Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9981581, 13 pages
https://doi.org/10.1155/2021/9981581

Research Article

Influencing Factors and Forecasting Statistics of Enterprise Market Sales Based on Big Data and Intelligent IoT

Zhen Guo[1] **and Tao Zou**[2]

[1]*Business School, University of Queensland, Brisbane Qld 4072, Australia*
[2]*School of Management, Northwestern Polytechnic University, Xi'an, 710000 Shaanxi, China*

Correspondence should be addressed to Tao Zou; zoutolele@mail.nwpu.edu.cn

Received 2 April 2021; Revised 27 May 2021; Accepted 25 June 2021; Published 26 July 2021

Academic Editor: Bo Rong

With the acceleration of economic development, enterprise management is facing more severe challenges. Big data analysis based on the intelligent Internet of Things (IoT) has a positive effect on the development of enterprise management and can make up for the shortcomings of enterprise management. In this paper, we develop a big data processing method based on intelligent IoT which can mine the factors that affect the company's market sales from the collected data. Then, we propose a KNN classification algorithm based on overlapping k-means clustering. This algorithm adds a training process to the traditional KNN algorithm, which can accurately classify data and greatly improve the efficiency of the classification algorithm. Numerical analysis results prove the effectiveness of the proposed algorithm.

1. Introduction

In recent years, with the rapid development of social economy and science and technology, big data analysis in Intelligent Internet of Things (IoT) has been widely applied to various industries and fields. It can help enterprises find problems existing in management and promote the improvement of enterprise management level. Applying big data analysis in intelligent IoT to enterprise management can adapt to the changes of internet enterprises, help enterprises better cope with various challenges, and lay a good foundation for the sustainable development of enterprises.

The development of modern science and technology also drives the development of intelligent IoT industry. Intelligent IoT is a concept that emerged in 2018. It refers to the system collects all kinds of information in real time through various information sensors (generally in the context of monitoring, interaction, and connection) and makes intelligent analysis of data through machine learning in terminal devices, edge domains, or cloud centers, including positioning, comparison, prediction, and scheduling. At the technical level, AI enables the Internet of Things to acquire the perception and recognition ability, and the Internet of Things provides the data of the training algorithm for AI.

As big data applications continue to penetrate into all walks of life around the world and take root, traditional data management methods no longer meet the data management needs of enterprises [1]. In the future, the surrounding environment of enterprises in business activities is unpredictable. To increase the risks and opportunities in its business activities, if it can predict the sales volume in the business activities, use big data technology to analyze the factors influencing the company's market sales, and formulate response strategies in advance, it can be better at resisting risks and transformation opportunities will ultimately increase the company's profit in the market and stabilize its leading position in business activities.

When applying big data technology in intelligent IoT to enterprise management, a large number of real data of enterprises can be extracted and processed and analyzed with big data technology, so as to provide reliable reference basis. Compared with the traditional data management system, the system architecture of the Internet of Things includes the following parts:

(1) The LAAS layer, which is the important data storage layer of the Internet of Things system, can select cloud for data storage to facilitate data query and utilization

(2) PaaS layer, mainly to provide the development languages and tools required by customers, such as Python, Hive, and Hadoop

(3) SaaS layer, mainly to provide the applications needed by customers, to facilitate the use of devices for client interface access, such as intelligent large screen, PC terminal, and common client interface. The data information of each business and each sales market can be monitored. Due to the different requirements for data information in data management of various enterprises, the application of big data analysis in intelligent IoT should be combined with its own situation, and the systematic analysis of the existing data resources of the enterprise should be carried out to help the enterprise find its own problems and find the best solution

Many foreign scholars have conducted research on the influencing factors of corporate market sales and forecast statistics and achieved good results. For example, Chemmanur et al. designed an attribute cleaning method for different categories of "dirty data" and proposed a method based on the tree-structured Bloom Filter algorithm cleans duplicate data. Massive data has been cleaned by multiple iterations to ensure the data quality of the following sales forecast analysis [2]. Singh and Mohanty proposed a new type of online data storage and processing model-online analytical processing. In terms of data analysis and processing functions, OLAP has greatly improved compared with OLTP and can meet various application needs of users [3]. Yadegaridehkordi et al. proposed a combined forecasting model, that is, analyzing the characteristics of a single forecasting model, and then combining different models according to a certain weight ratio, so as to give full play to the advantages of different models and improve the prediction accuracy of the model [4].

In the research of related scholars in our country, Wang and Han designed an extended radial basis function as the kernel function for the multidimensional and nonlinear characteristics of the product sales sequence and used an improved immune optimization algorithm to adjust the parameters carry out optimization, establish a set of support vector machine forecasting system, apply the system to car sales forecasting examples, and verify the feasibility of the system by comparing with BP neural network and general radial basis function forecasting accuracy [5]. Gu et al. use a background value optimization formula with adjustment factors to optimize the gray GM model and applies the established optimization model to the number of tourists and tourism income in Hangzhou's tourism industry to predict [6]. Xin et al. provide guidance and suggestions for the tourism industry to formulate sales strategies [7]. Xing et al. use principal component analysis to reduce the dimensions of 8 related clothing promotion factors, use particle swarm optimization algorithm to optimize the neural network, establish a clothing product sales forecast model, reduce the training time of the network, and improve the accuracy of the network prediction degree [8].

Facing a large amount of enterprise data acquisition, storage, and analysis work, we use big data technology in intelligent IoT to ease the workload of data analysts. This paper uses the big data analysis method in intelligent IoT to analyze the historical sales data of enterprises, and its time series presents the characteristics of dual trend changes. According to its characteristics, a combination forecasting model based on the trend method and seasonal index method is proposed, and MAE, RMSE, and MAPE forecast evaluation standards are used to forecast the combination the model is compared with the forecast effect of a single-trend forecast model and seasonal index model, and then, the optimal forecast model is selected.

2. Influencing Factors and Forecasting Statistical Model of Enterprise Market Sales Based on Big Data Analysis in Intelligent IoT

2.1. Intelligent IoT Architecture. It mainly includes three levels: intelligent devices and solutions, the OS layer, and infrastructure in intelligent IoT architecture, which are finally delivered through integration services, as shown in Figure 1. Intelligent equipment can achieve the data collection of view, audio, and pressure and perform the action of capturing, sorting, and handing. Usually IoT devices and solutions are provided to customers. This layer involves diversification of device forms. The OS layer is equivalent to the "brain" of intelligent IoT, which can mainly connect and control the device layer, provide intelligent analysis and data-handling capacity, and solidify the core applications for scenarios into function modules. This layer has high requirements on business logic, unified modeling, full-link technical capacity, and high-concurrency supporting capacity. The infrastructure layer provides the IT infrastructure of servers, storage, AI training, and deployment capabilities, etc. In the era of intelligent IoT, mass data generated in the production and life of people will be collected by sensors in intelligent IoT. In the era of big data, the individual behavior of consumers not only can be collected, quantified, and predicted but also consumers' personal opinions may change the operation of the business society.

