Scientific Research Publishing

Open Journal of Social Sciences, 2021, 9, 458-468
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Research and Case Analysis of Apriori Algorithm Based on Mining Frequent Item-Sets

Haoyu Xie

School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, China
Email: *20173302@aufe.edu.cn

How to cite this paper: Xie, H. Y. (2021). Research and Case Analysis of Apriori Algorithm Based on Mining Frequent Item-Sets. *Open Journal of Social Sciences, 9,* 458-468.
https://doi.org/10.4236/jss.2021.94034

Received: March 15, 2021
Accepted: April 26, 2021
Published: April 29, 2021

Abstract

In the background of the information age, the importance of data resources can be imagined, and the importance of the use of data resources means—data mining has also emerged. In the current situation, all industries are in a relatively equal stage, should make good use of data resources, use Apriori algorithm to mine association rules, formulate marketing strategies, promote sales growth and slow down the loss of national GDP. Countries can also use data mining to make predictions and support decisions. Briefly describe the basic concepts of data mining and association rules. Only with a basic understanding of data mining and association rules can we better understand the Apriori algorithm, why it is one of the most classic and influential algorithms. Understanding its essence and function, and suggesting suggestions for improvement by analyzing the ideas of the Apriori algorithm, understanding its process, advantages and disadvantages, and through the data collected from a supermarket in Bozhou City, Anhui Province, China; and understanding the application of the algorithm in the field of e-commerce and universities, and the effect it will achieve.

Keywords

Apriori Algorithm, Algorithm Examples And Applications, Association Rules, Data Mining

1. Introduction

1.1. Research Background and Purpose

China has fully entered the information age. Under the influence of the situation, this development has been promoted. People who did not have much contact with the Internet before have to start working online. And this is China, which has the largest number of netizens, and it generates more and more data resources. If used properly, it will generate more explicit or hidden benefits for China and reduce the GDP loss caused by the situation.

Data mining is a tool for obtaining knowledge and valuable information. Making good use of association rules and related algorithms can help the country make better predictions, provide decision support, and it is also a means to avoid outbreaks. In other industries, you can also formulate sales strategies based on customer needs to maximize your sales.

This topic is aimed at the Apriori algorithm, combined with related data, analyzes the algorithm, intersperses with my own understanding, and provides some suggestions on the improvement of the Apriori algorithm and its application in life.

1.2. Research Status in China and Abroad

The Apriori algorithm is an algorithm for mining association rules developed by Agrawal and Krishnan in 1994. After that, a variety of association rule algorithms improved based on the Apriori algorithm have also been produced.

The FP-growth algorithm proposed by Professor Han Jiawei in 2000 is one of them. It only needs to scan the database twice, which greatly saves the running time of the algorithm and does not generate candidate sets. Compared with the Apriori algorithm, the algorithm has a great improvement in algorithm efficiency, but it cannot find the association rules between data.

In recent years, Chinese research and applications on this issue have shown results in many fields. Such as, the personalized push in e-commerce which can satisfy consumers while expanding revenue; improvement of teaching work in colleges and universities; in terms of medical treatment, the degree of association between the disease and various factors, etc.

As shown in Figure 1, Figure 2, the visual analysis data obtained from CNKI can be seen that Chinese research in the Apriori algorithm field has grown rapidly since 2000 and reached its peak in 2009, but the research has not declined. The trend continues until now. In these years, the University of Electronic Science and Technology of China has published a large number of documents, and other universities have published a large number of related papers.

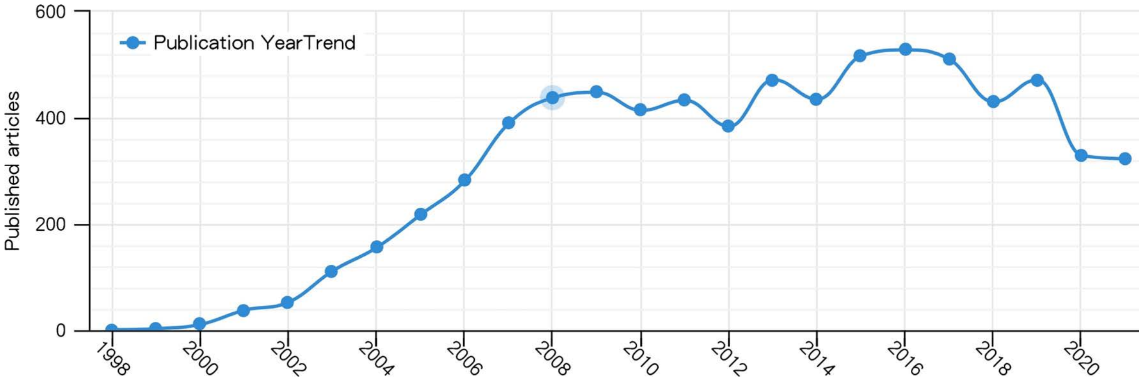


Figure 1.Recent trends of Chinese "Apriori algorithm".

