Research Article

Key Technologies of Media Data In-Depth Analysis System Based on Artificial Intelligence-Based Big Data

Yi Zheng

Research Center for Intelligent Engineering and Educational Application, Beijing Normal University, Zhuhai 519085, Guangdong, China

Correspondence should be addressed to Yi Zheng; zhengyi2000@bnu.edu.cn

Received 24 August 2021; Revised 22 October 2021; Accepted 27 October 2021; Published 1 December 2021

Academic Editor: Sang-Bing Tsai

Copyright © 2021 Yi Zheng. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

At present, big data related technologies are developing rapidly, and major companies provide big data analysis services. However, the big data analysis system formed by the combination method cannot sense each other and lacks cooperation, resulting in a certain amount of waste of resources in the big data analysis system. In order to find the key technology of the data analysis system and conduct in-depth analysis of the media data, this paper proposes a scheduling algorithm based on artificial intelligence (AI) to implement task scheduling and logical data block migration. By analyzing the experimental results, we know that the performance of LAS (Logistic-Block Affinity Scheduler) is improved by 23.97%, 16.11%, and 10.56%, respectively, compared with the other three algorithms. Based on real new media data, this article analyzes the content of media data and user behavior in depth through big data analysis methods. Compared with other methods, the algorithm model in this paper optimizes the accuracy of hot topic extraction, which has important implications for media data mining. In addition, the analysis results of the emotional characteristics, audience characteristics, and hot topic communication characteristics obtained by the research also have practical value. This method improves the recall rate and F value by 5% and 4.7%, respectively, and the overall F value of emotional judgment is about 88.9%.

1. Introduction

In the field of big data, many excellent products have been tested. Through the combination of these products, a variety of big data analysis systems can be formed. However, in the existing big data analysis system, the parallel processing layer and the data storage layer lack cooperation, which cannot guarantee the locality of tasks, make the system load unbalanced, and ultimately make the system resource utilization rate low. In the existing big data analysis system, the parallel processing layer and the data storage layer lack cooperation. The main reason is that, under the background of the big data environment, there is a large amount of data expansion and the collision of the system's CPU in data processing capabilities. Data processing is full, but a large amount of data is loaded, which affects the processing of the tasks currently being processed, which leads to unbalanced system load.

Artificial intelligence is the study of making computers to simulate certain human thinking processes and intelligent behaviors (such as learning, reasoning, thinking, planning, etc.). It mainly includes the principles of computer realization of intelligence, manufacturing computers similar to human brain intelligence, and making computers achieve higher level applications.

The deep data analysis system is a combination of computer, communication, and control. Liu et al. believe that the hydrostatic system is considered an indispensable support structure in heavy-duty machine tools. The calculation and analysis of hydrostatic bearings are always laborious [1]. Li et al. believe that the high-pressure hydrogenation heat exchanger is an important equipment in the refinery, but it has exposed the leakage problem caused by ammonium salt corrosion. Therefore, it is very important to evaluate the operating status of the hydrogenation heat exchanger. In order to improve the traditional method of

evaluating the operation status of the hydrogenation heat exchanger, they proposed a new method for evaluating the operation status of the hydrogenation heat exchanger based on big data. Aiming at the noise data that is ubiquitous in the industry, they proposed an automatic noise interval detection algorithm. Aiming at the problem that sensor parameters have huge and uncorrelated dimensions, a key parameter detection algorithm based on Pearson correlation coefficient is proposed. Finally, a system-based health scoring algorithm based on PCA (principal component analysis) is proposed to help field operators evaluate the health of heat exchanger hydrogenation [2]. Noguchi et al. believe that as 2020 approaches, Japan's tourism industry is expected to increase. In every possible situation, people need access to stable transportation and communication services and other social infrastructure. NTT is researching and developing solutions to meet these requirements by using big data technology. In their article, they introduced one such advanced, high-performance big data technology and described the results of a field test (Fukuoka test) of services provided to tourists visiting Japan [3]. The calculation and analysis of hydrostatic bearings are always laborious. The diagnostic evaluation system can evaluate the raw material and power consumption of different product companies to avoid the loss caused by the mismatch of raw materials and processes in the production process.

Although there are many solutions for big data analysis systems, the current big data analysis systems also face many challenges. One of them is about the problem of system load balancing. A large-scale big data analysis system is composed of dozens or even hundreds of servers. It is very important to provide efficient data analysis services.

