Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 131 (2018) 841–846

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology (ICICT-2018)

A Summary of Research on Frequent Itemsets Mining Technology

Wang Xin gang*

Administration Center for the DTH Service in China
No. 2, FuXingMenWai Street, XiCheng District, Beijing, China

Abstract

In recent years, communication technology and Internet technology are developing rapidly. In this context, people pay more and more attention to the importance of data, especially the implied information with timeliness and value, and accessing to such information relies on data mining. Data mining technology has many important branches. In recent years, data mining has been widely concerned about the association rules data mining technology is one of them, and its biggest advantage is the ability to derive valuable, but less significant, patterns from a large amount of information. The key and core problem of association rule data mining is frequent itemsets mining. Based on the research status of frequent itemsets mining technology, the research results and basic ideas of frequent itemsets mining algorithms are discussed and analyzed.

Keywords: Data Mining; Association Rules; Frequent ItemsetsType ;

1. Introduction

With the development of Internet technology and communication technology in recent years, the potential value of data has drawn wide attention from scientific research institutes, government enterprises, and other communities of the society, and data mining[1-4] has also become a hot topic in academia. Various organizations collect data and store them in large databases. The goal of data mining is to scan the data in the data warehouse through data mining algorithms, and classify and analyze the data to find out unknown patterns. Then further analyze the trend represented by the observed data, and even predict the development of the the things. Data mining technology can be widely applied to market analysis, business management, scientific research and other fields.

Benefit from the development of computer technology and the popularization of the Internet, communities, schools, government agencies, enterprises and other fields are constantly generating a large number of data, such as user search and browsing history of websites, securities transaction records, sensor network detection data,

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-*mail address:*wangxingang@huhutv.net

10.1016/j.procs.2018.04.276

communication records of social platforms etc. The essence of these data is the data stream, which is different from the traditional data. It is a kind of data form with fast and continuous transmission. This undoubtedly brings new challenges to the traditional data mining. In order to analyze and obtain the valuable data in the data stream, and then predict the trend of the data stream, the data stream mining has gradually become the focus of data mining technology.

At present, the commonly used data mining methods mainly include clustering, classification, association rule mining. The main goal of association rule mining is to find the insignificant correlations between a series of data, and then to analyze meaningful and valuable trends, patterns, and so on. It is a vital part of Knowledge Discovery in Database (KDD). Association rules are calculated by the frequency of patterns, and a pattern are the itemset. This article focuses on the frequent itemsets mining technology, which is the core of association rules mining.

As one of the most widely used data mining technologies, association data mining[5-7] is a very important field in data mining, and it is also a key issue in this field. The shopping basket analysis problem is a typical case of the technology. In this issue, by studying the different products that do not contain obvious relationships purchased by the same customer, analyzing the relationship between them, and then the analysis results will be applied to commodity recommendation, inventory arrangement, cargo location and so on. At the 1993 SIGMOD international conference, Agrawal et al. proposed the concept of association rule mining for the first time, which is used to facilitate the discovery and description of dependencies and feature rules between two or more variables. As the core of association rules mining, frequent itemsets mining has drawn extensive attention since the issue of shopping carts, and now there are many classic algorithms.

2. Traditional frequent itemsets mining algorithms

2.1. Apriori algorithm and its optimization algorithm

After years of development of data mining algorithms, there are two kinds of classical frequent itemsets mining algorithms that are widely used by people. One is the Apriori algorithm[8], which is one of the ten classical algorithms of data mining, proposed by Agrawal et al., who are the discoverers of association rules. The algorithm requires scanning all the data in the database to determine the support value of each data; the data whose support value is greater than the threshold is called frequent 1-itemset, thus the whole frequent 1-itemsets from the transaction data sets can be obtained; the algorithm then finds candidate 2- itemsets from the frequent 1-itemsets. By analogy, the algorithm can obtain a new candidate k-itemset from the frequent k-1 itemset generated in the previous step. Finally, the support value of the candidate itemset needs to be counted. In order to accomplish this step, the algorithm needs to repeatedly scan the data set and delete all the items whose support value is less than the threshold, and until no new frequent itemsets can been found, frequent k-itemset is obtained.

It can be seen from the implementation of the algorithm that the algorithm needs to repeatedly scan the data set multiple times. With the increase of the amount of data, the algorithm overhead increases significantly, which is one of the major problems of the algorithm.

