2017 3rd International Conference on Social Science and Technology Education (ICSSTE 2017)
ISBN: 978-1-60595-437-0

Data Mining Algorithm Implementation and Its Application in Parallel Cloud System based on C++

Jiangtao Geng[1], Xiaobo Xiong[1]

Abstract

This paper conducts the analysis on the data mining algorithm implementation and its application in parallel cloud system based on C++. With the increase in the number of the cloud computing platform developers, with the use of cloud computing platform to support the growth of the number of Internet users, the system is also the proportion of log data growth. At present applies in the colony environment many is the news transmission model. In takes in the rest transmission model, between each concurrent execution part exchanges the information, and the coordinated step and the control execution through the transmission news. As for the C++ in the data mining applications, it should firstly hold the following features. Parallel communication and serial communication are two basic ways of general communication. Under this basis, this paper proposes the novel perspective on the data mining algorithm implementation and its application in parallel cloud system based on C++. The later research will be focused on the code based implementation.

Keywords: Data Mining, Parallel Cloud System, C++, Implementation and Its Application

Introduction

The cloud computation is one kind of quite emerging business accounting model. It will calculate the duty to distribute on the resource pool which the massive computers will constitute and will enable each kind of application system according to need to gain the computation strength, the storage space and each kind of software service. More and more application developers are also turning to the cloud, to meet the need to accelerate the application development cycle and ensure the application of high stability and the high availability. With the increase in the number of the cloud computing platform developers, with the use of cloud computing platform to support the growth of the number of Internet users, the system is also the proportion of log data growth, how to collect these log data, how to store these massive data, how to help platform maintainers and application developers to analyze the data that are currently facing problems. In general a cloud computing environment will need to open the log collection and analysis ability into general services for the application of the cloud computing platform components and user use. The following challenges facing logging system in modern cloud computing environment which will serve as the basis of this research.

- High stability and high availability: logging system as the basis of the cloud computing system component needs to have high stability and availability.

[1]Guangzhou International Economics College, Guangzhou 510540, China

- Log data is very large: the requirements of the cloud environment in the log system can handle large amounts of data that can increase the amount of data in the case of good scalability.
- Log data sources are extensive: log data in the cloud computing environments comes from multiple distributed deployments of system platform components and application developers deployed applications that need to be able to collect logs for different sources and the different application developers want to view their application running log.

If uses the colony technology build high performance the server, between various servers each other system resources use factor has the very big disparity frequently causes various servers not to be able evenly to undertake the request which the user sends out, namely some servers are very busy, but some servers then very idle, and finally causes the colony overall performance to drop greatly that must solve this problem that must rely on the effective load equalization algorithm, the effective load equalization algorithm may the user request assigns reasonably for the backstage each server causes various servers to undertake the duty quite balanced, then enhances the colony system handling ability and the grade of service. Under this basis, this paper proposes the analysis on data mining algorithm implementation and its application in parallel cloud system based on C++. To begin, in the following figure one, we show the architecture of the data mining algorithms.

Figure 1. The Architecture of the Data Mining Algorithms.

The Proposed Methodology

Parallel System Architecture Demonstration. High-performance parallel computers can be divided into the four basic categories: (1) multi-vector processing system; (2) multi-processor based on shared memory; (3) large-scale parallel processing system based on the distributed

storage; (4) High-end computer connected through high-speed interconnection network from the cluster computer system.

At present applies in the colony environment many is the news transmission model. In takes in the rest transmission model, between each concurrent execution part exchanges the information, and the coordinated step and the control execution through the transmission news. The news transmission usually faces the distributional memory, but is also suitable for the sharing memory parallel machine. Message passing provides programmers with more flexible means of general control and expresses parallelism. Flexibility and the diversity of the control means are important reasons why messaging parallel programs can provide high execution efficiency. The message-passing model, and on the one hand, provides programmers with the flexibility and, on the other hand, the task of the exchanging complex information and coordinating control between the various parallel execution parts to the programmer, which to some extent will increase the programmers' burden, as which is the message programming model programming level of the main reasons for the low. When other nodes on this requested when locks, so long as its request scope requested with the before other nodes the scope does not conflict, the server may carry on the authorization to it, enables many nodes to be possible to visit the identical block data concurrently. But the traditional EL algorithm, as each lock scope all expands to the block boundary, in each block data could not by the many node concurrent visit, reduce greatly the system concurrency.