In recent years, artificial intelligence (AI) has become the key to growth in developed countries such as Europe and the United States and developing countries such as China and India [4]. The fields of neuroscience and artificial intelligence have a long and intertwined history [5]. In [6], the author investigates the current state of AI applications in healthcare and discusses its future. The main disease areas using AI tools include cancer, neurology, and cardiology. In [7], the author focuses on decentralized event detection, where sensor nodes use artificial intelligence, data fusion, and distributed pattern recognition performed locally in WSN to collaboratively detect events. In [8], the author emphasizes the most basic features of revolutionary technologies in the 5G era and discusses the relationship between AI and candidate technologies in 5G cellular networks. In [9], the author developed the theory of AI replacement work, which stipulates four intelligent methods required for service tasks: mechanical, analytical, intuitive, and empathetic. In [10], the author provides practical case studies and resource links for AI educators, as well as specific suggestions on how to integrate AI ethics into general AI courses and how to teach independent AI ethics courses. In [11], the author introduced a new artificial intelligence (AI) model for mapping flood sensitivity. The results show that the model is superior to all these models and that the model can be used for sustainable management of flood-prone areas [12]. In [13, 14], the authors evaluated the use of artificial intelligence

platforms on mobile devices to measure drug compliance in stroke patients during anticoagulation therapy. In [15, 16], the author gives a comprehensive review of the application of AI technology in improving the performance of optical communication systems and networks and also reviews applications related to optical network control and management. In [17, 18], the author explores the phenomenon of using artificial intelligence in higher education teaching to predict the future nature of higher education in a world where artificial intelligence is part of our university architecture.

Based on the analysis of the big data in-depth analysis system, this paper proposes a scheduling algorithm based on artificial intelligence (AI) that uses task scheduling and logical data block migration as implementation methods and verifies the algorithm through experiment and analysis verification [19, 20]. Based on real new media data, this article analyzes the content of media data and user behavior in depth through big data analysis methods. The algorithm model in this paper optimizes the accuracy of hot topic extraction and has important implications for media data mining. In addition, the analysis results of the emotional characteristics, audience characteristics, and hot topic communication characteristics obtained by the research also have practical value.

2. In-Depth Analysis Method of Media Data Based on Artificial Intelligence

2.1. Theoretical Basis for In-Depth Analysis of Media Data in the Context of Big Data

2.1.1. In-Depth Analysis of Big Data. Big data analysis is usually a complex process that contains semistructured and unstructured data for user behavior analysis and decision support [21]. Big data is almost everywhere. As an important asset for the survival of enterprises, it is very necessary to save a large amount of production data. But traditional techniques cannot handle such large data sets, so artificial intelligence is often used to process such data. With the help of ML (Machine Learning) and AI (Artificial Intelligence), complex analysis tasks completed on big data are faster than humans imagine. Artificial intelligence has such excellent working ability in data analysis, which is the main reason why artificial intelligence and big data are now inseparable. Artificial intelligence, machine learning, and deep learning learn from each data input and use these inputs to generate new rules for future business analysis [22].

2.1.2. Hadoop Database. Hadoop Database is a column-oriented distributed database with high reliability, high performance, and horizontal scalability, all relying on HDFS; HBase can also provide data support for multiple parallel processing frameworks, such as MapReduce and Spark. A data unit (Cell) in HBase contains a data record; each data unit is unique and is identified by a combination of RowKey, Column Family, Column Qualifier, and Timestamp. The row key is the unique identifier of each row of data, and the

column family is the division of data attributes. The column modifier represents the description of the column data. The column family and the column modifier together form the data column, and the time stamp represents the data insertion time [23]. Each cell uniquely identifies the data stored in the cell through a key consisting of a row key, column family, column qualifier, and time stamp and performs CRUD operations on the data [24].

2.2. Design of Affinity Scheduling Algorithm Based on Big Data Deep Analysis System

2.2.1. Big Data Analysis System Model. In order to describe the relationship between the work content of each component and each component in the big data analysis system more clearly, this paper uses the data model to model the entire system. For big data analysis systems, this paper proposes a discrete-time model [25]. Assume that the entire big data analysis system consists of n virtual machines, and the set of virtual machines is $V = \{v_1, v_2, \dots, v_n\}$. The data analysis task in the system is submitted to the system to run at different times and then is divided into multiple subtasks to be submitted to different working nodes to run. Here, a working node refers to a virtual machine set and a virtual machine in V . At time t , the set of m tasks running in the big data analysis system is denoted as $J_t = \{j_1, j_2, j_3, \dots, j_m\}$. For any task $j_i \in J_t$ in the set, $S_i = \{s_{i,1}, s_{i,2}, \dots, s_{i,pt}\}$ is used to denote the p_i stages of the task.

The storage model of the big data analysis system chooses the HBase storage structure as the modeling target. The entire database consists of k data tables, $D = \{t_1, t_2, t_3, \dots, t_k\}$. Let $T_i = \{l_{i,1}, l_{i,2}, \dots, l_{i,qi}\}$ denote the q_i logical data blocks that make up the T_i table, where logical data block $L_i = \{p_{i,1}, p_{i,2}, \dots, p_{i,wi}\}$ contains w_i physical data blocks [26].