Another problem is the setting of the threshold. If the threshold is too low, the algorithm will get a large number of frequent itemsets through mining, which not only increases the space-time complexity of the algorithm execution, but also is not beneficial to the extraction of effective information. On the contrary, if the threshold is too high, it may lose some valid information. In addition, the maximum length of the frequent itemsets increases with the increase of the average transaction length, and the process overhead of the final candidate itemset generation and the calculation of the candidate itemset support values also increase.

Aiming at the problem of Apriori algorithm, many optimization algorithms are produced. J.S. Park et al. proposed a DHP (Direct Hashing and Pruning) algorithm[9,10] to improve the screening speed of frequent 2-itemsets by using hash technology. S.Brin et al. proposed a dynamic itemset counting (DIC) algorithm using partitioning strategy[11], which divides database into blocks and marks each block. It uses a hash tree as the basic data structure to dynamically add and delete nodes according to the data flow. The efficiency is higher when the data in the data source is more uniform; on the contrary, its support is not very good for the source data with decentralized mining results.

A. Savasere et al. proposed a partition-based optimization algorithm Partition[12]. It is well known that when the source data is too large, it may not be able to be put into main memory at one time to execute the algorithm, and the efficiency will be low even if it is allowed. The algorithm divides the source data into several disjoint blocks, mining frequent itemsets for only one block at a time, and then merges them to calculate the support value. This not only effectively solves the problem of capacity, but also theoretically theoretically the mining algorithm can be performed on multiple blocks in parallel and improve efficiency.

Manmila et al.[13] and Toivonen et al.[14] used the idea of sampling for the minging of frequent itemsets. That is, selecting a subset of the source data as a sample for mining analysis according to a sampling strategy, which reduces the amount of computation effectively. However, the results are largely influenced by the sampling strategy and the sample distribution, which ultimately results in an approximation rather than an exact value.

With the development of Internet technology, people's social life and online shopping and other activities become more frequent, which results in a large amount of data. In response to the increasing amount of data, Agrawal et al. proposed the Count Distribution algorithm and the Data Distribution algorithm in parallel mode[17], the principle of which is similar to the Apriori algorithm. The principle of Count Distribution algorithm is to divide the candidate itemset into blocks first, so that they can be allocated to different modules for mining, and the various modules can exchange the support value count during the mining process. But the main problem is the low memory utilization. The Data Distribution algorithm effectively improves the problem under the premise of good communications between nodes.

2.2. FP-Growth algorithm and its optimization algorithm

The FP-Growth algorithm[15] is another basic algorithm proposed by J. Han et al., which is completely different from Apriori. In order to support this algorithm, it constructs a unique data structure FP-Tree, which is a relatively small tree data structure. Through this data structure a high degree of compression and generalization of the source data can be achieved. That is, the transaction data set is compressed to a frequent mode tree FP-Tree, then the data mining problem is transformed into a tree structure mining problem, and the original disk mining problem is transformed into a memory mining problem, which effectively reduces the time overhead of disk I/O when scanning data repeatedly.

Compared with the Apriori algorithm, the FP-Growth algorithm performs better in most cases, but from the principle of FP-Growth algorithm, it can be seen that the key of FP-Growth algorithm is to compress the original transaction data set. The algorithm efficiency increases with the increase of the compression ratio. Once the compression ratio is too low, the generated FP-Tree is very lush, and the efficiency of the algorithm will be significantly declined. In addition, the FP-Growth algorithm takes a long time to compress the source data set by constantly scanning the entire data set and building the FP-Tree in memory. When the original transaction data set is particularly large, the generated FP-Tree may not be all saved to memory, resulting in the algorithm does not work.

Aiming at the problem of the FP-Growth algorithm, Y.G. Sucahy et al. improved it and designed a high compression frequent pattern tree CT-Tree, which has better compression performance than the FP-Tree, and proposed the CT-PRO algorithm based on this point[16]. This algorithm effectively improves the mining efficiency by avoiding the recursive creation condition FP-Tree.

There are also two parallel improvement schemes based on the FP-Growth algorithm. One is the PFP-tree algorithm proposed by Javed et al.[18], which is based on distributed memory for parallelization improvement. The other is an improved MLFPT algorithm using shared memory[19] proposed by Osmar R. Zaiane et al.

3. Frequent itemsets mining algorithms in data stream

3.1. Data stream based mining

For the past few decades, batch processing has been the focus of data mining technology in research, and the data sets to be processed by algorithms in this model are limited. All data can be obtained as the algorithm is executed, and usually before the result can be obtained, the algorithm will repeatedly scan the data set multiple times. When facing large amounts of data, many methods attempt to sample the data to get a small processable data set, but

this approach requires that the data set must conform to some fixed statistical distribution and the size of the data set is fixed.

