The Development and Application of "One-Stop" Cluster Analysis Application System Under the Background of Big Data

Zhengju Song[1](✉) and Jun Li[2]

[1] Housing and Urban Rural Construction Administration Bureau of Dongying, Dongying 257000, Shandong, China
64121289@qq.com

[2] Disabled Persons' Federation of Kenli District, Dongying 257000, Shandong, China

Abstract. In the era of big data, the application level of data analysis and processing algorithm determines the presentation of data value. As one of the widely used exploratory analyses, clustering analysis has many ways to realize it, but users still face many difficulties in the actual application process. For this reason, this paper expounds the design and development of functional modules of clustering analysis and processing algorithm based on Hadoop framework and Spark platform, focusing on K-Means clustering algorithm and BIRCH clustering algorithm, and combining Java language development environment to complete the construction of "one-stop" clustering analysis application system. The system is designed in the form of Web application, and the specific operation steps involved in cluster analysis, such as data collection, data cleaning, data storage, data analysis and mining, are highly encapsulated. The special API interface is opened to the outside world, which can widely support all kinds of users. Through simple operation, the cluster analysis of massive data content can be completed, and the data analysis results can be obtained intuitively through data visualization. It not only greatly improves the work efficiency of data analysis, processing and calculation, but also improves the high cost and non-general situation of previous big data clustering analysis systems, and further expands the use dimensions and application scenarios of big data.

Keywords: Big data · Hadoop · Spark · Clustering application system

1 Introduction

At present, with the rapid development of digital information technology, China is gradually stepping into the digital economy era, and the rapid development of digital economy can not be separated from the support of digital technologies such as big data, cloud computing and artificial intelligence. As the core of a new round of scientific and technological revolution and industrial transformation, digital technology is the strategic focus of the country's persistent development at present and in the future. Among them, as a concept and trend of thought, big data originated from the computing field. With

G. Guan et al. (Eds.): ICBDSS 2022, AHCS 8, pp. 235–246, 2023.
https://doi.org/10.2991/978-94-6463-064-0_28

the rapid development and wide application of network information technology, it has formed a new system and pattern covering data infrastructure, data analysis, data application, data resources, open source platforms and tools, etc. In the meantime, the new generation of digital information technologies, such as collaborative cloud computing and artificial intelligence, have shown the trend of integration of digitalization, networking and intelligence, changed their single and auxiliary tool attributes, and become a new driving force to promote national social production and people's daily life.

Now, big data is in the stage of informatization 3.0, which is characterized by deep data mining and fusion application. The design and application of excellent data analysis and processing algorithm is the hot spot sought after by science and technology, and it is also the key to expand the value presentation of massive data and enrich application scenarios. As the most important part of big data analysis and mining technology, data analysis and processing algorithms can be divided into classification or prediction model discovery, data summarization, clustering, association rule discovery, sequential pattern discovery, dependency or dependency model discovery, anomaly and trend discovery, etc. As for massive data, it has the characteristics of low value density, variable data types and complex data situation, so it is necessary to choose appropriate data analysis and processing algorithms to complete data classification to improve data quality and mining efficiency.

As an unsupervised learning algorithm, clustering algorithm can be applied to the early stage of data exploration or mining, and exploratory analysis without prior experience, compared with classification algorithm, and it is also more suitable for data preprocessing in the case of large sample size [3]. However, the computational complexity and flexibility of cluster analysis are high, which is inseparable from the support of a large number of hardware devices and software systems, resulting in high use cost. Moreover, a large number of professional operations also set a higher threshold for ordinary users, lacking channels that can be used conveniently and efficiently. For this reason, this paper holds that, in view of the problems existing in the practical application of cluster analysis, with the help of the application advantages of Hadoop framework and Spark platform, K-Means and BIRCH are the core algorithms of cluster analysis, and a universal "one-stop" cluster analysis application system is constructed. As the main form of the system, JavaWeb can widely support ordinary users to complete the design, operation and analysis of cluster analysis tasks through simple operations such as condition selection, algorithm execution and result display. It not only reduces the difficulty of cluster analysis algorithm, but also improves the efficiency of data analysis and mining, and expands the depth and breadth of big data value application scenarios.