2.2.2. JLQ Algorithm Design (Join Local Queue). The task of the JLQ (Join Local Queue) scheduling algorithm is to distribute subtasks to the corresponding executor for execution, in order to simplify the model complexity. In the JLQ model, it is assumed that each worker node is configured as a single CPU; that is, only one subtask can be executed at a time. For a big data analysis system with n task execution nodes, JLQ generates a set consisting of n queues, with Q_i representing the i -th queue in the set, corresponding to the subtask set assigned to worker node S_i . At time t , vector $Q(t) = \{Q_1(t), Q_2(t), \dots, Q_n(t)\}$ represents all subtasks in the system at the current state. Vector $Q(t)$ updates internal data when two subtasks are issued and a subtask has been completed [27].

$$f_i(t) = \begin{cases} -1 \\ j \end{cases} \quad (1)$$

In this paper, $f_i(t)$ represents the execution status of node i 's current task, as shown in the above formula. When the value of $f_i(t)$ is -1 , it means that node i is in idle state at time t and can perform tasks. If node i is performing a task, the

value of $f_i(t)$ j indicates that the current node is executing a task from the j -th queue.

2.2.3. LAS Algorithm Design (Logistic-Block Affinity Scheduler). LAS (Logistic-Block Affinity Scheduler) scheduling algorithm is an independent module in the big data analysis system [28]. The main task is to determine the hot spots of the nodes by detecting the status of each working node in the system and, based on the results of the determination, analyze whether the migration of logical data blocks is required to ensure the locality of the program or reduce the possibility of hot spots. In the LAS algorithm, set $H(t) = \{h_1(t), h_2(t), \dots, h_n(t)\}$ is used to represent the load state of the n working nodes in the current system at time, where $h_i(t)$ represents the load state of node i at time t . The value of $h_i(t)$ is as shown below. When the value of $h_i(t)$ is 0, it indicates that the node is in a light load state, and the value of $h_i(t)$ 1 represents that the working node i is in a hot state [29].

$$h_i(t) = \begin{cases} 0 \\ 1 \end{cases}. \quad (2)$$

In order to clearly represent the relationship between subtasks and logical data blocks in the big data analysis system, set $U(t) = \{U_1(t), U_2(t), \dots, U_m(t)\}$ is used in the LAS algorithm model. At time t , there are a total of m logical data blocks in the distributed storage system. $U_i(t)$ in the model represents the logical data block t with the number i , where $U_i(t) = w_{i,1}(t) + w_{i,2}(t) + \dots + w_{i,n}(t)$ and n represents the total number of working nodes in the cluster, each addend in $U_i(t)$, and $w_{i,j}(t)$ represents the number of logical data blocks i used in working node j at time t .

2.3. In-Depth Analysis Model of Media Data Based on Emotion Judgment and Data Characteristics. The data analysis model means how to construct a structure in the data analysis process for specific implementation and establishment, including what data to collect, what content to extract as the basis for analysis, and which analysis methods, algorithms, and statistical items need to be more appropriate. The results of the analysis now make the data association or change trend clear at a glance, thereby more clearly showing the data development trend and logical association; you can also use the report summary method to present the data items that need to be highlighted, such as statistical sum, mean, proportion, ranking, best value, etc., which clearly show the association, statistics, and changes of data. They add color to data, quickly convey information, show the relationship between data more accurately, and highlight key points.

2.3.1. Model Inference. First, the update rule of the hidden variable Z^e is given, such as the formula

$$p(z_i^e = k | w, z_{-i}^e, \alpha^e, \beta^e) \propto \frac{n_{k,d,-i}^e + \alpha^e}{\sum_{k=1}^K n_{k,d,-i}^e + K\alpha^e} \times \frac{(n_{wk}^e + \beta^e)}{\sum_{w=1}^W n_{wk}^e + W\beta^e}. \quad (3)$$

The symbol “ $-i$ ” here means the count variable value after excluding the word whose index number is i in the user corpus. The symbol $n_{k,d,-i}^e$ indicates the number of subject terms k generated from the user document d_e in addition to the current allocation. Symbol n_{wk}^e indicates the number of times the subject word w is generated from the subject k . Next, sample the hidden variable l_e of the user sentiment index according to the following formula:

$$p(l_i^e = m | s^e, l_{-i}^e, \beta^e, \eta^e) \propto \frac{n_{sm,-i}^e + \eta_m^e n_{m,d}^e}{\sum_{s=1}^S n_{sm,-t}^e + S^e \eta^e N_{k,d}^e} \times \frac{n_{m,d}^e}{N_{k,d}^e}. \quad (4)$$

