Future Generation Computer Systems 88 (2018) 279–283

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Parallel processing algorithm for railway signal fault diagnosis data based on cloud computing

Yuan Cao [a,b], Peng Li [b,*], Yuzhuo Zhang [b]

[a] National Engineering Research Center of Rail Transportation, Operation and Control System, Beijing Jiaotong University, Beijing 100044, China
[b] School of Electric and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

HIGHLIGHTS

- Existing challenges for feature extraction and analysis of rail failure data are analyzed and summarized.
- A parallel processing algorithm based on cloud computing is proposed.
- Partitioning strategy of data flow is improved and bias classification algorithm is used to model and classify data.
- The algorithm is compared with the general one through practical examples.

ARTICLE INFO

Article history:
Received 27 March 2018
Received in revised form 21 April 2018
Accepted 17 May 2018

Keywords:
High speed railway
Cloud computing
Fault diagnosis
MapReduce

ABSTRACT

To explore the data processing of high-speed railway fault signal diagnosis based on MapReduce algorithm, the partitioning strategy of data flow was improved, and Bias classification algorithm was used to model and classify data. In MapReduce parallelization process, the data partition matrix T_k was stored in line segmentation, the computing load was distributed in every node of cluster, and the time consumption of mobile data matrix and the consumption of partitioned matrix were calculated. Results show that the algorithm proposed could reduce the amount of computation in the execution process, greatly reduce the memory space consumption, and improve the counting speed in railway signal system.

1. Introduction

With the further speed increase of China's high-speed railway and the continuous improvement of the railway information system, the conditions of collecting more railway running information are now available [1]. At present, the running high-speed railway train, through the deployment of a large number of sensors, collects a variety of data [2,3]. However, the traditional vibration data feature extraction and analysis technology is running on a single machine. This kind of technology, in the mass vibration data acquired by sensors, exposed the shortcomings of long processing time, various artificial intervention, and poor capability of processing big data file and so on [4,5]. The emergence of cloud computing technology provides a way of thinking to solve the above problems. Map Reduce is an effective parallel computing framework of processing big data, which is one of the main models of cloud computing, and can automatically assign tasks and realize task balance [6,7]. The working principle, operating mechanism and fault tolerance mechanism of Map Reduce calculation model are studied. In addition, combined with the characteristics of association rule generation algorithm, the traditional parallel algorithm is improved and the parallel optimization scheme of association rules algorithm based on Map Reduce is proposed. Moreover, the improved algorithm is used in the railway quality analysis and evaluation industry [8].

A parallel computing model proposed by MapReduce for Google can do parallel computing to millions of processors [9]. These calculations consume very low cost of users so that users do not need to be too much entangled in the distributed parallel computing technology [10]. The program can realize its distributed functions simply by Map function and Reduce function programming and deploying their own procedures to the cluster. In recent years, many researchers have proposed an improved algorithm for the shortcomings of parallel algorithms in practical applications. Hashem proposed an Apriori parallel improved algorithm introducing indexing structure. The algorithm improves the Apriori algorithm, and through MapReduce mechanism, conducts block processing of data [11]. It also increases the data index in each data block, so as to enhance the performance of the algorithm. In the implementation process, only the part of the data object affected is

* Corresponding author.
E-mail addresses: ycao@bjtu.edu.cn (Y. Cao), lipeng@bjtu.edu.cn (P. Li), 12120301@bjtu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.future.2018.05.038
0167-739X/

updated. Although the improved algorithm can effectively enhance the efficiency, there is still the problem of low precision of mining. Li regarded Hadoop as an open source cloud computing platform that can provide a simple clustering framework. Each computer node can cache the data in the local disk storage, and store respectively. In the implementation of the calculation, MapReduce of each computer can directly read the local data, so as to save the cost of network transmission.

2. Methodology

2.1. Problem description

For the management business of railway rail quality analysis, rail plays an important role in national traffic construction, and because of its uncertainty in natural environment, reasons that rails cause failures are various. For the management of maintenance section, it is of great significance to establish relevant supporting attributes from these fault attributes and dig out related decisions to support quality management analysis of managers. Through studying the related railway safety production quality analysis business, we find that in the traditional rail management process, the role of data decision support is not obvious [12]. In the field of rail quality analysis and evaluation, the research on fault analysis is relatively few. Combined with the characteristics of data mining, it is the key content of this chapter to clean, integrate and share the heterogeneous data of rail inspection.