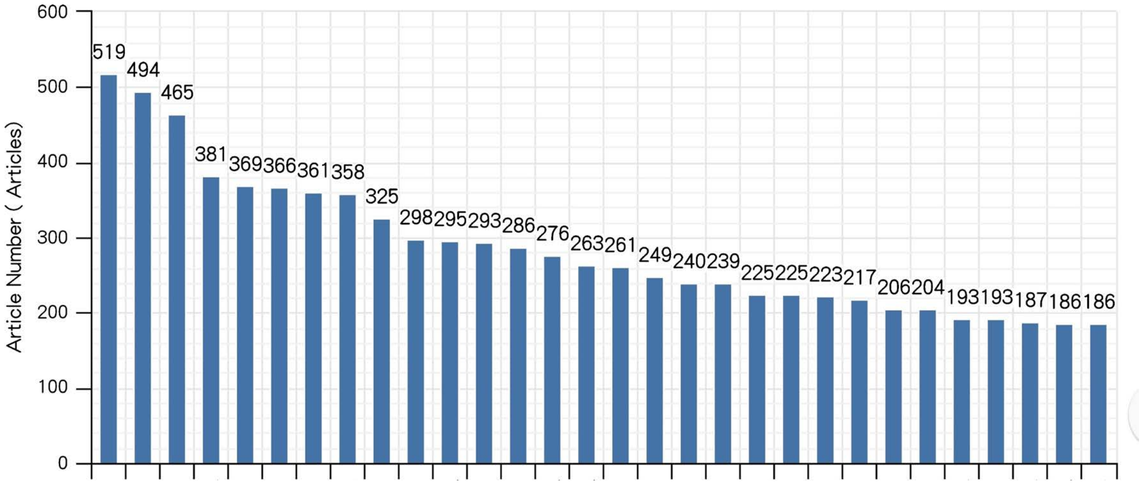


Figure 2.Number of documents published by Chinese universities on "Apriori algorithm".

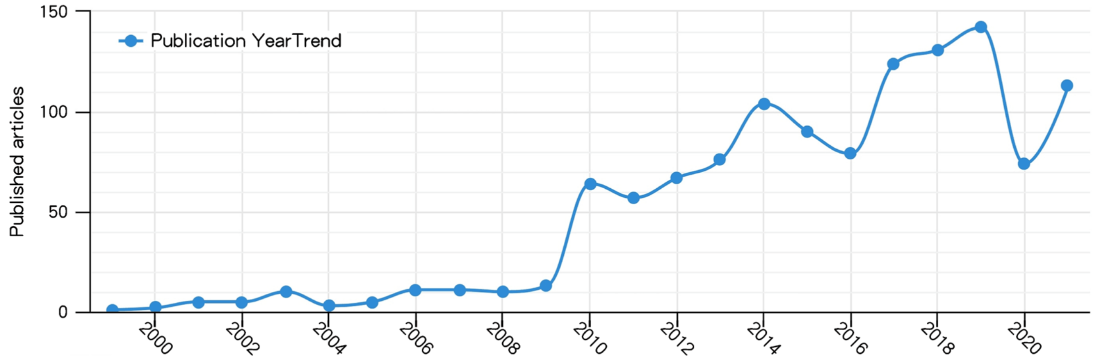


Figure 3.Recent Trends of "Apriori Algorithm" Abroad.

Following the proposal of the Apriori algorithm, many researchers abroad have successively proposed improved methods for the Apriori algorithm. For example, DHP algorithm and PCY algorithm. As shown in Figure 3 and compared with Figure 1, we can understand that the research on Apriori algorithm abroad is basically the same as that in China, but until 2009, the research on this has been tepid, with low attention, which is the opposite of Chinese research. After 2009, foreign research on the Apriori algorithm has gradually become hot topic, and there will be a growing trend in 2020.

