Research on the Mass structured Data Storage and Sorting Algorithm and Methodology for SQL Database under the Big Data Environment

Rong Wang[1]
[1] Hubei University of Science and Technology,
Xianning, Hubei Province,437100 China

Chunhui Wu[1, *]
[1] Hubei University of Science and Technology,
Xianning, Hubei Province,437100 China
*Corresponding Author：Chunhui Wu

Wenhua Dai[1]
[1] Hubei University of Science and Technology,
Xianning, Hubei Province,437100 China

Abstract. In this paper, we research on the research on the mass structured data storage and sorting algorithm and methodology for SQL database under the big data environment. With the data storage market development and centering on the server, the data will store model to data-centric data storage model. Storage is considered from the start, just keep a series of data, for the management system and storage device rarely consider the intrinsic value of the stored data. The prosperity of the Internet has changed the world data storage, and with the emergence of many new applications. Theoretically, the proposed algorithm has the ability of dealing with massive data and numerically, the algorithm could enhance the processing accuracy and speed which will be meaningful.

Keywords: SQL Database; Big Data Environment; Data Storage and Sorting; Mass Structure.

Introduction

With the improvement of today's social information degree, all walks of life are faced with massive data, these data are usually in the hundreds of GB or even tens of terabytes of level, and is growing at the speed of fast, and therefore these so-called huge amounts of data. Huge amounts of data storage and management is a hot issue nowadays. Traditional storage network architecture for the center with server in the face of a steady stream of data flow is ragged. People hope to find a new kind of data store model, independent of storage devices, and has a good scalability, availability, reliability, in order to meet the requirements of data storage in the future. With the data storage market development and centering on the server, the data will store model to data-centric data storage model. SAN is based on data storage center, by using scalable network topology structure, through a direct connection with high transmission rate of optical channel, provide the SAN any node within the multiple choice of data exchange between, and the data storage management focus on the relatively independent storage within the local area network. In a variety of optical channel transmission protocol gradually towards standardization and cross-platform cluster file system put into use after SAN will ultimately achieve in a variety of operating systems, the maximum data sharing and data optimization management, as well as the seamless expansion system.

The database under the operating system is managed in the form of files, the intruder can directly make use of loopholes of the operating system to steal a database file, or tamper with the database file content. On the other hand, the database administrator can access any data, often beyond the scope of their duties, also cause potential safety hazard. Therefore, the database of confidentiality issues include not only used in the process of transmission encryption protection and control illegal access, include protection for storage of sensitive data is encrypted, making even data leak or lost unfortunately, also hard to cause leaks. At the same time, the database encryption can be used by the user's own key encrypt your sensitive information, and don't need to understand the data content can't be normal, so as to ensure the safety of the users' privacy information. For database encryption will inevitably bring data storage and index, key distribution and management, data query and a series of problems. At the same time, the encryption can also significantly reduce the database access and operation efficiency. Inevitably there are conflicts between confidentiality and availability, the need to properly solve the contradiction between the two.

In the process of the development of the decision tree algorithm in data mining is gradually mature and perfect, but in the face of a large number of data sets, how to make use of its fast and accurately found hidden in the main classification rules, is still a problem worth studying. Due to the decision tree algorithm is used for machine learning algorithm, the traditional machine learning algorithms are usually based on main memory, a comparatively small amount of data processing, when faced with a large amount of data in the database for digging, shown the outstanding problem of the efficiency of the algorithm is a significant reduction in or even unable to run; Algorithm usually need the data provided in the form of data files, this is not only the need for more data before data mining conversion work, and not conducive to algorithm and database applications to achieve seamless integration, which to a large extent affected the data mining algorithm is practical. In the face of large amounts of data, did not make full use of database technology has advantages of data in the database operation. To construct the decision tree method is to use the recursive structure of top-down. Construction idea is, if the training sample set all of the examples is the same, will be as a leaf node, the node content is the category tag. Otherwise, according to certain strategy to select an attribute, according to the attribute values of the example set is divided into several subsets which make each subset of all the examples in the attributes have the same attribute values.

In this paper, we conduct research on the research on the mass structured data storage and sorting algorithm and methodology for SQL database under the big data environment. Storage is considered from the start, just keep a series of data, for the management system and storage device rarely consider the intrinsic value of the stored data. The prosperity of the Internet has changed the world data storage, and with the emergence of many new applications, information value rising, all of these requirements can be faster longer storage more data. IT is estimated that data storage to double every year, as the information carrier of the storage system is increasingly becoming the core of enterprise IT architecture and its performance, manageability etc. Will important influence on enterprise operation directly. With the rapid development of storage technology, storage the falling cost of hardware, but the rising of storage management costs. Therefore, a good storage management solution to reduce the total cost of ownership of the storage network is of great significance. So in terms of data storage, storage management software status is rising, many of the functions of storage products are replaced the hardware with the implementation of the software system, therefore, how to better use of the technology to achieve more efficient storage management is a problem in front of us. In the following figure one, we show the pattern of the data storage and sorting for databases.

