2018 International Conference on Computational Science and Engineering (ICCSE 2018)

Study on the Feature Mining in Big Data Environment based on Cloud Computing

Boxin Zhang, Xiaochen Zhao, Jingwen Ding

Tianjin Cadake Data Co., Ltd., Tianjin, 300393, China

Keywords: Cloud-based, Feature Mining, Big data Environment

Abstract: Based on the comparison between traditional data mining and big data mining, this paper discusses the connotation of big data mining, and proposes a big data mining architecture that integrates cloud computing and mining services, and adopts a multi-functional Hadoop big data mining platform. It analyzes the internal workflow of big data mining and analyzes its advantages and challenges, so as to provide reference for users' knowledge and application requirements of big data mining.

1. Introduction

Following the continuous development of the Internet, the Internet of Things, cloud computing and the popularization of intelligent terminals, massive and diverse data has exploded, which has prompted the arrival of the era of big data. As an important production factor, big data has become a strategic asset with great potential value, which promotes industrial upgrading and rise, affecting the transformation of scientific thinking and research methods. However, big data, while relying on its rich resource reserves and leveraging its powerful computing technology, also presents challenges. Massive, dynamic, and uncertain data exposes traditional data processing systems to storage and computational bottlenecks. At the same time, traditional data mining techniques are limited in their ability to quickly extract valuable information and knowledge from complex big data in real time. The function has been unable to meet the needs of users. Therefore, in the big data environment, a suitable technology, namely big data mining, is needed to cope with the current challenges. In view of the big data environment, the mining system developed with reference to the traditional data mining construction idea can not provide the service satisfactory to users, so it is essential to meet the construction and application requirements big data mining architecture. However, there are still few models available for reference, and some departments have proposed big data mining solutions based on business needs. Because the system is not portable and the internal components are poorly integrated, there are obstacles in its application.

2. Big data mining

Both have evolved as technology advances, the massive generation of data, and the need for valuable data. However, in terms of the advanced level of technology and the volume, complexity, and processing and analysis capabilities of data, traditional data mining does not have the enrichment of environmental conditions in the era of big data, in the context of databases, data warehouses, and Internet development. The development of data mining from independent, horizontal to vertical is realized. Big data mining is benefiting from the generation and development of technologies such as cloud computing, Internet of Things, and mobile intelligent terminals in the context of big data. It is aimed at the characteristics of big data and the problems faced by existing mining systems, and systematically integrates and improves them with advanced technology. Compared with the traditional data mining has been quite mature applications, algorithm research and system tool development, the research and application of its technology is still in the process of development. The mining of massive data is mainly based on the integration of related technologies based on cloud computing.

Because big data mining and traditional data mining face different data environments, there are

DOI: 10.25236/iccse.18.047

characteristic differences between the two processing objects. The data source of traditional data mining is mainly based on the generation of passive data of a certain range of management information systems, plus the active data generated by users in a few Web information systems. The data types are mainly structured data, plus a small amount. Semi-structured or unstructured data. In addition to the management information system and the Web information system, the data source of big data mining also includes simulation data automatically generated by sensing devices such as the information system. Compared with traditional data mining, big data mining has wider data sources, larger volume and more complex types. Correspondingly, the collection method is no longer limited to passive, the collection range is more comprehensive, the throughput is high, and the processing is real-time and fast. However, due to the low accuracy of the data, the redundancy and uncertainty of the data are high.

3. Cloud computing big data mining architecture

Faced with the shortcomings of traditional data mining in the era of big data, cloud computing, as a highly scalable, highly flexible, and virtualized computing model, provides the driving force for the storage capacity and processing speed of big data mining. The core technologies of cloud computing include distributed storage and distributed parallel computing. Among them, distributed storage is mainly distributed file storage and distributed database storage. The distributed file system represented by GFS has high scalability, high fault tolerance and high throughput. Most of them are suitable for large-scale, distributed and massive data concurrent access. It is not suitable for storing large amounts of small data files and there is a single point. Fault problems, etc., but some systems can store large amounts of small files such as Colossus, Haystack, and TFS (Tao File System). The distributed database includes transactional database and analytical database, see Table 2, which combines the high performance of parallel database and the high scalability of MapReduce. It can store structured, semi-structured and unstructured data to solve traditional data. Mining the storage issues facing them. At the same time, distributed parallel computing represented by MapReduce is easy to use and has good scalability. It is suitable for batch processing of massive data, which can effectively reduce computational complexity and improve computational efficiency. At present, some famous enterprises launch corresponding big data mining solutions based on cloud computing for their own business fields. For example, Google's new generation of search engine platform can realize massive file storage and real-time interactive analysis of big data. Microsoft will use Windows. Azure integrates with Hadoop and integrates with applications such as Business Intelligence BI to provide multi-platform, versatile, high-performance analytics mining services; IBM integrates Hadoop with systems such as stream computing and intelligent analytics for real-time data management And intelligent analysis; other companies on the open source Hadoop platform to improve the functionality, such as Yahoo developed Pig on the Hadoop platform, Facebook developed Hive, etc. Although the above big data mining architectures are different, they have no significant differences in the construction strategies that integrate cloud computing and mining functions. However, in the face of multi-domain data sharing and the applicability of the mining platform, the respective solutions need to learn from each other and further integrate.