1.3. Main Research Contents

This paper gives an explanation of the related concepts of data mining and association rules, and gives a brief description of the process and steps of the implementation of data mining and association rules, in order to better understand the process and related concepts of Apriori algorithm and the nouns in it.

In addition, this topic deeply analyzed the steps, properties, advantages and disadvantages of Apriori algorithm, and gives relevant examples to better explain the steps and implementation process of Apriori algorithm, and talk about the application of Apriori algorithm in some fields, as well as the results achieved.

2. Related Theories of Data Mining and Association Rules

2.1. Overview of Data Mining

Data mining is a process of digging out deep and valuable information for data users from large amounts of irregular data. Data mining is one of the important applications of database. Its function is to discover the hidden information of data.

Some people equate data mining with knowledge discovery, but the information from data mining may not necessarily form knowledge. In this respect, knowledge discovery can be regarded as the next operation or core component of the excavated data, collating, identifying and summarizing the excavated information, and finally forming knowledge.

The process of data mining:

First, select data: according to the actual needs, select the required data from the database.

Second, data pretreatment: simple processing of selected data, noise reduction, reduction of data redundancy and useless data, etc.

Third, data selection: according to the desired information, select the data that are simply processed.

Fourth, data transformation: the selected data is changed into Boolean data or unified into a data form. Boolean data is more suitable for the relevant algorithms of data mining.

Fifth, select the data mining method and algorithm: according to the target data type of mining, choose the appropriate mining method. Then according to the selection of data mining methods, combined with the characteristics of the data and actual needs, select the best data mining algorithm, such as Apriori algorithm, K-means algorithm.

Sixth, data mining: get hidden information.

Seventh, evaluation of data mode: delete the unnecessary information and leave the needed information through evaluation.

Eighth, representation of knowledge: make visual analysis and analysis of the information left behind to form easy-to-understand knowledge (Lin, 2017).

Method of data mining: The analysis methods of data mining mainly include the following five kinds: clustering analysis, classification and prediction, outlier analysis and association rule analysis (Liu, 2004).

2.2. Overview of Association Rules

Association rules reflect the interdependence and correlation between the two, and are used to dig out the correlation between valuable data items from a large number of data. Its purpose is to find frequent item-sets and strong association rules.

A typical case of association rules is the “Diapers and beer”. As shown below, in some supermarkets in the United States, when young fathers with new children went to buy diapers after work, nearly half of them also bought beer. Some stores that put diapers and beer next to each other had significantly higher sales than those that did not, and their sales have also increased after these supermarkets put diapers and beer next to each other (Jiao, 2013).

This is an application of association rules—shopping basket analysis. According to the goods in the customer's shopping basket, the rules are found and frequent item-sets is generated, that is, which goods will be purchased by the customer several times at the same time. These association rules can be used in merchants' marketing strategies, such as product promotion, product region division, etc.

3. Apriori Algorithm

It is an algorithm based on mining Boolean association rules. After each set of frequent item-sets is generated, the whole database is scanned and the association rules between data are mined from the generated frequent item sets, give us decision support.

3.1. The Idea of Apriori Algorithm

An item-set is a set of 0 or more items, and a frequent item-set is an item-set whose support is greater than the custom minimum support count (Shabtay et al., 2021).

Common evaluation criteria for frequent item-sets:

1) Support: It is one of the two basic parameters of association rules. It is the ratio of the number of transactions containing both x and y in All sample of dataset D to all transactions. If we have two data x and y that need to be analyzed for correlation, then the corresponding support degree is:

$$\text{Support}(x, y) = P(xy) = \frac{\text{num}(xy)}{\text{num}(\text{All Samples})}. \tag{1}$$

$$\text{Support}(X \Rightarrow Y) = P(X \cup Y) = \frac{\text{count}(X \cup Y)}{|D|}. \tag{2}$$

For example, a support rating of 28% means that "there is a 28% probability that an individual in the population will contain both X and Y".

2) Confidence: It is the ratio of the number of transactions including x and y to the number of transactions including y, namely conditional probability.

$$\text{Confidence}(x \Rightarrow y) = P(x \mid y) = \frac{P(xy)}{P(y)}. \tag{3}$$

$$\text{Confidence}(X \Rightarrow Y) = P(X \mid Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}. \tag{4}$$

Assuming that "52% of the terms containing X contain Y", the confidence is 52%.

