

## Hadoop-based ARIMA Algorithm and its Application in Weather Forecast

Leixiao Li<sup>1</sup>, Zhiqiang Ma<sup>2</sup>, Limin Liu<sup>3</sup> and Yuhong Fan<sup>4</sup>

<sup>1,2,3,4</sup> College of Information Engineering, Inner Mongolia University of Technology  
Huhhot, China

<sup>1</sup> llxhappy@126.com, <sup>2</sup> mzq\_bim@163.com, <sup>3</sup> liulimin789@126.com,  
<sup>4</sup> 342688785@qq.com

### Abstract

*This paper concentrates on the issue of weather data mining. We propose a ARIMA algorithm based on Hadoop framework, and implement an effective weather data analyzing and forecasting system. We present the procedure to parallelize the ARIMA algorithm in the Hadoop environment, and construct a scalable, easy-to maintain, and effective weather forecasting system. Several experiments are conducted and results show that the proposed system is highly effective in terms of data storage, management, as well as query.*

**Keywords:** Hadoop, ARIMA, Data Mining, Weather Forecast

### 1. Introduction

With the industrialization of the world, global warming and other climate issues become increasingly serious, abnormal weather is also emerging, these issues cause great economic and social losses to mankind, therefore human pay more and more attention to meteorology research. In recent years, experts and researchers conduct continuous research in the meteorological area and thus a large number of meteorological data are accumulated. Human summarize and acquire vast of meteorological forecasting knowledge through studying and summarizing experience from these documents. These forecasting knowledge do a great job in helping forecast and resistance of terrible weather. At the same time the computer technology develops contentiously. All these factors contribute to the informational construction of meteorological cause. The attainable meteorological data and the type of meteorological data are increasing contentiously. Meteorological sounding data are mainly gathered from surface meteorological observation stations and aerial observation stations, now there are more weather stations, the daily meteorological observation data which can be observed, acquired and processed are growing at an exponential speed [1]. Now these meteorological data are mainly stored in a message file or a database, and are mainly used for data analysis and the formation of weather forecasting, disaster forecasting and other information, providing decision support for other departments. Besides, meteorological data can also be saved as data, which can support related research in other areas. However, the storage cost of meteorological massive data is increasing and the effective management of these data becoming more important, which requires the computation and storage of data been properly utilized based on current available inexpensive computer cluster storage and computing environment to reduce cost, and effectively manage the current data, and can continuously accumulate and adapt to the storage of the increasingly more meteorological data [2].

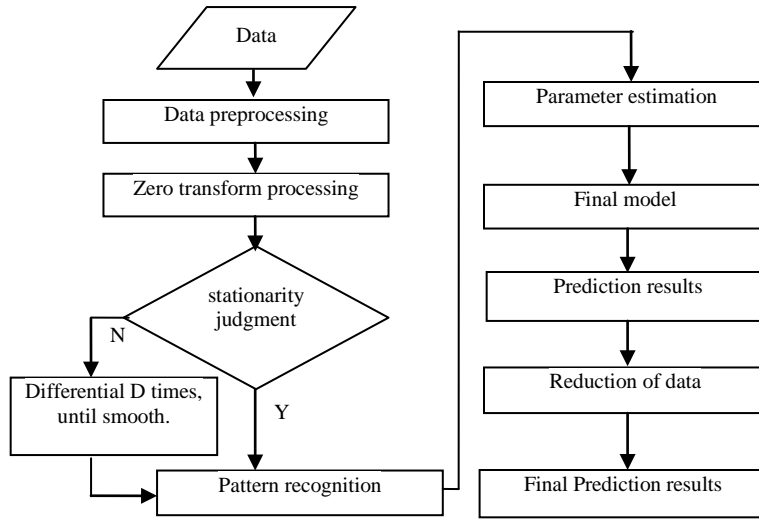
Prediction is one of the two basic goals of Data Mining. Data Mining is to dig out knowledge and rules, which are hidden and unknown, the user may be interested in or have potential value for decision-making, from the large amounts of data, these potential knowledge and rules can reveal the laws between data, we can take good advantage of these laws to predict, these methods are widely used in financial forecasting, commodity recommendation, market planning, *etc.* [3]. There are many kinds of technical methods of data mining, which mainly include: association rule mining algorithm, decision tree classification algorithm, clustering algorithm and time series mining algorithm, *etc.* [4]. Time series data mining is to extract information and knowledge from a lot of time series data, these information and knowledge are not known in advance for people but they are potentially useful and time-related, and for short-term, medium-term or long-term forecasts, guiding people's behavior such as society, economy, military and life[15]. In fact, almost all data in the meteorological field are time series data, and future meteorological data can be predicted better by means of the time series mining algorithm. The time series mining algorithm adopted in the article is ARIMA time series mining algorithm. The full name of the ARIMA (p,d,q) model is difference autoregression moving average model. The basic idea is that a group of orderly time series data formed over time is described by a corresponding mathematical model, and then future data are predicted according to the model and previous values and present values of the time series data [5].

