Research on Meteorological Data Mining Based on Cloud Computer and Hadoop

Lina Hu

Heilongjiang University of Technology, Jixi, 158100, China

Keywords: Meteorological data, Cloud computer, Hadoop

Abstract: The business goals of cloud computing and Hadoop are respectively the publicity of IT resources and the efficient processing of massive data. This paper designs a meteorological data mining algorithm based on cloud computing and Hadoop. It is not difficult to see from the experimental results that the computational efficiency and the integrity of the algorithm are better than those of the traditional algorithm when dealing with massive data, and the weather prediction results are more accurate.

1. Introduction

Meteorological data are widely used and play an important role in scientific research and national economic construction [1]. Meteorological data can not only provide information based on ground and high-altitude meteorology, but also contain a large number of meteorological laws in more disaster data and radar meteorological observation data, which provide data support for meteorological prediction. Large meteorological data contains abundant application and research value, and can provide various meteorological services, including internal operational services, scientific research services and public services of meteorological departments. Operational services within meteorological departments mainly refer to various services provided by meteorological department staff, such as meteorological data query, production of forecast products, meteorological data warehousing and other functions. Meteorological scientific research services mainly refer to the use of large meteorological data for numerical analysis, meteorological disaster risk assessment and other appreciation services. Meteorological services mainly refer to the processing and analysis of large meteorological observation data, which will benefit people's daily production, life and other industries, including oceans, agriculture, transportation and so on. Management, highly reliable storage, analysis, processing and retrieval are facing enormous technical challenges. The independent existence of meteorological data has little value, but often with other data, it will produce comprehensive results. It is of great significance to mine the relationship between meteorological data and other business activities so as to achieve accurate marketing and forecast. The accuracy of meteorological forecast has also increased from several hundred kilometers and tens of kilometers to several kilometers, which greatly increases the amount of calculation of the model. Therefore, how to efficiently mine the historical laws of Meteorology from these massive data and effectively store and process large meteorological data has become an urgent problem to be solved. Now, cloud computing and Hadoop technology can provide technical support for meteorological big data service [2].

2. Cloud Computer and Hadoop

2.1 Cloud Computer.

Cloud computing is a new business computing model. It distributes computing tasks on resource pools composed of a large number of computers, enabling various application systems to acquire computing power, storage space and various software services according to their needs [3]. Cloud is a virtual computing resource that can be self-maintained and managed. It has the following characteristics. Cloud computing does not need to use local devices, with the help of third parties to provide services. In data security, we adopt such measures as computing node isomorphism and data

DOI: 10.25236/dpaic.2018.027

multi-copy fault tolerance to ensure the security of its data, which can effectively avoid data loss or theft. Cloud computing virtualizes the underlying devices such as storage devices and servers on the basis of the network, and builds a virtualized resource pool. Users do not need to install any software, just through the browser can query a large amount of data information, access to various services provided by cloud computing. Advanced computing power is the main feature of cloud computing. Cloud computing is formally based on large data environment, and its computer power is not comparable with any other software. The computing power reaches 10 trillion times per second, which can fully meet any business needs of ordinary users. Cloud computing can achieve on-demand allocation and automatic growth of underlying resources in the process of data processing. Upper data can be matched with corresponding isolation applications on-demand to build a perfect IT architecture. Cloud storage technology enables a large number of different types of storage devices in the network to work together under the management of application software by cluster application grid technology and distributed file system technology to provide data storage and business access functions to the outside world. These characteristics of cloud computing make it possible to commercialize data storage, analysis and application, and also make data mining in cloud computing environment a research field with theoretical and applied value.