3) Lift: It is the ratio of the number of transactions containing x under the premise of including y to the total number of transactions occurring in x.

$$\text{Lift}(x \Rightarrow y) = \frac{P(x \mid y)}{P(x)} = \frac{\text{Confidence}(x \Rightarrow y)}{P(x)}. \tag{5}$$

Lift uses 1 as the target value to show the relationship between x and y. If the value is greater than 1, then $x \Rightarrow y$ is a valid strong association rule. Conversely, $x \Rightarrow y$ is an invalid strong association rule. When the value is equal to 1, however, there is a special case, that is, the x and y at independence, at the time $P(x \mid y) = P(x)$, so $\text{Lift}(x \Rightarrow y) = 1$ (Zhou et al., 2010).

Only custom minimum support, or a combination of custom support and confidence, can determine the frequent item-sets in the database.

Its core is to retrieve all frequent item-sets, and find all item-sets that are greater than or equal to the support by setting the minimum support count and iterating continuously.

3.2. Steps of Apriori Algorithm

The Apriori algorithm consists of two steps: connecting and pruning.

Connecting: The target is L_k (k is a constant). By connecting the sets in L_{k-1}, a set of candidate k item-sets, namely C_k, is generated. The condition that the two elements l_1 and l_2 in L_{k-1} can perform the concatenation operation $l_1 l_2$: the first k − 2 items of the candidate set in C_k are the same, and they are combined according to the lexicographic order, that is,

$$(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \cdots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-l) < l_2[k-1])$$

Pruning: $L_k \subset C_k$, C_k is a superset of L_k, C_k contains all frequent item-sets, but not all frequent item-sets in C_k. Therefore, it is necessary to scan the entire database, calculate the support of each k item-set, and retrieve L_k (Qian, 2006).

The steps of code design are as follows:

Input: Data set D, minimum support count α.

Output: the largest frequent k item-set.

1) Scan the entire database, arrange all the data in the data set, and get C_1, which is the candidate frequent 1 item-set. $k = 1$, frequent 0 item-sets are empty sets.

2) Mining frequent k item-sets:

a) Scan the database and filter the data sets larger than α.

b) Remove the item-sets whose support degree is lower than α in C_k, and get L_k, that is, frequent k item-sets. If L_k is an empty set, the result of the algorithm is L_{k-1}; otherwise, if there is only one item in the L_k, the item is the result of the algorithm. The algorithm ends.

c) When the item-set in L_k has two or more items, the connection generates C_{k+1}, and the algorithm continues.

3) Let $k = k + 1$, repeat step 2.

The code steps make the disadvantages of Aprior algorithm, scanning the database every iteration, appear on paper. This also leads to the low efficiency of the algorithm code when the database is large and there are many data to be mined (Zhu, 2014).

3.3. Properties of Apriori Algorithms

The two properties of Apriori algorithm can effectively reduce the search complexity of the algorithm, reduce the amount of computation, and improve the efficiency of mining frequent item-sets.

First, if X is a frequent item-set, then all subsets of X are frequent item-sets.

Second, if X is a non-frequent item-set, then all supersets of X are non-frequent item-sets (Yu et al., 2010).

3.4. The Pros and Cons of Apriori Algorithm

The space complexity of the algorithm is small, and the coding difficulty of the programmer is reduced. Through the nature of the algorithm and pruning operation, many repeated operations can be avoided, and the running speed of the algorithm is improved. If a given minimum support threshold is large, the number of database scans will be significantly reduced. It is suitable for sparse data, and has high ductility and high exploitability (Wang, 2010).

If there is more data in the transactional database, the number of items in the candidate set will increase exponentially, which will consume more memory and create a burden. However, when the candidate set is pruned, the database must be scanned, which will consume more time and reduce the efficiency of the algorithm.

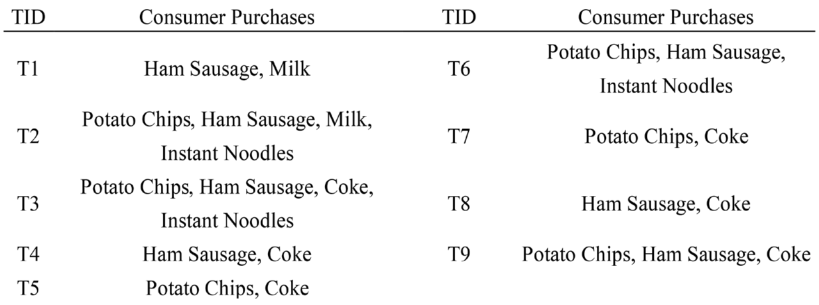
This results in an exponential decline in the efficiency of the algorithm as the size of the database increases.

