Computer Fraud and Security
ISSN (online): 1873-7056

Analysis of Subway Fault Data Based on Improved Apriori Algorithm

Shanshan He*, Jingjiao Chen
Anhui Sanlian University, Hefei 230000, China
**Corresponding Author.*

Abstract:
The subway is a high-capacity, fast and punctual urban rail transit that usually operates underground or on elevated tracks. It has the characteristics of large capacity, fast speed, high punctuality, good safety, environmental protection and energy conservation. However, some malfunctions may occur during the operation process. However, there is currently no good solution to the problem of undefined subway fault data and diverse fault types. This article improves the Apriori algorithm and compares it with the FP Growth algorithm. Python is used to simulate fault scenarios, and the association rules between fault data are obtained through the improved algorithm. Through performance comparison, it can be concluded that the improved Apriori algorithm can provide a good reference for subway fault analysis.

Keywords: CCS concepts, data mining, machine learning, machine learning algorithms, subway breakdown, data mining, association rules, Apriori algorithm

INTRODUCTION

The vast majority of urban rail transit systems are used to carry commuters in the city, and in many cases urban rail transit systems will be regarded as the backbone of urban transportation. Usually, the urban rail transit system is the way to solve the problem of traffic congestion in many cities. The Moscow Metro is one of the busiest in the world, with 8 million Muscovites taking an average of one ride per person per day and 44% of the city's total passenger transport. The Tokyo Metro is very close to the Moscow Metro in terms of mileage and passenger volume. The Paris metro already carries more than 10 million passengers a day. New York's subway has the world's fifth longest operating route, and the total number of daily passengers has reached 20 million, accounting for 60% of the city's transportation. Although the total length of Hong Kong's MTR is only 43.2 kilometers, its daily passenger traffic is as high as 2.2 million, and the peak is 2.8 million. By the end of 2020, Nanjing Metro has opened and operated a total of 10 lines 1, 2, 3, 4, 10, S1, S3, S7, S8, S9, with a total length of 394 lines. 3 km, with an average daily passenger volume of 2.18 million [1].

Association rule [2] is a rule that finds the relationship between items in a data set, and is usually used in shopping basket analysis, personalized recommendation systems, and product promotion. Typical examples of association rules include the "buy beer at the same time you buy diapers" phenomenon, which can help retailers optimize shelf layout and promotional strategies, thereby increasing sales. Association rule algorithm [3-7] is a popular algorithm for data analysis, among which FP-Growth algorithm [8,9] and Apriori algorithm [10-12] are the two most commonly used algorithms for association rule mining. Association rule mining allows us to discover the relationships between items in a dataset. It has many application scenarios in our daily lives, and "shopping basket analysis" is a common scenario. This scenario can discover the associations between products from consumer transaction records, and then bring more sales volume through bundled sales or related recommendations. So, association rule mining is a very useful technique.

Association rules reflect the interdependence and correlation between something and other things, and are commonly used in recommendation systems for physical stores or online e-commerce. By mining association rules from customer purchase record databases, the ultimate goal is to discover the inherent commonalities in the purchasing habits of customer groups, such as the probability of purchasing product B at the same time as purchasing product A. Based on the mining results, the layout and display of shelves can be adjusted, and promotional combination schemes can be designed to achieve sales growth. The most classic application case is<beer and diapers>.

Based on the difficulty of detecting and predicting subway faults, this paper optimizes Apriori algorithm and proposes an improved Apriori algorithm to mine association rules between subway data. The improved

algorithm is compared with the classical FP-Growth algorithm, and the results of the association rules of fault data are obtained, which provides a reference for subway fault analysis.

ASSOCIATION RULES PROBLEM DESCRIPTION

Data Mining

Data mining refers to the process of searching for hidden information from a large amount of data through algorithms. Data mining is a technique that analyzes each piece of data to find its patterns from a large amount of data. It mainly consists of three steps: data preparation, pattern search, and pattern representation. Data preparation is the process of selecting the necessary data from relevant sources and integrating it into a dataset for data mining;

As shown in Figure 1, data mining begins with data collection, which may be done through online collection or manual input techniques such as Python crawlers. The collected data is then stored in a database, commonly referred to as a data warehouse. The data format obtained during the collection process may not be correct, and further transformation of the data is necessary, such as forming multidimensional data, time-series data, etc. The final step is to process the data and transform it into a data mining problem, such as transforming it into association rules, classification, and other problems.

