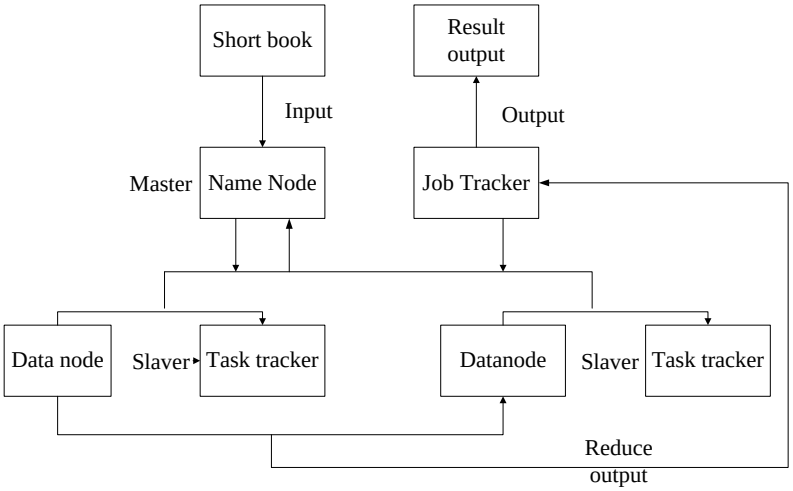

Figure. 6 A distributed parallel system architecture for short text

In the HDFS distributed storage system, the main node NameNode can only have one, and the slave node DateNode can be more than one. The MapReduce structure shows that there is only one JobTracker node, and the TaskTracker node can be more than one. In this experimental environment, we configure the IP-10 machine as Master. Of course, it can also be a slave node. As a matter of reason, the machine can also configure NameNode, DataNode, and JobTracker. The other four machines are Slave nodes. Therefore, it can be configured only as DataNode nodes and TaskTracker nodes.

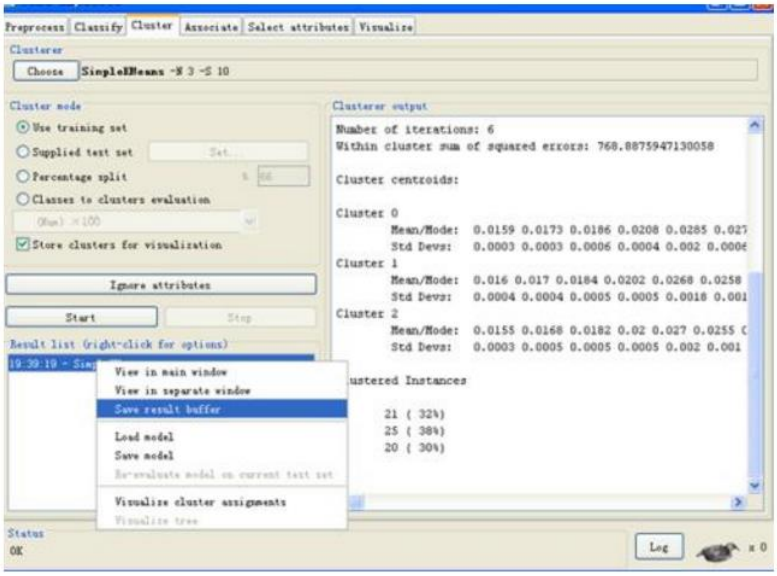

Figure. 7 Implementation of large scale short text processing system

5. Conclusion

The rapid development and popularization of mobile Internet technology has greatly changed the way of people's communication, such as Twitter, Facebook, campus BBS and SMS. The emergence of these platforms has changed the spread way of information. At the same time, these platforms usually update and publish information in the form of short text.

These short text messages have the characteristics of simple content, good focus, colloquial and prominent theme. At the same time, these short texts have the massive data. For example, Twitter releases more than 340 million pieces of information per day, while Sina micro-blog's daily information are more than 500 million pieces, and sometimes more than 500 million. How to dig out the attractive information from these large short text data is the focus of the research. Therefore, in this case, clustering technology and the Hadoop cluster model are generated. In this paper, the related research and the algorithm improvement are further developed.

References

[1] Zehner F, Sälzer C, (2016). Goldhammer F. Automatic coding of short text responses via clustering in educational assessment[J]. Educational and Psychological Measurement, 76(2): 280-303.

[2] Sung K H, Noh E H, Chon K H (2017). Multivariate generalizability analysis of automated scoring for short answer items of social studies in large-scale assessment[J]. Asia Pacific Education Review, 18(3): 425-437.

[3] Huang W, Li Z, Zhang L, et al (2016). Review of intelligent microblog short text processing[C]//Web Intelligence. IOS Press, 14(3): 211-228.

[4] Li L, Ye J, Deng F, et al (2016). A comparison study of clustering algorithms for microblog posts[J]. Cluster Computing, 19(3): 1333-1345.

[5] Gonzalez-Hernandez G, Sarker A, O'Connor K, et al (2017). Capturing the patient's perspective: a review of advances in natural language processing of health-related text[J]. Yearbook of medical informatics, 26(01): 214-227.

[6] Chen T T (2016). The congruity between linkage‐based factors and content‐based clusters—an experimental study using multiple document corpora[J]. Journal of the Association for Information Science and Technology, 67(3): 610-619.

[7] Vo D T, Ock C Y (2015). Learning to classify short text from scientific documents using topic models with various types of knowledge[J]. Expert Systems with Applications, 42(3): 1684-1698.

[8] Cao J, Cui H, Shi H, et al. Big Data: A Parallel Particle Swarm Optimization-Back-Propagation Neural Network Algorithm Based on MapReduce:[J]. Plos One, 2016, 11(6):e0157551.

[9] Cai Q, Gong M, Ma L, et al. Greedy discrete particle swarm optimization for large-scale social network clustering[J]. Information Sciences An International Journal, 2015, 316(C):503-516.

[10] Esmin A A A, Coelho R A, Matwin S. A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data[J]. Artificial Intelligence Review, 2015, 44(1):23-45.

[11] Gunasundari R. EMPWC: Expectation Maximization with Particle Swarm Optimization based Weighted Clustering for Outlier Detection in Large Scale Data[J]. International Journal of Control Theory & Applications, 2016, 9(36):517-531.

[12] Adibifard M, Bashiri G, Roayaei E, et al. Using Particle Swarm Optimization (PSO) Algorithm in Nonlinear Regression Well Test Analysis and Its Comparison with Levenberg-Marquardt Algorithm[J]. International Journal of Applied Metaheuristic Computing, 2016, 7(3):1-23.

[13] Almasi, M. H., Mounes, S. M., & Karim, M. R. (2015). Validating an improved model for feeder bus network design using genetic algorithm (ga) and particle swarm optimization (pso). Journal of the Eastern Asia Society for Transportation Studies, 11(125), 43-9.

[14] Kaur E J, Singh E J. Implementation of an Improved Path Selection Algorithm Using Particle Swarm Optimization (PSO) Technique[J]. Proceedings of SPIE - The International Society for Optical Engineering, 2015, 132(4):150–156.

[15] Esmin A A A, Coelho R A, Matwin S. A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data[J]. Artificial Intelligence Review, 2015, 44(1):23-45.