2022 International Conference on Big Data, Information and Computer Network (BDICN) | 978-1-6654-8476-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/BDICN55575.2022.00041

# Research of Big Data Information Mining and Analysis

## Technology Based on Hadoop Technology

Zhanchi Dong*
School of Data science, Melbourne University Melbourne, 3010, Australia
*Corresponding Author's Email: kalimosdzc@outlook.com

***Abstract*—Nowadays, with the continuous development of modern Internet and e-commerce technology, the network data is growing geometrically and the era of big data has come. The massive data information in the network contains a lot of industry resources, and how to properly mine the data information required has become a new issue for enterprises. In this paper, a new data mining and analysis technology is built based on the traditional Apriori mining algorithm and applying the latest Hadoop technology. This technology mainly focuses on the following two aspects. The first is the improvement of the traditional algorithm, which is mainly applied to Hadoop technology to upgrade the traditional data algorithm; the second part is the parallel analysis processing of mining data. The two major parts constitute the new big data mining analysis technology, which provides some references for the current logical data mining and network data applications.**

***Key words: Big Data; Information Mining; Data Analysis***

### I.INTRODUCTION

At the IDC conference in 2012, a large number of Internet experts and scholars pointed out that the big data and data analysis mining technology will be one of the core competitiveness of enterprise development in the future. In the same year, the U.S. government issued a big data development plan and considered the big data as the "new oil of the future". With the continuous development of the network at this stage, the big data has become a hot word, and the issue of how to effectively use the data "treasure" contained in big data has been widely discussed on the major media platforms. People hope that the business content in the information can be obtained through low-cost, efficient and fast mining of these massive and diverse data information, which can help them make the right marketing and business decisions [1].

In the face of huge data samples, the most important issue to achieve effective mining utilization is how to overcome the amount of calculation. Traditional data computing logic is complex, software and hardware requirements are extremely high, and system resources are seriously occupied. The market urgently needs a data information mining technology with powerful computing power and reliable data storage capability as a new data application tool, which is also a new direction for the future of big data and data mining research. For this industry market information, a new big data information mining and analysis technology based on the traditional Apriori mining algorithm and applying Hadoop technology framework is innovated and developed in the paper [2].

The core content of this research mainly consists of the following two aspects. One is the improvement of Apriori mining algorithm based on Hadoop technology, which includes the analysis of Apriori basic mining algorithm and the technology upgrading under the framework of Hadoop technology; the second part is the parallel analysis processing under collaborative filtering of big data, which includes data similarity calculation, co-word analysis and optimized recommendation, realizing the process from data mining and sorting to analysis and recommendation.

Association rules and data mining is the core topic in big data analysis, as well as the root of optimization and upgrading, the purpose of which is to find the connection between different data from the complicated data and obtain the data set required by users. In the big data environment, all kinds of data logic algorithms deplete a lot of content space and time, and the data mining recommendation strategy based on Hadoop technology proposed in this paper can effectively solve this problem through parallel algorithm, and provide customers with data services that meet their own characteristics. Facing the current situation that the total amount of big data is too huge, the data analysis mining algorithm proposed in this paper, can restrict most of the algorithms to a fixed sparse framework with its own utility matrix, which leads to the decline of its algorithms and solves the limitations of traditional algorithms.

### II.IMPROVEMENT OF APRIORI MINING ALGORITHM BASED ON HADOOP TECHNOLOGY

#### *A. Basic Process of Apriori Mining Algorithm*

The Apriori algorithm mainly uses the concept of support to increase the candidate set of data linking control to complete data simplification and finally achieve the purpose of controlling the option data set. The core principle of this algorithm is to focus all the data item sets into frequent subsets and use the cascading method to find iterations and complete data mining [3]. The specific formulation is to continuously iterate down from the frequent K-item set to mine the K+1-item set. The process is shown in the figure below.