Figure 1. Data generation process

Data Warehouse

In the 1990s, the concept of data warehouse first emerged, specifically defined as a subject oriented, integrated, time-dependent, and stable collection of data. Data warehouses have significant differences from traditional databases in that they can serve high-level decision-making. Data warehouses can not only collect, organize, and store large amounts of data from information workers, but also process and change historical data to obtain relevant information and data for decision-making analysis. This can make the decisions made by decision-makers more scientifically reasonable. In addition, a data warehouse is also a topic oriented database. Simply put, it can organize data according to a certain topic and process data information according to specific decision-making and analysis needs. And a data warehouse is also a database that contains historical data and information, which means that a data warehouse can not only be used for retrieval, but also for analyzing and processing the operational status and future development trends of the entire organization. In the basic architecture of a data warehouse, data sources can be specific data files or other data sources, which can serve a series of ordinary and traditional business databases.

The data warehouse has the following functions:

(1) Capable of providing various reliable and standardized reports and specifications chart. Moreover, due to the relatively wide range of data information sources within the data warehouse, it is retrieved from multiple transaction processing systems, which also enables the data warehouse to provide users with integrated information for the entire enterprise.

(2) A data warehouse can conduct multidimensional analysis based on the actual needs of users. By defining multiple attributes of entities as multiple dimensions, it helps users more convenient aggregation and processing of data information through different dimensional values comparative analysis, able to understand data information from different perspectives and levels and application.

(3) As the foundation of data mining technology applications. Data warehouse can to provide massive and accurate data information for the development of data mining and analysis work, ensure the reliability of data analysis and mining, and build upon existing data, realize effective prediction of future situations.

Association Rule Mining

Association rules can be used to discover the symbiotic relationship between items and extract specific information with potential value, which is a very important technology in data mining.

Association rules is a common unsupervised learning algorithm in data mining analysis. The main purpose is to find the rules and patterns of certain attributes in the data set, so as to obtain the association relationship of different attributes. Its expression is X→Y, where the term set X is the prerequisite, that is, the previous term; Item set Y is the corresponding association result, that is, the latter item. The item sets X and Y are both the item sets I (I= {I1, I2, ... Im}), and X has no intersection with Y, and item set I is the item set of transaction database D. These thresholds are artificially set based on mining needs. Association analysis is mostly divided into three categories: simple association, temporal association, and causal association [13]. Therefore, in order to make the mining results more valuable, relevant scholars often introduce other association analysis parameters to improve the association rule mining algorithm [14-15].

The association rule mining model is shown in Figure 2.