How to fully and effectively store, manage and use these massive meteorological data, effectively discover and understand the law and knowledge in the data to contribute to weather forecasting has attracted more and more Data Mining researchers' attention[6]. Combining with the characteristics of meteorological data, the article constructs the data mining platform based on the Hadoop framework, uses the time series mining algorithms for meteorological forecast and the forecast results are analyzed.

## **2. Algorithm Design for the Forecast of ARIMA Model**

### **2.1. Algorithm Flow Chart**

The article uses ARIMA prediction algorithm for weather forecast, and as almost all data in the field of meteorology are time-series data, the future weather conditions can be better predicted by means of the time [7]. The basic idea of the algorithm is that a group of orderly time series data formed over time is described by a corresponding mathematical model, and then future data are predicted according to the model and previous values and present values of the time series data. The algorithm has nine steps which are data preprocessing, zero transformation processing, stationarity judgment, pattern recognition, parameter estimation, the final model, prediction results, reduction of data and final prediction results, the flow chart is as shown in Figure 1.



**Figure 1. Algorithm Flow Chart of ARIMA Prediction Model**

## 2.2. Data Preprocessing

The article has carried out simple preprocessing on data, part of the data in the database is "32744", "32700", "32766", "300 + XXX" and other forms, and these data in all of the meteorological data represent the elements of space, all trace elements, all elements of the missing data and so on. As these values differ greatly from the real measurement data, which has a significant impact on the subsequent forecast, data preprocessing is necessary. The data preprocessing method adopted in this paper is to use the average value of the previous value and the latter value when faces with those values.

## 2.3. Zero Transform Processing

Zero transform processing enables the average of the data sequence to be zero, namely

$$zero[i] = src[i] - E(src)$$

, wherein

$$Ex_t = \frac{1}{n} \sum_{t=1}^n x_t$$

## 2.4. Stability Judgment

Stationary is the premise condition of ARIMA model, only a stationary time series can take the next step of models selection. Stationarity in general can be divided into strictly stationarity and wide sense stationarity.

For the strict stationarity:

the time series  $X_t$ ,  $t=0, \pm 1, \pm 2, \dots$ , if for any integret  $n$ , any  $t_1, t_2, \dots, t_n \in T$ , and  $t_1 + \varepsilon, t_2 + \varepsilon, \dots, t_n + \varepsilon \in T$ , the n-dimensional distribution functions are equal, namely:

$$F(x_1, \dots, x_n; t_1, \dots, t_n) = F(x_1, \dots, x_n; t_1 + \varepsilon, \dots, t_n + \varepsilon)$$

and then the time series is a strict stationary time series.

Numerical characteristics of the strict stationary time series are as follows:

- The mean value function  $EX_t = \mu$ ,  $\mu$  is a constant.
- The variance function  $D_x(t) = D[X(t)]$  is irrelevant to  $t$  and is a constant.
- The autocorrelation function

$$\rho_x(t_1, t_2) = \rho_x(t_2 - t_1) = \rho_x(k)$$

is irrelevant to the starting point and only relevant to the time interval  $k$ .

- The autocovariance function  $v_x(t_1, t_2)$  is only relevant to the time interval and irrelevant to the starting point.

For the wide sense stationarity: if the time series  $X_t$ ,  $t = 0, \pm 1, \pm 2, \dots$  meets the following requirements:

1.  $EX_t = \mu$  (constant),  $k = 0, \pm 1, \dots$ ;
2.  $EX_t X_{t+k}$  is irrelevant to  $t$ ,  $k = 0, \pm 1, \dots$ ;

and then the time series is a wide sense stationary time series.

