2014

ISSN : 0974 - 7435 Volume 10 Issue 17

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(17), 2014 [9562-9566]

# Research on data mining clustering algorithm in cloud computing environments

Xi Liu, Xiumei Zhang, Xiyong Li, Zhengguang Sun*
Department of Computer Science, Ping Xiang University, Pingxiang, (CHINA)

## ABSTRACT

Cloud computing is a model of business computing and it distribute computing tasks in a resource pool which constitutes by a large computers, so it can provide users with on-demand computing power, storage capacity and application service capabilities. The cloud computing provides cheap and efficient solutions for massive data storage and analysis. Data mining is finding potentially useful information and knowledge people do not know in advance from a large number of incomplete, noisy, fuzzy, random practical application data. And it played a guiding role in many areas of scientific research and business decisions, with far-reaching social and economic significance. The research on data mining clustering algorithm in cloud computing environments has an important theoretical significance and application value.

## KEYWORDS

Cloud computing; Data mining; Clustering algorithm.

© Trade Science Inc.

## INTRODUCTION

Today, due to the development of computer technology and storage technology and database technology, large amounts of data have been collected into a database computer. We are in a huge amount of data and rich, but knowledge is acquired from a considerable lack of data in the database information face hidden wealth of knowledge but cannot fully excavated and the use of age. In this regard, we urgently need a powerful data analysis techniques and tools that can be mass data analysis and processing, get one hidden role of information and knowledge, to provide an effective basis for decision support in all areas of society. Therefore, the data mining techniques have emerged, data mining plays an increasingly important role in the industrial and commercial fields.

Data mining is to find the data useful information hidden in the data, and provide support for decision-makers to make decisions, there are broad prospects for development. With the development of computer calculation system, a stand-alone treatment process to the cluster, and then use the Internet to form a supercomputer, making data mining processing capacity greatly. Data mining technology combines artificial intelligence, machine learning, pattern recognition many disciplines, statistical, database, visualization techniques, revealing data from a large number of implicit, previously unknown and potentially valuable information. Data mining as the world's leading information technology, it has attracted wide attention and research applications academia and industry.

Cloud computing is the full use of existing network resources and equipment, centralized network computing capacity, distributed parallel computing, to be combined with shared resources and makes the system security is guaranteed, greatly reducing the time and computing cost-saving resources to carry out large and complex task issues distributed parallel computing, systems integration management and self-maintenance and low cost me in one. In the face of very large scale data TB level or PB-level data mining, the use of parallel computing technology, cloud computing, will greatly reduce the time data processing and more efficient digging out useful information.

Cluster analysis is an important data mining, data analysis techniques, the collection of grouping physical or abstract objects become analytical process multiple classes by similar objects. It is an important human behavior. Target Cluster analysis is similar to the data collected on the basis of classification. Cluster analysis in business, geographic information, Internet applications, e-commerce, and so many fields have been widely used[1].

### The concept of cloud computing

Cloud computing is an emerging model of business computing. It distributed computing tasks in a large pool of computer resources constitute enable various application systems to obtain computing power needed storage space and a variety of software services. Definition of cloud computing has narrow and broad points.

Cloud computing refers to the delivery of narrow and use of IT infrastructure mode refers to the demand, and scalable way to get resources (hardware, platform, software) required by the network. "Cloud" of resources in the user appears to be infinitely scalable, and can be readily available, on-demand, any time extension, pay per use. This feature is often referred to as the use of water and electricity as the use of IT infrastructure. Cloud computing broadly refers to the delivery of services and usage patterns, refers to the demand, and scalable way to obtain the necessary services through a network. This service can be IT and software, Internet-related, it can be any other service.

This resource pool called "cloud." "Cloud" is a virtual computing resources that can self-maintenance and management, usually for a number of large server clusters, including computing servers, storage servers, broadband resources and so on. Cloud computing resources all together, automatically managed by the software, without human involvement. This allows application providers do not need to worry about tedious details, can be more focused on their business, is conducive to innovation and reduce costs.

Cloud computing is parallel computing, distributed computing and the development of grid computing, or that these commercial implementations of computer science concepts.

### The features of cloud computing

(1)It has a very large scale. "Cloud" of considerable size, Google cloud computing already has more than 100 million servers, Amazon, IBM, Microsoft, Yahoo and other "cloud" all have hundreds of thousands of servers. Enterprise private cloud typically have hundreds of thousands of servers. "Cloud" can give users unprecedented computing power.

(2)The virtualization. The cloud computing allows users at any location, using a variety of terminal acquisition applications. Resources requested from the "cloud", rather than a fixed tangible entity. Applications running somewhere in the "cloud", but in fact you do not need to know, do not worry about the specific location of the application to run. Only need a laptop or a cell phone, it can be achieved through the network service everything we need, even including such tasks supercomputing.