2.2. Use MapReduce to Simplify Massive Data. MapReduce adopts the master-slave mode, which is to set up a master node and multiple slave nodes to jointly complete the entire process of distributed computing [9, 10]. In the calculation process, data processing needs to be carried out through the cooperation of two stages of map (mapping function) and reduce (reduction function). Generally, the output of one stage is the input of the next stage, and the two require multiple coordination and cooperation.

2.2.1. JobTracker. JobTracker is mainly used to receive the application processing program, resource monitoring, and job scheduling submitted by the client. This part of the program mainly depends on the programmer to make some

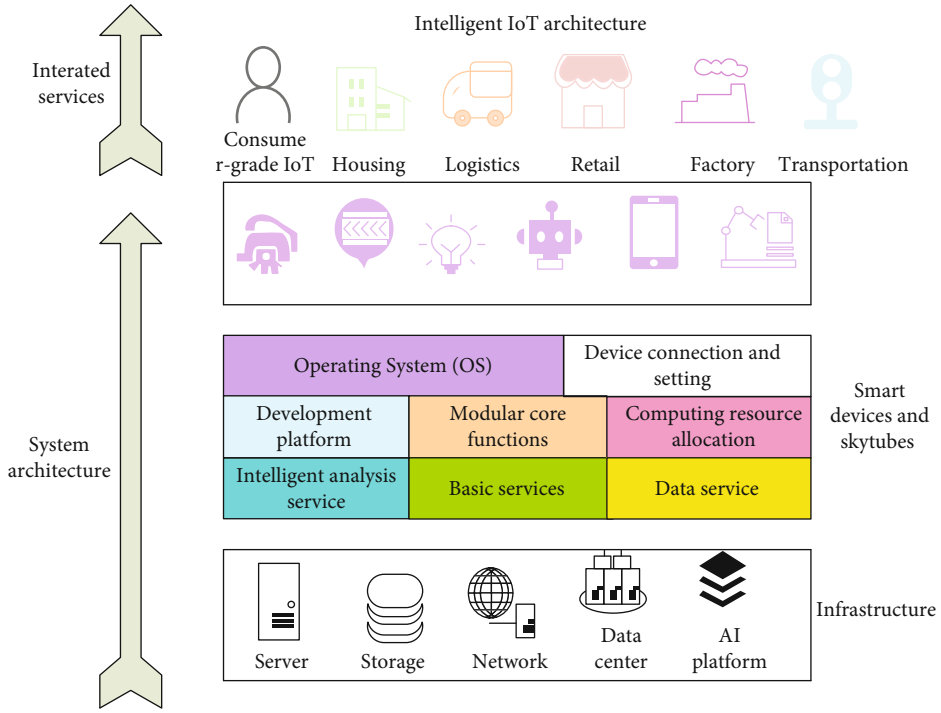

Figure 1: Intelligent IoT architecture.

complex algorithms. It monitors the running status of TaskTracker and jobs in the system by accepting the heartbeat information sent by TaskTracker. The good fault tolerance mechanism of JobTracker enables TaskTracker or jobs to run abnormally, and tasks running on TaskTracker can be aborted due to abnormalities the task is backed up and executed on other TaskTracker to ensure the stability and reliability of the system [11, 12].

2.2.2. Client. The client is mainly a layer where the user interacts with the MapReduce process. This layer mainly plays the role of input, passing the program that the user needs to perform operations to the JobTracker. During the operation, the operation process can be monitored through the console.

2.2.3. TaskTracker. The task scheduler is mainly used to coordinate specific calculations and responds to the JobTracker with information such as the time spent in the calculation process, the number of processing tasks, the occupied CPU, and memory, and at the same time, it processes the assigned tasks [13, 14]. Map and reduce are coordinated to complete. MapReduce uses split as the smallest processing unit of data and is used to store the corresponding data block to be processed. Each split will be processed by the corresponding Map Task.

2.3. Introduction to Machine Learning Classification Algorithms. Machine learning is essentially an approximation to the real model of the problem. Among them, the supervised classification algorithm has been widely used in many business scenarios. There are many ways to solve classification problems. The basic classification methods mainly include decision trees [15], naive Bayes [16], support vector machines [17], K-nearest neighbors [18], and artificial neural networks [19].

The decision tree algorithm is a method of approximating the value of a discrete function. It is a typical classification method. It first processes the data, uses induction algorithms to generate readable rules and decision trees, and then uses decisions to analyze new data. Naive Bayes classification is a classification method based on Bayes' theorem and the assumption of the independence of characteristic conditions. It originated from classical mathematical theory and has a stable mathematical foundation and classification efficiency. A support vector machine is a supervised learning method, which can be widely used in statistical classification and regression analysis. K-nearest neighbor algorithm, referred to as KNN (k-nearest neighbor), is also a relatively simple classification and prediction algorithm. For selecting the K training data that are most similar to the data to be classified and predicted, the results of the K data or the classification labels are averaged and the mode is taken to obtain the results or classification labels of the data to be classified and predicted. The artificial neural network, abbreviated as neural network or quasineural network, is a mathematical model or calculation model that imitates the structure and function of biological neural network and is used to estimate or approximate functions. The neural network is calculated by connecting a large number of artificial neurons. In most cases, the artificial neural network can change the internal structure on the basis of external information and is an adaptive system.

2.4. KNN Classification Algorithm Improved KNN Data Mining Algorithm and Time Series Prediction Algorithm Research

2.4.1. KNN Algorithm Classification. Select the K samples with the smallest distance from the sample to be classified as the K-nearest neighbors of X and finally judge the category of X based on the K-nearest neighbors of X.

