Research on Shopping Mall Sales Based on Apriori Association Rule Mining Algorithm

Quan Jiang[1], Ming-le Ma[2], Liang-yu Dong[3*]
2510975497@qq.com, mamingle1129@qq.com, 35516722@qq.com

[1]College of Science, Dalian Jiaotong University, Dalian, China

[2]College of foreign languages, Dalian Jiaotong University, Dalian, China

[3]College of Computer and Communication Engineering Dalian Jiaotong University Dalian, China

Abstract—The placement of goods in a shopping mall may seem insignificant, but in fact it has a significant impact on sales and customer experience. In order to improve the sales of goods in shopping mall, taking the sales data set of shopping malls as the research object, the Apriori algorithm based on association rule mining is used to carry out association analysis on the sales data of shopping malls, and the algorithm model is established according to the minimum support and minimum confidence. Comprehensive analysis of which goods should be adjusted to increase sales and improve customer happiness index. Experimental results show that the Apriori algorithm is used to analyze the results, compared with the original display of goods, the probability of customers to buy naturally increases by about 300%. The application of the model provides a solution for the shopping mall to improve sales and improve customer experience.

Keywords- Mall sales; Association rule mining; The Apriori algorithm; Correlation analysis

1 INTRODUCTION

In recent years, with the rapid development of science and technology, the term "big data" has become well known. In many fields such as society and technology, the amount of data has exploded. In the face of massive and complex data, people increasingly find that data is an important resource for human development, valuable data can promote enterprises, even the country and even the whole society to achieve rapid and sustainable development [1], but the proportion of data that can be used in the massive amount of data is very small, it is difficult for people to rely on "intuition" to find value from these data, so the analysis and mining of data is becoming more and more important.

In 1993, R. Agrawal et al. [2] first proposed the theory of association rule mining. In 1995, at the Annual Conference on Computers in the United States, the concept of data mining was proposed. There are many commonly used data mining methods, such as decision trees, Bayesian networks, and association rule mining [3]. Since its development, correlation rule mining has become an important branch of data mining, which is widely used in bioinformatics, finance, retail and Internet and other fields. JI Wenlu, WANG Hailong [4] and others

FFIT 2022, October 28-30, Shenzhen, People's Republic of China
Copyright © 2023 EAI
DOI 10.4108/eai.28-10-2022.2328459

mentioned in the review of the recommendation method based on the association rule algorithm that the recommendation algorithm based on the association rule can find the deep relationship between the recommended items, and the accuracy rate will continue to improve with the accumulation of data. The classic Apriori algorithm is commonly used to mine the association rules between data, which is an important method for association rule mining. CUI Yan and BAO Zhiqiang [5] elaborated on the Apriori algorithm in the review of thesis association rule mining, mentioning that the algorithm has good performance.

Nowadays, there are many kinds of goods, but customers will not increase their purchases because of the variety of goods, and the variety will cause customers to be tired of choosing. For some goods, customers often forget to buy, such as customers with children at home who buy the table often do not remember to buy the anti-collision sponge stuck in the corner of the table, and if the customer buys the table, the merchant recommends the anti-collision sponge to the customer, at this time, the customer will most likely buy the anti-collision sponge. Therefore, in order to obtain greater profits, understand what goods customers have purchased, and which goods will be purchased, analyze the sales data of the mall to find the law, which is of great research significance for increasing sales.

Based on the fact that the current merchants have increased sales to obtain higher profits, and customers have the purchase demand for related products, this paper mainly uses the Apriori algorithm mined by the correlation rules to explore the implicit laws between commodities and commodities [6]. By analyzing the data of a large number of customers to buy goods, we help customers make purchases, and at the same time optimize the layout of shopping mall goods, thereby helping shopping malls to improve profits.

2 PREPROCESSING OF DATA

Data mining is an intersectional emerging discipline that combines theories and technologies in statistics, database technology, machine learning, high-performance computers, artificial intelligence, and multiple fields [7]. When solving practical problems, it is necessary for professional database systems to provide effective technical support, excavate valuable data and information from different angles, and provide effective support for decision makers. But in the real world, most of the data obtained has incomplete, inconsistent dirty data, which makes it impossible to directly mine data or the results of mining do not meet the desired value. In order to improve the quality of the results produced by data mining, data preprocessing technology came into being.

Data preprocessing is the cleaning of raw data that has not yet been processed, the elimination of duplicate values, and so on. The data in this article comes from the sales database of an offline shopping mall. The purpose of data preprocessing of this data set is to transform the data into a form that meets the requirements of the Apriori association rule mining algorithm.