After the traditional EL algorithm each time obtains the block scope the lock all to read the block data local cache. After installing Linux, configure the Linux network. TCP/IP network protocol is installed on all nodes microcomputer, and all node computers are set to the same workgroup. Then the node microcomputer plans to install MPI software and stored parallel program hard disk partition or set the folder as a shared document. The following figure shows the sample.

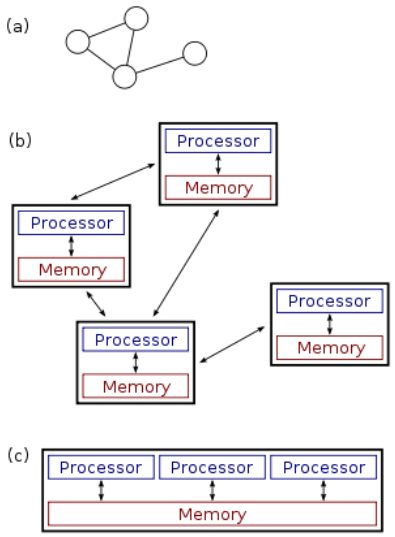

Figure 2. The Parallel System Architecture Illustration.

The Cloud System Architecture Demonstration. Cloud computing network is defined as: cloud computing is the development of distributed computing, parallel computing and grid computing or commercial implementation of these concepts. The cloud system backstage uses the virtual machine by the massive colonies the way, through the high speed internet

interconnection, composes the large-scale hypothesized resource pool while these hypothesized resources may manage and the disposition independently, with the data redundancy way guarantees hypothesized resources high usability, and has distributional characteristics and so on memory and computation, high extension, high usability and user friendliness. Cloud computing has the following main features.

- The cloud computing system is composed the airplane group by the massive commercial computers to provide data processing service to the user. Along with the computer quantity increase, the system appears the wrong probability big increase.
- Due to the special fault tolerant measures of the cloud that can be used to form a cloud node extremely cheap, the cost of data center management automation centralized management cloud so that a large number of companies without the burden of high versatility, the cloud resource utilization rate is greatly improve the traditional system.
- The traditional load equalization algorithm aims at is the duty size is the equal situation, also balanced only to connects the number or the turning on duty number carries on balanced, but calculates platform this regarding the cloud is obviously insufficient, because calculates under the platform in the cloud, the duty size is not fixed, moreover sometimes between the duty can differ in a big way that needs to consider is not merely between the server end performance balance, but also needs to consider the duty the size is imbalanced.
- Cloud computing systems often rely on the basis of many distributed components to provide services. Cloud computing system itself components will produce large amounts of log data, such as, load balancer, application container operation log etc. And the application of the cloud will produce large amounts of log data every day, usually for streaming data, such as, web page browsing, query, and so on.

The MapReduce model becomes abstractly the distributional operation Map and the Reduce two steps, thus realizes the highly effective distributional application. Hadoop MapReduce is a use simple software frame, took an entire quantity data the batch run system, Hadoop by its volume of goods handled in a big way, automatic fault-tolerant and so on the merits, obtained the widespread use in the magnanimous data processing. In ours system, may use in the historical diary data which saves to base in carries on the analysis. Clouds are virtual computing resources that can be self-maintained and then managed, usually for large server clusters, including compute servers, storage servers, and broadband resources. Here the cloud refers to: participate in cloud computing computer collection, relative to the client in terms of a relative concept.

Data Mining Algorithm and C++. In the development of computer languages, type systems are used to define how values and expressions in programming languages are categorized into many different types, how they are operated, and how these types interact with each other. The purpose of the type system is to prevent runtime errors occur during execution. The source code level debugs refers on foot debugs the control movement the unit is on foot in the higher order language source code line of code. Each kind debugs the function to include three kinds generally on foot: Jumps the human, the jump, jumps out that can be then summarized as the follows.

