Construction of multi tier distributed computing data mining system in cloud computing environment

Xia Wendong[1], Liu yuanfeng[2], Chen deli[1]
1. College of Computer, Jiaying University, Meizhou Guangdong , 514087, China
2. Guangdong Ji Tong Information Development Co., Ltd, Guangzhou Guangdong, 510632, China
Email: 89537987@qq.com

Key words: cloud computing; distributed; data mining; system building; big data

Abstract. multi-source data mining system based on cloud computing service mode is designed in combination with three-layer and four-mode cloud computing service level system targeting such problems as underutilized multi-source distributed data and low data processing efficiency. Firstly, the technical flow of multi-source distributed data mining is elaborated; secondly, the cloud computing service level system is introduced to the system building and multi-source distributed data mining system frame in cloud computing environment is designed; finally, main features of this system are introduced in details, which lays foundation for further study on multi-source distributed data mining system.

Introduction

Since 1980s, the distributed technology has gradually won people's attention [1]. In recent 20 years or more, the study on application of distributed technology in China is mainly conducted with the guidance of geological thinking while the distributed application level largely lags behind the development of spatially distributed technology, which is distinctly manifested by: large amount of useful information in accumulated multi-source satellite distributed data are not fully mined and utilized while the reception and processing of huge amount of data as well as information reception rely on better data processing technology and mining system [2]. Therefore, the design and realization of multi-source distributed data mining system and data processing efficiency have become the hot spot issues for full utilization of distributed information.

Data mining

Data mining is one kind of advanced information extraction technology, which is mainly able to extract the hidden predicative information from large database [4]. Data mining technology can not only recognize the potential mode among data, but mine the valuable information and technology for customers so as to further instruct practical application and study [5].

Data mining methods are generally summarized based on the target of data to be mined by multi-source rocker data mining system; meanwhile, for further application of subsequent multi-source data mining system of this system, one interface is set specially in the part of data mining methods, i.e. user-defined data mining algorithm, which can increase its expandability and multifunctionality.

To sum up, the data mining methods in this paper are divided into five categories as follows:

(1)Comprehensive knowledge mining: usually the functions are mined preliminarily in the first place and the mined comprehensive knowledge can be vividly displayed to customers in graphic way in combination with visualized technology.

(2)Relevant knowledge mining: it mainly covers three types, i.e. association, clustering and classification, among which the common algorithms for clustering are mainly divided into five categories based on division, hierarchy, density, network and model; common methods for classification are decision-making tree, neural network, genetic algorithm and evolutionary theory, Bayesian classification, support vector machine, association classification, analogy learning, rough

set, fuzzy set, etc.

(3)Anticipated speculative knowledge mining: it mainly includes neural network, machine learning, classic statistical approach, etc.

(4)Special abnormal knowledge mining: it is mainly used to reveal the special law that the matters deviate from normal state, which is divided into sequential exception analysis, outlier analysis, special rule detection, etc.

(5)User-defined data mining algorithm: the user can save the algorithm programmed by himself or other function algorithms into the system platform via this setting so that they can be used, improved and sold in future.

Cloud computing

About cloud computing, Tim O' Reilly once pointed out in literature [1] that the cloud computing is one basis for next generation of computing. It is a network, i.e. the world of all computation platform, where all things which we now consider as computers are only one device connected to large computer we build [1]; it is pointed out by Tuncay Ercanlll in literature [2] that cloud computing is one efficient computing mode integrating many functions like expandability, use of virtual resources and sharing by users so that it can be used by different kinds of users without having to understand the background knowledge for this service. Wu et, al. once mentioned in literature [3] that cloud computing is super-computing mode based on virtualization technology, taking the network as the key carrier, mainly focusing on providing various kinds of service and performing cooperative work combining large scale and expandable distributed computing resources. It is considered in this paper by combining the opinions put forward in literature above that the cloud computing is one shared service mode providing relevant operation and functions of information technology based on network and with virtualization technology.

The cloud computing is usually divided into three layers which are application layer, platform layer and infrastructure layer [1]. However, generally the problem that the large amount of data is not used fully exists during the application of satellite's multi-source distributed data. Today, the cloud computing will provide in-depth data analysis and mining for customers, which is called RdaS (the distributed data is taken as service) in this paper. To sum up, the three-layer service level of computer clouding is divided into four service modes, i.e. IaaS, PaaS, SaaS and RdaS.

(1)Infrastructure layer: it mainly takes the infrastructure as the service (IaaS) mode. IaaS is at the lowest level, which serves as standardized service providing relevant storage and computing capabilities on the network. Virtualization is its typical representative technology; hardware, storage and database are the commonest basic service of this service mode.

(2)Platform layer: it takes the platform as the service (PaaS), which is the abstract encapsulation of development environment and valid service load, i.e. the server platform or development environment is taken as service to be provided for customers or systems.

(3)Application layer: it mainly includes two kinds of service modes: the mode taking software as service (SaaS) and the mode taking the distributed as the service (RdaS). This is the highest level in three-layer service of the cloud computing. SaaS provides the software for different customers or systems via internet and the users do not have to buy software independently but use the webpage-based software by leasing from provider; besides, you do not have to maintain the software in use; RdaS, however, takes the original multi-source distributed data and the processed intermediate data as the service to be provided for customers.