Numerical characteristics of the wide sense stationary time series are as follows:

- The autocorrelation function  $\rho_k$ ;
- $E|X(t)|^2 < +\infty$
- the mean square error function

$$\psi_x^2(t) = E|X(t)|^2 < +\infty$$

- the variance function

$$D_x(t) = D[X(t)] = \psi_x^2(t) - (EX(t))^2 < +\infty$$

Generally only wide sense stationarity is required. There are also a variety of Stationarity test methods, such as: observation, run method [16]. The article's stationarity test method is the ADF test method of unit root test method.

## 2.5. Pattern Recognition

Pattern recognition is to determine the stationary sequence is suitable for which models of AR(p), ARMA(p, q), or MA(q) and preliminarily determine the model order, that the preliminary determination of p value, q value, or p and q value[8]. Its judgment is based on the sample mean, the autocorrelation coefficients and the partial autocorrelation coefficients of the stationary sequence.

First calculate the autocorrelation coefficients  $\{\hat{\rho}_k\}$  and partial autocorrelation coefficients  $\{\hat{\phi}_{kk}\}$ . Autocorrelation coefficient is calculated as:

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

Wherein n is the number of data sequence, k is the lag phase,  $\bar{Y}$  is the average of the sequence.

Autocorrelation coefficient represents the degree of correlation between the time series and its after lag time periods, and its range is  $-1 \leq \hat{\rho}_k \leq 1$ , and the closer  $\hat{\rho}_k$  is to 1, the higher the degree of correlation is.

Partial autocorrelation coefficient is conditional correlation between  $Y_t$  and  $Y_{t-k}$  when  $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}$  of the  $Y_t$  sequence are given. The formula is as follows:

$$\hat{\varphi}_{kk} = \begin{cases} \hat{\rho}_1 & k=1 \\ \left( \hat{\rho}_k - \sum_{j=1}^{k-1} \hat{\varphi}_{k-1,j} \times \hat{\rho}_{k-j} \right) / \left( 1 - \sum_{j=1}^{k-1} \hat{\varphi}_{k-1,j} \times \hat{\rho}_j \right) & k=2,3,\dots \end{cases}$$

Second, select the calculation model according to the truncation of the autocorrelation coefficient and partial autocorrelation coefficient of the sequence, and preliminarily determine the model order.

$$BIC(k) = N \ln(\hat{\sigma}^2) + k \ln N$$

1. For each q, calculate  $\hat{\rho}_{q+1}, \hat{\rho}_{q+2}, \dots, \hat{\rho}_{q+M}$  (M is taken as  $\sqrt{n}$  or  $n/10$ , M is taken as  $\sqrt{n}$  in this article), and investigate whether the number which meet

$$\hat{\rho}_k \leq \frac{1}{\sqrt{n}} \sqrt{1 + 2 \sum_{i=1}^q \hat{\rho}_i^2}$$

or

$$\hat{\rho}_k \leq \frac{2}{\sqrt{n}} \sqrt{1 + 2 \sum_{i=1}^q \hat{\rho}_i^2}$$

(the article selected the latter) accounts for 95.5% of M. If the number accounts for 95.5% of M and if  $1 \leq k \leq q_0$ ,  $\hat{\rho}_k$  are significantly different from zero, and after  $\hat{\rho}_{q_0}$  that  $\hat{\rho}_{q_0+1}, \hat{\rho}_{q_0+2}, \dots, \hat{\rho}_{q_0+M}$  are near zero, you can approximate  $\{\hat{\rho}_k\}$  is the  $q_0$  steps is truncated, or for the tail, the model selected as MA (q),  $q_0$  is initially identified model order. This article select the setting value = 2/sqrt (N) as  $q_0$ ,  $\hat{\rho}_k$  is k which is the first less than of the value (value = 2/sqrt (N)).

2. For each p, or a sequence satisfy investigated whether the number which meet  $|\hat{\varphi}_{kk}| \leq \frac{1}{\sqrt{n}}$  or  $|\hat{\varphi}_{kk}| \leq \frac{2}{\sqrt{n}}$  accounted for 95.5% of M, if satisfied, and approximate  $\{\hat{\varphi}_{kk}\}$  is  $p_0$  steps truncated, or for the tail. At this time, the selection model is AR (p), and  $p_0$  is the initially determined order.

3. If sequence  $\{\hat{\rho}_k\}$  and  $\{\hat{\varphi}_{kk}\}$  are not truncated, the selection model is ARMA (p, q).