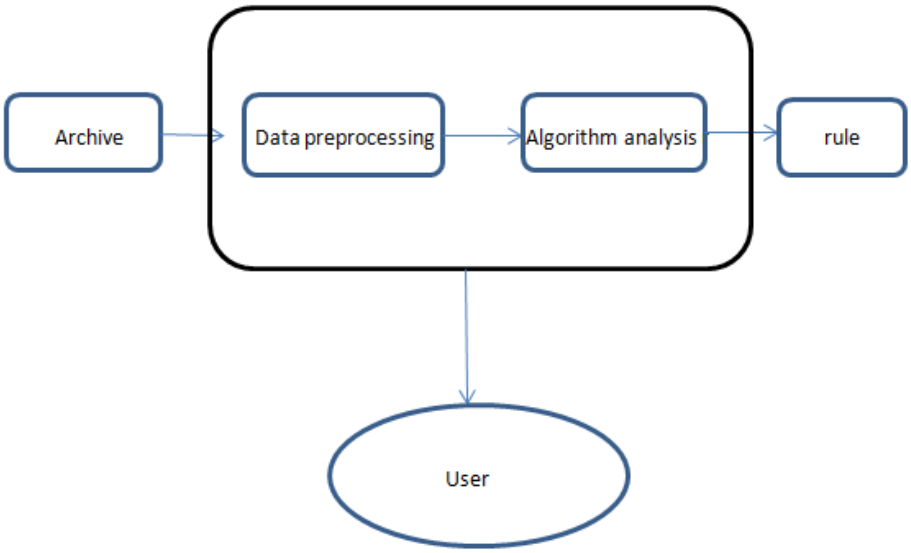

Figure 2. Basic model of association rule mining

1)Association Rules

It is a technique used in data mining to discover interesting relationships between data items. It is commonly used in fields such as shopping basket analysis and product recommendation to help identify patterns in user purchasing behavior. Association rules are an implication in the form of X → Y, where X and Y are respectively referred to as the antecedent and consequent of association rules

2)Support

The frequency of patterns appearing in a rule, which refers to the percentage of occurrences of itemsets in a transactional database compared to the total number of transactions in the database. If s% of transactions in the transaction database contain XY, then the support of association rule XY in D is called s%. Let I= {i1, i2, ..., im} be the set of all items, and D be the event. The transaction database consists of a series of TID transaction groups with unique identifiers Cheng. Let X ⊆ I be a set composed of items, called an itemset. If project set X contains k projects, it is called a k-item set. Assuming X ⊆ I, the support (X) in itemset X is a matter. The proportion of transactions in transaction database D that support X accounts for the percentage of transactions, denoted as:

$$\text{Support(X)}=\frac{Support(X)}{DataBase(D)} \quad (1)$$

3)Confidence

Refers to the strength contained, that is, the proportion of transactions containing X and XY at the same time. If the support degree of X is support (x), then the trust degree of the rule is support (XY)/support (X), expressed as conditional probability P (Y | X).

$$\text{Confidence}(X \rightarrow Y) = \frac{Support(XUY)}{Support(X)} \quad (2)$$

IMPROVED APRIORI ALGORITHM

Apriori Algorithm

The Apriori algorithm is mainly used to discover association rules by constructing longer itemsets from known frequent itemsets, which are referred to as candidate frequent itemsets. The algorithm first calculates all candidate itemsets C1 and finds the frequent itemset L1 from C1; Then, based on L1, determine candidate itemset C2 and find all frequent itemsets L2 from C2; Then, based on L2, determine candidate itemset C3 and find all frequent itemsets L3 from C3. Repeat this process until a higher dimensional frequent itemset cannot be found.

Apriori algorithm is a classic data mining algorithm mainly used to discover frequent itemsets and association rules in a given dataset. This algorithm was first proposed by Rakesh Agrawal et al. in 1993. The initial motivation was proposed for the shopping basket analysis problem, with the aim of discovering the connection rules between different products in the transaction database. These rules can characterize customers' purchasing behavior patterns, and for merchants, they can be used to guide the scientific arrangement of procurement, inventory, and shelf design. The name Apriori algorithm comes from the algorithm's use of prior knowledge to compress the search space and improve algorithm efficiency.

The Main Idea of Improving Apriori Algorithm

In this paper, Apriori algorithm is improved to meet the requirement of subway fault data association rule mining. The improved Apriori algorithm can obtain the association rules between frequent item sets by traversing the data once.

Step 1: TID identification for each transaction item;

Step 2: Once through the database, delete irrelevant transaction items, and count the deleted items to get the set of interest B;

Step 3: Self-connect, generate candidate items;

Step 4: Set the intersection operation to obtain the associated transaction items.

The steps to improve Apriori are as follows:

Step 1: Delete irrelevant transaction entry records. Let the total number of transaction entries be m and the traversal database be D. When Dx (x= 1, 2, ... m). When count=1, Dx is deleted, the number of deleted transactions is counted as 1, and so on through the loop to get a new database D'. Let the set of interest be B, if D'x (x= 1, 2, ..., n), B∉D'x, then delete D'x and iterate through the loop to get a new dataset D ".

Step 2: Mining frequent item sets. Each transaction item is counted to obtain candidate 1-item set, where items ≥min_ sup constitute frequent item set L1.