In order to solve these problems, many algorithms are derived to improve the mining efficiency based on Apriori: FP-growth algorithm, hash-based technology, thing reduction technology, partition technology, etc. (Song, 2010).

3.5. Example of Apriori Algorithm

Figure 4 shows the purchase records of some consumers extracted from the transaction database of a retail store in Bozhou County, Anhui Province, China, so that the minimum support count (min_support) is 2. The process of mining frequent itemsets is as follows:

The supermarket consumer purchase record database D is shown in Figure 4.



TID	Consumer Purchases	TID	Consumer Purchases
T1	Ham Sausage, Milk	T6	Potato Chips, Ham Sausage, Instant Noodles
T2	Potato Chips, Ham Sausage, Milk, Instant Noodles	T7	Potato Chips, Coke
T3	Potato Chips, Ham Sausage, Coke, Instant Noodles	T8	Ham Sausage, Coke
T4	Ham Sausage, Coke	T9	Potato Chips, Ham Sausage, Coke
T5	Potato Chips, Coke		

Figure 4. Supermarket consumer purchase record database D.

Step 1: scan the entire dataset to get I = {Potato Chips, Ham Sausage, Coke, Milk, Instant Noodles}, and all the data obtained at this time is used as the 1 candidate set C_1. Compare the support count of all items with min_support, and strip out the data that is less than the threshold. In this case, all items should be greater than or equal to the min_support, and the frequent item-set L_1 should be obtained.

Candidate set C_1 and frequent item-set L_1 are shown in **Figure 5**.

Step 2: use the items left by item-set L_1 to perform the connection operation I_1I_2. And scan the whole data set again, count the data of two projects at the same time, calculate the degree of support, and form 2 candidate set C_2. Similar to step 1, compare min_support and strip out unqualified item-sets (hereinafter referred to as "compare and deletion") to obtain frequent item-set L_2.

Candidate set C_2 and frequent item-set L_2 are shown in **Figure 6**.

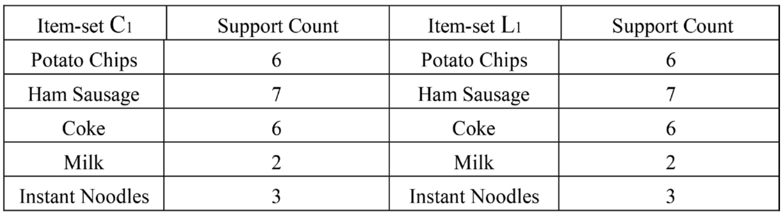
Step 3: use the items left by item-set L_2 to perform the connection operation, then the project contains three commodities, and scan the whole data set again, count the data of three items at the same time, calculate the degree of support, and form three candidate sets C_3. At this time, C_3 contains many non-frequent item-sets, which need to be pruned to form candidate set P_3, which is compared and deleted to get frequent item-set L_3.

Candidate set C_3 is shown in **Figure 7**, candidate set P_3 and frequent item-set L_3 are shown in **Figure 8**.

Step 4: repeat the above steps, after C4 pruning, P4=∅. The algorithm ends, all frequent item-sets are retrieved.

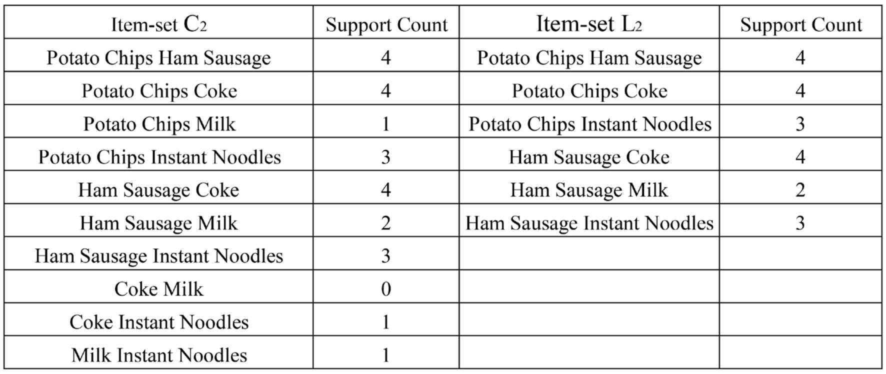
3.6. Application of Apriori Algorithm

Apriori algorithm is involved in various fields, including e-commerce, university