Finally, the model order, when the model selection has been completed, need to re-determine the order of the model, namely, model checking, selected the final model order. Order determination methods used in this article have two kinds, AIC criterion and BIC criteria.

1. Akaike's information criterion

Akaike's information criterion is the minimum information standards, its function is:

$$AIC(p, q, \mu) = \ln(\hat{\sigma}_k^2) + 2k/N$$

Among them ( $k = p + q + 1$ ),  $\sigma_k$  is residual, N is the sequence length; with the method of the Akaike's information criterion set order for:

$$AIC(k_0) = \min_{1 \leq k \leq M(N)} AIC(k)$$

Select a different p and q, calculate the corresponding value of AIC, the minimum p, q values of AIC is the determined model order, along with the increase of p and q, in formula (1) the first term decreasing and the second term rising, the article selected Akaike's information criterion is the AR (p) model and the ARMA (p, q) to make model order, wherein when the model is selected as the AR (p), where the upper limit of p is the initial order of model selection[9].

2. Bayesian Information Criteria

Bayesian Information Criteria function is:

$$BIC(k) = N \ln(\hat{\sigma}^2) + k \ln N$$

Among them,  $\sigma_k$  is the residual,  $k = p + q + 1$ , similarly with the Akaike's information criterion, the p, q is the determined order of the model when p、q is the minimum value of BIC.

**2.6. Parameter Estimation**

The method of parameter estimation adopted by the article is the least squares method, its fundamental principle is:

When the sequence is represented as the following linear model:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + e_i, i = 1, 2, \dots, n$$

Among them,  $y_1, \dots, y_N$  as the values to be predicted,  $x_{i1}, \dots, x_{in}$  as the known independent variables,  $\beta_1, \dots, \beta_n$  as the parameters to be estimated,  $e_i$  is the residual which is uncorrelated and zero mean, this formula can also be written as:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}$$

Or

$$Y = X \beta + e$$

If sum of squared errors is minimum , namely

$$Q(\beta) = Q(\beta_1, \beta_2, \dots, \beta_n) = \sum_{i=1}^N (y_i - \beta_1 x_{i1} - \dots - \beta_n x_{in})^2 = \sum_{i=1}^N e_i^2$$

Based on moment N all the least squares for observation are:

$$\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nn} \end{pmatrix}^T \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

## 2.7. Model and Forecast Evaluation Index

1. According to the above method to determine the time-series analysis model , using the corresponding prediction algorithm to forecast and analyze, the predictive models are:

- Autoregression(AR) model

The expression of the AR model is:  $y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t$

Wherein  $\phi_1, \dots, \phi_p$  are parameters to be estimated,  $\varepsilon_t$  s error or white noise.  $y_t$  is a predicted value, and  $y_{t-1} \dots y_{t-p}$  is a detected value. Predicted values in a coming period of time can be calculated by means of the recursion operation. The model does not include a moving average part.

- Moving average (MA) model

The expression of the MR model is:

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}$$

The MA model does not include an autoregression part.

- Autoregression moving average (ARMA) model

The expression of the ARMA model is:

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_q \varepsilon_{t-q}$$

## 2. Evaluation index of the predicted results

This article uses the mean absolute percentage error (Mean Abs. Percent Error MAPE) and mean absolute error (Mean Absolute Error MAE) as the evaluation index.

Mean Abs. Percent Error MAPE is

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

Mean absolute error is:

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|$$

Among them,  $y_t$  as the actual value,  $\hat{y}_t$  as the predicted value,  $n$  is the sequence length.

## 2.8. Hadoop-based ARIMA Algorithm

Currently, the meteorological observation data faces many problems in the field of data storage and data management, main problems are: the growth rate of the volume of data and the type of data attributes is high, data response speed is required to be high, data are required to have good stability and high safety, and be easy to use and easy to maintain. To solve the above problems, this article constructs a meteorological data storage and mining platform based on the Hadoop framework, and implements the ARIMA model prediction algorithm. Hadoop is an open source software platform of Apache, large volumes of data can be stored and managed in the platform, also it is relatively easy to write and run applications for massive data processing. An HDFS which is short for Hadoop Distributed File System is realized. The HDFS is good in fault-tolerant property and is designed to be arranged on low-cost hardware. The HDFS is of a Master / Slave structure and comprises a NameNode and a plurality of DataNodes, wherein the NameNode is responsible for management of metadata, file blocks and namespace of the HDFS. It monitors request, processes request and heartbeat detection The NameNode is the master server, and the DataNodes are responsible for practical storage and management of data [10].