(3)High reliability. "Cloud" using multiple copies of data fault tolerance, isomorphic interchangeable compute nodes and other measures to protect the service and high reliability, the use of cloud computing and reliable than using the local computer.

(4)The versatility. Cloud computing is not for a specific application, in the "cloud" can be constructed under the support of the ever-changing applications, with a "cloud" can support different applications running simultaneously.

(5)The high scalability. "Cloud" size can be dynamically scalable to meet the needs of applications and user scale growth.

(6)The on-demand service, "cloud" is a huge pool of resources available on demand; cloud can be as billing, like running water, electricity, gas.
(7)It is extremely cheap due to the "cloud" special fault tolerance measures can be used to form an extremely inexpensive node cloud, "cloud" automated centralized management makes a lot of business without the burden of an increasingly costly data center management book, "cloud" Universal makes utilization of resources than traditional systems dramatically, so users can fully enjoy the "cloud" of low-cost advantage, often as long as a few hundred dollars to spend a few days time to complete the previously required thousands of dollars, a few months time to complete task.

## The definition of data mining

Data mining is a lot of, incomplete, noisy, fuzzy and random data extracted from implicit in them, people do not know in advance, but is potentially useful information and knowledge. With the rapid development of information technology, the amount of data accumulated in the rapid growth of people, hundreds of dollars in TB, how to extract useful knowledge from vast amounts of data has become a problem that must be solved. Data mining is to adapt to this need emerged and rapidly developed data processing techniques.

Data mining is a key step in knowledge discovery. It is the use of specific algorithms to extract patterns and knowledge from the data. Such knowledge or information is implicit, previously unknown and potentially useful knowledge extraction performance concept, rules, laws, patterns and other forms. Data mining is a set of technologies and applications, or a method for large-capacity data and data relationships between study and modeling of collections. Its goal is to large volumes of data into useful information and knowledge. Its structural data mining objects from the source to the semi-structured and non-structured data sources, including relational databases, object-oriented databases, relational databases spatial reasoning databases, multimedia databases, temporal databases, text databases, image databases, and audio and video data sources. A data mining algorithm usually consists of the following elements: model, priority criteria and search algorithms.

## The functions of data mining

Data mining is used to specify data mining tasks looking mode type. In general, data mining tasks can be divided into two categories: description and prediction. Descriptive data mining tasks describe the general nature of the data in the database. The task of predictive data mining is to make predictions current inference.

(1)The description of concept/class: characterization and distinction. Class/concept description refers to the summary, concise and accurate way to describe the various categories and concepts. This description may be obtained by the data characteristics and data distinguished.

Data characterization is a summary of the general features or characteristics of the target class data. Typically, a user-specified class data collected through a database query. The effective methods of data collection and features include: simple data based on statistical summary metrics and graphs, based on the volume of data cube OLAP operations and attribute-oriented induction technology.

General characteristics of the data distinguish the target class is a data object with one or more of the general characteristics of the type of object contrast are compared. Target class and contrasting class specified by the user, and the corresponding data query through a database search. A method for the data is similar to the method for distinguishing characteristics of the data.

(2)The correlation analysis. The purpose of association analysis is to outline some of the data generated, for example, to find relationships derived relationship between a subset of data or some data with other data. The most common technique is the use of association rules. Computing association rules depends on identifying the relevant data appear frequently in data sets. Given by the user minimum support, find all frequent item sets that meet the support of not less than the minimum support all projects subset. In fact, these frequent item sets may have contained relationship. Generally, only care about the so-called largest collection of frequent item sets are not included in other frequent item sets. Find all frequent item sets is the basis for the formation of association rules.

(3)Classification and prediction. The concept of classification is to find a category description, information that represents the entirety of such data, i.e., the connotation of the class described in this description and the structural model represented by the general rules or decision tree model. Classification is the use of the training data set is obtained by a certain algorithm and classification rules. Classification rules can be used to describe and predict.

Prediction is the use of historical data records automatically deduce the promotion given an overview of the data, and thus to predict future data. Typically use mathematical and statistical methods to identify property and related properties to be predicted, and the property value estimate based on an analysis similar to the distribution of data

(4)Cluster analysis. Cluster analysis is based on its feature clustering or classification of things, the so-called feather flock together, and found the law and typical patterns. Through subsequent clustering, data sets will be converted to class set, the same kind of data with similar values of variables and variables of different types of data values do not have a similarity.

Clustering and classification and prediction of different classification and prediction is for training data, however, clustering is not known in advance how much the target database contains the class situation, all of the records sought to merge different classes.

(5)Outlier analysis. Database may contain some data objects, the general behavior or model they are inconsistent with the data. These data objects are outliers. Most data mining methods will be considered outliers or unusual noise and discarded. However, in some applications, the rare thing may be more interesting than the normal events occur. Outlier mining data analysis called outliers. You can assume a data distribution or probability model, using statistical tests to detect outliers; or using distance metric, the distance to any cluster of objects as outliers. Based on the difference between the deviation by a method the main characteristics of the study group of the object to identify outliers, or instead of using a statistical distance measure.