2 Overview of Key Technologies

2.1 Big Data Technology

The big data (mega data) can be called huge data, which refers to the massive, high growth rate and diversified information assets that need new processing modes to have stronger decision-making, insight and process optimization capabilities. It is the inevitable outcome of the development of network information technology, and it is also a new stage of the informatization process after long-term accumulation. The characteristics of huge

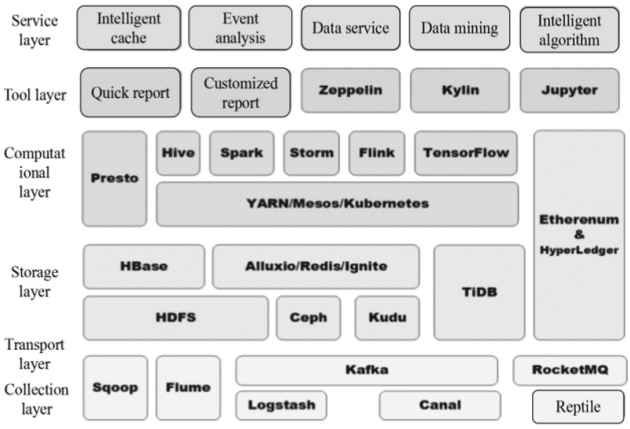

Fig. 1. Big Data technology stack

scale, various data types, fast processing speed, low value density and complex content determine the application direction and development trend of big data, and also pose new challenges to traditional data processing methods and processing thinking in terms of acquisition, storage, management and analysis.

The big data technology, that is, the big data processing technology, focuses on the whole process support for the acquisition of the value of big data from the aspects of collection, storage, analysis and mining, and application presentation. There is an inseparable relationship between each link and process, which also carries the path of the whole data flow, and is also a high collection of many data processing means. With the wide application of big data, the development speed of big data technology is constantly accelerating, and many technical means have gradually gathered and formed a systematic and ecological big data technology stack, as shown in Fig. 1, which is a big data technology stack.

a. **Hadoop**
As the most widely used distributed storage and parallel computing architecture in big data processing, Hadoop can support users to use cluster system to complete the storage, analysis and processing of massive data. As compared with the traditional stand-alone server, Hadoop architecture has outstanding advantages in data extraction, writing and loading, which can effectively improve work efficiency, enhance system stability, improve fault tolerance and reduce system operating costs. The core of Hadoop architecture is distributed file system (HDFS) and distributed computing programming framework (MapReduce), and it also contains components such as HBase, Zookeeper, Pig, Hive and YARN. Many functional components not only enrich and strengthen the functions of Hadoop architecture, but also laterally reflect Hadoop's encapsulation of the underlying details, which can make users pay more attention to the design and development of business logic.

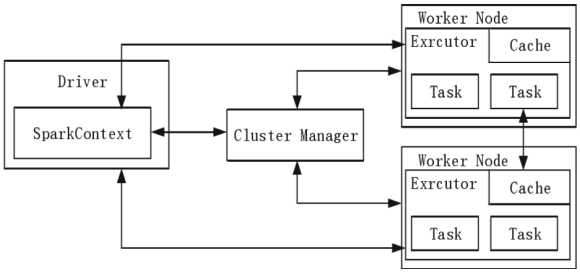

Fig. 2. Spark architecture diagram

b. **Spark**
The Spark, as an open source big data computing engine, is a distributed computing framework based on Yarn. As compared with the batch computing mode represented by MapReduce, Spark framework can support both batch and stream processing computing modes, and effectively simplify the parallel programming running on the cluster system. From the essential point of view, Spark framework inherits the linear extension and high fault tolerance of MapReduce framework, expands and abstracts the bottom operations of resource scheduling, task flow and node communication, and transforms it into a direct API interface to process distributed data, which greatly saves the computing time, improves the resource occupation of the system under the computing state, significantly improves the working speed of iterative operation, and can maintain the stability of the system under high concurrency operation. As shown in Fig. 2, the composition architecture diagram of Spark, in which Driver is the main driver, which is mainly responsible for the construction and transmission of SparkContext information; Work Node is the work node of Spark computing task, which contains many Executor processes, and a single process has corresponding computing task. And Cluster Manager is responsible for the overall resource management of Spark.

2.2 Cluster Analysis Algorithm

Cluster Analysis, also known as cluster analysis, as an unsupervised learning technique, can mine hidden data distribution rules from seemingly irregular data, and it is a classification method widely used in the field of data mining. The core feature of clustering analysis is that the automatic attribution of data can be completed by similarity judgment only by virtue of the characteristics of data without grouping rules. The whole cluster analysis process involves three links, namely similarity comparison, cluster method selection and data result simplification. With regard to some special algorithms, data simplification can be completed automatically, which can reduce the whole process to two links. During the similarity comparison, the similarity degree of each data object under a certain characteristic condition will be expressed according to the distance function or kernel function, and it will be used as the basis for judging whether they can belong to the same “cluster”. The selection of clustering methods will be based on different application scenarios and actual clustering analysis requirements.