With the research and development of computer technology in recent years, more and more means of communication such as GPS data, mail, sensor networks, mobile phone communication are becoming more and more popular. so the data mining algorithms are facing more dynamic environment nowadays. Data in this environment comes endlessly over the time, forming a data stream.

Data stream is a large number of infinitely variable data sequences that arrive continuously and change at any time. Because it is different from the traditional data, the data mining for data stream has the following characteristics[3-6]:

- Because the data stream is constantly arriving and the total amount is unlimited, it cannot be all stored in database or on disk devices as the traditional data. Therefore the data stream mining must be performed in real time.
- The total amount of data is unknown, so the mining strategy can not be improved and changed in advance according to the size of total data amount.
- Because the data stream arrives constantly and the data is updated all the time, the accurate result of a set of data can not be obtained at one time as the traditional data, and the mining result of the data stream is always approximate.
- The data in a data stream is not reproducible, and since the data stream cannot be stored, there is no guarantee of multiple access to the same data.

3.2. Structure Frequent itemsets mining algorithms in data stream

In the previous section, the classic data mining algorithm for traditional data and the improved solutions to increase its efficiency are introduced,. including some parallel mining solutions. However, due to the inherent characteristics of the data stream, not all the obtained data can be stored, the original data mining algorithm is not well suited to the data stream. In order to solve this problem, domestic and foreign researchers have proposed the sequence-based and time-based window models, and implemented the algorithms based on two models, which are well applied to the frequent itemsets mining based on data stream.

Giannella et al.[21] proposed an FP-Stream data structure based on the FP-Growth algorithm to keep frequent patterns in memory. The algorithm performs piecewise batch processing on data and saves the historical data in the FP-Stream by using a tilt time window. The advantage of this algorithm is that historical data can be queried at any time, but it also brings about such problems as large cost of new data processing, large number of intermediate results, difficult to expand the algorithm and so on, and it is difficult to adapt to the environment of high-speed data stream.

G. S. Manku et al. proposed the Sticky Sampling and Lossy Counting algorithm[22]. The algorithm only needs to scan each transaction once, and during the process, the algorithm count the itemsets whose occurrence frequency exceeds the specified threshold. The algorithm results in an approximation of the error within the user-acceptable range. The Lossy Counting algorithm accepts two user input parameters: one is the support value threshold s, which satisfies $s \in [0,1]$; the other is the error value ε, which satisfies $\varepsilon \in [0,1]$ and $\varepsilon \ll s$. The algorithm conceptually divides the arriving data stream into different blocks and assigns an ID to each data block starting from 1. For a data item e, assume that it has a frequency of fe in the data stream so far. Frequency items are stored in a data structure T, T is a set, and each item in the set contains three elements (e, f, Δ), where e is the data item itself, f is the estimated frequency of the item, and Δ is the maximum error of the estimated frequency. The advantage of this algorithm is that the list of estimated frequency of an itemset can be outputted at any time. However, the outputted result of this algorithm is an approximation with a controllable error. At the same time, the algorithm is insensitive to time, neglects the fact that transaction weights will change in different periods, and occupies more memory as the execution time of the algorithm increases.

Y. Chi et al. proposed a transaction-sensitive frequent itemset mining algorithm based on sliding window: Moment[23], which uses the memory-based prefix tree structure CET (Closed Enumeration Tree) to achieve the maintenance work of the frequent itemsets and potential frequent itemsets in the current window. In general, the

itemsets in the sliding window do not change their original state due to the sliding of the window, and only a small number of itemsets change from infrequent to infrequent or from infrequent to infrequent. For nodes with unchanged state, what needs to be done is only to modify the count of support values, and the itemsets with changed state must be on the boundary of the tree. The main advantage of the Moment algorithm is the ability to obtain the accurate results for any frequent itemsets in the current window. Later A-Moment algorithm[24] was also proposed to improve the mining efficiency of the Moment algorithm by means of approximate induction.