978-1-6654-8476-3/22/$31.00 ©2022 IEEE
DOI 10.1109/BDICN55575.2022.00041

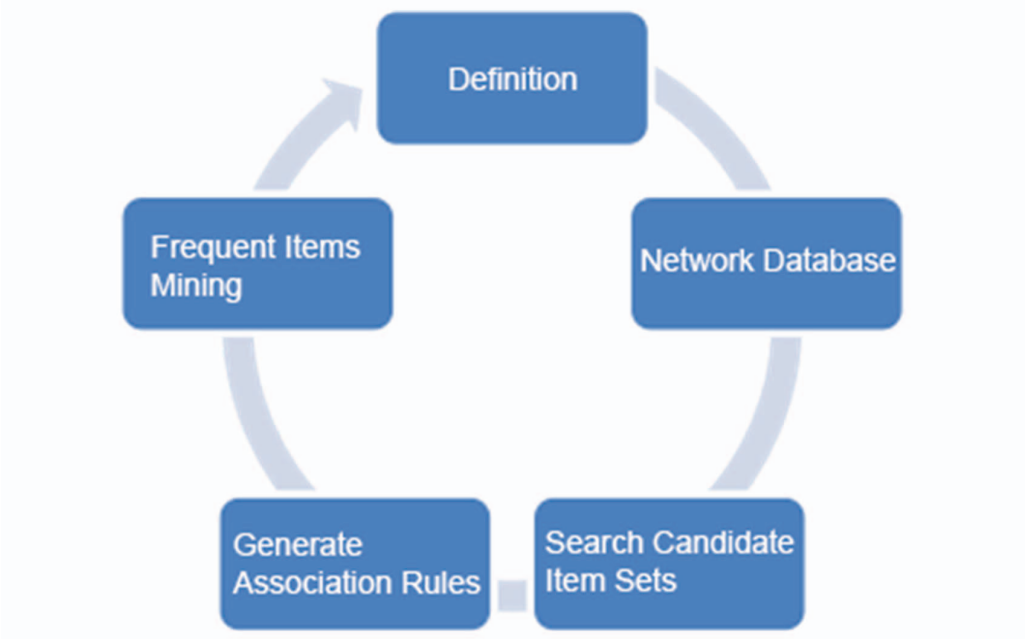


Figure 1. Apriori Algorithm Flow

From the above figure, it can be seen that the whole data analysis process mainly encapsulates the task of finding candidate items and association rules display, i.e., it includes the search of frequent item sets and the generation of data rules [4]. However, with the diverse development of network data, different frequent item search strategies also have obvious differentiation.

*1)Process Analysis of Frequent Item Search*

*a)*The different unit data elements are represented as a set of candidate items (generally as C1) for the data list. Then further selection begins by ranking the support of all data items. Find the user's minimum support, remove it from the options set, and finally generate a new items set L1.

*b)*After linking L1 according to the data classification, at least 2 sets of data candidate items can be obtained. Through database iteration, the support of different candidate items set can be determined, and finally compared with the minimum support. All data items with greater support than the minimum are retained to construct the frequent item set L2.

*c)*Through the above method for data stratification and data iteration, all frequent item sets can be found. That is, after finding frequent K item sets, stop the operation when the K+1 item set is found to be 0 [5].

*2)Analysis and Generation of Association Rules*

The above frequent item search process can be regarded as the core component of the whole algorithm, which determines the efficiency of the algorithm. After obtaining the data frequent item set with the above algorithm, the corresponding rules can be analyzed and generated. The rule can be obtained directly, because all frequent item data have been judged in the first step to be higher than the minimum support of the data item.

### *B. Apriori Mining Algorithm Upgrade under Hadoop Technology*

The above process analyzes the core algorithm of Apriori algorithm, and this section upgrades the traditional Apriori algorithm by using Hadoop technology and makes it more adaptable to current data mining of association rules for big data.

*1)Analysis of Parallel Algorithm Applications*

The process improvement for parallel rules of Apriori algorithm in this research is shown in the following figure:

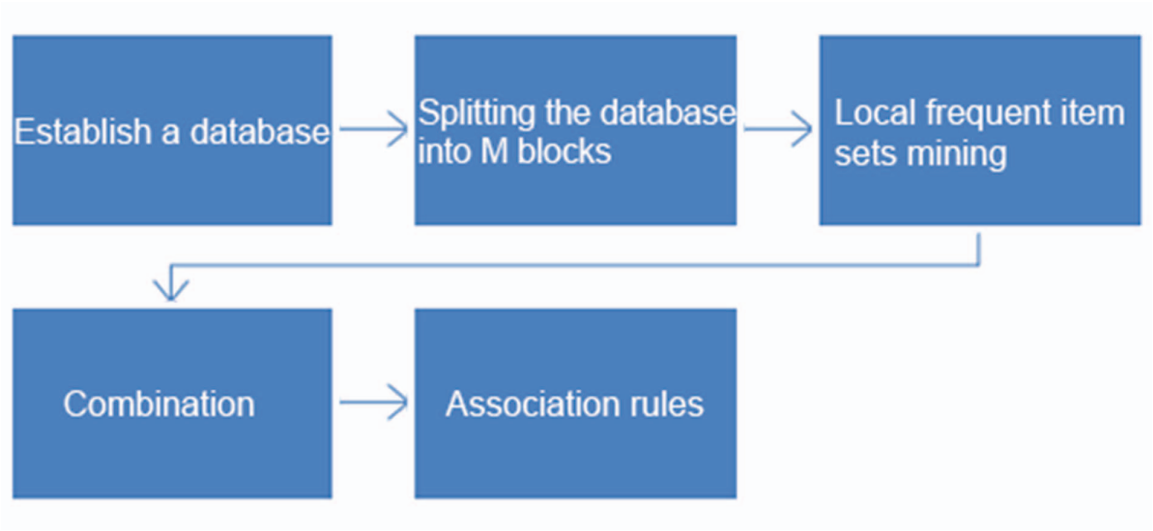


Figure 2. Hadoop Technology Improvement Flow

It can be seen from Figure 2 that the improved mining algorithm also contains 2 core processes, and the biggest improvement point is the difference in the search process of frequent item sets.

(1) Frequent Item Set Optimization

a. Firstly, all the data are aggregated into a unified database for data splitting, generally in n blocks, and then all the databases are created separately as Mapper task sequences. The data are parsed into key-value relative forms by InputFormat, where the data secret key is the transactional ID record, and VALUE is used as the specific data link.

b. Execute the Mapper task sequence process, call the map method to filter the data blocks separately, execute each frequent item set in different Mapper tasks by the Apriori algorithm analyzed in the above section, call the total K item sets as frequent item sets under local integration, and finally use combine to merge different Map tasks, manage the items and output the merged results at the same time.

c. Integrate the local frequent items obtained from different working nodes in the above steps to build the Reduce task, thus obtaining the global candidate item set K. (During the above execution, the frequent item set k under different data blocks called by applying the Apriori algorithm needs to be concatenated with the local frequent item set and finally merge again by using combine), and input the merged notation over to the MAP task to obtain the final data result.

d. Integrate the above data items. Specifically, the Reducer data program can be executed to iteratively sort out the database D, count the minimum support for the whole database of candidate items, compare them centrally, leave all the data options and aggregate them within the frequent item set.

(2) Sorting of Data Association Rules

After applying the above procedure, the highest-specified frequent 3-item set is obtained. The confidence of all data subset rules are shown in the following table. Let the highest frequent item set (3 items) be I1I3I4, and the subsequent non-empty subsets be { I1，I3，I4}. There are the following associations.

TABLE I. SUPPORT AND CONFIDENCE OF ASSOCIATION RULES

| Association Rules | Support | Confidence |
|---|---|---|
| I1 belongs to I3≥I4 | Support(I1 belongs to I3≥I4)=0.4 | Confidence(I1 belongs to I3≥I4)=0.4 |
| I1 belongs to I4≥I3 | Support(I1 belongs to I4≥I3)=0.4 | Confidence(I1 belongs to I4≥I3)=0.4 |
| I3 belongs to I4≥I1 | Support(I3 belongs to I4≥I1)=0.4 | Confidence(I3 belongs to I4≥I1)=0.4 |
| I1≥I3 belongs to I4 | Support(I1≥I3 belongs to I4)=0.4 | Confidence(I1≥I3 belongs to I4)=0.4 |
| I3≥I1 belongs to I4 | Support(I3≥I1 belongs to I4)=0.4 | Confidence(I3≥I1 belongs to I4)=0.4 |
| I4≥I1 belongs to I3 | Support(I4≥I1 belongs to I3)=0.4 | Confidence(I4≥I1 belongs to I3)=0.4 |

From the association rules shown in the above table, it can be seen that the data have the strongest association rules when the data support and confidence are 0.4 and 0.6 respectively. With the above association rules, the intrinsic connection between different data items can be determined, and the foundation for the subsequent parallel improvement can be laid.

### C. *Parallel Improvement*

The Hadoop platform is innovatively applied in this design to split the timely tasks, and then the data blocks are divided into different areas by distributing nodes through different data platforms. Finally, after all the calculations are finished, the main program will divide the work tasks of each node by request. Even if there is some data not distributed, the main program can still perform task skipping to avoid problems between different

nodes and generate abnormal results. The improved algorithm characteristics are as follows:

Firstly, relying on the traditional Aprior algorithm, the different candidate data item sets are selected separately to achieve data partitioning through frequent item set mining after parallel data is achieved. In the process of data frequent item set search, the Mapper process is executed to determine the local mining and analysis process of different nodes to improve the computational efficiency.

