Research on the Technology Realization of Network Mass Information Processing Platform

Qiang Zhao¹,*, Sha Liu¹b*

¹Dean's Office, Wuhan Donghu University, Wuhan City, Hubei Province, 430212, China

¹*College of economics, Wuhan Donghu University, Wuhan City, Hubei Province, 430212, China

E-mail: zhaqiang1486wdy@wdu.edu.cn

bE-mail: liusha@wdu.edu.cn

Abstract: In recent years, With the rapid spread and development of the Internet, especially the huge reserve of information and convenient search function, it attracts more and more of the community to the Internet users ranks. But these need supports of high-performance computer technology. Therefore, in order to enhance the capabilities of Internet vast amounts of information intelligent parallel processing, and promote the comprehensive development of the overall system of information storage, management, integration and intelligent processing, data mining, information retrieval and utilization functions, I proposed a massive information processing platform based on high-performance data mining. The one hand, it analyzes the key technologies to achieve the platform; the other hand, it gives a brief introduction to its hotspot applications. Therefore, it provides a reference and guidance with real-world applications for the technology of high performance and highly efficient network of mass information processing platform.

1. Introduction

In the current era of rapid information expansion, people are producing, disseminating, searching, and applying various information content at an unprecedented speed. Especially in the past two decades, the total amount of information generated by the entire society has far exceeded the total amount of information generated in one million years since the birth of mankind. From the perspective of Internet media, according to a report released by the China Internet Network Information Center in 2009: By the end of 2008, compared with 2009, the total number of web pages in my country has increased by 90%, reaching 16 billion, of which the number of bytes It also exceeds 460TB. With the current acceleration of social informatization and the continuous improvement of the level of network technology, this number is showing an increasingly accelerating development trend. In this context, the technical realization of the network's massive information processing platform has become a core issue that many network workers or technicians think about. In fact, before systematically studying this issue, it is necessary for us to briefly summarize the information characteristics of Internet media.
2. Overview of Information Features in Internet Space

2.1 The amount of information is large, the type is large, and the growth rate is fast
At present, the scale and types of information in cyberspace are expanding rapidly. As of July 2008, the number of pages indexed by Google's website has exceeded one trillion, and it is still growing.

2.2 Information generation is fast and easy to change
As we all know, data information in cyberspace is mainly dependent on interactive, real-time, and integrated high-speed network systems and computer hardware equipment. It is formed with obvious instantaneous characteristics, and information can be easily realized by only relying on the mouse. Copy and change. Since information on the Internet is stored in the form of data, it is easy to change.

2.3 The information collection is messy and difficult to find
Because the information capacity in the Internet space is extremely large and diverse, the current information in the Internet space lacks effective and reasonable organization, and its disorderly storage characteristics greatly reduce the efficiency of people's retrieval, acquisition, and use of information. Therefore, this is also one of the core issues that need to be resolved in the current establishment of a massive information processing platform on the Internet.

2.4 Network information varies from good to bad
At present, my country's online media presents a decentralized and liberalized development model, with low entry barriers, especially the lack of effective supervision, which makes relevant departments unable to effectively control the release and dissemination of information. For example, the Internet is full of repetitive and false information, and even violent, reactionary, and pornographic information, which greatly interferes with the normal use of information by netizens.

2.5 Network information has obvious effect of public opinion
With the rapid popularization and development of the Internet, especially the improvement of various application functions such as instant chat, BBS forums, and Weibo, people are becoming more and more accustomed to publishing their daily trivia, what they have seen, and what they say in cyberspace. Views and opinions on specific issues. This leads to the obvious effect of public opinion on network information. Especially in various emergencies, massive amounts of network information will often bring huge public opinion pressure to government departments, thereby promoting the open, timely and transparent resolution of related issues.

It can be seen from this that the special information characteristics of network media in information storage, management, data mining, real-time processing, intelligent processing, audio and video data, network text and other technical issues, put forward a higher level of technical realization of the network massive information processing platform. Challenge.