The symbol $n_{sm,-i}^e$ indicates the number of times the emotion word s is generated from the emotion index $l_e = m$ in addition to the current assignment. Symbol $n_{m,d}^e$ indicates the number of times that the sentiment index of the sentiment word in the user document d_e is m . The symbol indicates the total number of occurrences of all words in user document $N_{k,d}^e$. After a series of Gibbs sampling, the values of user document-topic distribution, user topic-word distribution, and user topic-emotional distribution can be calculated approximately according to the obtained parameters. The specific calculation method is shown in the following three formulas:

$$\begin{aligned} \theta_{k,d}^e &= \frac{n_{k,d,-i}^e + \alpha^e}{\sum_{k=1}^K n_{k,d,-i}^e + K^e \alpha^e}, \varphi_{wk}^e = \frac{n_{wk}^e + \beta^e}{\sum_{w=1}^W n_{wik}^e + W^e \beta^e}, \pi_{sm}^e \\ &= \frac{n_{sm,-t}^e + \eta^e}{\sum_{s=1}^S n_{sm,-t}^e + S^e \eta^e}. \end{aligned} \quad (5)$$

2.3.2. Emotional Score Acquisition. The model can be used to calculate the sentiment value of each sentiment word (between -1 and 1). The closer the sentiment value of the word is to 1 , the more likely it is to be a positive vocabulary; otherwise it is more likely to be a negative vocabulary. According to this model, the sentiment score for each topic is shown in the formula

$$Q(z, l) = \frac{1}{2} \left[\sum_{w=1}^{N_{wk}} p(w|z = k) \cdot Q_w + \sum_{s=1}^{N_{sk}} p(s|l = k) \cdot Q_s \right], \quad (6)$$

$Q(s)$ and $Q(w)$ are the sentiment scores of a specific word. $Q(s)$ represents the sentiment word score, and $Q(w)$ represents the text word score. E represents the overall emotional orientation of a topic.

3. Experimental Data and Evaluation Methods

3.1. Acquisition of Media Data Sets. This article uses two parts of data. The collected data mainly comes from media data, which is mainly used to measure the proportion of nearby activities. Then there is the friend information data of social media users. This part of the data will become a data

source for measuring the connection between the target user and the local people. By using the two parts of the data, we can collect the information we need very well. The media data set contains the following attributes: tweet number, tweet text, time stamp, name and number of the location, and longitude and latitude of the location. For Twitter users, the data set contains the following attributes: user ID, user name, user description, user location, number of friends, and number of tweets. The data set contains 23.45 million tweets and a total of 819,000 users. The earliest tweet in the data set was sent on July 11, 2018, and the last tweet was posted on December 1, 2019. Therefore, the total time interval of the case study data set used is 505 days.

Table 1 shows the data source characteristics of the original data set and the sample data set. The sample data set contains 304,000 tweets totaling 10,000 users, and about 12% of the tweets have coordinates. And the average number of tweets, proportion of coordinated tweets, and time interval of each user in the sample data set are basically consistent with the original data set.

3.2. Experimental Environment of Deep Data Analysis System. In this paper, the distributed environment composed of Spark and HBase is used as the big data analysis system used in the experiment. The entire system is deployed on the Alibaba Cloud ECS server. A master node is used in conjunction with 4 slave nodes. The LAS module is placed on the master node to run, and there is no separate virtual machine setup. The virtual machine hardware settings are shown in Table 2 and the software version is shown in Table 3 throughout the experiment.

Each section in the big data analysis system uses CentOS 6.5 as the virtual machine operating system, as shown in Table 2. The CPU model of the physical machine where the master node and the slave node are located is the same. The five nodes all use 4-core CPU and 8 GB memory configuration. The difference is that, in addition to the 40 GB hard disk used by the system, the master node mounts an additional 50 GB hard disk, and each slave node mounts an additional 200 GB hard disk for environment deployment and experimental data storage. The software environment used in the experiment is shown in Table 3. The entire system adopts the stable version combination of Hadoop 2.6.0, HBase 1.0.3, and Spark 1.4.0 as the big data analysis system used for experimental testing, and the basic operating environment JDK version is 1.8.0.

3.3. Evaluation Index of Data Depth Analysis System

3.3.1. Evaluation Index of Data Distribution. In order to compare the performance of various algorithms in the same environment in the experiment, this paper presents an important data distribution evaluation index: distribution deviation. The distribution deviation is expressed by λ , and the value is in the interval $[0, 1]$, which represents the balance between the distribution of all logical data blocks in the database and the Spark data analysis task.

TABLE 1: Data set description.

	Original data set	Sample data set
Amount of users	819,000	10,000
Number of tweets	23,450,000	304,000
Average number of tweets per user	28	30
Tweet scale with coordinates	12.56%	12.04%
The earliest time to send a tweet	2018/7/11	2018/7/11
The latest time to send a tweet	2019/12/1	2019/12/1

TABLE 2: Node hardware configuration.

Master node hardware configuration	Processor model	Intel(R) Xeon(R) CPU E5-2630 @ 2.50 GHz
	Number of processor cores	4
	RAM	8 G
	Hard disk	40 + 50 GB
Slave node hardware configuration	Processor model	Intel(R) Xeon(R) CPU E5-2630 @ 2.50 GHz
	Number of processor cores	4
	RAM	8 G
	Hard disk	40 + 200 GB

TABLE 3: Experimental environment software version.