(1) Algorithm Flow. Calculate the distance between the sample to be classified and each training sample: the distance function in the KNN algorithm generally has Euclidean distance:

$$d_{\text{euc}}(x,y) = \left[\sum_{j=1}^{d}\left(x_j - y_j\right)^2\right]^{1/2} = \left[(x-y)(x-y]^T\right]\frac{1}{2},$$
$$d_{\text{mah}}(x,y) = \sqrt{(x-y)\sum(x-y)^T}. \tag{1}$$

Manhattan distance:

$$d_{\text{mah}}(x-y) = \sum_{j=1}^{d}\left|x_j - y_j\right|. \tag{2}$$

Chebyshev distance:

$$d_{\text{che}}(x,y) = \max_j\left(\left|x_j - y_j\right|\right). \tag{3}$$

As well as Min's distance, average distance, and geodetic distance, among these distances, the Euclidean distance is often used because of its simplicity.

(2) Selection of Prediction Algorithm. Because the sales of cigarettes need to be predicted, and the sales data of cigarettes is not a continuous time, but a time point, it is more appropriate to choose a time series model when choosing a forecasting algorithm, and the time series can establish a relationship that includes dynamic dependencies. Based on the data model, the trend of future data can be observed from historical behavior information [20, 21].

2.4.2. Improve Data Mining Algorithm. This paper introduces the training process of clustering after partitioning into the KNN algorithm. That is, the big data is divided into equidistant blocks, and then, the data is clustered on each block of data. In this way, for big data, dividing the data into many blocks can effectively reduce the requirements for computer memory, so that the KNN classification algorithm can be used in a big data environment.

First, the big data is divided into blocks. According to different requirements, it can be divided into blocks according to different blocking methods. Given n data samples $\{x_1, x_2, \cdots, x_n\}$, find K clusters Class center $\{a_1, a_2, \cdots, a_n\}$, so that the sum of squared distances between each data sample and its nearest cluster center is the smallest. This sum of squared distances is called the objective function W_N, and its mathematical expression is formula (4), the data is recorded as data matrix formula (5), the difference between data and data uses dissimilarity matrix formula (6), according to the characteristics of the data in this article, the objective function W_N is transformed into formula (7):

Table 1: Demographic factors and the trend of cigarette sales.

	Permanent residents	Floating population	Total employees	Product sales	Sales amount
2016	573	66	383	200	86
2017	615	69	416	248	235
2018	689	104	447	307	462
2019	805	497	601	408	863

$$W_n = \sum_{i=1}^{n}\min_{1\le j\le k}\left|x_i - a_j\right|^2, \tag{4}$$

$$\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}, \tag{5}$$

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \cdots & \cdots & \cdots & 0 & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}, \tag{6}$$

$$W_N\sum_{i=1}^{n}\min_{1\le j\le k}\left|\sqrt{\left(x_{ik} - a_{jk}\right)^2}\right|^2. \tag{7}$$

In formula (5), each column represents a data attribute, and each row represents a piece of data. In formula (6), $d(m, n)$ represents the degree of difference between the mth data and the nth data. When the difference between the two data is smaller, the value of $d(m, n)$ will also be smaller [22, 23].

2.4.3. Time Series Stationarity Test. According to the calculation formula of unit root in MyEclipse, if the sequence is nonstationary, there is $\beta = 1$, and if the sequence is stationary, then $\beta < 1$. Now, assuming that the sequence is nonstationary, let $\beta = 11$ into formula (8), (9) and calculate the value of DF:

$$s\left(\widehat{\beta}\right) = \frac{\sqrt{(1/T-1)\sum_{t-2}^{T}\widehat{u}_t^2}}{\sqrt{\sum_{t-2}^{T}y_{t-1}^2}}, \tag{8}$$

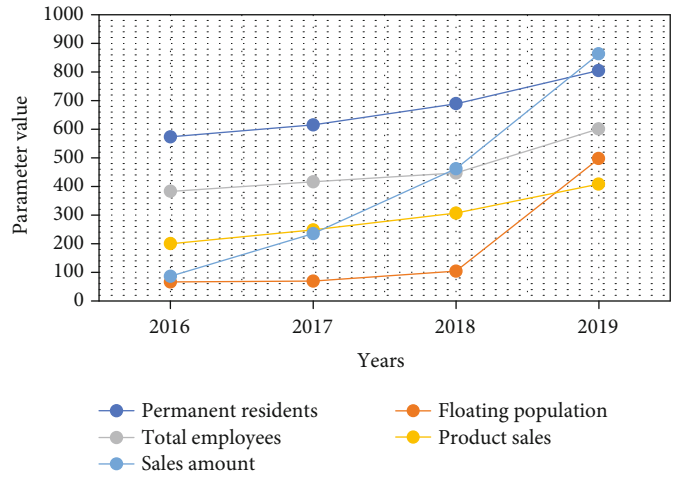

Figure 2: Demographic factors and the trend of cigarette sales.

$$\mathrm{DF}(X) = \frac{\widehat{\beta} - 1}{s\left(\widehat{\beta}\right)}. \tag{9}$$

The running result is the unit root test value of the original data. According to the value, it can be judged whether the original data is stable. When the value does not exist or is particularly small, the original data can be judged to be a stationary sequence; otherwise, it is not stable [24, 25].

2.4.4. Time Series Smoothing Processing. Because the original sequence is a nonstationary sequence, the difference calculation is performed on the original sequence for stationary processing. The calculation formula is

$$\begin{aligned} \Delta x_t &= x_t - x_{t-1}, \\ \Delta^2 x_t &= \Delta x_t - \Delta x_{t-1}, \\ \Delta^d x_t &= \Delta^{d-1} x_t - \Delta^{d-1} x_{t-1}, \end{aligned} \tag{10}$$

Among them, t is the time point. If there are periodic fluctuations in the time series, then the data should also be subjected to a seasonal difference operation. The seasonal difference processing operation can clear the periodicity of the time series. The calculation formula is

$$\Delta_s X_t = X_t - X_{t-s}. \tag{11}$$

2.5. Second Moving Average Method. Establishing a forecasting model for the second moving average method is the key to forecasting using this method. The forecasting model of the second moving average forecasting method is shown in formula (12):

$$Y_{t+1} = a_t + b_t \times T, \tag{12}$$

$$a_t = 2M_t^{(1)} - M_t^{(2)}, \tag{13}$$

$$b_t \frac{2}{n-1}\left(M_t^{(1)} - M_t^{(2)}\right). \tag{14}$$

The T in the above formula represents the expectation that starts at time t and moves backward. $M_t^{(1)}$ is the last moving average in the first moving average sequence obtained by calculation. $M_t^{(2)}$ represents the last moving average in the second moving average sequence [26, 27]. Correspondingly, the formulas for calculating the primary and secondary moving average are as follows:

$$M_t^{(1)} = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-n+1}}{n}, \tag{15}$$

$$M_t^{(2)} = \frac{M_t^{(1)} + M_{t-1}^{(1)} + \cdots + M_{t-n+1}^{(1)}}{n}. \tag{16}$$