2.2 Hadoop.

As a distributed computing platform based on Linux platform, Hadoop has two core technologies. They are HDFS distributed file system and MapReduce parallel computing framework, which provides an open source framework for developers. Users can store large amounts of data distributed through their own needs, and provide a reliable multi-node backup to prevent the system from paralysis due to the failure of a node. At the same time, users can redevelop on the basis of MapReduce framework according to their own needs to realize parallel computing in the processing process, which can greatly improve the efficiency of data computing and analysis to meet the needs of data computing in the Internet era. At the same time, in order to speed up the efficiency of data sharing, Hadoop supports data movement between nodes, and through the load balancing mechanism can ensure the dynamic balance of data distribution among nodes. As one of the core technologies of Hadoop, HDFS is a distributed file system that provides high throughput, is the cornerstone of distributed storage and management of massive data, and is the open source implementation of GFS. HDFS uses efficient distributed algorithm to store a large number of backup data scattered to multiple data nodes and can effectively store data scattered in the cluster. At present, it has become the flagship file system of Hadoop. MapReduce is mainly used for parallel computing of large-scale data sets above TB/PB level. Map and Reduce divide a huge computing task into several independent small tasks and distribute them to each node of the computer cluster for computing processing. Finally, the calculation results of each node are merged and integrated to obtain the final calculation results. MapReduce aims to enable programmers who are not familiar with parallel programming to write distributed programs quickly, and to use the powerful computing power of distributed computer cluster to accomplish huge computing tasks [4].

3. Meteorological Data Mining Based on Cloud Computer and Hadoop

3.1 Experiment Environment.

This experiment uses two servers and one disk array. The server is configured with 32GB memory, 500GB hard disk and 8-core CPU. The capacity of the disk array is 5TB. The network environment has a bandwidth of 100 megabytes. Three-node cluster is built in the experiment. Virtual 3 PCs in the server, configure 4 GB memory, 100 GB hard disk, 2 2 core CPUs. Ubuntu version 12.04 is selected as the operating system and JDK version 1.6.30 (Linux version) is selected as the Java environment. The Hadoop platform version used in cluster building is Hadoop 1.0.3. The IDE environment is Eclipse 3.7. The platform of Hadoop product development and operation is Linux system, which has been validated on the cluster system composed of 2000 nodes. Because distributed operation has not been fully tested on Windows platform, Windows platform is only supported as a development platform, so we build the platform on Ubuntu 12.04. In addition, Hadoop cluster supports single-machine mode, pseudo-distributed mode and fully distributed mode. Our experimental environment chooses fully distributed mode. The Hadoop cluster used in the experiment consists of three nodes, the master as Nome Node and Job Tracker, node1 and node2 as Data Node and Task Tracker. They communicate through ethernet. Open the / etc / hosts file as root, add IP address and host name, and record one line at a time. Noe1 and Noe2 need to be set up in the same way. Hadoop installation must first install Java version 1.5. X or more, and after successful installation, it needs to configure environment variables. Secondly, the remote Hadoop daemon needs to be managed during the operation of Hadoop, so SSH Server should be installed. After installation, SSH needs to be password-free configuration, mainly in order to implement the execution of instructions between machines without entering a password. We ensure that after all the necessary software is installed on each node in the cluster, we can install and configure the Hadoop cluster.

3.2 Experiment Process.

We have done a good job in data preparation, using the daily data set of climate data provided by the China Meteorological Science Data Sharing Service Network at the China Ground International Exchange Station. Boolean association rules deal with discrete and categorized values, which show the relationship between these variables. For the mining of association rules of meteorological data, our idea is to discretize the data, transform its attributes into hypotheses, divide each factor into several levels, and mark each factor with a symbol. At 20-20 hours, the precipitation is marked by symbol J, the average wind speed is marked by F, the average temperature is marked by W, the average vapor pressure is marked by Y, the average relative humidity is marked by S, and the small evaporation is marked by Z. The parallel FP-Tree algorithm is written in Java language on Eclipse platform, and the user interface is designed. The algorithm uses the data set in HDFS as input and writes the operation results into HDFS. The specific experimental process is as follows. First, the data set to be mined is uploaded to Hadoop Distributed File System (HDFS). When the program uses the data, it reads from HDFS. At the same time, the output file name is set so that the system can store the operation results in HDFS. Before data mining, the user sets the minimum support, and then generates frequent itemset for the data set to be mined. The generated results are saved from the system to HDFS and displayed in the user interface. Hadoop port 50030 is used to view the running process and platform parameters. After the program runs, the experimental results can be obtained from the background through HDFS or directly from the user interface to verify the experimental results. Change different parameters, repeat experiments, record the experimental results, and analyze its performance. The rules are extracted from the actual data, the support and confidence are calculated, and the meteorological data are predicted.

