# Research on Data Mining Algorithm Based on Hadoop Cloud Platform

**Wang Cuiping, Lu Jiasheng**

Rizhao Polytechnic, Rizhao, Shandong, 276826, China

**Keywords:** Cloud platform; Data mining; Decision tree algorithm

**Abstract:** In the information age, the increase of information content provides convenience for our life. However, the increasing amount of information makes the data scale even larger. Therefore, people have higher and higher requirements on data mining technology. Data mining has become more complex and difficult. This paper first explains the concept of data mining technology and cloud platform. Then it analyzes the Hadoop cloud platform data mining mode. Finally, this paper studies and analyzes how to deal with SPRINT decision tree algorithm of big data.

## 1. Introduction

Since the 1960s, information technology and database technology have developed rapidly. Originally, we could only use them for simple files, but now it has become a powerful and complex database system. A large number of advanced databases, such as web-based databases and data warehouses, began to appear at the end of the last century. With the increase of social information, the scale of data is increasing rapidly. Data structures have also become more complex and diverse. This makes everybody to data processing technology requirement is higher. Therefore, how to extract valuable information from these "data graves" becomes an important issue. In order to solve this problem, people invented data mining technology. It's like digging gold out of ore.

## 2. Data mining and cloud computing

### 2.1 Data mining

Data mining is to find previously unknown, implicit and potentially useful information and data from a large amount of data [1]. It is a process of searching useful information hidden in a large amount of information through algorithms. Many people also see data mining as a synonym for another term, "knowledge discovery (KDD)," while others see data mining as part of KDD.

### 2.2 Cloud computing

The definition of cloud computing is not uniform, it is more diverse, there is no uniform standard definition, but it is all the same. Cloud computing is a business model that provides computing power. Key technologies that implement cloud computing include virtualization and distributed systems. Currently, server virtualization and application virtualization are the two main commonly used core technologies [2]. Distributed storage and distributed computing are the two main commonly used core technologies in distributed systems [3].

Today's data scale is huge, but useful information is scarce. We need to use data mining technology to obtain useful knowledge and rules from the original data. Traditional technology has become inadequate in the face of huge data volume; the emergence of cloud platforms provides us with new conditions. Therefore, combining the two, this paper proposes a data mining algorithm based on cloud computing platform to solve the problems faced by traditional data mining algorithms.

Hadoop platform itself belongs to distributed computing in the core technology of cloud computing. It is a distributed computing platform generated under the background of big data. The Hadoop platform works in parallel, which improves processing efficiency. Therefore, when we use Hadoop platform to realize the idea of traditional mining algorithms, we must make traditional data mining algorithms adapt to the processing requirements of big data. At the same time, we can take

  DOI: 10.25236/csam.2019.031

advantage of Hadoop platform's advantages in distributed processing to improve processing performance. This undoubtedly makes sense.

## 3. Data mining under Hadoop platform

In order to apply the data mining algorithm to Hadoop, we need to deal with three problems: global, random write operation and life cycle. After many experiments, we found that the persistence method of database is an ideal way to solve the above problems. Figure 1 shows the data mining pattern under Hadoop platform[4]. Massive data sets are stored HDFS and scanned by Hadoop; We extract necessary information with classical data mining modules, which are decomposed and embedded into corresponding map and reduce modules of Hadoop. The objects we extract, such as linked lists, trees, and graphs, are automatically persisted to the database. JPA and simple JDBC implement ORM objects in data mining. In this way, we can achieve knowledge accumulation between tasks.

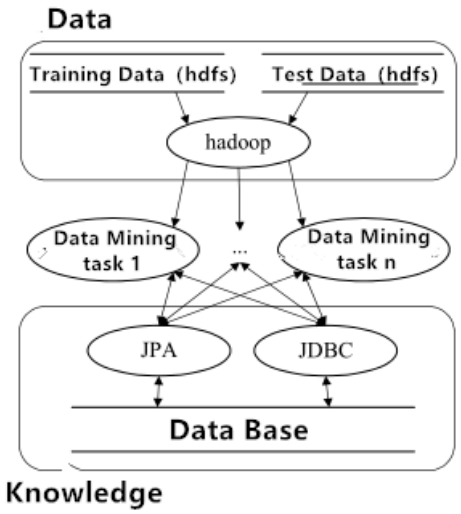


Figure 1 Data mining pattern under Hadoop platform

In order to ensure the stability of computation, distributed database is suitable for massive high-dimensional data mining. MySQL cluster is such a scheme, which is designed to remove any single point of failure of the system. It's also an open source project; These two properties make it ideal for open source data mining when combined with Hadoop. Implementing a Hadoop cluster is a tedious task that programmers spend a lot of time maintaining. Vega cloud can provide clusters that can be deployed automatically like Hadoop clusters, for example; The logged-in user can operate the virtual machine as if it were a physical machine.