(6)Evolution analysis. Data evolution analysis is the law or trends describe the behavior of objects change over time, and its modeling. This analysis includes the time-related data in addition to characterize, differentiate, association, classification or clustering, including time-series data analysis, sequence or cycle pattern matching and analyze data based on similarity[2].

## The research status of data mining clustering algorithm in cloud computing environment

After ten years of efforts of a generation, and now the data mining technology research has made remarkable excellent results. For KDD research mainly revolves around the theory, technology and applications in three aspects. Most researchers use effective techniques is to integrate a variety of theories and methods in order to achieve better purpose. Currently, the latest developments in data mining study abroad mainly in the knowledge discovery process further exploration and research. In the application of the algorithm is mainly reflected in the development of commercial data mining software tools to solve problems from a single isolated problem-solving process for the establishment of steering the overall system, its main client software for large banks and insurance companies, and so the sales industry. United States as the world's most prosperous data mining technology research areas, occupies a central position in its research and exploration.

Compared with foreign and domestic research on data mining has many shortcomings, the late start and the development of immature, is currently in development started to normalize stage. The latest developments include: integration of rough sets and fuzzy set theory applied to knowledge discovery process integration; theoretical model of Chinese text mining and implementation techniques; using the concept of text mining; trying to build a collection of theoretical system, to achieve massive data processing data classification; structure construct intelligent expert systems; fuzzy system identification method and fuzzy system knowledge model.

## The problems of data mining clustering algorithm in cloud computing environment

(1) Scalability is not strong. Many clustering algorithms work on data collection in hundreds of data objects which works well, but the practical application of data mining projects are usually a few more examples of millions of objects for analysis, and now rarely suitable for handling large clustering algorithm data collection, and can only handle numerical data, the class attribute data often appear in the data mining analysis cannot be achieved[3].

(2)It lacks of ability to handle different types of properties. Many types of applications may have a lot of data, such as numeric, binary type, property type, etc. However, many clustering algorithms designed only adapt to the numeric type, so the treatment is not effective for most applications. Even some of the existing clustering algorithms can handle these different types of data analysis, but cannot handle large data sets.

(3)It needs better prior knowledge for decision of input parameters. Require the user to enter specific parameters, such as the hard k-means algorithm and fuzzy k-means algorithm are required to enter the desired number of clusters k clusters before most clustering algorithms during operation. Moreover, these input parameters in practice are often difficult to determine. Further, in general the results of cluster analysis for the input parameter are very sensitive. This requires the user to input parameters to determine a priori way to give users a certain amount of work and the burden, while it is not the kind of algorithm for unsupervised learning the true sense.

(4)It cannot identify clusters of arbitrary shape. Clustering algorithms will typically use the Euclidean distance or Manhattan distance to measure the similarity of data, based on the distance metric algorithms tend to have similar structures found in globular clusters scale and density. However, in practical applications, a cluster may be any shape, so that a good clustering algorithm must be able to effectively and accurately identify clusters of arbitrary shape.

(5)The ability to handle noisy data is weak. Most of the data are included in reality there are isolated points and noise. If the algorithm for such data-intensive, it may result in reduced quality of the clustering results. Therefore, the clustering algorithm must be able to remove or filter noise and discrete values.

(6)Lack of clustering validity studies for the class attribute data.

For cluster analysis, the validity can often translate into optimal number of categories k decisions. And before the relevant clustering validity of research, mostly focused on the analysis of the effective type of data, data mining for common generic type of data, there is no effective way clustering validity analysis.

(7)For large data distributed file system data mining support is inadequate.

In recent years, "big data" concept was born, distributed systems and data processing technology is improving and has been widely used. At the same time, in many data mining applications, many users information or business information are located in different databases or data files to the Internet, for example, quite structured data, which gives parallel processing technology provides a lot of opportunities.

# CONCLUSIONS

Cluster analysis as an important branch of data mining functions, it is a non-supervised pattern recognition and has a number of theoretical basis and algorithm and achieved encouraging research results. However, in a cloud computing environment, there are still a lot of problems with clusters research. With the growing complexity of the soaring amount of information and data objects, clustering analysis faced with more new content and challenges. This requires the introduction of a new improved method of clustering, and proposed new theories and methods to adapt to new applications.

## REFERENCES

**[1]** Zuocheng Wang, LIxia Xue, Yongshu Li; Spatial data mining knowledge map visualization. Journal of Application Research of Computers, **14(3)**, 253-255, 42 **(2006)**.

**[2]** Xiaoli Qi; An Improved Fast clustering algorithm and parallel research. Journal of Lanzhou University, **18(9)**, 23-25 **(2009)**.

**[3]** Rogers James, Barbar et al; Detecting spatio-temporal outliers withkernelsand statistical testing. ETRI Journal, **53(10)**, 66-69 **(2012)**.