Common clustering methods include partition clustering, hierarchical clustering, density clustering, grid classification and model classification. For this study, we focus on the universality of clustering analysis algorithm, and divide clustering into hierarchical clustering and focus on the analysis in the selection of clustering methods.

a. **Partitioning and Clustering**
The purpose of clustering is to divide the data object into many different regions, and the region is the similarity judgment feature, and the region is the cluster, and the number of regions is represented by k. Taking K-Means as the representative, after receiving the input of the sample data set, the number of clusters K is determined, and the center point of each cluster is set, so that the distance between other points in the cluster and the center point is smaller than that of other clusters. Through repeated iterative calculation, the clustering center is updated continuously until the criterion function reaches the convergence condition [7]. The final criterion function of K-Means algorithm is shown in Formula 1, where J represents the smallest sum of squares of errors between data points and center points in a class cluster, C represents the class to which each data point belongs, μ represents the center point in a class cluster, and X represents data points. And in this system, the K-Means algorithm will be realized in Java code.

$$J(\mathrm{c}, \mu) = \sum_{i=1}^{n} \left\| x_i - \mu_{\mathrm{e}^{(i)}} \right\| \tag{1}$$

b. **Hierarchical Clustering**
The purpose of hierarchical clustering is to complete the storage of data objects in the data set through the tree structure, and to complete the classification description by referring to the hierarchical division of the tree structure. There are two ways of hierarchical clustering: top-down and bottom-up. Top-down can decompose data sets layer by layer in the process of clustering analysis. On the contrary, the bottom-up method merges the data sets of layers. The BIRCH algorithm is the representative. After receiving the data set, all the data will be scanned, and the initialized CF tree will be established. The dense data will be clustered and the sparse data will be isolated. Then, the abnormal CF nodes will be filtered continuously. In the screening process, other clustering algorithms can also be used to cluster the leaf nodes of CF tree, so as to further eliminate the unreasonable node structure. As shown in Fig. 3, the basic structure of CF tree of BIRCH algorithm is shown. BIRCH algorithm is also implemented in Java language in this system.

2.3 Development Process

According to the application requirements of the above related application technologies, complete the configuration and deployment of the "one-stop" cluster analysis application system development environment. The development content of the system is divided into two parts. One is the cluster setting of cluster analysis and processing function under Hadoop architecture. Two, under the Java development environment, the Spring framework is adopted to complete the development of the server side of the system.

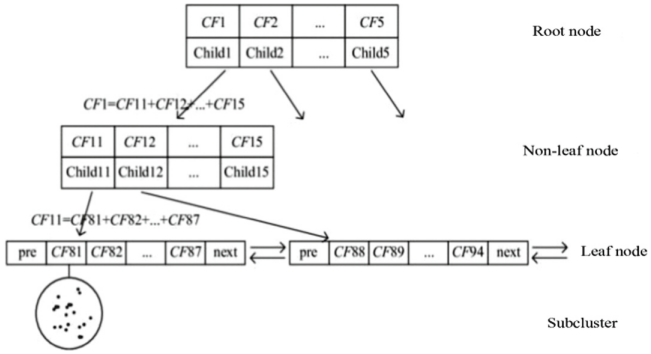

Fig. 3. Basic structure of CF tree of 3: BIRCH algorithm

The Hadoop cluster architecture needs a lot of hardware and software as support, the operating system is Linux, the version is CentOS 6.7(x86_64), and the JDK version is jdk-8u91-linux(x64). According to the application requirements of the system, Hadoop cluster will be set to five nodes, named Master, Slave1, Slave2, Slave3 and Slave4, with Master as the master node and Slave as the slave node. The version of Hadoop is 2.7.7, which is installed in each node, and components such as Yarn, HDFS, Zookeeper and HBase are also deployed in each node. And for the deployment of Spark, the version selection is Spark-2.2.0-bin-hadoop2.7, which also needs to be completed synchronously on the Master and Slave nodes. For the realization of K-Means algorithm and BIRCH algorithm under Spark platform, it needs to go through the steps of injecting dependency, setting attributes, reading and analyzing data, clustering training, and outputting results. The key execution code of K-Means algorithm is shown in Fig. 4.