3. The basic architecture of the network's massive information processing platform
The information processing platform designed by the author is mainly based on high-performance computer technology. It includes unified view middleware, distributed parallel database, parallel data mining service, parallel computing environment, cluster high-speed interconnection and so on. This has thus formed a well-functioning information processing system that can deeply analyze and mine various data and information, and improve the overall information processing, information retrieval, public opinion analysis, trend research and judgment capabilities in cyberspace. As shown in Figure 1, the overall architecture of the platform mainly includes data acquisition, storage and organization, business analysis layer, data integration layer, user interface layer and other parts.
In this system, the data mining module is the core part of the entire massive information processing platform. Based on the open mining algorithm library, it provides a comprehensive analysis, processing, retrieval and effective application platform for network mass information including business analysis modes such as hotspots, traceability, statistics, linkage analysis, behavior analysis and mining. Specifically, the data acquisition part is mainly to access the platform through the Internet data gateway, and enter different types of data information into a specific data organization after a series of data cleaning and format unification processing. This ensures the validity and robustness of these data information in the subsequent processing. The function of data organization is mainly to conduct online processing activities on data streams. It includes information recognition, text extraction, fast scanning, feature extraction, data filtering, file deduplication, information classification and many other functions. Feature extraction mainly forms document feature data, and text extraction mainly forms derivative text data information. Data storage uses a distributed parallel database to process three levels of granularity in parallel to improve the data processing capability of the system. Based on distributed middleware, parallel processing activities between multiple parallel data can be realized, and multiple nodes and multithreading on nodes can be processed in parallel within the parallel database. The data information formed after specific data organization uses unified view middleware for data storage and management. Finally, within the system, the unified view middleware enables upper-layer applications to transparently and directly access the underlying distributed parallel database through data dictionary and policy management services. In addition, through a specific unified view middleware, the data integration layer extracts the original data information from the distributed parallel database, and integrates and stores the data information in a subject-oriented manner. This improves the efficiency and performance of data mining and data analysis of the entire system. Specifically, the data integration layer mainly includes various functions such as data cleaning and loading, and integrated data modeling with analysis as the main goal. Among them, the integrated data modeling with analysis as the main goal is based on different mining applications, through a topic-oriented multi-dimensional data model to integrate data information. The model can organize massive data information from multiple angles and levels, and provide views of materialized effects with different granularities. This realizes real-time query of data information from macro to micro, in order to ensure different granularities and a full range of data mining and data analysis.

In addition, the main function of the business analysis layer is to conduct parallel data mining, which includes two parts: specific data mining business and open data mining algorithm library. Among them, the specific data mining business mainly includes hot spot analysis, linkage analysis, traceability analysis, behavior analysis, statistical analysis, behavior mining, etc. The open data mining algorithm library includes clustering, classification, association, text mining, sequence patterns and so on. The user interface layer is to provide network information searchers and users with special automatic background tasks, mining task guides, custom tasks, mining analysis visualization, user
screening and evaluation, and many other technical services.

4. Analysis of the key technologies of the network massive information processing platform

As shown in Table 1, the article builds a network massive information processing platform based on high-performance data mining. This requires the realization of the following key technologies:

| Field | Platform | Key technology |
|-------|----------|----------------|
| Storage and management of a small amount of network data | Data storage layer | High-speed interconnection of parallel databases based on InfiniBand RDS protocol |
| Real-time processing activities of network data information | Data organization, data storage layer, data integration layer | Distributed Parallel Database Unified View Middleware |
| Network data mining | Business analysis layer | Parallel data mining based on database |
| Data retrieval of network audio, video and text | Data integration layer, business analysis layer | Web Text Data Mining |

4.1 High-speed interconnection of parallel databases based on InfiniBand RDS protocol

The system designed by the author is mainly to store and manage massive amounts of information on the network through a distributed parallel database. Each parallel database contains multiple database server nodes to share storage to form a cluster database system. Each huge distributed parallel database system is composed of multiple parallel databases. Based on this, it can smoothly realize the rapid storage and management of hundreds of terabytes of massive data information in cyberspace. Therefore, the distributed parallel database is also based on the parallel database. From the perspective of technical implementation, it affects or even determines the scalability and parallel processing efficiency of parallel databases, or the key to parallel processing performance is the node interconnection bandwidth, delay, and processor overhead. At present, with the continuous expansion of computer database scale, open interconnection technology. This, such as Gigabit Ethernet, is far from being able to meet the current needs of computer parallel database node interconnection. We must reconstruct a computer parallel database system with a large number of nodes to improve the information storage and management capabilities and efficiency of the entire platform.

As we all know, InfiniBand is mainly an open and advanced interconnection standard defined by the InfiniBand industry association. This is a channel-based I/O system that uses a switching structure. Reliable Datagram Sockets Protocol, namely Reliable Datagram Sockets is the upper layer protocol of InfiniBand. It has the characteristics of low latency, low load, and high bandwidth. It provides a more reliable datagram service in the IB network to support the application of the UDP protocol.