Multi-source data mining system in cloud computing environment

A. System structure design

Based on above analysis, one multi-source distributed data mining system frame in cloud computing environment is designed in this paper, as shown in Fig. 3. This system mainly consists of four sub-systems, one database cluster and one server. While using the cloud computing service mode, this system takes the tested distributed data as one shared resource service. This not only

enables the system to better realize the multi-source distributed data mining task, but makes the system improve and update itself continuously via open system so as to improve the data mining efficiency.

Different users use different kinds of terminal devices to be connected with distributed data mining system via internet and each user uses his own account and password to log in the system. All the operation of user in three subsystems, i.e. data management sub-system, mining algorithm subsystem and data mining sub-system, is under the monitoring and control of account management subsystem under this system.

(1)Distributed data mining sub-system

This part is the main part of the system, which uses the data mining process to make knowledge discovery for the data provided by the system and user data and complete the set mining task. It mainly covers eight modules: understanding and definition issue, multi-source distributed data search and extraction, distributed data purification, distributed data mining engine, distributed data algorithm engine, running data mining algorithm, evaluation of results and refinement of data and problems, application of results.

(2)System account management subsystem

System account management system is mainly for use of relevant system service by different users and recording the statement of income and expenditure of other user accounts in details at the same time. Different users mainly include management users and common users. The management users, under the monitoring of this sub-system, use distributed data to make processing and get useful information; common users share distributed data, mining algorithm and various kinds of database resources by paying certain fee and get relevant rewards by sharing other useful distributed data information.

(3)Data mining algorithm subsystem

This subsystem is mainly used to manage the typical algorithm and algorithm model needed by rocker data mining subsystem and it can help realize the function of the “user-defined mining algorithm” in rocker data mining subsystem.

This subsystem is set separately with items like system classic mining algorithm, user-defined mining algorithm, algorithm use record, algorithm sales record, algorithm sharing and help. System classic mining algorithm refers to the classic mining algorithm set inside the system and introduction of its relevant functions; the user-defined algorithms refer to the mining algorithms defined by users themselves or other improved algorithms according to their different needs; the historical record of algorithm refers to the use, search and other relevant situation of the algorithm inside this system.

B. System features

Targeting specific application problems with multi-source distributed data processing, the system designed in this paper expands the three-layer service mode of the cloud computing and takes the distributed data itself as one kind of resources to realize the shared service, which improves the utilization rate and processing efficiency of the multi-source distributed data and overcomes part of drawbacks of the data mining system. Compared with other systems, the system proposed in this paper has following features:

(1)Strong openness, which help the system keep upgrading and expanding. Different kinds of users can use different terminal devices to utilize the obtained distributed data via Internet and make relevant operation, which greatly improves the system openness; meanwhile, the function of user-defined data mining algorithm set by the system enables the user to make improvement and sharing based on the algorithm provided by the system, which not only keeps expanding the types and performance of mining algorithms but also boosts rapidly the study on the algorithm itself.

(2)Convenient system use and fast operation. The system provides various kinds of help functions for the users, including detailed description about mining algorithm, analysis of distributed data and evaluation of results, which enables users of different levels to master the processing of distributed data mining in a faster way.

(3)Low system operation cost and high cost performance. Different user types are designed by

the system, including management users and other users. The management users can finish the data mining purpose with various functions provided by multi-source distributed data set system obtained by the satellite and manage the data use by other users, etc.; other users share this system resource by paying certain fee. Use of service mode in cloud computing environment by this system not only reduces the system operation cost, but also improves the utilization rate and processing efficiency of the system.

Conclusions

Multi-source distributed data resource is introduced into the environment of cloud computing in this paper. Three-layer and four-mode service system targeting distributed data mining system is put forward after analysis of cloud computing system structure, based on which the multi-source distributed data mining system in cloud computing environment is designed. The subsystems of this system provides flexibly relevant data mining service for different kinds of users by combining with relevant cloud service mode and in way of loose-coupled association. To certain extent, this system solves the problem of low utilization rate of the distributed data and low processing efficiency with data sharing and distributed data storage and processing technology and it provides better solution for distributed data mining figures with its better expandability and openness.

Acknowledgement

Guangdong Provincial Department of science and technology in the province of 2014 frontier and key technology innovation (2014B010117002); Guangdong Provincial Academy of Sciences comprehensive strategic cooperation (2013B091500060); Guangdong professional town small and micro enterprise service platform construction project (2013B040500010)

Reference

[1] Malensek M, Pallickara S, Pallickara S. Minerva: Proactive Disk Scheduling for QoS in Multitier, Multitenant Cloud Environments[J]. IEEE Internet Computing, 2016, 20(3):19-27.

[2] Bi J, Zhu Z, Tian R, et al. Dynamic Provisioning Modeling for Virtualized Multi-tier Applications in Cloud Data Center[C]// IEEE, International Conference on Cloud Computing. IEEE Computer Society, 2010:370-377.

[3] Ryan T, Lee Y C. Multi-Tier Resource Allocation for Data-Intensive Computing [J]. Big Data Research, 2015, 2(3):110-116.

[4] Singh S, Chana I. QRSF: QoS-aware resource scheduling framework in cloud computing[J]. The Journal of Supercomputing, 2015, 71(1):241-292.

[5] Ming T. SLA-based optimisation of virtualised resource for multi-tier web applications in cloud data centres[J]. Enterprise Information Systems, 2015, 9(7):743-767.

[6] Xhagjika V, Navarro L, Vlassov V. Enhancing Real-Time Applications by Means of Multi-tier Cloud Federations[C]// IEEE, International Conference on Cloud Computing Technology and Science. IEEE, 2015:397-404.