As for the development environment of JavaWeb application, the operating system is Windows 10.0, the JDK version of development kit is 1.8.0_74, the Web server is Apache Tomcat 9.0, the Java integrated development tool is IntelliJ IDEA 2019, the project management tool Maven 3.5.0, and the database is MySql 8.0. In IntelliJ IDEA, choose to create a maven-archpetype-webapp, and after completing the settings.xml configuration, add all kinds of dependencies under the pom.xml file, including J2EE, Mysql, Spring Framework and other jar packages. And then choose to add Spring framework, and complete the establishment of controller, dao, pojo and service. When finished, complete the corresponding configuration under web.xml. Where <context:annotation-config/> represents the setting for starting spring, and <mvc:annotation-driven/> is the configuration annotation driver. The overall environment of system development, the configuration of related software and tools, and the technical feasibility of the whole project of “one-stop” cluster analysis application system are determined through the introduction of the above key technical theories.

```
import org.apache.spark.{SparkContext, SparkConf}
import org.apache.spark.mllib.clustering.{KMeans, KMeansModel}
import org.apache.spark.mllib.linalg.Vectors
object KmeansTest {
  def main(args: Array[String]) {
    val conf = new SparkConf().setAppName("K-Means Clustering")
    val sc = new SparkContext(conf)
    val rawTrainingData = sc.textFile("/Users/august/Desktop/data/Kmeans/data_training")
    val parsedTrainingData =
   rawTrainingData.filter(!isColumnNameLine(_)).map(line => {
        Vectors.dense(line.split(",").map(_.trim).filter(!"".equals(_)).map(_.toDouble))
      }).cache()
    val numClusters = 8
    val numIterations = 30
    val runTimes = 3
    var clusterIndex: Int = 0
    val clusters: KMeansModel =
      KMeans.train(parsedTrainingData, numClusters, numIterations, runTimes)
    println("Cluster Number:" + clusters.clusterCenters.length)
    println("Cluster Centers Information Overview:")
    clusters.clusterCenters.foreach(
      x => {
        println("Center Point of Cluster " + clusterIndex + ":")
        println(x)
        clusterIndex += 1
      })
    val rawTestData = sc.textFile("/Users/august/Desktop/data/Kmeans/data_training")
    val parsedTestData = rawTestData.map(line => {
      Vectors.dense(line.split(",").map(_.trim).filter(!"".equals(_)).map(_.toDouble))
    })
    parsedTestData.collect().foreach(testDataLine => {
      val predictedClusterIndex:
      Int = clusters.predict(testDataLine)
      println("The data " + testDataLine.toString + " belongs to cluster " +
        predictedClusterIndex)
    })
    println("Spark MLlib K-means clustering test finished.")
  }
  private def isColumnNameLine(line: String): Boolean = {
    if (line != null && line.contains("Channel")) true
    else false
  }
}
```

Fig. 4. Key code of executing K-Means algorithm on 4: Sprak platform

3 The Needs Analysis

3.1 System Requirements Analysis

The "one-stop" cluster analysis application system will aim at many difficulties existing in the process of data mining by ordinary users who lack the professional technical ability of big data, take advantage of Hadoop cluster architecture and Spark computing engine, and take JavaWeb application as the medium to provide a convenient and efficient comprehensive application solution for conventional data cluster analysis.

The system will support ordinary users and administrators to obtain unique account information through user registration, and complete the login and use of the system after authentication. According to the platform application requirements of different user roles, the system will complete the corresponding permission allocation and management. For ordinary users, the system will give you permission to import data, create

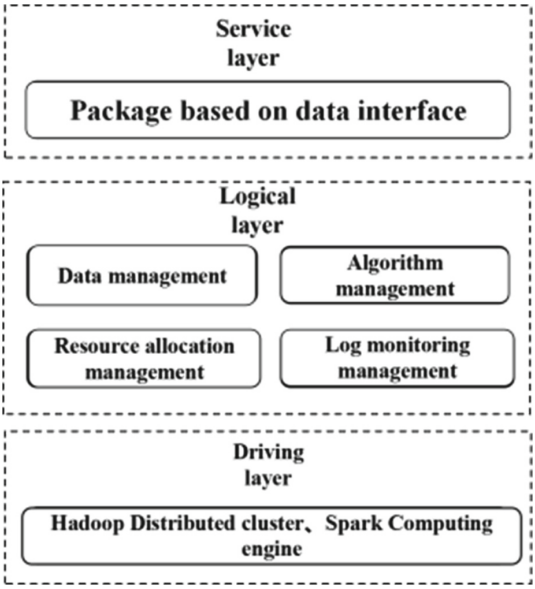

Fig. 5. Overall architecture of the system

new projects, view results, save and export, etc. And the administrator users can maintain and upgrade the user information, algorithm base, data cache and other parts.