The generated frequent item set L1 is self-joined, the candidate 2-item set is generated, and the set intersection operation is performed on it to get the transaction TID set. Frequent item set L2 is composed of items ≥min_ sup.

Calculate the modules of Lk| Lk|, |Lk| when ≤k, the operation terminates and the frequent item set L is obtained; otherwise, step B is repeated.

Mining association rules. Calculate the support degree and confidence degree, analyze the association relationship between variables, summarize some regularity between variables, and generate association rules.

Pseudo Code for Improving Apriori Algorithm

The improved Apriori algorithm only needs to traverse the database once, resulting in frequent updates. The basic idea of the association rule results between itemsets is to traverse the database and obtain association rule results.

Input: Dataset D '', small support min_ stup;
Output: frequent itemset L_k;
The processing procedure is as follows:
$L_1 = find_{frequent1} -$ itemsets (D″);
$C_2 = L_1 \infty L_1$
L_1 = items in $C_2 \geqslant$ min_ sup;
For(k = 3; $L_{k-1} \neq \emptyset$; k ++);
Prunel(L_{k-1});
$L_x \in L_k$, $L_y \in L_k$;
if (L_x [1] = L_y [1] $\wedge L_x$ [2] = L_y [2] $\wedge \cdots \wedge L_x$ [k − 2] =L_y [k − 2] $\wedge L_x$ [k − 1] < L_y [k − 1]);
If (k − 1) − subsets of c $\notin L_{k-1}$
Then delet c from C_k
$C_k = c \cup C_k$;
L_k = New_ quick_ support_ count(C_k, TID_ Set)
Answer = $U_k L_k$;
New_ quick_ support_ count(C_k, TID_ Set)
For all itemsets c $\in c_k$
C. TID_ Set = L_{k-1}.TID_ Set $\cap$ L1. TID_ Set
C. Sup = Length(C_k.TID_Set
If C.Sup < min_ sup
Delet C from C_k
L_k = {C $\in C_k$ | C.Sup $\geqslant$ min_ sup}
Prunel(L_k)
For all itemsets $L_1 \in L_k$

CASE ANALYSIS

The original data of subway Line 3 was selected for experiment, the corresponding minimum support degree and running time of the three algorithms were compared, and different values of the algorithm parameters were taken for simulation experiment, and the association rules in subway fault data were obtained according to the analysis of the experimental results. Simulated hardware environment :Core (TM) i5 CPU @2.10GHz 8.0 GB ORAM

Fault Data Import

This paper collected the data of Metro Line 3, Line 4, Signal failure data, 400pieces of data from line 3 are selected as a case study. The types of signal faults in the data are divided into ATP faults, equipment faults, benchmarking faults, and compact faults. There are 5 kinds of fault, such as fault, platform door linkage fault, and each type is divided into 4 fault levels such as A, B, C, and D.

Data Preprocessing

In this case, data preprocessing includes data screening and data change. The data comes from a subway company, and the list of valuable data for case analysis needs to be screened, and the invalid option column should be removed. Since the data in this database is a continuous numerical variable, the Apriori association rule mining algorithm cannot process it, so it is necessary to classify the data attributes and discretize the data. The 24 hours of a day are divided into 12 sections with 2 hours as a unit interval. 1, 2, 3... twelve. After data preprocessing, the final data set is formed according to the fault level, as shown in Table 1.

Table 1. Table 1 final modeling data set

Time	On-board ATP failure	Signal Equipment failure	Benchmarking Fault	Train tightening Failure	Platform door Linkage failure
1	X1	C2	P3	Q4	Z5
1	D1	D2	P3	Q4	Z5
3	X1	C2	P3	Q4	Z5
3	D2	D2	P3	Q4	Z5
4	D1	A2	P3	Q4	Z5
4	D1	C2	P3	Q4	Z5
..	...	...	...	...	...
8	X1	Y2	P3	C4	Z5
8	C1	D2	P3	C4	Z5
8	D1	D2	P3	C4	Z5
...	...	...	...	...	...
10	C1	C2	D3	Q4	B5
10	D1	Y2	P3	Q4	Z5
11	C1	Y2	P3	Q4	Z5
11	D1	Y2	P3	Q4	Z5

Model Establishment

According to the process of improving Apriori algorithm, a model is created, different values of algorithm parameters are analyzed, and an algorithm with high matching degree is selected and parameters are set according to the analysis results to carry out the mining of association rules between subway fault data and time.