4. Concluding remarks

This paper mainly introduces the research of frequent itemsets mining in data mining technology. With the continuous expansion of data sets, traditional data analysis methods are difficult to study them accurately and efficiently, so the value of data mining has been constantly highlighted. Therefore, data mining technology is currently undertaking the task of modeling and prediction, correlation analysis, cluster analysis and anomaly detection. The important research topic in the field of data mining is the association rules mining. After this problem was raised by Agrawal et al., more and more improved algorithms based on Apriori and FP-Growth algorithms emerged. Some researchers began to improve the FP-Tree structure, and others used hashing, sampling and dividing methods for improvement. There are also many researchers who expanded the algorithm from the perspective of parallel computing, so that the algorithm no longer relies on the computation performance of a single node. In recent years, the rapid development of Internet and communication technology has led to continuous arrival and change of data, and the data mining algorithms need to adapt to the dynamic environment gradually. Therefore, the research of frequent itemsets mining algorithms based on data stream has drawn wide attention.

References

1. Jiawei H,Kamber M. Data mining: concepts and techniques.San Francisco, CA,itd:Morgan Kaufmann, 2001, 5.
2. Fayyad U M,Piatetsky-Shapiro G,Smyth P,et al.Advances in knowledge discovery and data mining.The MIT Press,1996.
3.PAN Yun-he, WANG Jin-long, XU-congfu. State-of-the-art on frequent pattern mining on data streams. Acta Automatica Sinica, 2006,32(4):594–602.
4. Li H F,Lee S Y. Mining frequent itemsets over data streams using efficient window sliding techniques.2009,36(2):1466–1477.
5. Li H F,Lee S Y,Shan M K. An efficient algorithm for mining frequent itemsets over the entire history of data streams. First International Workshop on Knowledge Discovery in Data Streams.2004.744-746(1):129-132.
6. Tan P N,Steinbach M,Kumar V.Introduction to data mining[M].Beijing:Pearson Education Ltd., 2006.
7. Agrawal R,Imieliński T,Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD international conference on Management of data. 1993. pp.207-216.
8. Agrawal R,Srikant R,others.Fast algorithms for mining association rules.Proceedings of the 20th International Conference on Very Large Data Bases. 1994. pp.487-499.
9. Park J S,Chen M S,Yu P S.An effective hash-based algorithm for mining association rules. ACM,1995.
10. Park J S,Chen M S,Yu P S.Efficient parallel data mining for association rules. Proceedings of the fourth international conference on Information and knowledge management.1995:31–36.
11. Brin S,Motwani R,Ullman J D,et al. Dynamic itemset counting and implication rules for market basket data. ACM SIGMOD Record.1997:255–264.
12. Savasere A,Omiecinski E R,Navathe S B.An efficient algorithm for mining association rules in large databases.Georgia Institute of Technology,1995.
13. Mannila H,Toivonen H,Verkamo A I.Efficient algorithms for discovering association rules. AAAI workshop on Knowledge Discovery in Databases.1994:181–192.
14. ToivonenH,et al. Sampling large databases for association rules. 22th International Conference on Very Large Data Bases. 1996:134–145.
15. Han J,Pei J,Yin Y.Mining frequent patterns without candidate generation.ACM SIGMOD Record.2000:1–12.
16. Sucahyo Y G,Gopalan R P.CT-PRO: A Bottom-Up Non Recursive Frequent Itemset Mining Algorithm Using Compressed FP-Tree Data Structure.IEEE Icdm Workshop on Fimi .2004: 212–223.
17. Agrawal R,Shafer J C.Parallel mining of association rules[J].1996(6): 962–969.
18. Javed A,Khokhar A.Frequent pattern mining on message passing multiprocessor systems.Distributed & Parallel Database.2004,16(3):321–334.
19. Zaïane O R,El-Hajj M,Lu P.Fast parallel association rule mining without candidacy generation. IEEE International Conference on Data Mining. 2001:665–668.
20.Woo J.Apriori-Map/Reduce Algorithm.The International Conference on Parallel and Distributed Processing Techniques and Applications.2012.pp.1-5.

21. Giannella C,Han J,Pei J,et al.Mining frequent patterns in data streams at multiple time granularities. Data Mining Next Generation Challenges & Future Directions. 2003,212:191–212.
22. Manku G S, Motwani R. Approximate frequency counts over data streams. 28th international conference on Very Large Data Bases. 2002:346–357.
23.Chi Y,Wang H,Yu P S,et al.Moment: Maintaining closed frequent itemsets over a stream sliding window. Fourth IEEE International Conference on Data Mining. 2004:59–66.
24. LIU Xu, MAO Guo-jun, Sun Yue, LIU Chun-nian. An algorithm to approximately mine frequent closed itemsets from data streams. Acta Electronica Sinica, 2007.35(5):900–905.