## 4. SPRINT decision tree algorithm

Decision trees are powerful and popular classification and prediction tools. Unlike "black box" sorting algorithms such as neural networks, it has no rules. It is a method based on tree graphs and has rules[5]. To make it easier to understand, the rules can be expressed in words. The origin of decision tree is CLS. In 1983, the ID3 decision tree algorithm was proposed by Quinlan, J.R. Quinlan, J.R. is a leading expert in machine learning. Then the C4.5 algorithm was formally proposed in 1993. The structure of a decision tree is more like a flowchart, where the root nodes are at the top of the tree and present an inverted tree. Starting with the uppermost root node, each internal node of the tree is classified as an attribute[6]. Each branch represents a classification output, and each leaf node (the end node) holds a class label. The path from the root node to each leaf node has its own specific path, which is an expression. It is used to classify data rules. Most of the decision tree algorithms include two steps: spanning tree and pruning. Spanning trees are the main link, accounting for at least 95 percent of the work[7]. Typical decision tree, begin from the root node split, according to the node splitting measurement standards to choose the best split attribute,

split the tree nodes and leaf nodes; Then, each tree node calculates its own optimal splitting attribute to split the nodes, until all nodes are leaf nodes and complete the construction of the whole decision tree[8].

When generating the decision tree, we should constantly split the nodes by calculating the attribute measurement of each node. They're incredibly computationally intensive and now most of the decision tree algorithms are memory resident algorithms, such as ID3, C4.5, CART and other algorithms. All of these algorithms require some or all of the data set to be kept in memory. However, if the information processed by these algorithms is a large data set, a lot of memory overflow will occur. This situation largely limits the access of data mining to large data, especially when the amount of data processed by people becomes larger and larger. As a result, SPRINT is one of several decision tree algorithms developed by data mining researchers for large data sets.

### 4.1 Algorithm introduction

SPRINT algorithm is a decision tree algorithm with high scalability for processing large data sets proposed by John Shafer and Rakesh Agraw in 1996. Compared with ID3, C4.5 and other most memory-resident algorithms, the main improvement of SPRINT algorithm is to describe all features of data set through two data structures, namely attribute list and histogram. For a large data set, only its class histogram has been kept in memory, while property list as needed (for a list of attributes for the current processing) commonly kept in memory, and other property list is stored in the disk, thus solve the data mining process large data when the memory leak problem.

The SPRINT attribute list is a simple concept that splits the attribute dimensions of a dataset into a single attribute list to store the dataset. Class histogram is mainly used to calculate the optimal segmentation property of each node, and the property list stored in memory is determined by it.

### 4.2 Calculate the resulting step

The process of SPRINT algorithm to generate the decision tree is not much different from the traditional decision tree generation process. A simple decision tree generation process is as follows:

(1) Split the data set into columns to form various property lists, which represent all the characteristics of the data set. If it is a continuous value attribute, the sort is performed.

(2) Create root node, root, and attach all property lists to this node.

(3) Calculate the corresponding Gini index and the corresponding split point of records for all the property lists of this node;

(4) Compare Gini indexes of all attribute class indexes and select the best segmentation attribute to split the node N1 and N2;

(5) Recursively call the tree building algorithm to generate the decision tree N1 and N2 until the data set of this node belongs to the same class of standard or the number of samples is less than the set threshold.

## 5. Summary

Experiments on Hadoop platform can prove that the parallelization of PRINT algorithm has a good cluster acceleration ratio and scalability, and the computing speed of data sets of different sizes has been greatly improved. However, with the increase of cluster nodes, the speedup ratio and scalability ratio are slightly reduced, which is mainly due to the corresponding data set size, and the degree of parallelization has been maximized [9].

## References

[1] Qiao Shimin, Yang Zhaohui. Overview of cloud computing technology [J]. Research on the application of information technology, 2011:26. P. Klingstom □

[2] Xu Dongchai, Bao Cunhou, etc. Based on the concept of cloud computing network modeling and simulation platform - "cloud simulation platform" [J]. Journal of system simulation, 2009, 21 (17): 5292-5299.

[3] Li Baihu, Zhang Lin, Wang Shilong, et al. Cloud manufacturing -- a new mode of service-oriented networked manufacturing [J]. Computer integrated manufacturing system (s1006-5911), 2010, (01): 1-7,16.

[4] Hua Xiang, Kang Fengju, Tian Xuewei et al. Research on private cloud framework of visual simulation [J]. Journal of system simulation, 2011, 23(08): 1652-1656.

[5] Huang Anxiang, Feng Xiaowen, Li Jingsong, et al. Aviation training simulation architecture based on cloud computing platform [J]. Journal of system simulation, 2011, (S1): 106-109. (in Chinese)

[6] Pan Tianming. Research on parallelization of decision tree algorithm based on Hadoop platform [D]. Shanghai: east China normal university: 2012.

[7] Zhang Yongyong. Research and implementation of automatic text classification based on Hadoop [D]. Harbin: Harbin Institute of Technology: 2012

[8] Liu Mulin, Zhu Qinghua. Research on association rule mining algorithm based on Hadoop -- taking Apriori algorithm as an example [J]. Computer technology and development, July 2016 (26):1-5

[9] Xin Daxin, Qu Wei. Research on cloud computing algorithm based on Hadoop [J]. Electronic design engineering, February 2013 (3):33-39