3.3 Performance Evaluation.

The running time of the algorithm decreases obviously with the increase of the number of nodes. If the computing resources are sufficient, the algorithm can improve the computing speed more obviously. It can be seen that adding computing nodes appropriately when dealing with very large data sets can obviously improve the computing speed, and the algorithm has good scalability. Throughout the three groups of implementation results, we can see that the expected design is basically achieved. When MapReduce-based parallel FP-Tree programs deal with large-scale data sets, the improvement of parallelism is closely related to the improvement of performance. Because Hadoop platform is built on the operating system, the task scheduling cycle is larger than the process scheduling cycle of the operating system, because the processing of small data sets requires less computation and serious performance waste, so it takes longer time than running on a single machine. At the same time, the allocation of node capacity will also affect the performance of the algorithm. Reasonable configuration of cluster nodes has obvious effect on improving platform performance. In the experiment of changing the support degree, we can see that the change of the support degree has a certain impact on the performance of the algorithm. The larger the support degree, the faster the running speed, but it may have a certain impact on the accuracy of the generated rules. In practical application, we should find an optimum degree of support so that the operation speed and accuracy can reach an ideal balance state.

3.4 Meteorological Prediction.

Prediction experiments use data as real data. After data mining, the rule set extracted can be used as the rule base of prediction, and the support and confidence of rules can be obtained. Because the data collected in this experiment are all real data, but because mining association rules with support and trust cannot distinguish whether the results are true or not, it is necessary to consider a problem of relativity of computing rules, the possibility that correlation V appears simultaneously with time and various events. Completely independent possibilities as the ratio of possibilities, V less than 1, are negatively correlated and meaningless. V equals 1, which means that events are independent of each other, and V > 1 is positively correlated and meaningful. Data mining is an important part of meteorological data processing. It is a process of discovering various models and outlines from known data sets. In fact, data mining is a cyclic and repetitive process. Firstly, some analysis tools are used to check the data and analyze the data from a certain aspect, and some modifications may be made to the data. Then go back to the beginning and apply other analytical tools to get a different or better result. Nevertheless, data mining is not a random application of analytical methods, but a process of carefully arranging and taking into account what is the most useful and appropriate. The process of data mining can be summarized as follows: problem definition, data cleaning and integration, data selection and transformation, data mining algorithm execution, and evaluation and representation of results. In the process of integration and convergence, cloud/client processing of business system can make business application of meteorological department realize mobile interconnection step by step, so as to get rid of the limitation of deskside system and use mobile intelligent terminal to realize mobile business. The process of integration also provides an opportunity for combing and optimizing business processes to optimize business processes.

4. Conclusion

The meteorological big data service provides the direction and goal of meteorological service for meteorological workers, and it is also an important branch in the process of national informatization. Making meteorological service products with large meteorological data can benefit people in all aspects of daily life. Cloud computing and Hadoop technology can be effectively applied to the processing and analysis of large meteorological data. The service and application of large meteorological data in cloud environment can effectively improve public satisfaction with meteorological services.

Acknowledgements

This paper is the result of the Planning Project of Science and Technology of Jixi City named Research on Meteorological Data Mining Based on Cloud Computer and Association Rules Mining Technology.

References

[1] Zhang Jie, Xue Shengjun. Application of the Services of Meteorological Big Data in Cloud Computing [J]. Journal of Anhui Agricultural Sciences, 2016, 44(5): 298-301.

[2] Zhao Bingxue, Wang Lei, Cheng Dongya. Comparison of Spatial Interpolation Method for Meteorological Data and Distribution Characteristic in Anhui Province [J]. Research of Soil and Water Conservation, 2017, 24(3): 141-145.

[3] Chen Qing, Yang Ming, Xiao Yun, et al. Application of Cloud Data Storage Technology in Meteorological Data Storage [J]. Computer Applications and Software, 2018, 35(8): 124-127+158.

[4] Jia Songlin, Wang Ying, Xue Lei, et al. Design and Implementation of the Meteorological Data Element Administration System [J]. Advances in Meteorological Science and Technology, 2018, 8(1): 123-126+157.