First, the dataset is imported into Python. According to the data set of customers as the standard for dividing the shopping basket, the goods purchased by the same customers are put into the corresponding shopping basket for subsequent association rule mining. Then the data exploration analysis of the sales dataset of the mall is carried out. This is done by importing the dataset into Python, then calling the duplicated and sum functions to check for the number of

duplicate values. After preliminary analysis, a total of 43367 data were obtained in the dataset, each data containing customer names and product names, and the number of duplicate values was 0, with a total of 9835 customers, 169 kinds of goods, and 9835 shopping baskets. Products include citrus fruits, margarine, instant soups, semi-finished bread, and more. The top 15 best-selling items in the mall are shown in Fig 1.

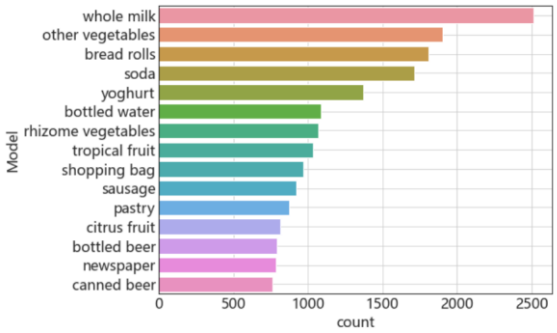

Fig 1. 15 best-selling items

3 ALGORITHM APRIORI

3.1 Definitions

Itemset. It is a collection of items. The itemset count is defined as. If itemset A contains k elements, then this itemset A is called a k -itemset. For example, a,b form a set $\{a,b\}$, where a,b are items and $\{a,b\}$ is itemset, which is called 2-itemset.

Support. Support for an association rule $A \to B$ refers to the probability that both the itemset A and the itemset B will occur at the same time. If the probability of the two appearing at the same time is large, it means that they are related, and vice versa is less related. Its mathematical expression is as follows.

$$support(A \to B) = P(A \bigcup B) \tag{1}$$

Confidence. The confidence level of the association rule $A \to B$ refers to the probability that the itemset B will occur if A occurs. Similar to support, the larger the value of the confidence level, the more it indicates that the two are intimate. Its mathematical expression is as follows.

$$confidence(A \to B) = \frac{P(A \bigcup B)}{P(A)} \tag{2}$$

If both $support = (A \to B)$ and $confidence(A \to B)$ meet the threshold of minimum support and confidence, the association rule is said to be a strong association rule. Support and confidence are two important indicators to measure association rules.

Lift. The lift of an association rule $A \rightarrow B$ is the ratio of the probability of the itemset B occurring under the condition that the itemset A occurs to the probability of the itemset B occurring alone. Lift greater than 1 and higher indicates a higher positive correlation between the two. Lift is less than 1 and the lower it is, the opposite is true. Its mathematical expression is as follows.

$$lift(A \rightarrow B) = \frac{P(B \mid A)}{P(B)} \tag{3}$$

Candidate itemset. The itemset obtained by downward merging. It is defined as C_k.

Frequent k-itemset. Itemset that meet the minimum support threshold of itemset A is called frequent k-itemset, that is, itemset with support greater than or equal to a specific minimum support, expressed as L_k.

3.2 Algorithm description

Apriori algorithm uses an iterative method of layer by layer search. Its core idea is to generate candidate items and their support through connection, and then generate frequent item sets through pruning [8]. The specific implementation process of the algorithm is as follows:

1) First scan the transaction database, count the number of occurrences of each item in the item set in the database, compare it with the set support, delete the items less than the support threshold, and the remaining items form a frequent item set, which is recorded as L_1.

2) Continue to iterate, and connect L_{k-1} with items including itself to generate candidate set C_k.

3) Prune C_k, scan the transaction database, count the support of each itemset in C_k, eliminate itemsets that do not meet the set support threshold, and obtain frequent k-itemset, which is recorded as L_k.

4) Repeat steps 1) and 2) until no new frequent itemsets are generated.

The algorithm flow chart is shown in Fig 2.

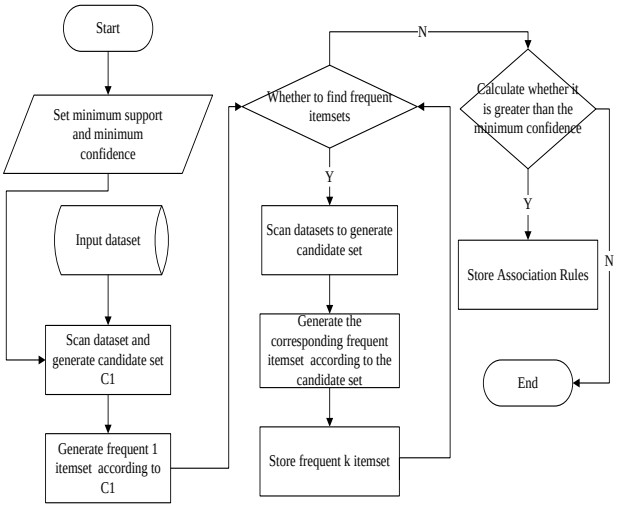

Fig 2. Algorithm flow chart

The following is a partial dataset of sales in a shopping mall.

