Research on Distributed Vertical Frequent Pattern Mining Method Based on Metadata Integration

Li Sun[1], Li Guo[2], Huan Tian[3]*

[1]Longqiao College of Lanzhou University of Finance and Economics;

[2]Lanzhou modern vocational College

[3]Department of Electronics and Information Engineering,Lanzhou Vocational and Technical College

*Corresponding Author's email: 120839127@qq.com

Abstract. In order to better integrate the information industry into people's life and work, and even in social development, how to mine data has become a hot issue. Metadata is a kind of data about data. Mining metadata helps data application and storage. Therefore, it is very important to find an efficient and intelligent data mining method. This paper introduces the metadata and its integration technology. On this basis, it introduces the distributed vertical frequent mode, and introduces its use in the process of mining metadata, providing a new working idea for the staff engaged in related industries.

1. metadata and its integration technology

Metadata is data about data. Metadata is a variety of descriptions of data. The content of the description mainly includes data source, data accuracy, data quality, data processing process, and data update and information maintenance [1]. The concept of metadata was originally introduced. First, it is to operate the database more efficiently and conveniently, and improve the efficiency and result optimization of database update and maintenance; Secondly, the introduction of metadata can assist the computer industry to provide professional skills for other industries and better integrate computer technology into other industries.

Metadata is widely used, and all walks of life have relevant research on metadata. Therefore, metadata has certain differences in different industries, which is one of the fundamental characteristics of metadata. Another feature of metadata is that the metadata itself must be responsible for the data, and it is possible to accurately describe the data in the fullest sense. At present, in the information industry and computer technology, the use of metadata can improve the efficiency of data access and retrieval, and can also achieve deep data mining, processing and processing of data [2].

At present, the integration technology of metadata has become a research highlight in the field of data mining and machine learning. It has become one of the four important research directions in machine learning. It can be seen that the integration technology of metadata is of great significance. Compared with the metadata itself, the integration of metadata can further improve the value of metadata. However, the integration of metadata will also cause large data problems, and the space requirement for storing data will increase. Therefore, it will find the best. The scientific approach to integrating metadata for integration is of great importance in the field of metadata research.

2. Distributed vertical frequent mode

Today's social information technology is highly developed. Data integration and mining provide powerful data support for the development of information technology, and it is the technical support for information technology application in all walks of life. Metadata integration requires many different types of data interactions, complementing each other. Data mining is the discipline that provides basic data for data integration. Distributed vertical frequent mode is one of the widely used methods in data mining.

The meaning of distributed in metadata mining is to divide the whole data into multiple independent individuals. The distributed vertical frequent mode is to divide the data into several different individuals or subsets according to different classification forms in massive data. Then the individual or subset of data with the greatest importance is mined, and finally the frequent item set output is formed [3].

Frequent itemsets are defined in the database discipline as: Kn(n=1,2,...) is n items, K={K1, K2,..., Kn} is a set of items, and D is a transaction database. The support number of the item set S in the transaction database indicates the number of transaction items including the item set S in the transaction database, which is recorded as Scount, and the support degree of S in the transaction database refers to the frequency of occurrence of S in the transaction database. Recorded as S.sup. If the support degree of S is greater than or equal to the given minimum support threshold Minsup, the item set S is a frequent item set in the transaction database, which will be mined in the distributed vertical frequent pattern mining of subsequent metadata integration.

The main object of distributed vertical frequent mode mining is frequent itemsets. The distributed vertical frequent mode is used to search for massive data, and one of the data is mined out, and the data frequently appearing with it is mined together. The item set is filtered out and analyzed as a result. At present, there are two main algorithms for distributed vertical frequent mode, namely Apriori algorithm and FPGrowth [4].

The Apriori algorithm first constructs a data candidate set in the data, and performs mining in these data candidate sets. This algorithm needs to repeat the steps multiple times, and it has a long history. Therefore, when the data amount is large, the Apriori algorithm is used. Less efficient and not suitable for use. The first step of the FPGrowth algorithm is to build the FP-tree, and then use the recursive algorithm to mine the data in the FP-tree. This algorithm has only two steps, the efficiency is very high, and the data requires less storage space and is widely used.

Distributed vertical frequent mode prohibits the exclusion of data with significant influence from frequent itemsets. At the same time, it is required to rebuild frequent itemsets in frequent itemsets. It should also ensure independence and similarities between frequent itemsets.

3. Distributed vertical frequent mode

In the field of computers, the speed of development of computer equipment and technology can be expressed in terms of "Moore's Law". In order to achieve better development and cooperate with computers, metadata integration is progressing in line with the development speed of Moore's Law.

In the computer industry, big data and cloud computing are two emerging industries and disciplines with broad development prospects. Metadata is the foundation of these two disciplines. With the increasing demand for data in the information industry, traditional The data mining method can not meet the user's requirements for metadata. The distributed vertical frequent pattern mining method can meet the requirements of modern data mining work. Data mining is to find efficient, practical and representable data information from massive, less complete, noisy, fuzzy, and random data sets. Data mining is an interdisciplinary subject. Mining techniques include expertise in multiple disciplines. The distributed vertical frequent mode mining method has the characteristics of high reliability, online and elastic scalability, and can provide intrinsic relationship and application value between different data, which can provide convenient, fast, rapid and high for data mining practitioners in decision-making. Quality data [5].