In the face of multivariate heterogeneous data in railway track quality inspection data, by analyzing the characteristics of these data, the first problem to be solved is data storage format and storage strategy. The algorithm is set up based on the data structure. The primary problem of improving the algorithm is the optimization of its data structure. According to the fault data format of railway rail, after analysis and summary, the main features are summed up as follows:

Firstly, the increment of rail failure data is relatively large and the data of railway track quality detection is the data real-time acquired by sensors in the operation process. These data are generated in the train operation in accordance with mileage, the increment is relatively large, and the data performance cannot be guaranteed.

Secondly, the railway rail testing data have the characteristics of great repeatability and many data dimensions. Because railway rails are easily affected by natural conditions, there are many reasons for the failure data [13]. In order to ensure the accuracy of data, we need to consider various dimensions to establish data models when grading the work area.

Thirdly, the rail fault data format is complex and the data noise is much. The railway track test data have different data formats because different instruments. In addition to the data transmitted by sensors, there is "artificial multiplication", "track inspection car" and other manual operation data. The data have different lengths, the format of data is various, data noise is much, and the structure is relatively complex.

2.2. Data partitioning strategy for data flow

The generation strategy of parallel data flow association rules can be mainly divided into four parts: data division strategy of data flow, parallel transmission strategy of data flow, task processing in MapReduce framework and data frequent item sets mining process. In this paper, the partition strategy of data flow is improved, and the Bias classification algorithm is used to model and classify the data. In the process of MapReduce parallelization, the partition matrix T_k of data is stored in line segmentation [14]. The computing load is distributed in every node of cluster, and the time consumption of moving data matrix and the consumption of partitioned matrix are calculated.

There are many problems in the data classification of data flow, such as lack of consideration in the candidate patterns, low perception and so on. There are also many algorithms in the partition of data flow, in which K-neighbor classification algorithm is a simple algorithm [15]. It can store the classification method of support vector machine combined with the nearest neighbor algorithm, as well as all available examples and classification methods based on similarity measure. Each instance is divided into a master node and its neighbor nodes. According to the Euclidean formula, the distance between two points is determined as Euclidean:

$$Euclidean = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1}$$

In the above formula, x and y represent arbitrary vectors in the Euclidean space, and k is an arbitrary real number.

2.3. Process of mining frequent item sets for data

The rapid and high-dimensional growth of data is facing huge challenges for big companies like Facebook, Google, Amazon and YAHOO. These companies need to quickly execute TB and PB level data to meet their user requirements and query operations. A parallel computing model proposed by MapReduce for Google allows users to perform parallel algorithm in large cluster machines of commodity. Moreover, programmers can achieve their distributed functions only through the Map function and Reduce function programming and deploying their own procedures to the cluster, not being too much entangled in the distributed parallel computing technology. The core of MapReduce lies in its Map function and Reduce function. Map tasks and Reduce tasks form MapReduce tasks, which can parallel and compute millions of processors, which cost users very low cost. Among them, the Map() method accepts a key–value pair $\langle k_1, v_1\rangle$ as input, and sends out a new key–value pair $\langle k_2, v_2\rangle$ and user logic as intermediate output [16]. The Reduce() process combines all key values of the same key and generates the final output.

The data flow frequent item algorithm A-SON (Apriori SON) algorithm based on the MapReduce is a priority selection algorithm. We use the traditional SON algorithm to effectively reduce the CPU and I/O load. And we first of all choose the item sets that do not exceed the threshold range in the parallel processing. The core idea is that the input data flows are divided into multiple blocks in the sliding window. Each block is regarded as a sample data, and all blocks use MapReduce for parallel processing and run algorithm in the block. The ratio of each block to the whole file is set to p, the support threshold is s, and the frequent item sets of each block are stored in the disk space. When all block processing is completed, the frequent item sets generated by each block are merged into candidate sets.

Through the definition of data flow frequent term algorithm, the A-SON algorithm is specifically implemented as follows:

Initialize the data set $I = \{Xi\}, Xi \in R^D$ and set the data set X to generate discernibility matrices $[M_{ij}]_{mxn}$ and n_{jk}, and matrix $P = MN$:

$$p_{ik} = \sum_{j}^{1} m_{ij} n_{jk} \tag{2}$$

First of all, multiple device nodes, MapReduce, are opened to scan the partitioned data sets, respectively. Map and Reduce are set as follows: from the device node Map function: the support threshold s of data subset is reduced to ps, and a key–value pair is set as $(F, 1)$ set, where F is a frequent item set in the sample and