3.2 Overall Design

The "one-stop" cluster analysis application system will be based on JavaWeb application design, adopt B/S architecture, and complete the overall design and development with Spring framework in Java language environment. The server-side design adopts the layered design mode, which is divided into three layers as a whole, namely, the driver layer, the logic layer and the service layer. The overall architecture of the system is shown in Fig. 5. [10] Among them, the driver layer consists of Hadoop framework and Spark computing engine, which can provide the basic operation foundation for the whole system. And the logic layer is responsible for data management, algorithm management, resource allocation management, log monitoring management, etc. The service layer encapsulates the logic layer comprehensively, and defines the service interface based on HTTP communication protocol, so as to facilitate users to call from the front-end interface.

4 Detailed Function Realization

4.1 Ordinary Users

The system users can log in and use the system from browsers of different equipment terminals. When an ordinary user enters the account password, the system will log in

```
public class HDFSTest {
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        conf.set("fs.defaultFS", "hdfs://127.0.0.1:9000");
        FileSystem fs = FileSystem.get(conf);
Path("/Users/zhangsf/bigdata/spark-2.4.7-bin-hadoop2.7/logs/spark-zhangsf-org.apache.spark.depl
oy.worker.Worker-1-ever.local.out"),
        new Path("/zhangvalue/input/1111local.out"));
        fs.close();
        FileInputStream in = new FileInputStream(
"/Users/zhangsf/bigdata/spark-2.4.7-bin-hadoop2.7/logs/spark-zhangsf-org.apache.spark.deploy.w
orker.Worker-1-ever.local.out");
        FSDataOutputStream out = fs.create(new Path("/zhangvalue/input/1111local.out"));
        byte[] b = new byte[1024 * 1024];
        int read = 0;
        while ((read = in.read(b)) > 0) {
            out.write(b, 0, read);
        }
    }
}
```

Fig. 6. Key code of data import function implementation

and verify according to the user's identity certificate. This kind of verification method is completed by Basic Auth (HTTP basic authentication), which can be completed quickly and supports various types of client browsers.

When the ordinary users successfully log in, they will first enter the homepage interface. In this interface, users can intuitively get a lot of instructional information. The content includes the introduction of cluster analysis, the application of various algorithms, some actual cases, etc. This part of content can strengthen the understanding and cognition of ordinary users on cluster analysis and data mining, and can also provide necessary knowledge services for users' subsequent cluster analysis. In addition, the system will also present the operation demonstration content of the whole process of cluster analysis project, which includes a detailed description of the combination of pictures and texts, as well as a video tutorial, so as to further help ordinary users without professional knowledge and skills to quickly master the "one-stop" cluster analysis application system.

In the data import module, users can import all kinds of data sets or data contents into the system, and form the original data to provide the necessary foundation for subsequent cluster analysis. This function depends on Java's design of data stream and data interface, in which the FileInputStream () method is needed to create the data storage path of HDFS. The key code is shown in Fig. 6.

Under the new project module, users can create new cluster analysis projects according to their actual needs. First of all, you need to enter basic information such as project name, project introduction, permission statement, and execution period in turn. Secondly, the sample size of cluster analysis is determined, and users can select key fields according to the uploaded data to complete the preliminary analysis, extraction and cleaning of the original data. Then, select the corresponding clustering algorithm, and