Name of software	Hadoop	HBase	Spark
Version number	2.6.0	1.0.3	1.4.0

In order to calculate the value of λ , we first need to introduce three important matrices: A , B , and D . The following formula is of matrix A expression, where n represents n working nodes in the system, and k represents k tables in HBase. The temporary expression $C_{i,j,l}$ indicates whether a logical data block l in HBase belongs to table j and is distributed on worker node i .

$$A = (a_{i,j})_{[n \times k]}, a_{i,j} = \sum_{l=1}^{p_j} C_{i,j,l}, C_{i,j} = \begin{cases} 1 \\ 0 \end{cases}. \quad (7)$$

The matrix B expression is shown in the following formula. The k value has the same meaning as the matrix A , and m represents m data analysis programs in the system. The temporary expression $E_{i,j,l}$ indicates whether the l th stage of the data analysis task in the system needs to access the data table i , as shown in the formula. Therefore, each element $b_{i,j}$ in the matrix B represents the data access relationship between the table numbered i and the table numbered j in the system. The larger the value of $b_{i,j}$, the closer the data dependence.

$$B = (b_{i,j})_{[k \times m]}, b_{i,j} = \sum_{l=1}^{q_j} E_{i,j,l}, E_{i,j} = \begin{cases} 1 \\ 0 \end{cases}. \quad (8)$$

Matrix D is the product of matrix A and matrix B ; see the formula below. The matrix D represents the connection between the data analysis program in the cluster and the data distribution on each node.

$$D = (d_{i,j})_{[n \times m]} = A \times B. \quad (9)$$

3.3.2. Algorithm Evaluation Index. During the experiment, three scheduling algorithms were selected for comparative experiments: HFS, JSQ-MAX Weight, and LTF. In the experimental analysis process, in order to more intuitively complete the performance comparison between the comparison algorithm and the logical data block affinity scheduling algorithm, this article uses query execution time (QET) as an important comparison condition for the comparison experiment. QET represents the program execution delay, that is, the time from the start of sending a data request to the acquisition of all data, as shown in the following formula. QET_{other} represents the query time of the comparison algorithm, and QET_{LAS} represents the query time of the affinity scheduling algorithm. If $\text{Imp}(\text{other})$ is greater than 0, it means that the efficiency of the logical data block affinity scheduling algorithm is higher than that of the comparison algorithm. Otherwise, it means that the efficiency is lower than the comparison algorithm.

$$\text{Imp}(\text{other}) = \frac{QET_{\text{other}} - QET_{\text{LAS}}}{QET_{\text{other}}} \times 100\%. \quad (10)$$

4. Results and Discussions

4.1. Comparison of Media Data Query and Analysis Results under the Background of Big Data Based on Artificial Intelligence

4.1.1. Simple SQL Query. This article selects Q12 as a comparative use case for simple SQL query operations. Q12 query description is as follows: given the year, month, and category of media data, find out all users who have searched for media data of the specified category on the web site for that month and have used this type of media data in the following three months.

The execution time of Q12 in various data distribution conditions and different algorithms is shown in Figure 1. It can be seen from the comparison and analysis in Figure 1 that, under simple SQL query operations such as Q12, the distribution deviation λ has a certain impact on the operation time and the execution time of the algorithm. In the three comparison algorithms, the data query time becomes longer with the increase of the distribution deviation λ , indicating that the original distribution state of the data will affect the execution efficiency of the three comparison algorithms. The execution time of the proposed algorithm is not affected by the distribution deviation λ and is shorter than the execution time of the three algorithms. It proves that, in a simple SQL query program, the performance of the proposed algorithm is significantly improved.

4.1.2. Distributed SQL Query and Analysis. Choose Q2 as a comparative example of distributed SQL query and analysis. Q2's query description is as follows: given a type of media data, find the products that are often browsed online with the media data and ranked in the top 30 page views, and the user session timeout is set to 60 min.

The query time of Q2 in each algorithm is shown in Figure 2. As can be seen from Figure 2, among various algorithms, the JSQ-MaxWeight algorithm generally performs the longest in distributed SQL query and analysis operations, and there are large data fluctuations. The HFS algorithm and LTF algorithm gradually increase with the distribution deviation λ , and the query time is similar. There are still cases where the value of the distribution deviation λ becomes larger and the algorithm execution time becomes longer. The experimental results of the logical data block affinity scheduling algorithm are relatively stable; that is, when the distribution deviation λ value is 0, the data query and processing time are similar to the other three algorithms. However, as the distribution deviation λ becomes larger, the polyline of the execution time before and after is relatively smooth. The execution time of each data distribution is the shortest of the four algorithms.

4.1.3. Data Analysis Realized by Custom Function. Q1 is a typical data analysis implemented by a custom function. The query about Q1 is described as follows: Find out 100 types of media data that are often browsed together in the specified media data.