- Through for, while loop, break, continue statement jump to other lines of code, such exports as L (Label) class exports, export address set is denoted by L. These jump instructions in C are eventually encoded as jumping fingers in the BWDSP chip instruction set.

- Run the bank code to program address directly, the export are called N (Next), its export address for N. Program is running after the current line, will naturally run commands in the next line, so the address of the beginning of the next line instruction is an export address bank code. For part of the C language code, and not the next row address for export.
- Return statement through the jump to the return address of this frame, such exports known as R (Return) class export, return address is recorded as R. A function may have multiple return statements, a line of C language code may also have multiple functions return. But because the program flow in the function call stack in the top of the stack, the function returns the first scene must return the frame, so that the function of the current PC in the multiple return statements actually return to the same address.

As for the C++ in the data mining applications, it should firstly hold the following features. Parallel communication and serial communication are two basic ways of general communication. Parallel communication is the use of multiple data lines, each parallel at the same time the transmission of multi-bit data, such as the printer interface 8 data lines at the same time to send data to send one byte at the time. Multi machine is mainly used in the function of distributed system (such as computer control system and machine tool setting tool base management and other systems), the parallel control system of multi machine (such as real-time image processing and data acquisition etc.), local network system (such as communication control etc.). By understanding the features of the C++, we can then propose the integration model of C++ and data mining as the follows.

- Cantor set. Cantor in 1883 first proposed a one-dimensional space of self-similar structure. Take a line segment (0,1), divide it into three equal parts, and then remove one of them, leaving each segment and three equal parts and remove the middle section, so continue to do so, leaving all the segments on the Constitute the Cantor set and it is clear that Cantor set constitutes an infinite level of self-similar structure.
- Silbinski gaskets. Take an equilateral triangle, it is divided into 4 four equilateral triangles with the same size and dig to the middle one, for the rest of the three triangles, each divided into four small equilateral triangle and dig to the middle one. Such points down, the resulting images constitute an infinite level of self-similar structure, called gaskets or arrow design.
- Binski, carpet. A square is divided into nine small squares and dug one of the middle, the remaining eight squares followed by the same method. So go on, finally the self-similar structure of an infinite hierarchy, called the Seychelles binski carpet.

Conclusion

This paper conducts the analysis on data mining algorithm implementation and its application in the parallel cloud system based on C++. More and more application developers are also turning to the cloud, to meet the need to accelerate the application development cycle and ensure the application of high stability and the high availability. With the increase in the number of the cloud computing platform developers, with the use of cloud computing platform to support the growth of the number of Internet users, the system is also the proportion of log data growth. Under this basis, this paper then proposes the C++ integration model with the data mining approaches. In the future, we will then make the code based verification of the proposed model.

References

[1] Rutkowski, Leszek, et al. "A new method for data stream mining based on the misclassification error." *IEEE transactions on neural networks and learning systems* 26.5 (2015): 1048-1059.

[2] Wang, Fei, and Jimeng Sun. "Survey on distance metric learning and dimensionality reduction in data mining." *Data Mining and Knowledge Discovery* 29.2 (2015): 534-564.

[3] Breier, Jakub, and Jana Branišová. "Anomaly Detection from Log Files Using Data Mining Techniques." *Information Science and Applications*. Springer Berlin Heidelberg, 2015. 449-457.

[4] Wu, Zeling, and Haoxiang Wang. "Super-resolution Reconstruction of SAR Image based on Non-Local Means Denoising Combined with BP Neural Network." *arXiv preprint arXiv: 1612.04755* (2016).

[5] Wang, Tong, Huijun Gao, and Jianbin Qiu. "A combined adaptive neural network and nonlinear model predictive control for multirate networked industrial process control." *IEEE Transactions on Neural Networks and Learning Systems* 27.2 (2016): 416-425.

[6] Salvati, L., et al. "Unveiling soil degradation and desertification risk in the Mediterranean basin: a data mining analysis of the relationships between biophysical and socioeconomic factors in agro-forest landscapes." *Journal of Environmental Planning and Management* 58.10 (2015): 1789-1803.

[7] Caballero, Daniel, et al. "Modeling salt diffusion in Iberian ham by applying MRI and data mining." *Journal of Food Engineering* (2016).