Formulas (15) and (16) represent the time series observation values to be predicted, and $M_t^{(1)}$ and $M_t^{(2)}$ represent the primary and secondary moving average values of period t, respectively and n is the spanning dimension of this calculation. The basic prediction formula of the exponential smoothing prediction model is

$$S_t = 2Y_t + (1-a)S_{t-1}. \tag{17}$$

In formula (17), when S_{t-1} represents time t, the actual value Y_{t-1} at that time corresponds to the smooth value, and S_{t-1} represents the smooth value corresponding to the actual value Y_{t-1} at time $t-1$. The parameter a in the formula is a weight value, which is also called a smoothing constant under normal circumstances, and the value range is [0,1].

2.6. Design of Enterprise Marketing System Based on Hadoop

2.6.1. System Architecture. The design of the enterprise marketing system is to build a Hadoop-based data processing platform as a data management center and provide massive data storage and processing support to implement a Hadoop-based enterprise marketing system [28, 29].

(1) Data Source Layer. The main job of the data source layer is to collect data. The data source of the enterprise marketing

Table 2: Personal spending power and the trend of cigarette sales.

	Per capita consumption amount	GDP per capita	Per capita disposable expenditure	Per capita consumption expenditure	Average salary
2016	737	7215	3665	2654	7976
2017	1654	7567	4317	3475	8242
2018	1964	7988	4784	3864	8527
2019	2268	8365	5521	4269	8731

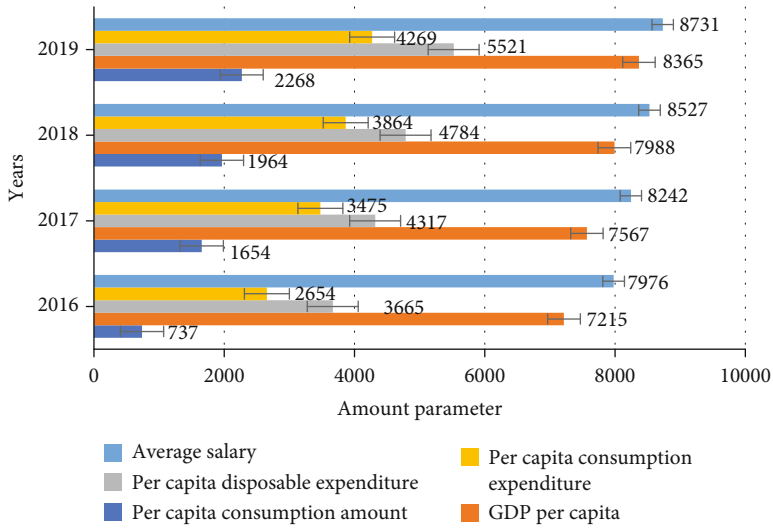

Figure 3: Personal spending power and the trend of cigarette sales.

system in this article includes internal data and external data. The data collection method is mainly through the National Bureau No. 1 Project. Downstream data includes the production and sales information of more than a dozen industrial companies across the country, as well as the company's purchase, sales, and inventory data in various markets across the country. Salespersons report market data and import external systems through the Web Services interface.

(2) Data Transmission Layer. There are two main ways of data transmission; one is through ETL middleware, and the other is a data transmission interface through enterprise applications. Generally, small-scale data provides a data transmission interface, and the data can be transmitted directly through the interface, while for large-scale data extraction, data is directly extracted by connecting to the database through ETL middleware.

(3) Data Processing Layer. Since the source data transmitted from the data source layer is not only the finest granularity but also the amount of data is very large, and there are many "dirty data," the transmitted data needs to be preprocessed before storage including data cleaning and processing. Due to the huge amount of data, data processing is performed by the Hadoop platform to achieve dimensionality reduction and aggregation of massive data and simplify the data on the basis of satisfying model analysis and maintaining data integrity and accuracy.

Table 3: Comparison of sales forecast results and actual values based on trend model.

Month	Forecast sales	Guess sales	Actual sales
201801	528739	562284	572639
201802	217479	318374	302746
201803	337468	337595	347290
201804	352648	347284	342699
201805	313947	359837	368270
201806	301748	364829	352890

(4) Data Storage Center. The data storage center of the system in this paper is coordinated by the Hadoop distributed storage platform and the relational database. The Hadoop platform is built to store massive data, process massive data, and transmit data to the system [30]. After the data is exported from the Hadoop platform to the relational database, the relational database performs real-time analysis or mining on it to ensure data consistency during processing. Using Hadoop and relational databases to work together to process computing tasks, separate the processing of massive data from the processing of real-time data, so that large-scale data operations will not affect the operating efficiency of the marketing system, and also make the entire system easier for expansion, more stable.

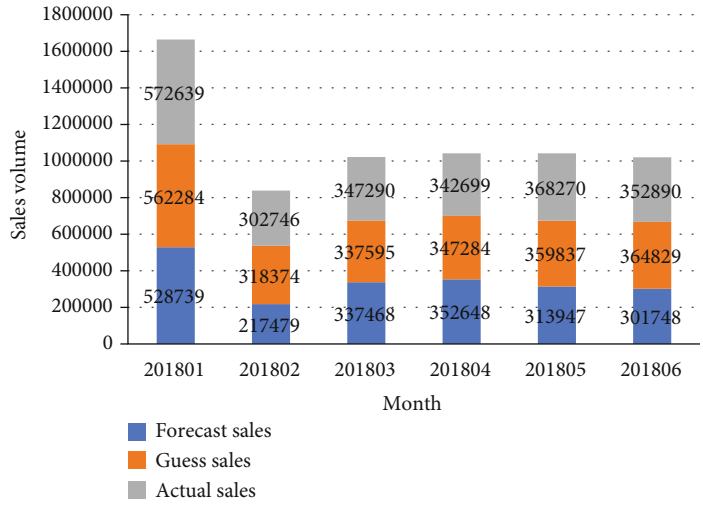

Figure 4: Comparison of sales forecast results and actual values based on trend model.

(5) Data Analysis Layer. The data analysis logic is mainly realized through the stored procedures of the relational database and the MapReduce calculation model. Since the data sets of small data scale are designed to be stored in the relational database, the analysis of this type of data can be realized through the stored procedures of the relational database. Large-scale data sets are stored in HDFS, and the data model needs to be implemented by writing a MapReduce calculation model.