The execution time of Q1 in each algorithm is shown in Figure 3. It can be seen from the analysis in Figure 3 that the peak value appears in the line chart during the change of the distribution deviation λ in the three comparison algorithms. When the distribution deviation λ value is 0.5 to 0.6, the program execution time of the three comparison algorithms becomes significantly longer, and the results of the JSQ-MaxWeight algorithm fluctuate greatly. Among the four algorithms, the data affinity scheduling algorithm is relatively stable, is less affected by the data distribution, and has the best improvement effect.

FIGURE 1: Q12 execution time comparison chart.

FIGURE 2: Q2 execution time comparison chart.

FIGURE 3: Q1 execution time comparison chart.

4.2. Comparison of Media Data Testing and Adaptability Analysis Results under the Background of Big Data Based on Artificial Intelligence. The media data set uses the data format and collection time as the table division scheme, which represents all the data collected on May 29, 2019. In fact, the tasks X1, X2, and X3 performed by the data set are all the same data analysis program but are only used for data analysis performed on different dates, and the data accessed is irrelevant to each other. The test results of media data under the background of

big data based on artificial intelligence are shown in Figure 4, respectively, the test results of the media data set and the comparison results of the polyline. Through calculation, the performance of the LAS algorithm is 23.97% higher than the HSF algorithm, 16.11% higher than the JSQ-MAX algorithm, and 10.56% higher than the LTF algorithm. It can be seen from the chart analysis that in the big data analysis system for multitask execution, various algorithms fluctuate to a certain extent as the distribution deviation λ changes from 0 to 1. In general, the larger the value of the distribution deviation λ , the longer the program execution time. When the distribution deviation λ is 0, the effect of the logical data block affinity scheduling algorithm and the LTF algorithm is almost the same, but as the distribution deviation λ becomes larger, the logical data block affinity scheduling algorithm is smaller than the LTF algorithm in execution time. The LTF algorithm is better than the HFS algorithm and the JSQ-MaxWeight algorithm. The HFS algorithm program has the longest execution time and the affinity scheduling algorithm program has the shortest execution time. The scheduling results are the best among all algorithms, and the performance of the big data analysis system is fully utilized.

The test results of the media data in-depth analysis system are shown in Figure 5. Through the above analysis, it can be seen that the logical data block affinity scheduling algorithm has a good performance advantage in a large data analysis system in which multiple data query operations and multiple applications coexist. When the distribution deviation λ value is 0, the logical data block affinity scheduling algorithm and the optimal algorithm among the three comparison algorithms have little difference in execution time. However, with the increase of the distribution deviation λ , the execution time of the three comparison algorithms obviously increased or fluctuated. The data affinity scheduling algorithm has a small change interval during the execution of a single application, and there is a certain degree of increase when multiple applications are parallel, but the relative change is not large. When the distribution deviation λ value is 0, it means that each table in HBase satisfies load balancing in units of tables; when the distribution deviation λ value is 1, it means that all tables corresponding to the task are on the same node. From the above analysis, it can be seen that the effect of the data affinity scheduling algorithm when the data is in the best distribution state is not much different from the Spark default scheduling algorithm HFS. But it is applicable to the situation where the data distribution is relatively uneven.

4.3. In-Depth Analysis Results of Media Data under the Background of Big Data Based on Artificial Intelligence

4.3.1. Emotional Judgment. Figure 6 compares the results of this model and the basic method. After calculation, in the judgment of the emotional characteristics of new media content, the accuracy rate of the method in this paper is 87.7% for positive emotion judgment, 91.1% for negative emotion judgment, and 88.1% for neutral emotion judgment. The overall accuracy rate of emotional judgment is about 89.0%, and the accuracy rate has increased by 4%. The

FIGURE 4: Comparison of data set execution time.

FIGURE 5: Comparison of execution time of the data set.

judgment recall rate for positive emotions is 88.1%, the recall rate for negative emotion judgments is 88.4%, the recall rate for neutral emotion judgments is 89.8%, the overall recall rate for emotion judgments is about 88.8%, and the recall rate is increased by 5%. The F value for positive emotion judgment is 88.0%, the F value for negative emotion judgment is 89.7%, the F value for neutral emotion judgment is 88.9%, and the overall F value for emotion judgment is about 88.9%, an increase of 4.7%. After testing, our model based on emotional judgment and data characteristics has better judgment results than other research results, which improves the accuracy of judgment, recall rate, and F value, and can meet the relevant requirements of the application.

Figure 7 reflects the distribution ratio of sentiment tendency of the article headlines of today's Toutiao platform and Yidian information platform. From Figure 7, we can see that nearly 34% of the article titles of a little information platform show negative emotions, nearly 46% show neutral emotions, and nearly 20% show positive emotions. Today's Toutiao platform's article titles have 32% negative emotions, 44% neutral emotions, and 24% positive emotions. It can be seen from this that although the sentiment span of the article headlines of today's Toutiao platform is large, the sentiment distribution is more scattered. But overall, the headlines of today's headlines are more positive.