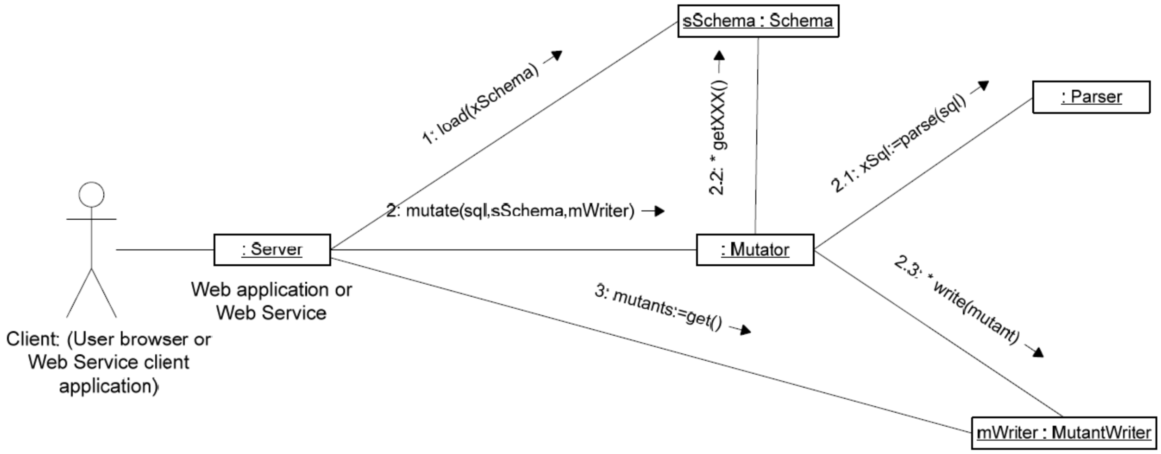

Fig. 1The Pattern of the Data Storage and Sorting for Databases

Our Proposed Methodology

The Big Data Environment for Database. Task of data mining is the discovery of patterns which can be found hidden in the data model generally fall into two categories: descriptive model and the model of forecasting model to describe is the fact that exist in the current data do specification description, depict common features of the current data; Prediction model in time for the key parameters, for the time series data, according to its history and the current value to predict its future. Statistical database is a special type of database, it is compared with the general database have a lot in common, but there are also many uniqueness. And general database, in statistical database can store a lot of information, including confidential information. In a typical database, as long as not in violation of the security requirements of the database, the user can get a record of the information by questioning. Clustering is a data items grouped into multiple classes, or clusters, data differences between classes should be as large as possible, the difference of data within a class should be as small as possible, to minimize the similarity between classes, maximize the patterns of similarity principle and classification of class, clustering is to divide the category of the unknown, it is a kind of doesn't depend on the predefined classes and class label with unsupervised learning of the training data set, without the background knowledge, including class by system according to certain performance index automatically determine the number of the classes.

Model of regression function definition and classification of similar, the main difference between discrete forecast classification model, and the regression model using continuous predictive value in this view, the classification and regression are forecasting problem. Another problems of data perturbation method is based on the disturbance after data query response may be different from the original data based query response, such as the disturbance caused by the deviation. In other words, the disturbance will change the part of the original database or more statistical characteristic attributes. The relationship between the attribute is often many institutions to take advantage of a lot of important decisions. Decisions based on the error deviation information usually can cause very serious consequences. Now many of the perturbation method is generally does not make the database all attributes of the relationship between the keep the same level before, during, and after disturbance. The following figure shows the database structure under the big data environment.