The core idea of Hadoop framework is Map/Reduce, wherein the Map/Reduce is a programming model that can be used for calculation of mass data, and is a fast and efficient task scheduling model, it will divide a large task into many subtasks of fine-grained, these subtasks can be dispatched among spare processing nodes with higher processing speed, so as to handle more tasks. Thus, slow processing nodes are avoided to shorten the time of the completion of the whole task [11].

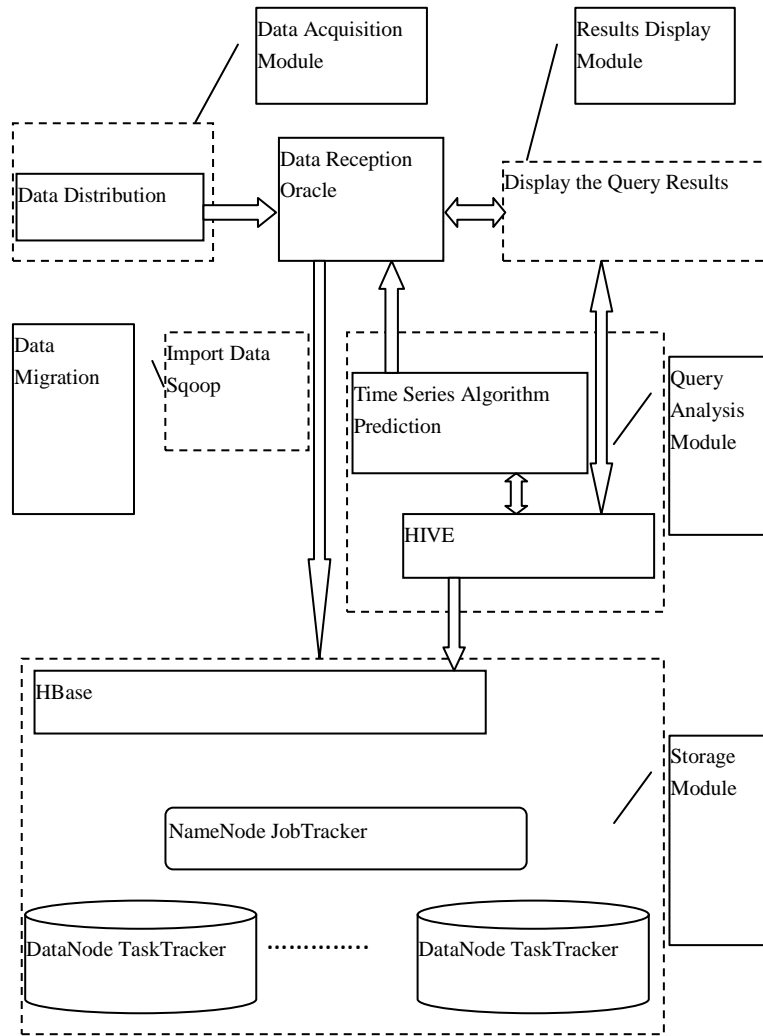
Correlation technologies of Hadoop comprise of Hadoop Database (HBase), Hive and Sqoop. HBase is a distributed storage system which is highly reliable, effective in performance, column-oriented and flexible. A large-scale structured storage cluster can be constructed on a low-cost PC Server by means of the HBase technology. Hive is a data warehouse tool based on Hadoop, and is capable of mapping structured data files onto a data base chart, providing a complete sql inquiry function, and converting sql statements to Map/Reduce tasks for operation [12]. Sqoop is a tool used for transferring data between Hadoop and data in a relational database, and is capable of importing the data in a relational database such as MySQL, Oracle and Postgres to the HDFS of the Hadoop and importing the data of the HDFS to the relational database [13].

## 3. Experiment Design and Result Analysis

### 3.1. The Weather Forecasting Platform

Based on the above analysis and design, the overall functional structure weather forecasting system based on the Hadoop Framework can be designed as five parts which are data acquisition module, results display module, query analysis module, storage module and data migration[17], as shown in Figure 2.





**Figure 2. The Whole Function Structure Diagram of Weather Forecast System**

**3.1.1.** Data Acquisition Module provides visual Server Application Programming Interface of data distribution and data reception. Meteorological data can be published manually, data can be acquired with access of meteorological data acquisition equipment, the small received data are first stored in an Oracle database, when small data are accumulated to a certain number, the small data will be transferred into the storage module, transferred data will be automatically deleted.