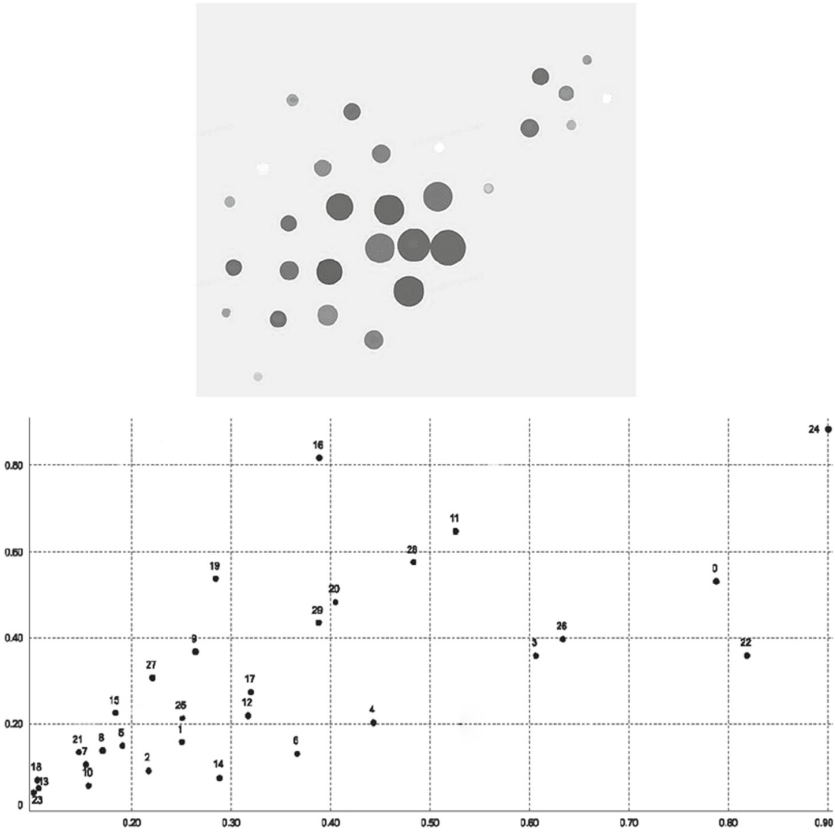

Fig. 7. Scatter chart and detailed information table of cluster analysis

click Cluster Analysis to execute after completion, and the system will automatically execute according to the user's choice.

The user can see the cluster scatter diagram, the cluster detail table and the report information of each key field under the View Results module, as shown in Fig. 7. Among them, each belt in the scatter diagram represents a cluster, the total number of points is the number of clusters K set by the user, the size of points represents the number of samples, and the distance between points represents the similarity of clusters.

The system supports users to save each cluster analysis result and export charts in the form of pictures under the save and export module. On the other hand, ordinary users can query and view all cluster analysis items, which further improves the practicability of the system.

4.2 Administrators

For administrators, the system will support the functions of viewing and managing general user information. On the other hand, the system supports administrators to constantly

update and optimize the clustering analysis algorithm in the system, so as to better meet the changing needs of users and ensure the long-term stable operation of the system.

5 Conclusion

To solve the difficulties faced by many ordinary users in cluster analysis and data mining, the construction of "one-stop" cluster analysis application system can realize the execution function modules of K-Means, BIRCH and other clustering algorithms with the help of Hadoop framework and the data analysis and processing capabilities of Spark platform, and take JavaWeb application as the presentation form, thus effectively providing convenient and efficient comprehensive application solutions for conventional data cluster analysis. It not only reduces the difficulty of cluster analysis algorithm, but also improves the efficiency of data analysis and mining, and expands the depth and breadth of big data value application scenarios. During the follow-up research, we will continue to expand the realization of more clustering algorithms and integrate them into the algorithm library of the system to provide better clustering analysis services for more users.

Bibliography

1. Yao Xuechao. The New Trend of China's Big Data Industry Development [J].Software and integrated circuit.2022.01
2. Cheng Xueqi. Liu Shenghua. Thoughts on the new system of big data analysis and processing technology [J]. CAS Bulletin.2022.01
3. Ji Qiang. Sun Yanfeng. Review of Deep Clustering Algorithm [J].Journals of Beijing University of Technology.2021.08
4. Peng Yu . Pang Jingyue. The Big Data: Connotation, Technical System and Prospect [J].The Journal of Electronic Measurement and Instrument.2015.04
5. Ma Yu. The Research and Application of Clustering Algorithm Based on Spark Platform [D].Hefei University of Technology.2020.04
6. Zhang Yonglai. Zhou Yaojian. Overview of clustering algorithms [J].Computers application.2019.04
7. Wang Zilong. Li Jin. The improved K-means algorithm based on distance and weight [J].The Computer Engineering and Application.2020.10
8. Shang Jiaze. An Weipeng. The Improvement and Analysis of BIRCH Algorithm Based on Threshold [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition).2020.06
9. Wang Yue. Zhang Lei. The Research of Enterprise Web Project Architecture Design Based on Spring [J]. software .2019.06
10. Liu Yang. Research on K-Means clustering algorithm based on Hadoop cloud computing platform [D].Harbin University of Science and Technology.2017.03