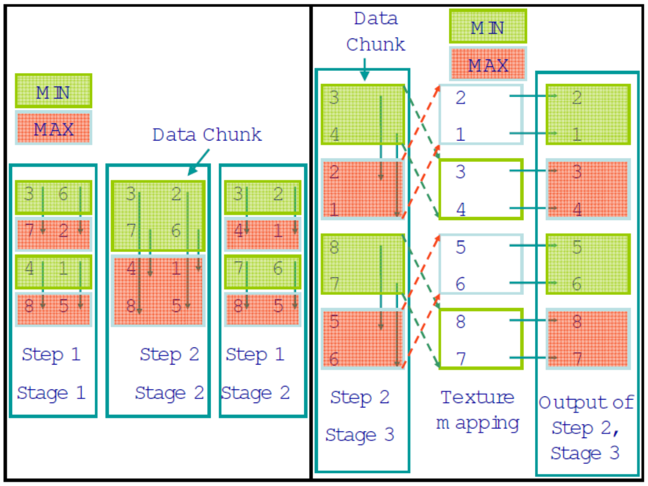

Fig. 2The Database Structure under the Big Data Environment

The Mass structured Data Storage and Sorting. Query optimization has been the most programmers' research question, because this is the most time-consuming work to the operation of the database. Although different query plan can be a SQL statement execution a speed difference is not big, but for visits to the larger web system, may this statement is performed thousands of times a day, accumulated time difference is big. In addition, from the perspective of cost optimization, each SQL statement should be separate optimization and the programmers to the optimization of SQL statements can often make the query efficiency exponentially. Moreover, the fewer the number of database query, read the database storage disk number the less, can prolong the service life of the server. Before we clear the optimization goal, we should first solve why the SQL query optimization problem, in the actual cases, often have some defects such as database design, a sharp increase of user visits, and increase the system function modules of the impact of the database structure and function demand. For a web system, there are two concepts must be attention: throughput and response time of user requests. This is a pair of contradictions two aspects, and in the same system, we hope that the throughput and response time of user requests can reach the best.

The database management systems to complete the query operations are connecting the query optimization strategy based on cost considerations. The first to use brute force method is a list of all possible join order query, again from the selection of the lowest cost estimate strategy. Due to the number of query execution strategies along with the augment of parameters and the connection number according to the law of index increased, therefore, even if adopted heuristic algorithms perform a search, choose the best performers of the system overhead also will rapid growth. Obvious query operations had had a serious impact on the system efficiency.

With the expansion and application of database and data is increasing, the performance of the database query problem more and more serious and these factors are common in the point that programmers use against the optimizer processing of the query. For the same query, the same query can be expressed through different ways to implement, such as using the subquery is changed to use the connection operation, and different expressions for database response speed brings the serious influence. SQL operations are commonly associated with the index entry of system object and the faster the speed of the system to access the page, the performance of the SQL statement is better. The purpose of query optimization is to use the smallest query cost the same query results.

The Performance of the Combined Methodology. With the help of database management system, directly to the storage of fault data samples of data query database efficiently. First properties together by calculation of each condition attribute information entropy drop size and get attributes of the split decision tree selection order. Entropy drop and the greater the description attribute of information gain, the greater the classification for it is the key. The attributes sorted according to the entropy drop size, by querying the database recursive have to build a decision tree for classification rules and the rules in the rule base, when meet the termination conditions exit the program. The last will get the classification rules library rules stored in database, which guides the new classification of sample classification. Algorithm to realize the whole process of all can with the help of the DBMS based on the existing database query is complete, the middle need to save data as the text processing, greatly improving the efficiency and feasibility of the algorithm, and put forward the effective use of the properties of the selection method gradually constructing classification rules tables, can make the algorithm grow along with the number of tuples of good scalability.

Conclusion

In this paper, we research on the research on the mass structured data storage and sorting algorithm and methodology for SQL database under the big data environment. Classification is on an event or a set of objects, through the training data set of mining classification model can be obtained, using the classification model to analyze the existing data, you can also use classification model to predict the future. Decision tree is a kind of typical classification algorithm, what can get similar under what conditions will be the result of the rules. Query optimization has been the most programmers' research question, because this is the most time-consuming work to the operation of the database. Our method combines the general features of the mass data and the corresponding algorithms. We will test the effectiveness and the accuracy of the proposed method.

Acknowledgement

（1）National Natural Science Foundation Project，61373108；
（2）Natural Science Foundation of Hubei Province,2015CFC778；
（3）Science and Technology Research Project of Department of Education of Hubei Province, 20082803；
（4）Humanities and Social Sciences Research Project of Department of Education of Hubei Province，2012D125。

References

[1] Zhang, Zhou F, Xu Z, et al. Distributed Storage and Processing Method for Big Data Sensing Information of Machine Operation Condition[J]. Journal of Software, 2014, 9(10).

[2] Chen Z K, Yang S Q, Tan S, et al. The Data Allocation Strategy Based on Load in NoSQL Database[J]. Applied Science Materials Science & Information Technologies in Industry, 2014.

[3]Binnig C, Salama A, Zamanian E, et al. XDB: a novel database architecture for data analytics as a service[C]// Big Data (BigData Congress), 2014 IEEE International Congress on. IEEE, 2014.

[4] Tambe P. Big data investment, skills, and firm value[J]. Management Science Journal of the Institute for Operations Research & the Management Sciences, 2014, 60(6):pags. 1452-1469.