2.6.2. System Technical Architecture. This paper uses the J2EE enterprise-level application development framework and adopts the stable and convenient B/S operation mode and component development technology to design and develop the system architecture [31]. Data transmission uses XML format as the data transmission standard for each interface and realizes the application integration and data integration of each system through message middleware, and realizes the integration of data collection, data storage, and data preprocessing. Building a Hadoop-based data processing platform as the system's data management center [32], which is low-coupled with other running hardware devices, processes, and stores massive amounts of data at high speed, and can allocate computing and storage resources to other running systems, and use the Hadoop platform working with relational data, it separates the processing of massive data from business logic and analysis operations, reduces the coupling degree of analysis and calculation at the hardware level, and greatly improves the analysis and calculation performance and stability of the system.

3. Influencing Factors and Forecasting Statistical Experiments of Enterprise Market Sales Based on Big Data Analysis in Intelligent IoT

3.1. Experimental Subjects and Data Collection. This article selects the monthly sales volume of a certain brand of cigarettes from China Tobacco in this province from 2016 to 2019 as the analysis data set. The time series of the brand's sales volume has a strong upward trend and also has periodic phenomena, so the time series has the characteristics of dual trend changes, namely, seasonal volatility and overall trend variability, the desired effect will not be achieved. In this paper, the single-term model of the trend method and the seasonal index method will be fitted to forecast, respectively, and then, the linear combination forecasting model of the two will be established to forecast the sales, and finally the forecast results will be compared.

Table 4: Comparison of sales forecast results and actual values based on seasonal index method.

Month	Forecast sales	Guess sales	Actual sales
201801	538478	562284	572639
201802	286491	318374	302746
201803	361937	337595	347290
201804	358505	347284	342699
201805	347190	359837	368270
201806	365281	364829	352890

3.2. Forecast Method

(1) Seasonal index method the calculation method and steps are as follows:

(a) Calculate the average of the same quarter over the years. Suppose the average of the same quarter over the years is r_i, $i = 1, 2, 3, 4$. A total of 12 quarterly time series in three years are represented as $y_1, y_2, \cdots, y_{12}$, then you can get:

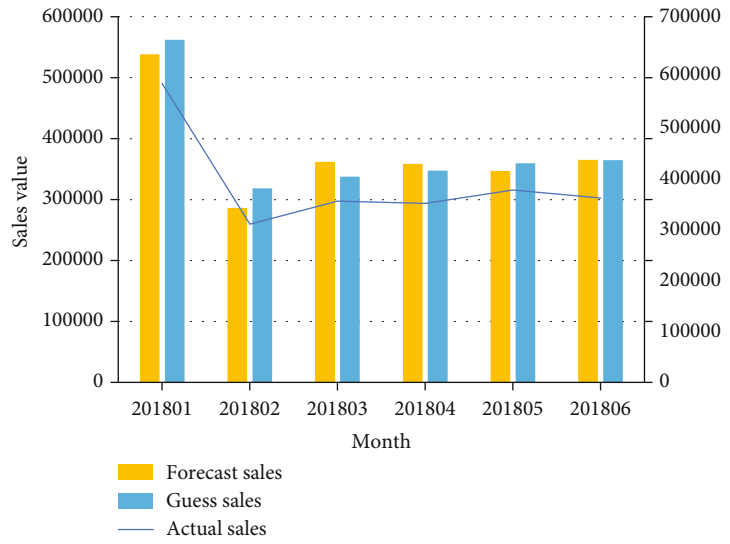

Figure 5: Comparison of sales forecast results and actual values based on seasonal index method.

$$r_1 = \frac{1}{n}(y_1 + y_5 + \cdots + y_{4n-3}),$$
$$\cdots \quad (18)$$
$$r_4 = \frac{1}{n}(y_4 + y_8 + \cdots + y_{4n})$$

(b) Calculate the average of each season. Let $\bar{y}_t$ denote the quarterly average of year t, $t = 1, 2, \cdots, n$, there are

$$\bar{y}_1 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4),$$
$$\cdots \quad (19)$$
$$\bar{y}_n = \frac{1}{4}(y_{4n-3} + y_{4n-2} + y_{4n-1} + y_{4n})$$

(c) Adjust the seasonal index of each season. Theoretically, the sum of seasonal indices should be 4, but due to calculation errors in practice, the sum of seasonal indices is greater than or less than 4, so it needs to be readjusted. The adjustment formula is

$$a_i = \frac{n}{\left(\sum_{j=1}^{n} \bar{a}_j\right)} \bar{a}_j \quad (20)$$

When using the seasonal index method to forecast time series, it should be noted that the time series should not have an obvious linear trend; otherwise, the forecast accuracy will be greatly reduced.

(2) Long-term trend

This article uses the least square method to find the parameters in the linear trend formula. The core idea of the least square method is to use a straight line to approximate the historical data in the past. The mathematical language

Table 5: Comparison of sales forecast results and actual values of the combined forecasting model.

Month	Trending method to predict sales	Combination model predicts sales	Actual sales
201801	452738	547282	572639
201802	246371	337625	302746
201803	382736	336481	347290
201804	248474	321746	342699
201805	407634	346289	368270
201806	336585	312648	352890

is the actual observation value of the time series of the object model y_i and the predicted value $\hat{y}_i$ in the linear trend model have the smallest sum of squared deviations, that is, the value of $\sum(y_i y\wedge_i)^2$ is the smallest. The least square method is used to determine the value of the parameter. The specific derivation process is omitted here. The calculation formulas for parameters a and b are

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i, \bar{x})^2},$$
$$\hat{a} = \bar{y} - \hat{b}\overline{x.} \quad (21)$$

(3) Decomposition prediction model

After determining the seasonal index and long-term trend with the decomposition method, the two key factors, the new forecast value of the cigarette sales model can be calculated according to formula (22).