FIGURE 6: Emotional judgment results of media data. (a) Basic sentiment dictionary. (b) The model of this article.

FIGURE 7: Comparison of the emotional tendency of the headlines of "today's headline" and "little news."

FIGURE 8: Content richness analysis of each new media platform.

4.3.2. Data Characteristics. Figure 8 reflects the calculation results of the richness of the content of each new media platform. We arranged the calculation results of the articles on the two platforms in descending order, using the article serial number as the horizontal axis and the content richness value as the vertical axis. When the content richness calculation result is greater than 1, the content richness is higher than the average level; when the content richness calculation result is less than 1, the content richness is lower than the average level. It can be seen from Figure 8 that the content richness of today's headlines and little information platform articles conforms to the average distribution. About 37% of the article content belongs to the relatively rich content, and the remaining 63% of the article content belongs to the relatively rich content. At the same time, comparing the distribution of the two curves in the figure, we can find that the content richness of today's Toutiao platform articles will be slightly higher than a little information.

Figure 9 reflects the results of user engagement calculations for each new media platform content. We arranged the calculation results of user engagement of the two platforms in descending order, using the article serial number as the horizontal axis and the user engagement value as the

FIGURE 9: Analysis of user participation of various new media platforms.

vertical axis. When the user participation calculation result is greater than 1, the user participation degree of the content is higher than the average level; when the user participation calculation result is less than 1, the user participation degree of the content is lower than the average level.

We can see from Figure 9 that the user participation of today's headlines and little information platform articles conforms to the long-tail distribution. About 20% of the content of the article belongs to the content with high user

engagement, and they tend to attract 80% of user reviews and user interactions. The remaining 80% of the content of the article belongs to the content with low user participation, and they tend to divide the remaining 20% of the comment volume and interaction level. The conclusion of this analysis is in line with the law of practice. The more popular reviews and interactive articles will attract more people to comment on the interaction, and the topic is easy to ferment. At the same time, comparing the distribution of the two curves in the figure, we can find that the user participation of Toutiao platform today is closer to that of a little information platform.

5. Conclusions

Based on the analysis of the big data in-depth analysis system, this paper proposes a scheduling algorithm based on artificial intelligence (AI) that uses task scheduling and logical data block migration as implementation methods and verifies the algorithm through experiment and analysis verification. By conducting multiple sets of experiments in the existing big data set, the experimental results are analyzed. The performance of the LAS algorithm is 23.97% higher than the HSF algorithm, 16.11% higher than the JSQ-MAX algorithm, and 10.56% higher than the LTF algorithm. Obtain the scheduling algorithm to improve the execution time of the program in single-task execution and multitask parallelization, respectively, and find the applicable scenario of the scheduling algorithm by analyzing the results during the operation of multiple programs.

Based on real new media data, this article analyzes the content of media data and user behavior in depth through big data analysis methods. Compared with sentiment analysis based on the basic sentiment dictionary, the analysis accuracy is optimized. Compared with other methods, the algorithm model in this paper optimizes the accuracy of hot topic extraction, which has important implications for media data mining. In addition, the analysis results of the emotional characteristics, audience characteristics, and hot topic communication characteristics obtained by the research also have practical value. This method improves the recall rate and F value by 5% and 4.7%, respectively, and the overall F value of emotional judgment is about 88.9%.

Data Availability

No data were used to support this study.

Conflicts of Interest

The author states that this article has no conflicts of interest.

Acknowledgments

The author received no financial support for the research, authorship, and/or publication of this article.