Figure 2 shows the variation of running time corresponding to different algorithms, along with minimum support

The running time of different algorithms decreases, and when the support degree reaches 3.5%, the running time is the least.

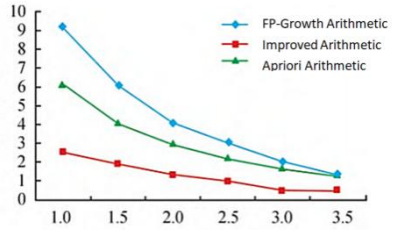

Figure 3. Comparison of minimum support before

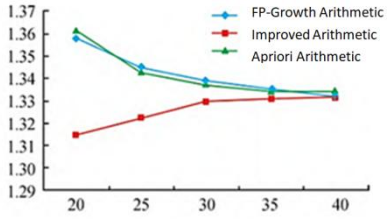

Figure 4. Comparison of minimum confidence before and after improvement

As shown in Figure 3, after the support degree increases, the algorithm time before and after the improvement decreases, but it can be seen that the running time of the improved algorithm is smaller than the algorithm time before the improvement.

As shown in Figure 4, with the increase of confidence, the time difference between algorithms is not large, but when the confidence is small, the running time of the improved Apriori algorithm is much longer than that of the FP-Growth algorithm.

To sum up, the efficiency of the improved Aprori algorithm is higher than FP-Growth, so the minimum support degree of 6% and minimum confidence degree of 75% are adopted in this paper for simulation, and the operation interface is shown in Figure 5.

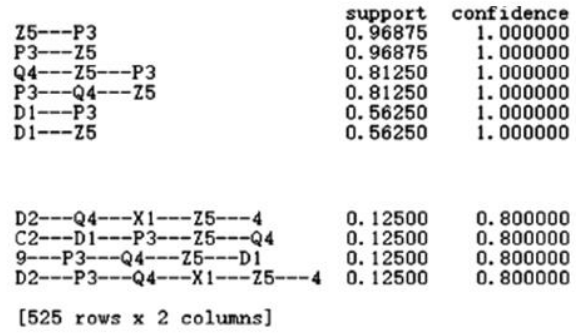

Figure 5. Output result

Model Analysis

According to the above operation results, the study obtained 600 association rules, which indicated that D2, X14, representing Class D failure of signal equipment and class X1 failure of on-board ATP equipment occurred in the time period of 6:00-8:00, the support degree is 12.5%, and the confidence degree is 80%. Table 2 shows part of the subway fault association rules.

Table 2. Mining of association rules for subway faults

Serial number	Association rule	Support degree	Confidence degree
1	Platform door linkage Z-level fault → Benchmarking fault P3⇒ associated occurrence	0.968 75 0. 968 75	1 1
2	Benchmarking fault P3→ Platform door linkage Z-level fault ⇒ occurs		
3	Train tightening device Q fault → platform door linkage Z fault → benchmarking fault P3⇒ associated occurrence	0.81250	1
4	Train tightening device Q fault → platform door linkage Z fault → benchmarking fault P3⇒ associated occurrence	0.81250	1
5	Vehicle ATP system Level D fault → Benchmarking fault P3⇒ associated occurrence	0.56250	1
6	Vehicle-mounted ATP system level D fault → platform door linkage level Z fault ⇒ related occurrence	0.56250	1
7	Class D fault of on-board ATP system → Class Q fault of train tightening device → Class X fault of on-board ATP system → Class Z fault of platform door linkage ⇒ All the faults occurred between 6:00 and 8:00	0.12500	1
8	Class C fault of signal equipment → Class D fault of on-board ATP system → Benchmarking fault P3→ Class Z fault of platform door linkage → Class Q fault of train tightening device ⇒	0.12500	1
9	Benchmarking fault P3→ Train tightening device Q fault → Platform door linkage Z fault → Platform door linkage Z fault → Platform door linkage Z fault → on-board ATP system D fault ⇒ The faults occurred between 16:00-18:00	0.12500	0.8
10	Class D fault of on-board ATP system → Benchmarking fault P3→ Class Q fault of train tightening device → Class X fault of on-board ATP system → Class Z fault of platform door linkage ⇒ The faults all occurred in the time period from 6:00 to 8:00	0.12500	0.8