Table 1 partial dataset of sales in a shopping mal

Customer id	***Commodity id***	***Customer id***	***Commodity id***
1	a, c, e	6	b, c
2	b, d	7	a, b
3	b, c	8	a, b, c, e
4	a, b, c, d	9	a, b, c
5	a, b	10	a, c, e

Assuming that the minimum support is 0.2, the specific process in which the Apriori association rule algorithm runs is shown in Fig 3.

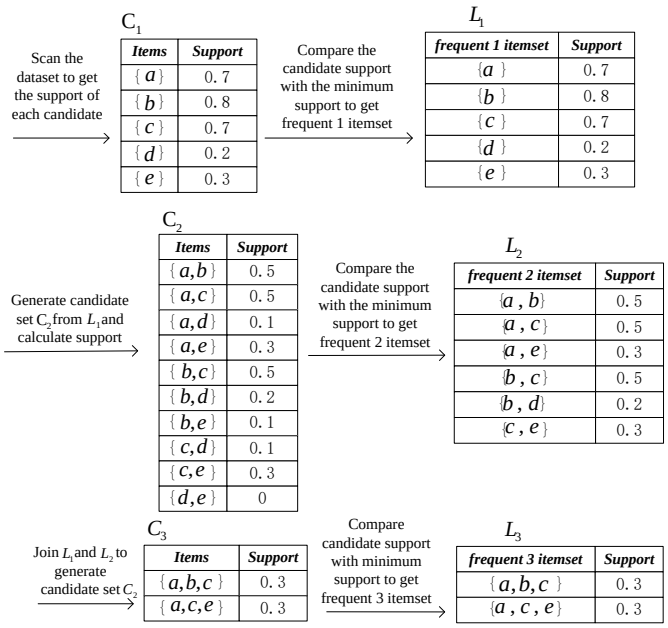

Fig 3. Algorithm implementation process

4 RESULTS OF APRIORI ALGORITHM

4.1 Result

The minimum support set in this paper is 0.01 and the minimum confidence is 0.1. The shopping mall sales data set is imported into python. The imported data set is analyzed by using the Apriori association rule algorithm. The results are as follows.

Table 2 complementary commodities

	Lhs	***Rhs***	***Support***	***Confidence***	***Lift***
1	Beef	Rhizome vegetables	0.017387	0.331395	3.040367
2	Curd	Sour cream	0.010473	0.196565	2.742150
3	Kernel fruits	Tropical fruit	0.020437	0.270161	2.574648
4	Butter	Sour cream	0.010168	0.183486	2.559698
5	Onion	Other vegetables	0.014235	0.459016	2.372268

In the table, Lhs refers to the product purchased by the user. Rhs refers to another product recommended according to the product purchased by the user. For example, the support degree of No. 1 in the table is 0.017387, which means that the probability of beef and rhizome vegetables appearing in a shopping ticket at the same time is 0.017387. The confidence level in No. 2 is 0.196565, which means that the probability of purchasing sour cream oil when purchasing curd is 0.196565. Lift refers to the degree of improvement, which is popularly

understood as that when merchant recommend rhizome vegetables to a customer on the premise of buying beef, the probability of the customer buying is about 300% of the probability of the customer naturally buying.

Table 3 mutually exclusive commodities

	Lhs	***Rhs***	***Support***	***Confidence***	***Lift***
1	Canned beer	Bread rolls	0.011286	0.145288	0.789888
2	Citrus fruits	Soda	0.012811	0.154791	0.887668
3	Soda	Rhizome vegetables	0.018607	0.106706	0.978964
4	Margarine	Soda	0.010168	0.173611	0.995607
5	Bottled beer	Whole milk	0.020437	0.253788	0.993237

Table 4 recommended products for beef purchase

	Lhs	***Rhs***	***Support***	***Confidence***	***Lift***
1	Beef	Rhizome vegetables	0.017387	0.331395	3.040367
2	Beef	Other vegetables	0.019725	0.375969	1.943066
3	Beef	Yoghurt	0.011693	0.222868	1.597601

Table 5 recommended pastry for purchasing goods

	Lhs	***Rhs***	***Support***	***Confidence***	***Lift***
1	Kernel fruits	Pastry	0.010676	0.141129	1.586290
2	Whole milk	Pastry	0.033249	0.130123	1.462587
3	Soda	Pastry	0.021047	0.120700	1.356665
4	Bread rolls	Pastry	0.020946	0.113875	1.279956

4.2 Discussion

Table 2 is a table of complementary commodities, that is, the lift of commodities greater than 1. In this paper, five results with the greatest lift are selected and shown in the table. From the No. 1 in Table 1, it can be concluded that the probability of purchasing rhizome vegetables after the customer has purchased beef is about 33%. At this time, if the rhizome vegetables are placed near the beef, the sales of rhizome vegetables can be promoted to the greatest extent and the customer's shopping experience can be enhanced.