For data, Data mining is a work with strict workflow, including data cleaning, data transformation, data mining development, data mining quality assessment and mining results knowledge representation of the eight main processes. The metadata mining process based on the distributed vertical frequent pattern mining method also needs to collect data information, centrally manage data of different types, different provenances and different characteristics, and formulate rules to represent the data set, which is useless or less relevant. The data, the data is converted into the required format or the data format is unified, and then the distributed vertical frequent pattern mining method is used for data mining according to the information in the data, and the quality of the mined metadata is evaluated according to requirements, and finally the element is The data is presented and applied to other areas. The following is a framework for distributed vertical alternating pattern mining algorithms based on metadata integration, as shown in Figure 1.

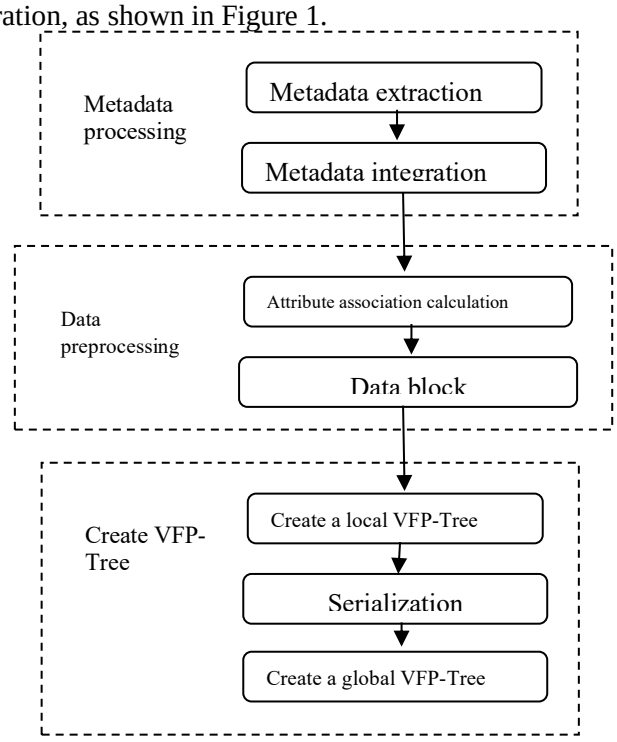

Figure 1 Algorithm framework

The above is the basic process of data mining. In the distributed vertical frequent pattern mining method of metadata integration, many researchers continue to propose new mining algorithms. Based on the Apriori algorithm, the data is scanned by inductive means. This method usually only needs to perform one scan, but it can accurately find frequent items in massive data, and then pick out valuable data for analysis. Metadata. There are also data mining algorithms based on the FP-Tree algorithm, which form a frequent item set through data acquisition of different phases. Another algorithm is to segment the data in the FP-grow algorithm and gradually mine the frequent itemsets in the data. This algorithm allows users to obtain the frequent itemsets needed online, and the algorithm mines Frequent itemsets are of high quality.

Because of our different environments and various factors such as innate genes, each person has a different personality and therefore has a personalized character for the needs. The distributed vertical frequent pattern mining method of metadata integration can perform data mining for different personalities, and exert data strengths to achieve user satisfaction. Since the reform and opening up, the living conditions of the people have been greatly improved, the requirements for quality of life have been significantly improved, and the private custom-made industry has been loved by more and more people. The distributed vertical frequent pattern mining method is used to update and maintain the customer's metadata can greatly reduce the cost of the business, while better serving customers.

The staff engaged in private ordering will estimate the customer's needs in advance according to the customer's requirements or usual hobbies. Through Data mining, it will help to improve the compliance of employees' prediction results with customer needs.

4. Summary

With the continuous improvement of the quality of life and the continuous development of society, the amount of data generated by human beings is increasing, and the management and application of data has great commercial value and social value. In the era of big data, the distributed vertical frequent pattern of metadata integration can better adapt to the needs of the big data industry and improve the effectiveness of metadata integration.

References

[1] Yin Jiena. Research on Distributed Vertical Frequent Pattern Mining Method Based on Metadata Integration[D].Liaoning University, 2014.
[2] Jiang Bing. Research on distributed closed frequent pattern discovery method based on MapReduce [D]. Harbin Institute of Technology, 2011.
[3] LIANG Fei, ZHU Yifeng, HE Yanxiang. Distributed frequent pattern mining algorithm using grid service[J].Computer Engineering and Applications,2004,40(7):179-181.
[4] YE Feiyue. Distributed Frequent Pattern Mining Algorithm Based on Adaptive Hash Chain[J].System Engineering and Electronics technology,2005,27(3):560-564.
[5] MA Ke, LI Lingjuan, SUN Dujing. Distributed Parallel Data Stream Frequent Pattern Mining Algorithm[J]. Computer Technology and Development, 2016(7): 75-79.