CONCLUSION

Aiming at the problems such as the variety of subway fault data and the difficulty in defining the degree of influence, this paper establishes the algorithm model of association rules mining. An improved Apriori calculation is proposed

Law; Select 5 kinds of fault data, will each faults are classified into four levels, such as A, B, C, and D, and the association rules between faults are mined. The results show that the improved Apriori algorithm improves the efficiency of data processing and the reliability of subway fault association rules mining.

ACKNOWLEDGMENTS

This work was supported by Research on the application of data warehouse and data mining technology in rail transit AFC System(2023AH051701) and Research on Traffic Congestion Identification Model Based on Big Traffic Data in Hefei (2024AH050514).

REFERENCES

[1] Fan X, Muyan L. New Space for City Communication: A Study on Culture Transmission by Nanjing Metro. Cross-Cultural Communication, 2021, 17(1): 30-34.

[2] Park J S, Chen M S, Yu P S. An effective hash-based algorithm for mining association rules. Acm Sigmod Record, 2016, 24(2) : 175 — 186.

[3] Liu X Y, Niu X Z. Fournier Viger Philippe Fast Top-K association rule mining using rule generation property pruning. Applied Intelligence, 2020, 51(04): 2077 — 2093.

[4] Djenouri Y, Djenouri D, Habbas Z, et al. How to exploit high performance computing in population-based metaheuristics for solving association rule mining problem. Distributed and Parallel Databases, 2018, 36(02) : 369 — 397.

[5] Zhang Chun, Zhou Jing. Research and Application of EMU Fault Association Rule Mining Optimization Algorithm. Computer and Modernization, 2017(09): 74-78.

[6] Zhou J. Research on the correlation analysis technology of the operation and maintenance efficiency of the key components of the EMU. Beijing Jiaotong University, 2017.

[7] Zhou B. Research on association rules mining method of massive engineering data based on Hadoop. Beijing Jiaotong University, 2016.

[8] Zhu Xingdong, Zhang Siyu, Wang Zheng. Mining method of association rules for aircraft fault maintenance records. Journal of Ordnance Engineering, 2019, 40 (07): 164-169

[9] Russell I, Markov Z. An Introduction to the Weka Data Mining System ACM Sigcse Technical Symposium on Computer Science Education ACM, 2017: 742 — 742.

[10] Zhou Kai, Gu Hongbo, Li Aiguo. An improved algorithm for mining Apriori Algorithm based on Association rules. Journal of Shaanxi University of Technology (Natural Science Edition), 2018, 34 (05): 40-44.

[11] Sharmila S, Vijayarani S. Association rule mining using fuzzy logic and whale optimization algorithm. Soft Computing, 2021, 25(02) : 1431 — 1446.

[12] Agapito G, Guzzi P H, Cannataro M. Parallel and distributed association rule mining in life science: A novel parallel algorithm to mine genomics data. Information Sciences, 2018(prepublish). DOI: 10. 1016 /j. ins. 2018. 07. 055.

[13] Qin X G. Research on signal system equipment maintenance strategy based on big data risk analysis. Beijing Jiaotong University, 2018.

[14] Bai Yingying, Shen Chenchen. Improved Apriori algorithm based on Association rule Mining Electronic Technology and Software Engineering, 2017(03): 203-204.

[15] Zeng Zixian, Gong Qingge, Zhang Jun. Improved association rule mining algorithm -- MIFP-Apriori algorithm. Science Technology and Engineering, 2019, 19 (16): 216-220.