1) Support the platform layer. As a resource and power support for big data mining, the platform creates a powerful and resource-rich cloud environment by combining promiscuous big data with multiple cloud-based support processing technologies. This kind of cloud environment can not only provide data, hardware, software and other resources to the outside world, but also can calculate the way of moving to data to inject powerful power into the process of preprocessing, analysis and mining of multi-source complex data. 2) Functional layer. This layer automatically performs intelligent analysis mining based on user needs and preferences. Among them, tools such as analysis and mining rely on the cloud platform for efficient storage and computing capabilities with high scalability and scalability. 3) Service layer. Big data mining automatically interacts with the service provider and user through the client. The mining results are presented to users in the form of services through visualization, data source and other technologies. Overall, big data mining presents the model of cloud services, that is, the functional layer, the service layer and the platform layer are

mutually integrated and interdependent. The three form a variety of analysis, mining and display with powerful computing and storage capabilities as the core. A mining cloud with functional integration, real-time analysis and mining of big data in the cloud, the results are provided to users in the form of infrastructure as a service (IAAS), platform as a service (PaaS) and software as a service (SaaS) .

4. Data storage and its calculation analysis

The traditional data mining storage management is based on relational database systems such as data warehouse, operational database system and file system. It mainly uses row storage to statically and certain structured data with E-R (entity and contact) or multi-dimensional data. Model storage, storage is more passive and access mode is random. The specific mode is generally defined by the system, and the flexibility and scalability are poor. The requirements of transaction ACID (Atomicity, Consistency, Isolation, Durability) are high, and fault tolerance is not. Strong. In addition to traditional data storage, big data mining includes distributed storage, storage structure, semi-structured and unstructured data. Storage policies are mainly based on column storage or row-column mixed storage, and the mode is generally implemented externally, usually does not support ACID features and supports BASE (Basically All Available, Soft State, Eventually Consistent) features and has limited functionality supported by relational databases. For example, Google's Bigtable uses column storage and stores data in the new data model Ordered Table. The model is flexible and simple, and has strong scalability, but the data consistency and compatibility relational data model have problems. In this regard, the Spanner system effectively combines high scalability with ACID features by supporting simultaneous replication and visual sharding across data centers and providing SQL user interfaces. In addition, for uncertain data, big data storage has corresponding uncertain database management systems, uncertain data lineage management techniques, etc. data is stored in an uncertain relationship model, storage methods are direct and strictly sequential, and can be based on memory. Instead of building a summary data structure on disk, it implements direct storage processing of dynamic, indeterminate data.

Compared with traditional data mining, the centralized batch processing mode of data movement calculation, big data mining uses distributed computing to integrate large data into multiple parallel computing modes. For a small amount of static data with less dimension, traditional data mining presents higher query analysis performance due to repeated and precise query methods, OLAP's strong flexibility and faster processing and analysis capabilities. However, in the face of massive data with large dimensional attributes and huge data cubes, traditional OLAP cannot be analyzed automatically, and SQL-based query languages are difficult to express complex analysis models to be constructed, so their query analysis Quality and efficiency can be severely affected. However, big data mining is a fusion of system functions for the poor scalability of traditional analysis tools and weak analysis functions of existing cloud platforms, improving the distributed parallel computing capability of the original analysis and mining and the analysis capabilities of the supporting platform. . Deep integration of R analysis software with Hadoop, and integration of traditional mining algorithms with existing algorithms based on Hadoop. For dynamic graph data, a memory-based distributed data management system can support low-latency query processing. For the data stream, it adopts a method oriented to the sliding window model, and performs a single approximation and direct processing directly through the probabilistic dimension index. Hadoop-based Apache Mahout converts classic algorithms into MapReduce mode to improve algorithm throughput and performance, and supports semi-structured or unstructured data such as music, video, etc., and collaborative filtering and content analysis in an automated interactive manner. At the same time, in addition to the traditional query language SQL, big data mining has corresponding query languages, such as HiveQL, Pig Latin and other dedicated APIs, which have flexible scalability, but the query performance is low and the resource utilization rate is not high.

5. Conclusion

The emergence of big data has brought about a rich and diverse range of potential value resources, both in terms of traditional data management methods and scientific thinking methods. In the face of massive, complex and uncertain dynamic data, traditional data processing methods are faced with severe challenges in terms of computing power and storage capacity. Their scalability and flexibility cannot meet the real-time processing requirements of big data. Cloud computing provides powerful computing and storage power for big data processing, while big data mining provides an opportunity for deep integration of big data and cloud computing.

References

[1] PHRIDVIRAJ M S B, GURURAO C V. Data mining—past, present and future—a typical survey on data streams [J]. Procedia Technology, 2014 (12): 255-263.

[2] Zhao Youlin, Deng Zhonghua, Lu Yingxi, et al. Data mining cloud service analysis research [J]. Intelligence Theory and Practice, 2012, 35 (9): 33-36, 44.

[3] JI Xiaokang, MA Xiuli, HUANG Ting, TANG Shiwei. Continuously extracting high-quality representative set from massive data streams [J]. Lecture Notes in Computer Science, 2013, 8346 (1): 84-96.

[4] TODD D P, YUNG R C, YOSHIMURA A. Using performancemeasurements to improve MapReduce algorithms [J] . Procedia Computer Science, 2012 (9) : 1920-1929.

[5] Liu Peng, Wu Zhaofeng, Hu Guyu, et al. Big data—a profound change that is taking place [J]. ZTE Technology, 2013, 19 (4): 1-7.