**3.1.2.** Storage Module is responsible for data storage of Metadata and entity data, and provides data backup. Hbase is storage database of entity data and the metadata, HDFS is underlying storage container, HDFS is not limited by data type and can be any type of data. Small data in the data acquisition module accumulated to a certain amount will be deposited in the storage module Hbase on a regular basis.

**3.1.3.** Query Analysis Module includes two parts of the truthful data reading and the establishment of forecast data. The truthful data reading is achieved mainly by Hive which is

a data warehouse tool. The Hive supports the SQL-statement and enables query to be easy. Hbase does not support the SQL-statement, developers are required to learn Hbase-supported languages specially in the development process, which is very inconvenient. Hive also offers external data query management API (Server Application Programming Interface), Hive can automatically compile SQL-statement which is submitted by the user into Map/Reduce form to execute, Map/Reduce is suitable for large data processing, can significantly improve the response speed, and return query results [14]. Establish forecast data is prediction function of future data, the function is realized through ARIMA algorithm, The data of past few years are used to make predictions about data in the coming 15 days, the prediction results are stored in forecasted statement of Oracle database of acquisition module, and users are enabled to enquire about the weather forecast easily.

**3.1.4. Results display Module.** For ordinary users results which are returned from the query analysis module will be displayed in this module in a visualized mode; For administrators it not only can display the query results but also can display a distributed file system structure, some management operations can be conducted to file system structure, and can manage database tables.

**3.1.5. Data Migration Module.** This module is used for transferring data from the Oracle database to HBase database through Sqoop, the execution is self-timing.

**3.2. The Establishment of a Data Set**

Experimental data is ground meteorological data which include eight properties, namely site, date, daily mean temperature, daily mean humidity, daily mean vapor pressure, daily atmospheric pressure, daily maximum temperature and daily minimum temperature, as shown in Table 1.

**Table 1. Hbase Database Table**

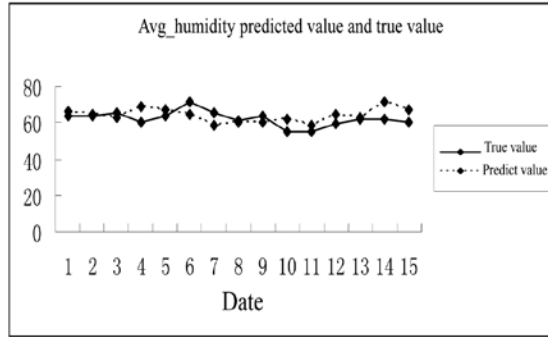
Row key	Time-stamp	Column Family					
		temperature			pressure		humity
attribute	T	AT	MAXT	MINT	AVP	AP	AH

In the Table I, AT as avg\_temperature (average temperature), MAXT as max\_temperature (maximum temperature), MINT as min\_temperature (minimum temperature), AVP as avg\_water\_vapor\_pressure (average vapor pressure), AP as atmospheric\_pressure (atmospheric pressure), AH is avg\_humity (average humidity). Attribute as a database table Rowkey, Timestamp is automatically assigned when there are written in Hbase Hbase. Temperature, pressure, humity are clusters of three columns. Under each column cluster also includes several columns, temperature include three columns, AT, MAXT, MINT. Pressure include two columns AVP, AP. humity only include the AH.

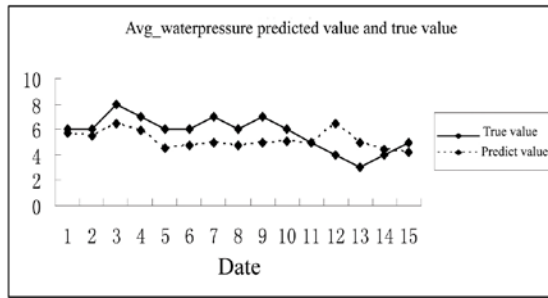
**3.2. Experiment Design and Result Analysis**

This article adopts daily mean water pressure and daily mean humidity data of a station over the past 10 years to conduct the forecasting experiment and predict the data of the coming 15days, and uses the above weather forecasting system which enables the ARIMA model prediction algorithm to be realized to make predictions. After each step of ARIMA model prediction algorithm, the final model of the daily mean vapor pressure is ARIMA (2, 1, 2), and the final model of daily mean humidity is ARIMA(4, 0, 3). These two models are used

to predict daily mean vapor pressure and daily mean humidity of the coming 15days, the comparison between predicted results and real data is shown in Figure 3 and Figure 4.