$$X_t = T_t \times S_t \times C_t \times I_t, \quad (22)$$

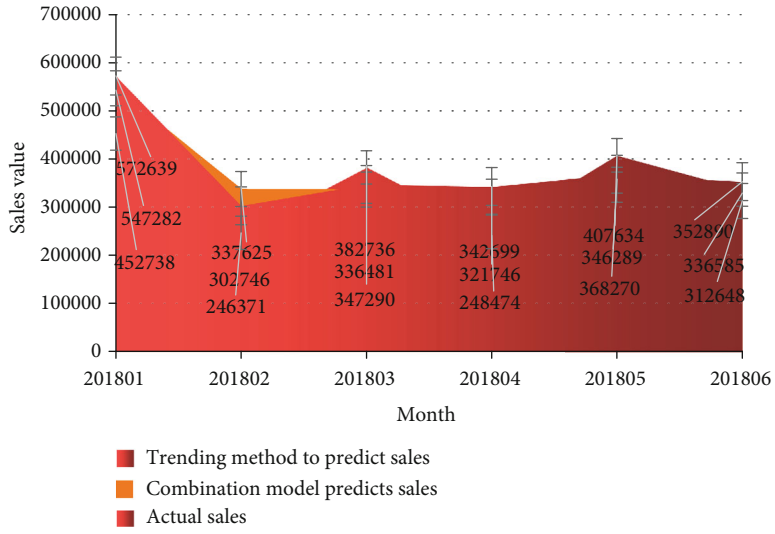

Figure 6: Comparison of sales forecast results and actual values of the combined forecasting model.

As the factors in the model are determined, the random fluctuation I and the cyclic index C have been reduced. After simplification, the predicted value is obtained.

4. Based on Big Data Influencing Factors of Enterprise Market Sales and Forecast Statistical Experimental Analysis

4.1. Macroscopic Factors Affecting Cigarettes. As shown in Table 1, among the demographic factors in this area are the resident population, floating population, and other influencing factors, cigarette sales, cigarette sales, and other data values and data trends.

As shown in Figure 2, there are many factors affecting cigarette sales. All the factors affecting the product are analyzed by big data technology. After a large number of calculations and modeling, the product sales amount is mainly reflected in the number of permanent residents and economic the efficiency is getting higher and higher over time.

4.2. Personal Consumption Ability and Product Sales Trend. As shown in Table 2, the data association trends between factors are per capita consumption expenditure, average wages of employees, and cigarette sales in the region.

As shown in Figure 3, according to big data, the main reason for affecting product sales is the city's per capita living expenses, followed by per capita GDP in 2016-2019. Due to the limitations of data statistics, there is a lack of regularity. In this case, traditional mathematical induction statistical methods cannot be used.

4.3. Sales Forecast Analysis. We use this model to obtain the sales forecast value of the key brand from January to June 2018 and compare it with the actual sales value and compare it with the sales value of the same period in 2017. The unit is box. The results are shown in Table 3.

As can be seen in Figure 4, the sales forecast model based on the trend method has a good forecast of the seasonality and periodicity of the monthly sales of cigarettes, but the relative error of the forecasted sales in a certain two months is still more than 10%. The prediction effect of extreme values in the time series is not ideal, and we can continue to improve on the basis of this prediction.

Table 6: Comparison of prediction effects of three prediction models.

Predictive model	MAE	RMSE	MAPE (%)
Trend model	4628.3	5835.2	13.28
BP neural network model	18636	20351	7.37
Combined forecasting model	9635	113634	3.12

4.4. Forecast by Seasonal Index Method. After repeated training, the number of hidden layer nodes is determined to be 5, and the trained seasonal index method model is used to predict the test data set. The results obtained are shown in Table 4:

It can be seen intuitively from Figure 5 that the relative error of the sales forecast model based on the seasonal index method is relatively stable and the error is small. Therefore, this paper adopts a seasonal index method model based on the forecast based on the trend method model to modify the forecast value of the trend method model to improve the accuracy of the forecast.

4.5. Establish a Combined Forecasting Model Based on Trend Method and Seasonal Index Method. First, the linear structure part of the time series is fitted with a trend method model to obtain the predicted value $\widehat{L}_t$. The first order and seasonal differences have been performed on the time series before modeling. The purpose is to eliminate the trend of the time series and reduce the time series. The seasonal index method is used to identify the nonlinear part e_t of the time series, and the prediction result $\widehat{N}_t$ is obtained. After repeated trials and comparisons, this paper constantly adjusts the

Table 7: Algorithm performance comparison.

Sample	Algorithm		Random block KNN		Classic KNN	
	Correct rate	Time	Correct rate	Time	Correct rate	Time
1	0.9315	58.23	0.8642	128.39	0.9362	382.28
2	0.9163	413.56	0.9017	983.01	0.9276	3728.28
3	0.9063	502.42	0.8437	1273.52	0.9117	5583.63

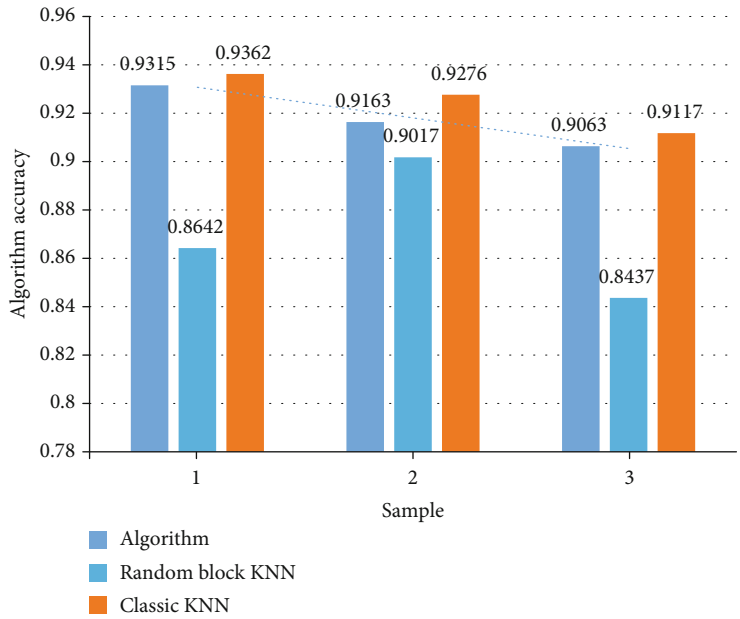

Figure 7: Algorithm performance comparison.

number of nodes, and finally determines the input node is 12, the hidden layer node is 5, and the output node is 1 structure. The predicted results are shown in Table 5.

It can be seen intuitively from Figure 6 that the relative error of the forecast results of the combined forecasting model is less than 5%, which is also lower than the relative error of the forecasting results of the trend forecasting model and the seasonal index forecasting model, that is, the forecasting effect is better than that of a single model. Through the comprehensive application of the two models of trend method and seasonal index method, they can give full play to their respective strengths to achieve the purpose of improving the forecasting effect.