Table 3 is a table of mutually exclusive commodities, that is, the lift of commodities less than 1. From No. 1 in Table 3, it can be concluded that if the bread rolls is placed near the shelf of canned beer, the possibility of customers' purchase will be reduced. Therefore, shopping malls should place canned beer and bread rolls in different positions to increase the sales of bread rolls and avoid creating a bad shopping experience for customers.

Table 4 shows the list of recommended products for beef purchase, that is, in order to maximize the overall sales after customers purchase beef, the probability of recommending Rhs column products to them is higher than that customers naturally purchase. After purchasing meat,

customers often purchase vegetables, which can be verified from table 3, which meets the needs of modern people for a healthy life. According to the table, shopping malls can place rhizome vegetables, other vegetables and yogurt near beef to maximize sales.

Table 5 shows the table of recommended pastry for purchasing goods, that is, after customers purchase the goods in the Lhs column of the table, they recommend the pastry in the Rhs column to customers. At this time, the probability of recommending customers to purchase is higher than that of purchasing alone, which can promote the sales of pastry. Unsalable goods in shopping malls are a common phenomenon. If the pastry in this shopping mall are unsalable, it can be analyzed in Table 4 to recommend pastry to customers who have purchased nuts, whole milk, soda and bread rolls, so as to alleviate the unsalable problem of cakes.

To sum up, shopping malls can adjust the placement of their shelves according to table 2, table 3 and table 4, which can not only make it more convenient for customers to buy, improve the happiness index of customers' shopping, but also improve the sales volume of shopping malls. If a commodity in the shopping mall is unsalable, the data similar to table 5 can be obtained through the algorithm. That is, after the customer has purchased which products will purchase unsalable products, so as to adjust the commodity layout, alleviate the unsalable problem of the commodities in the shopping mall and reduce the losses of the shopping mall. The experimental data are applied to a shopping mall for practice. It is found that the actual results are roughly consistent with the above output results during the observation period.

5 CONCLUSIONS

This paper takes the market sales data set as the research object, uses the Apriori algorithm based on association rule mining to analyze the market sales data, and establishes the algorithm model according to the minimum support and minimum confidence. Comprehensively analyze which goods should be adjusted in the mall, and play the role of recommending goods when customers buy, so as to increase the sales of the mall and improve the customer happiness index. The experiment shows that the probability of customers natural purchase after Apriori algorithm analysis is about 300% higher than that before. This shows that the algorithm has good performance. The application of the model provides a solution for the mall to improve sales and customer experience. In future work, it will be considered to make further analysis for customers of different ages and genders.

REFERENCES

[1] ZHOU Yu, CAO Yingnan, WANG Yongchao. Overview of Data Processing and Analysis Algorithms for Big Data[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021,53(05):664-676.

[2] AGRWAL R, SRIKAN R. Fast algorithms for mining association rules in large databases[C] // Proceedings of the Twentieth International Conference on Very Large Databases, Santiago, Chile: [s.n].1994, 9 :487-499 .

[3] LI Ying, TANG Yong. Research on the Data Mining Algorithm Based on Association Rules and Similarity [J].Journal of South China Normal University (Natural Science Edition), 2021, 53(05):121-127.
[4] JI Wenlu, WANG Hailong, SU Guibin, LIU Lin. Review of Recommendation Methods Based on Association Rules Algorithm[J]. Computer Engineering and Applications, 2020, 56(22): 33-41.
[5] CUI Yan, BAO Zhiqiang. Survey of association rule mining[J]. Application Research of Computers, 2016, 33(2): 330-334.
[6] Qisman M,Rosadi R,Abdullah A S. Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data (case study of Mizan computer retail stores) [J]. Journal of Physics: Conference Series,2021,1722(1).
[7] Pinto da Costa Joaquim Fernando,Cabral Manuel. Statistical Methods with Applications in Data Mining: A Review of the Most Recent Works[J]. Mathematics,2022,10(6).
[8] TIAN Ke, CHANG Huajun. Research on Improving Velocity Measurement Accuracy of Radar Based on Apriori Association Rule Algorithm [J/OL]. Journal of Gun Launch & Control:1-6.