**Figure 3. Comparison Chart of Daily Average Relative Humidity**



**Figure 4. Comparison Chart of Water Vapor Pressure**

Intuitive analysis of experimental results shows that with the increase of the prediction step length of the two sequences the predicted effect is getting worse. This article selects mean absolute percentage error (Mean Abs. Percent Error MAPE) and mean absolute error (Mean Absolute Error MAE) as evaluation indexes for accurate analysis, the results are shown in Table 2, Table 3.

**Table 2. Error Analysis Table of Daily Average Relative Humidity**

STEP	MAE	MAPE
1	2.60	0.041
2	1.75	0.027
3	2.00	0.031
4	3.60	0.058
5	3.50	0.056
6	4.05	0.063
7	4.40	0.068
8	4.00	0.062
9	3.96	0.062
10	4.29	0.069
11	3.87	0.068
12	4.00	0.070
13	4.03	0.066
14	4.41	0.072
15	4.60	0.075

**Table 3. Error Analysis Table of Daily Average Water Vapor Pressure**

STEP	MAE	MAPE
1	0.30	0.050
2	0.40	0.067
3	0.77	0.107
4	0.85	0.119
5	0.98	0.146
6	1.03	0.157
7	1.17	0.176
8	1.19	0.181
9	1.28	0.193
10	1.24	0.188
11	1.14	0.173
12	1.25	0.211
13	1.30	0.243
14	1.23	0.233
15	1.21	0.228

As shown in the MAPE column, the error of daily average relative humidity does not exceed 10%, the effect is good and the daily average vapor pressure prediction is poor. Start from twelfth step of prediction, the error exceeds 20%, but overall does not exceed 25%. As shown in the MAE column and MAPE column, as the prediction step increases, basically, the prediction error of daily average relative humidity and the prediction error of daily average vapor pressure are both on the rise, which indicates that there are some bugs in the prediction algorithm and multi-step prediction model and further improvement is needed.

#### 4. Conclusions

This paper designs the Meteorological data storage and mining platform based on the Hadoop-based ARIMA algorithm. The platform is based on the distributed file system HDFS, and combine the distributed database HBase, the data warehouse management tool Hive, the distributed databases, the relational database data migration tool Sqoop and other tools. The data mining prediction algorithm-ARIMA time series prediction algorithm is also integrated into the system. The platform has the ability of mass storage of meteorological data, efficient query and analysis, weather forecasting and other functions.

#### Acknowledgment

Authors are very grateful for the funding support from Natural Science Foundation of Inner Mongolia, China(The number of Item is : 2012MS1008), and the Scientific Research Project of Colleges and universities in Inner Mongolia, China(The number of Item is: NJZY11087).

#### References

- [1] Y. W. Dou, L. Lu, X. Liu and Daiping Zhang, "Meteorological Data Storage and Management System", *Computer Systems & Applications*, vol. 20, no. 7, (2011) July, pp. 116-120.
- [2] C. Zhang, W.-B. Chen, X. Chen, R. Tiwari, L. Yang and G. Warner, "A Multimodal Data Mining Framework for Revealing Common Sources of Spam Images", *Journal of multimedia*, vol. 4, no. 5, (2009) October, pp. 313-320.