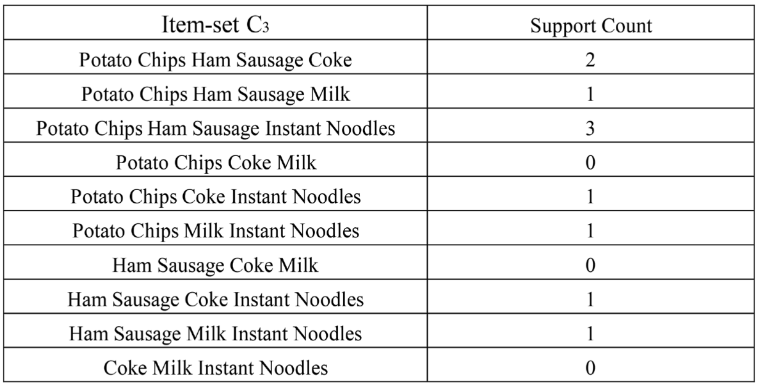
Item-set C_1	Support Count	Item-set L_1	Support Count
Potato Chips	6	Potato Chips	6
Ham Sausage	7	Ham Sausage	7
Coke	6	Coke	6
Milk	2	Milk	2
Instant Noodles	3	Instant Noodles	3

Figure 5. Candidate set C_1 and Frequent item-set L_1.



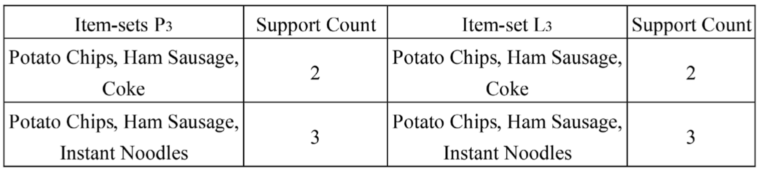
Item-set C_2	Support Count	Item-set L_2	Support Count
Potato Chips Ham Sausage	4	Potato Chips Ham Sausage	4
Potato Chips Coke	4	Potato Chips Coke	4
Potato Chips Milk	1	Potato Chips Instant Noodles	3
Potato Chips Instant Noodles	3	Ham Sausage Coke	4
Ham Sausage Coke	4	Ham Sausage Milk	2
Ham Sausage Milk	2	Ham Sausage Instant Noodles	3
Ham Sausage Instant Noodles	3		
Coke Milk	0		
Coke Instant Noodles	1		
Milk Instant Noodles	1		

Figure 6. Candidate set C_2 and Frequent item-set L_2.



Item-set C_3	Support Count
Potato Chips Ham Sausage Coke	2
Potato Chips Ham Sausage Milk	1
Potato Chips Ham Sausage Instant Noodles	3
Potato Chips Coke Milk	0
Potato Chips Coke Instant Noodles	1
Potato Chips Milk Instant Noodles	1
Ham Sausage Coke Milk	0
Ham Sausage Coke Instant Noodles	1
Ham Sausage Milk Instant Noodles	1
Coke Milk Instant Noodles	0

Figure 7. Candidate set C_3.



Item-sets P_3	Support Count	Item-set L_3	Support Count
Potato Chips, Ham Sausage, Coke	2	Potato Chips, Ham Sausage, Coke	2
Potato Chips, Ham Sausage, Instant Noodles	3	Potato Chips, Ham Sausage, Instant Noodles	3

Figure 8. Candidate set P_3 and Frequent item-set L_3.

information system, big data platform and so on.

Application of apriori algorithm in e-commerce: In the aspect of e-commerce, through mining association rules for a large amount of data, we can achieve the purpose of understanding customers' consumption habits, and can more effectively carry out market analysis, market prediction and the selection of target customers to achieve personalized service, so that enterprises can carry out targeted sales activities and improve sales efficiency and market competitiveness (Wu, 2012).