This article uses historical monthly sales data of a certain brand of cigarettes from 2016 to 2019 to establish a trend-based sales forecast model, a seasonal index-based forecast model, and a combined forecast model to predict the monthly sales in 2018. Table 6 shows the comparison of the prediction effects of the three prediction models by the MAE, RMSE, and MAPE prediction evaluation standards.

Table 6 shows that from the three indicators of MAE, RMSE, and MAPE, it can be seen that the evaluation indicators of the combined model are the lowest among the three models, so the prediction effect is the best. In the original data, there are both linear factors and nonlinear factors, so a single forecasting model, whether it is a trend forecasting model or a seasonal index forecasting model, cannot achieve the ideal forecasting effect. The combined model of the trend method and the seasonal index method can synthesize the advantages of a single model, better dig out the complex linear and nonlinear features behind the data, and also improve the prediction accuracy of the model.

Table 8: Comparative analysis of cigarette sales volume forecast and actual results.

Time	Trend forecast	Seasonal index forecast	Decomposition prediction	Actual value
2018.6	15.8	15.37	15.31	16.24
2018.7	15.9	16.35	16.47	17.24
2018.8	17.1	15.83	16.74	16.68
2018.9	17.0	15.35	16.43	17.19
2018.10	16.8	15.43	17.53	16.83
2018.11	16.6	14.79	17.43	17.52

4.6. Comparative Analysis of Algorithm Performance. Based on the above classification results, these sample data are processed into feature vector values. The KNN classifier is used in MATLAB to classify the three data samples, and the

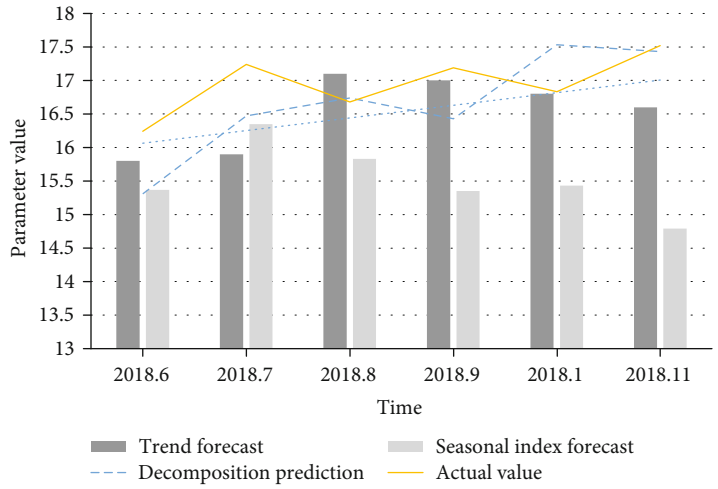

Figure 8: Comparative analysis of cigarette sales forecast and actual results.

number of correct classification results of the data classified by the three classification algorithms is counted. Calculate the accuracy rate obtained in each classification algorithm in each sample and use the highest accuracy rate as the accuracy rate of the classification algorithm in the sample classification result of the classification algorithm. Another performance test standard is the classification time on each sample. The results are shown in Table 7.

As shown in Table 7 and Figure 7, it can be concluded that the classification accuracy of the algorithm proposed in this paper is 0.41% ~6.2% higher than the random block algorithm, and 0.7% ~1.8% lower than the traditional KNN algorithm. It is 55.15% ~63.17% faster than the random block algorithm, and 83.28% ~90.55% faster than the traditional KNN algorithm. As the data increases, the speed will increase more obviously.

4.7. Comparative Analysis of Forecast Results. According to the data in the above chart, the value obtained by the long-term trend method predicted by the formula, and the value obtained by the decomposition method, the forecast data for the next few periods are, respectively, predicted. Finally, we compare these data with the actual sales results of the cigarette market. For comparison, the data details are shown in Table 8.

From Figure 8, we can see that the error analysis between the forecasted value of the three forecasting models of trend method, seasonal index method, and decomposition method and actual sales volume, we can see that the time series decomposition method established by the multiplication model performs best. The average error rate is about 2% and the fluctuation is small, followed by the seasonal index method, and the trend method has the worst performance of about 4% and the fluctuation is large. For the tobacco industry, the decomposition prediction model can fully meet its forecasting needs, so as to guide industrial companies to produce cigarettes and commercial companies to sell cigarettes based on the predicted values.

5. Conclusion

The KNN classification algorithm based on overlapped *k*-means clustering proposed in this paper is still lower than the traditional KNN algorithm in classification accuracy. This situation comes from the algorithm's clustering process, which affects the accuracy of the classification algorithm. If the effect of clustering can be improved in the future, then the accuracy of classification will catch up with the traditional classic KNN algorithm, especially if other excellent clustering algorithms can be introduced or the conditions and methods of partitioning can be changed according to the data. The effect and efficiency of the algorithm can be further improved.

Using the big data method in intelligent IoT to analyze the historical sales data of the company, its time series presents the characteristics of dual trend changes. According to its characteristics, a combined forecasting model based on the trend method and seasonal index method is proposed, and MAE, RMSE, and MAPE forecasting evaluation standards are used to compare the combined forecasting model. A comparative analysis with the forecasting effect of a single-trend forecasting model and a seasonal index model proves that the combined forecasting model is better than a single model.

This paper studies the construction of a Hadoop massive data processing platform, an in-depth study of its key technologies HDFS and MapReduce, analyzes its working mechanism, connects with the actual situation of the enterprise, analyzes the feasibility of technology implementation and environment construction, and uses the Hadoop platform to sell mass sales first data information undergoes preprocessing such as data cleaning, dimensionality reduction, and structural standardization and then provides these processed data to a relational database to perform data analysis and processing of related businesses. This provides support for sales forecasting models based on massive data processing.

Data Availability

This article selects the monthly sales volume of a certain brand of cigarettes from China Tobacco in this province from 2016 to 2019 as the analysis data set.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is not supported by a special funding but by University of Queensland.

References

[1] J. Chen, Z. Lv, and H. Song, "Design of personnel big data management system based on blockchain," *Future Generation Computer Systems*, vol. 101, pp. 1122–1129, 2019.

[2] T. J. Chemmanur, G. Hu, and J. Huang, "Institutional investors and the information production theory of stock splits," *Journal of Financial and Quantitative Analysis*, vol. 50, no. 3, pp. 413–445, 2015.