- [3] C. Li, M. Zhang, C. Xing and J. Hu, "Survey and Review on Key Technologies of Column Oriented Database Systems", *Computer Science*, vol. 37, no. 12, (2011) February, pp. 1-8.
- [4] M. Zhang, "Application of Data Mining Technology in Digital Library", *Journal of Computers*, vol. 6, no. 4, (2011) April, pp. 761-768.
- [5] C.-W. Shen, H.-C. Lee, C.-C. Chou and C.-C. Cheng, "Data Mining the Data Processing Technologies for Inventory Management", *Journal of Computers*, vol. 6, no. 4, April (2011), pp. 784-791.
- [6] Z. Danping and D. Jin, "The Data Mining of the Human Resources Data Warehouse in University Based on Association Rule", *Journal of Computers*, vol. 6, no. 1, (2011) January, pp. 139-146.
- [7] J. Jiang, B. Guo, W. Mo and K. Fan, "Block-Based Parallel Intra Prediction Scheme for HEVC", *Journal of Multimedia*, vol. 7, no. 4, (2012) August, pp. 289-294.
- [8] S.-Y. Yang, C.-M. Chao, P.-Z. Chen and C.-Hao, "SunIncremental Mining of Closed Sequential Patterns in Multiple Data Streams", *Journal of Networks*, vol. 6, no. 5, (2011) May, pp. 728-735.
- [9] Z. Fu, J. Bai and Q. Wang, "A Novel Dynamic Bandwidth Allocation Algorithm with Correction-based the Multiple Traffic Prediction in EPON", *Journal of Networks*, vol. 7, no. 10, (2012) October, pp. 1554-1560.
- [10] Z. Qiu, Z.-W. Lin and Y. Ma, "Research of Hadoop-based data flow management system", *The Journal of China Universities of Posts and Telecommunications*, vol. 18, (2011) February, pp. 164-168.
- [11] J. Cui, T. S. Li and H. X. Lan, "Design and Development of the Mass Data Storage Platform Based on Hadoop", *Journal of Computer Research and Development*, vol. 49, no. 12, (2012) May, pp. 12-18.
- [12] P. Sethia and K. Karlapalem, "A multi-agent simulation framework on small Hadoop cluster", *Engineering Applications of Artificial Intelligence*, vol. 24, no. 7, (2011) May, pp. 1120-1127.
- [13] H. Yu, J. Wen, H. Wang and L. Jun, "An Improved Apriori Algorithm Based on the Boolean Matrix and Hadoop", *Procedia Engineering*, vol. 15, (2011) July, pp. 1827-1831.
- [14] B. Dong, Q. Zheng and F. Tian, "Optimized approach for storing and accessing small files on cloud storage", *Journal of Network and Computer Applications*, vol. 35, no. 6, (2012) May, pp. 1847-1862.
- [15] G. Mao, "Theory and Algorithm of Data Mining", Beijing: Tsinghua University Press, (2007), pp. 121-142.
- [16] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh and D. A. Wallach, "Bigtable: A distributed storage system for structured data. Proc.of the 7th USENIX Symp.on Operating Systems Design and Implementation, (2006), pp. 205-218.
- [17] S. Ghemawat, H. Gobiuff and S.-T. Leung, "The Google File System", *Proc. of the 19th ACM Symp on Operating Systems Principles*, (2003), pp. 29-43.

## Authors



**Leixiao Li**, He was born in July 1978, and acquired master's degree in the field of Computer Application Technology from China Inner Mongolia University of Technology in July 2007, Since September 2007, he taught in Department of Computer Science of Information Engineering College in Inner Mongolia University of Technology. The main research areas include Cloud Computing, Data Mining, Software Modeling, Analysis and Design, Web Information Systems. In recent years, he has published more than 10 papers about teaching and research in the core journals and presided over a provincial scientific research project and two university research projects.



**Zhiqiang Ma**. He was born in July 1972, and graduated HoHai University in China in July 1995. He joined Inner Mongolia University of Technology. He acquired a master's degree in the field of Computer Application Technology from Beijing Information Science & Technology University in July 2007. He won the outstanding master's thesis. He became associate professor and master's tutor in May 2010. The main research areas include Cloud Computing, Data Mining, Search Engine and Chinese Word Segment. In recent years, he has published more than 10 research papers in the core journals and presided over a provincial scientific research project and one university research projects.



**Limin Liu.** He was born in October 1964, and acquired master's degree in the field of automatization from China Tsinghua University in July 2002, Since September 1995, he taught in Department of Computer Science of Information Engineering College in Inner Mongolia University of Technology, he obtained the title of professor in December 2005. He was a senior member of China Computer Society, his main research areas include Cloud Computing, Data Mining, Software Modeling, Analysis and Design. In recent years, he has published more than 10 papers about teaching and research in the key journals and presided over multiple provincial scientific research projects and multiple university research projects.



**Yuhong Fan.** She was born in Nov. 1987. July 2010, he obtained his bachelor's degree in computer science and technology from Shanxi Datong University. She studied in College of Information Engineering in Inner Mongolia University of Technology from Sept. 2010 to July 2013. In July 2013, she got the master degree of computer application technology. The main research areas include Data mining, Cloud Computing and Java Application System.