Application of Apriori algorithm in universities: The application of Apriori algorithm in colleges and universities is analyzed in two points. First, it is the use of association rules algorithm, including Apriori algorithm in teaching evaluation. By using the data obtained from the students' evaluation of the curriculum at the end of the semester, the association rules are analyzed from different factors, such as the interaction between students and teachers, the use of multimedia courseware, teaching content and style, to help the teaching management department to make decisions, and to provide feedback to substitute teachers to help teachers improve their teaching quality and efficiency (Zhang, 2017).

The second is the employment field of graduates. Now colleges and universities advocate personalized autonomous learning, school elective courses and Chinese universities MOOC and other online learning platforms, but also provide teaching services, college students can learn what they are interested in both online and offline. Therefore, in addition to their majors, college students also have other skills. Through correlation analysis, they can push more and more employment opportunities for graduates who are about to find jobs.

Third, the data mining of students' life behavior, because of students' self-esteem, some poor students will give up applying for poverty grants, so they cannot achieve a good financial aid effect. However, according to the credit card records of students' campus cards, we can know the living standards of students and subsidize them through these data.

The fourth is the application in the detection of academic misconduct. By scanning the articles and judging the repetition rate, we can avoid academic fraud and ensure the quality of scientific research work.

4. Conclusion

Nowadays, the related algorithms of association rules have been successful in many fields. Through this topic, we can have a deeper understanding of the specific process and core of Apriori algorithm, and better understand the value of association rules. The main work of this paper includes the following points.

The main contents are as follows:

1) The basic theories of data mining and association rules are briefly described and the related processes of them are analyzed.

2) The Apriori algorithm is analyzed deeply. Through the example of mining association rules, the relevant steps and achievable results of the algorithm are explained.

3) The application of Apriori algorithm is simply analyzed, such as mining potential customers in e-commerce and teaching evaluation analysis of colleges and universities.

4) Suggestions for improving the Apriori algorithm: first, by sorting the generated candidate set, the subsequent candidate set can be formed by scanning the previous candidate set and ending when the min_support threshold is scanning. Pruning can be carried out to reduce the amount of computation and improve the efficiency of the algorithm; second, when scanning the database, remove the data of the items whose number of items is less than the candidate set, that need to be checked to improve the efficiency of scanning the database.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Jiao, D. W. (2013). *Research and Improvement of an Association Rule Algorithm*. Ph.D. Thesis, Changchun: Changchun University of Science and Technology.

Lin, P. (2017). *Research and Application of Two Improved Key Rule Mining Algorithms*. Ph.D. Thesis, Hefei: Hefei University of Technology.

Liu, H. (2004). *Research on the Theory and Method of Knowledge Discovery and Its Application*. Ph.D. Thesis, Dalian: Dalian University of Technology.

Qian, D. Y. (2006). *Research on Association Rules Algorithm in Data Mining*. Ph.D. Thesis, Tianjin: Tianjin University.

Shabtay, L., Fournier-Viger, P., Yaari, R., & Dattner, I. (2021). A Guided FP-Growth Algorithm for Mining Multitude-Targeted Item-Sets and Class Association Rules in Imbalanced Data. *Information Sciences, 553,* 353-375. https://doi.org/10.1016/j.ins.2020.10.020

Song, X. D. (2010). *Research on Key Technologies of Enterprise Group Data Warehouse System.* Ph.D. Thesis, Dalian: Dalian University of Technology.

Wang, D. (2010). *Research and Implementation of Apriori Algorithm for Association Rules Based on Cloud Computing*. Ph.D. Thesis, Nanchang: Nanchang University.

Wu, W. B. (2012). *Apply the Data Mining Technology of Apriori Association Rule Algorithm to Mine E-Commerce Potential Customers*. Ph.D. Thesis, Zhejiang: Zhejiang University of Technology.

Yu, M., Hu, M., King, K., Hu, L., & Zhao, K. (2010). Telecom Network Alarm Application of Association Rule Algorithm. *Journal of Jilin University (Information Science Edition), 28,* 264-269.

Zhang, Q. H. (2017). *Research on Association Rules Algorithm for Big Data*. Ph.D. Thesis, Xi'an: Xi'an University of Science and Technology.

Zhou, J. Z., Liu, J., Yu, C. Y. (2010). Research and Application of Data Mining Based on Web Log. *Science, Technology and Engineering, 10,* 2762-2766.

Zhu, Y. (2014). *Research on the Improvement of Apriori Algorithm Based on Array*. Ph.D. Thesis, Harbin: Harbin normal University.