Parallel database IB, RDS network environment is mainly composed of four parts: host channel adapter, IB switch, database application support software, and subnet management. RDS greatly improves the scalability of parallel databases and the performance of applications. Compared with IPOIB, its CPU occupancy rate is reduced by about 50%, and compared with UDP protocol, its delay is also reduced by half. The advantages of RDS over Gigabit Ethernet mainly lie in its ease of use, low latency and low processor occupancy, high bandwidth and high availability, reliable packet transmission without discarding or retransmission, and so on. Based on the InfiniBand RDS protocol, a high-speed interconnected eight-node RAC parallel database experimental environment constructed by the author. It uses the TPC-H benchmark benchmark test to compare the performance between the Oracle RAC database IB RDS protocol and the Gigabit Ethernet interconnection. The total running time of the three TPC-H typical queries, the latter increased by about 33% over the former.
4.2 Distributed parallel database unified view middleware
Through the unified view middleware, the distributed parallel database can integrate many parallel databases into a relatively large distributed parallel database system. The unified view middleware plays an extremely important role in the entire system. Generally speaking, unified view middleware systems and distributed parallel databases mainly include client APIs, unified view middleware services, statistical backup modules, system security, policy management services, database access, and so on. In this database system, the unified view middleware can prompt upper-layer applications to transparently access the underlying distributed parallel database. This provides an efficient development interface for the network's massive information processing platform to deal with upper-layer applications. In addition, the software uses many important technologies such as parallel query optimization, SQL syntax analysis, multi-level load balancing, and fault tolerance. This greatly guarantees the reliability, availability, high speed, and concurrent performance of the entire system.

Parallel loading and parallel query services are the core functions of the unified view middleware. It accomplishes these functions consistently through the data dictionary. Through actual practice, it can be found that its overall performance can meet the needs of accurate and real-time processing of massive information on the network.

4.3 Parallel data mining based on database
Data mining is to extract or mine valuable information content that can meet the needs of users from a large amount of network data information. This is also one of the core technologies of the network's massive information processing platform.

Parallel data mining based on database is mainly to build an open mining algorithm library at the bottom. It includes clustering, classification, association, text mining, sequence patterns, anomaly recognition, important attributes, feature extraction, etc. Its advantage lies in facilitating network information management and data preparation, automatic parallel mining and calculation, reducing data movement, making full use of SQL functional services, improving data security, and so on. The network information processing platform developed by the upper layer includes a variety of business analysis methods such as hot spot analysis, statistical analysis, and linkage analysis. Its advantage lies in supporting parallel computing and multiple parallel database mining, supporting Windows/Linux platform, high-speed and concurrency performance and so on.

4.4 Web text data mining
Network text is the most common way of presenting network information. Its data mining is an extremely important basic function of the network's massive information processing platform. It mainly includes two important contents: storage and management of text and safe retrieval. Generally speaking, the former is based on the file system, which organizes and stores full-text text through a three-tier model structure of logical storage, physical storage, and user view. Among them, in physical storage, the entire system supports multi-level and multi-directory storage based on categories such as region, time, and subject. Logical storage is mainly to store the index information of the text library, which uses multi-level file indexing technology. Based on the file management provided by the computer operating system, it uses various data structures to store the physical storage locations of the files in a hierarchical manner. The user view takes logical storage as the core, and manages different text information through convenient and efficient text indexing technology.

Generally speaking, full-text search is a typical distributed text search tool. Its core functions include obtaining network text information according to user needs and performing text analysis and extracting the text to remove duplicate information. It divides the text into words, establishes a partial index, and finally forms an index library. It sends query requests to other nodes or queries the index library, etc. Practice has proved that the information processing platform can perform intelligent parallel processing of information and in-depth analysis and mining of characteristic information. It has extremely high scalability, and to a certain extent can meet the needs of the rapid growth of Internet information.
References
[1] HeFangyuan, The basic research about the massive data processing technology, Silicon Valley, 2009 (8), P59-60
[2] QiLixin, Content-oriented network to massive information processing platforms and systems research, China's media technology, 2006 (9), P68-72
[3] LiXiaoyu, The search about the using of Multi-Agent in network vast amounts of information, Science and technology information, 2009 (14), P24-27
[4] WangLiangying, The search about the technology of massive information resource storage and sharing, Information systems engineering, 2011 (11), P31-34
[5] LiaoYi, LiuDong, Asynchronous called to the SCADA system of massive information processing method, Automation of Electric Power Systems, 2006 (20), P24-27