[3] N. Singh and S. R. Mohanty, "A review of price forecasting problem and techniques in deregulated electricity markets," *Journal of Power and Energy Engineering*, vol. 3, no. 9, pp. 1–19, 2015.

[4] E. Yadegaridehkordi, M. Hourmand, M. Nilashi, L. Shuib, A. Ahani, and O. Ibrahim, "Influence of big data adoption on manufacturing companies' performance: an integrated DEMATEL-ANFIS approach," *Technological Forecasting and Social Change*, vol. 137, pp. 199–210, 2018.

[5] Y. Wang and X. Han, "Research on power system load forecasting based on classification of influence factors," *Journal of Computational and Theoretical Nanoscience*, vol. 13, no. 12, pp. 9798–9803, 2016.

[6] D. Gu, S. Khan, I. U. Khan, and S. U. Khan, "Understanding mobile tourism shopping in Pakistan: an integrating framework of innovation diffusion theory and technology acceptance model," *Mobile Information Systems*, vol. 2019, Article ID 1490617, 18 pages, 2019.

[7] T. Xin, C. Kai, and L. I. Gang, "Influencing factors analysis and trend forecasting of China's carbon emissions——empirical study based on STIRPAT and GM (1, 1) models," *Journal of Northeastern University*, vol. 36, no. 2, pp. 297–300, 2015.

[8] Z. Xing, H. Guo, and Q. Fu, "Analysis of influencing factors of rainfall in irrigation area and combining rainfall forecasting," *Transactions of the Chinese Society for Agricultural Machinery*, vol. 46, no. 8, 2015.

[9] J. Šindelář, "Investigation of factors influencing employee performance," *International Journal of Organizational Analysis*, vol. 24, no. 2, pp. 340–368, 2016.

[10] K. Yang, "The construction of sports culture industry growth forecast model based on big data," *Personal and Ubiquitous Computing*, vol. 24, no. 1, pp. 5–17, 2020.

[11] R. Yang, L. Yu, Y. Zhao et al., "Big data analytics for financial Market volatility forecast based on support vector machine," *International Journal of Information Management*, vol. 50, no. Feb., pp. 452–462, 2020.

[12] E. W. K. See-To and E. W. T. Ngai, "Customer reviews for demand distribution and sales nowcasting: a big data approach," *Annals of Operations Research*, vol. 270, no. 1-2, pp. 415–431, 2018.

[13] S. Li and B. Wang, "Hybrid parrallel Bayesian network structure learning from massive data using MapReduce," *Journal of Signal Processing Systems*, vol. 90, no. 8-9, pp. 1115–1121, 2018.

[14] P. G. Kulkarni and S. R. Khonde, "An improved technique of extracting frequent itemsets from massive data using MapReduce," *International Journal of Engineering and Technology*, vol. 9, no. 3S, pp. 400–406, 2017.

[15] S. S. Gavankar and S. D. Sawarkar, "Eager decision tree," in *2017 2nd International Conference for Convergence in Technology (I2CT)*, pp. 837–840, Mumbai, India, 2017.

[16] D. Wang, D. Yuan, and C. Miao, "Sparse naïve Bayes base on entropy correlation for GPR image denoising," in *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*, pp. 167–171, Xi'An, China, 2020.

[17] Z. Liu and L. Bai, "Evaluating the supplier cooperative design ability using a novel support vector machine algorithm," in *2008 12th International Conference on Computer Supported Cooperative Work in Design*, pp. 986–989, Xi'an, China, 2008.

[18] J. Vieira, R. P. Duarte, and H. C. Neto, "kNN-STUFF: kNN STreaming unit for Fpgas," *IEEE Access*, vol. 7, pp. 170864–170877, 2019.

[19] J. H. Andreae, "Brains, neural networks and expert systems," in *Proceedings 1993 The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, pp. 3-4, Dunedin, New Zealand, 1993.

[20] Y. Ma, X. Meng, and S. Wang, "Parallel similarity joins on massive high-dimensional data using MapReduce," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 1, pp. 166–183, 2016.

[21] A. T. Azar and A. E. Hassanien, "Dimensionality reduction of medical big data using neural-fuzzy classifier," *Soft Computing*, vol. 19, no. 4, pp. 1115–1127, 2015.

[22] F. N. Afrati, S. Sharma, J. R. Ullman, and J. D. Ullman, "Computing marginals using MapReduce," *Journal of Computer and System Sciences*, vol. 94, pp. 98–117, 2015.

[23] S. K. Zhang, G. Y. Shi, Z. J. Liu, Z. W. Zhao, and Z. L. Wu, "Data-driven based automatic maritime routing from massive AIS trajectories in the face of disparity," *Ocean Engineering*, vol. 155, pp. 240–250, 2018.

[24] H. Duan, Y. Peng, G. Min, X. Xiang, W. Zhan, and H. Zou, "Distributed in-memory vocabulary tree for real-time retrieval of big data images," *Ad Hoc Networks*, vol. 35, pp. 137–148, 2015.

[25] D. Wang and J. Liu, "Optimizing big data processing performance in the public cloud: opportunities and approaches," *IEEE Network*, vol. 29, no. 5, pp. 31–35, 2015.

[26] C. Sreedhar, N. Kasiviswanath, and P. Chenna, "A survey on big data management and job scheduling," *International Journal of Computer Applications*, vol. 130, no. 13, pp. 41–49, 2015.

[27] L. Ren, L. Zhang, F. Tao, C. Zhao, X. Chai, and X. Zhao, "Cloud manufacturing: from concept to practice," *Enterprise Information Systems*, vol. 9, no. 2, pp. 186–209, 2015.

[28] Byung, Ho, Jung, Dong, Hoon, and Lim, "RHadoop," *Journal of the Korea Society of Computer and Information*, vol. 22, no. 4, pp. 9–16, 2017.

[29] A. Pandey, "Simplilearn big data hadoop review," *PC Quest*, vol. 32, no. 4, pp. 33–33, 2019.

[30] M. Grossman, M. Breternitz, and V. Sarkar, "HadoopCL2: motivating the design of a distributed, heterogeneous programming system with machine-learning applications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 3, pp. 762–775, 2016.

[31] K. McDermott, "Achieving data liquidity across health care requires a technical architecture," *Bulletin of the Association for Information Science and Technology*, vol. 43, no. 1, pp. 19–22, 2016.

[32] Z. Lv, D. Chen, and A. K. Singh, "Big data processing on volunteer computing," *ACM Transactions on Internet Technology (TOIT)*, 2020.