References

- [1] Z. Liu, Y. Wang, L. Cai, Q. Cheng, and H. Zhang, "Design and manufacturing model of customized hydrostatic bearing system based on cloud and big data technology," *International Journal of Advanced Manufacturing Technology*, vol. 84, no. 1–4, pp. 261–273, 2016.
- [2] L. Li, X. Liu, R. Du, and P. Xu, "Application of big data technology in evaluation of operating status of high-pressure hydrogenation heat exchanger," *China Petroleum Processing and Petrochemical Technology*, vol. 20, no. 3, pp. 17–23, 2018.
- [3] K. Noguchi, Y. Sato, and H. Shiohara, "Advanced, high-performance big data technology and trial validation," *NTT Technical Review*, vol. 14, no. 2, pp. 1–7, 2016.
- [4] X. Leng, G. Chen, Y. Jiang, J. Zhang, and Y. Cao, "Data specification and processing in big-data analysis system for monitoring and operation of smart grid," *Dianli Xitong Zidonghua/Automation of Electric Power Systems*, vol. 42, no. 19, pp. 169–176, 2018.
- [5] X. Leng, G. Chen, J. Bai, and J. Zhang, "General design of smart grid monitoring operation big data analysis system," *Dianli Xitong Zidonghua/Automation of Electric Power Systems*, vol. 42, no. 12, pp. 160–166, 2018.
- [6] B. Zhang, X. Wang, and Z. Zheng, "The optimization for recurring queries in big data analysis system with mapreduce," *Future Generation Computer Systems*, vol. 87, pp. 549–556, 2018.
- [7] F. Song, "Application of discourse interaction data analysis system in teaching evaluation for TCSOL," *Journal of Language Teaching and Research*, vol. 8, no. 5, p. 922, 2017.
- [8] X. Wang, "The construction and realization of big accounting data analysis system based on cloud computing platform," *Clinica Chimica Acta*, vol. 42, no. 2, pp. 570–574, 2017.
- [9] D. Liao, "University logistics management data analysis system based on business intelligence," *Revista de la Facultad de Ingenieria*, vol. 32, no. 14, pp. 635–640, 2017.
- [10] C. Marques, J. Castanheira, and A. Rocha, "The ligo data analysis system," *Climate Dynamics*, vol. 42, no. 7–8, pp. 1887–1903, 2017.
- [11] P. K. M. Kanaujia, M. Pandey, M. Pandey, and S. Swarup Rautaray, "A framework for development of recommender system for financial data analysis," *International Journal of Information Engineering and Electronic Business*, vol. 9, no. 5, pp. 18–27, 2017.
- [12] P. Tejaswi, M. Bhavana, P. S. Sreeram, and J. S. Kiran, "Text data analysis computer aided automated assessment system: a survey," *Journal of Advanced Research in Dynamical & Control Systems*, vol. 11, no. 2, p. 6, 2019.
- [13] H. Fang, K. Zhou, Y. Zou, W. Deng, X. He, and J. Zhou, "A diagnosis optimization system for grain processing based on multiple data analysis algorithms," *Systems Science & Control Engineering*, vol. 7, no. 2, pp. 96–107, 2019.
- [14] S. Valsamidis, S. Kontogiannis, T. G. Theodosiou, and I. Petasakis, "A web e-voting system with a data analysis component," *Journal of Systems and Information Technology*, vol. 20, no. 1, pp. 33–53, 2018.
- [15] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, vol. 95, no. 2, pp. 245–258, 2017.
- [16] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain intelligence: go beyond artificial intelligence," *Mobile Networks and Applications*, vol. 23, no. 2, pp. 368–375, 2018.

- [17] F. Jiang, Y. Jiang, H. Zhi et al., "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.
- [18] P. J. M. Havinga, N. Meratnia, and M. Bahrepour, "Artificial intelligence based event detection in wireless sensor networks," *University of Twente*, vol. 85, no. 6, pp. 1553–1562, 2017.
- [19] R. Li, Z. Zhao, X. Zhou et al., "Intelligent 5g: when cellular networks meet artificial intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, 2017.
- [20] M.-H. Huang and R. T. Rust, "Artificial intelligence in service," *Journal of Service Research*, vol. 21, no. 2, pp. 155–172, 2018.
- [21] E. Burton, J. Goldsmith, S. Koenig, B. Kuipers, N. Mattei, and T. Walsh, "Ethical considerations in artificial intelligence courses," *AI Magazine*, vol. 38, no. 2, pp. 22–34, 2017.
- [22] K. Chapi, V. P. Singh, A. Shirzadi et al., "A novel hybrid artificial intelligence approach for flood susceptibility assessment," *Environmental Modelling & Software*, vol. 95, pp. 229–245, 2017.
- [23] S. Price and P. A. Flach, "Computational support for academic peer review," *Communications of the ACM*, vol. 60, no. 3, pp. 70–79, 2017.
- [24] D. L. Labovitz, L. Shafner, M. Reyes Gil, D. Virmani, and A. Hanina, "Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy," *Stroke*, vol. 48, no. 5, pp. 1416–1419, 2017.
- [25] K. W. Johnson, J. Torres Soto, B. S. Glicksberg et al., "Artificial intelligence in cardiology," *Journal of the American College of Cardiology*, vol. 71, no. 23, pp. 2668–2679, 2018.
- [26] J. Mata, I. De Miguel, R. J. Durán et al., "Artificial intelligence (ai) methods in optical networks: a comprehensive survey," *Optical Switching and Networking*, vol. 28, pp. 43–57, 2018.
- [27] S. A. D. Popenici and S. Kerr, "Exploring the impact of artificial intelligence on teaching and learning in higher education," *Research and Practice in Technology Enhanced Learning*, vol. 12, no. 1, p. 22, 2017.
- [28] S. Namasudra and P. Roy, "PpBAC," *Journal of Organizational and End User Computing*, vol. 30, no. 4, pp. 14–31, 2018.
- [29] R. Kanniga Devi, M. Gurusamy, and P. Vijayakumar, "An efficient cloud data center allocation to the source of requests," *Journal of Organizational and End User Computing*, vol. 32, no. 3, 2020.