## III. Parallel Analysis and Processing of Big Data Collaborative Filtering

The above section is mainly based on Aprior algorithm, the Hadoop technology is used to upgrade the traditional Apriori algorithm to realize data mining, while data analysis and parallelization processing through the steps of data similarity calculation and co-synthesis analysis is achieved in this section to improve the comprehensive ratio of data application.

### *A. Data Similarity Calculation under Hadoop Optimization*

With applying the Hadoop technology platform, the similarity calculation method contained in traditional data mining is optimized, and the Map Reduce analysis procedure with complete similarity calculation process is applied in the calculation model as follows:

(1) Converge the data of different users into a unified database D, and use the utility matrix to complete the data level division. For example, set the utility matrix as U, the splitting results are U1, U2 and U3, and the target user is User.

(2) Split the data divided by the above steps and build a data set (key, value), where key is the data user number and value is the data vector option corresponding to different user, and then the new data sets (key, value) can be obtained by adjusting the form of data map and data calculation.

(3) After obtaining the new set (key, value) by the above calculation, the data similarity is calculated in the form of the Reducer process calculation. The core principle is to get the data degree value by pre-distance calculation. The smaller the values are, the more similar the records are. After the results processing, a new set of values (key, value) can be obtained. At present, the key indicates the relationship between the data target user and the current user, and the value indicates the final result value.

(4) The processing results of the Reducer are obtained a further Srot processing, the process is mainly for data similarity rearrangement. With the Java collection line TreeMap program, the output operation can be completed, and the final data elements can be obtained.

### *B. Improved Co-word Analysis under Hadoop Optimization*

After calculating the similarity, a co-word analysis is also required. The specific approach is to populate the data utility matrix. Because the above section has been introduced about the data correlation search, it is not repeated here. The design is designed to improve the speed of data processing, the whole utility matrix is loaded into the memory values for calculation, and then the data filling is completed based on the improved co-word of Hadoop technical framework. From the analysis process, it can be seen that the utility matrix is characterized as follows:

(1) The utility matrix is the square matrix of the data item;

(2) The utility matrix is completely symmetric;

(3) The diagonal element of the matrix is 1.

Based on the above characteristics, the symmetric matrix element compression is applied inthe design to store the upper or lower triangular part for all kinds of data. Through calculation, it can be determined that matrix filling needs to establish m (m+1)/2 data unit spaces. Because the diagonal element of the matrix is 1, the storage unit can be reduced by m. Then the improved data recommendation process based on Hadoop technology is as follows:

Step1: Establish the connection between big data users and data items, which can build the utility matrix;

Step2: Based on the utility matrix, improve the co-word analysis through the Hadoop technology framework, and fill the data with blank elements to get the final utility matrix;

Step3: According to the final utility matrix, improve the data similarity, then the data corresponding to user recommendation can be obtained. The data mining and analysis sorting can be achieved.

## IV. Conclusion

Data analysis and sort mining in the big data environment has a large amount of data computation, and the traditional algorithm has been difficult to meet the current data mining needs. In order to effectively solve this problem, the Hadoop technology is taken as the core in the design, and a new data mining analysis technology is built to achieve big data information mining and recommendation. The first is the mining under data iteration; the second is the parallelization of user information processing recommendation. After the actual test, the efficiency of data mining and application can be effectively improved through the researched data mining technology, which is worth promoting and learning.

## References


[1] Zhu Yueqin, Tan Yongjie, Zhang Jiantong, et al. Technical Framework for Geological Big Data Integration and Mining Based on Hadoop [J]. Journal of Geodesy and Geoinformation, 2015, 44(0z1):152-159.

[2] Chen Qi. Characteristic Analysis Research of Power Big Data Based on Hadoop [D]. North China Electric Power University (Beijing), 2016.

[3] Liao Jinggui. Research and Implementation of Association Rule Mining Algorithm for Big Data Based on Hadoop [D]. South China University of Technology.

[4] Cui Haijiang. Research and Implementation of Big Data Analysis and Mining Technology for Communication Industry Based on Hadoop [D]. Beijing University of Posts and Telecommunications, 2016.

[5] Wang Lihong, Liu Ping and Yu Guanghua. Research of Big Data Analysis System for Trade with Russia Based on Hadoop [J]. Computer Knowledge and Technology, 2018, 014(001):20-22.