Application of Apriori Association Rules Algorithm and Big Data Technology in Transportation Field

Zhiqi Guo [1, a *], Yi Guan [2, b], Jiacong Zhao [1, c]

[1,2,3]Dalian Neusoft Institute of Information, Dalian 116000, Liaoning, China

[a]guozhiqi@neusoft.edu.cn; [b]guanyi@neusoft.edu.cn; [c]zhaojiacong@neusoft.edu.cn

Keywords: Data mining; Apriori algorithm; Big data; Traffic management

Abstract. With the increasing demand for material and cultural life, the holdings of automobiles are increasing day by day, and the per-capita holdings are also decreasing. With the appearance of various traffic problems, traffic accidents frequently occur. It poses a serious threat to the property and life safety of the vast number of residents. The traffic congestion problem has caused many inconveniences to us, delaying a lot of precious time, reducing the happiness index of the people, and solving traditional problems, such as widening roads and increasing traffic. Control personnel, etc., Have not been able to meet the needs of the current stage very well. In recent years, the rise of data mining technology and big data technology has enabled us to see the hope of solving the problem.

The rapid development of computer technology and the use of modern hardware technology to collect data have led to a large amount of data collection and processing in many fields. Data mining technology appears in a timely manner, and the generated data is summarized, and a large number of decision-making conclusions are obtained. Traffic pressure has played a big role. Combined with big data technology, the data is fully covered. The sampling method is not used, which makes the data mining conclusion more accurate and more targeted, and achieves accurate problem solving. This paper mainly introduces the Apriori algorithm to process the collected data to find the association rules for traffic accidents, provide decision support for traffic managers, and alleviate various traffic problems that are now emerging.

Research Background

With the rapid development of China's economy, the number of private cars has risen sharply. The Traffic Management Bureau of the Ministry of Public Security issued on July 16, 2018. As of the end of June, the number of motor vehicles in the country reached 319 million. The traffic pressure in cities is getting bigger and bigger. In almost all major cities, there are serious traffic jams. Before the National Day, there have been frequent traffic jams in Beijing. According to WHO estimates, more than 200,000 people die in traffic accidents every year in China. The number of people killed or injured in traffic accidents in China has ranked first in the world for 10 consecutive years[1], bringing irreparable harm to the people's body and property. How to effectively prevent traffic accidents and improve traffic safety are issues that we need to consider urgently. Traditional methods such as widening roads and increasing traffic control personnel can no longer meet the current traffic conditions in China.

Research Methods

The idea of traditional methods Apriori algorithm. The basic idea of the Apriori algorithm is to first find all the frequent itemsets, and then generate strong association rules from the frequent itemsets. These rules must satisfy the minimum support and the minimum confidence. Searching all frequent itemsets requires multiple searches of the transaction database, which is a major factor affecting the performance of the associated algorithm. The Apriori algorithm uses the k–1 frequent itemsets to generate candidate k-frequency itemsets, but the candidate frequent itemsets are usually very large. For example, in the product defect management analysis, the item set composed of m items may be

  DOI: 10.25236/icess.2019.003

generated 2^m-1 Frequently selected itemsets and 3^m-2^{m+1}+1 Association rules。However, in general, most of these rules do not satisfy the conditions of strong association rules. This problem becomes the bottleneck of association rule mining. Therefore, reducing the size of the candidate set and then scanning the transaction database, it is necessary to calculate the support of the candidate set. If the longest frequent item set is n, then n +1 transaction database scans are required. Therefore, how to efficiently find frequent itemsets is a key issue in association rule mining.

Apriori algorithm uses "any subset of frequent itemsets must also be frequent or the superset of infrequent itemsets must be infrequent" a priori property to reduce the search space of frequent itemsets [2]。

As shown, it is the item set of {i1, i2, i3, i4}, which enumerates all possible itemsets. Assuming that {i2, i3, i4} is a frequent item set, then all its subsets {i2}, {i3}, {i4}, {i2, i3}, {i2, i4} and {i3, i4} are Frequent. On the other hand, if {i1, i2} is infrequent, all its supersets {i1, i2, i3}, {i1, i2, i4} and {i1, i2, i3, i4} are infrequent.

Assume that the items in the frequent item set L_{k-1} are arranged in order, The k-frequent item set apriori-gen(L_{k-1}) with k-1 frequent item set generation candidates needs to generate and trim the candidate frequent itemsets. This step needs to avoid generating too many unnecessary, repetitive candidate frequent itemsets.

The first step: self-join Lk-1
insert into Ck
select p.item1, p.item2, …, p.itemk-1, q.itemk-1
from Lk-1 p, Lk-1 q
where p.item1=q.item1, …, p.itemk-2=q.itemk-2, p.itemk-1 < q.itemk-1
Step 2: Trim
forall itemsets c in Ck do
forall (k-1)-subsets s of c do
if (s is not in Lk-1) then delete c from Ck
Example of generating a candidate set:
L3={abc, abd, acd, ace, bcd}
self-join : L3*L3
abc and abd get abcd
acd and ace get acde
prune:
ade not in L3，delete acde
C4={abcd}

Apriori's Application in Traffic Management

Table 1　Traffic accident influencing factors

Accident number	License number	influences i1	influences i2	influences i3	influences i4	influences i5
1	005	i1		i3	i4	
2	028	i1	i2	i3		i5
3	032		i2	i3		
4	098	i1		i3	i4	i5
5	105	i1			i4	
6	159		i2	i3		i5
7	170			i3		
8	190	i1			i4	i5
9	197	i1	i2			
10	199	i1	i2		i4	i5

Assume that the minimum support number is 3 and the minimum confidence is 60%.

The available frequent itemsets are {i1,i4,i5}

Calculate the confidence separately:

{i1}->{i4,i5}=3/7

{i4}->{i1,i5}=3/5

{i5}->{i1,i4}=3/5

{i1,i4}->{i5}=3/5

{i1i5}->{i4}=3/4

{i4,i5}->{i1}=3/3

The strong association rules are {i4}->{i1,i5}, {i5}->{i1,i4}, {i1,i4}->{i5}, {i1i5}->{i4}, {i4 ,i5}->{i1}

Therefore, the relationship between the impact factors of traffic accidents can be obtained, so that it can be concluded that avoiding the conditional factors can avoid the occurrence of the result factors, thus avoiding the occurrence of traffic accidents, and also providing some decisions for traffic control. in accordance with.

Limitations of traditional methods. From the above example, it can be seen that the accident data is sampled in 10 groups of 5 dimensions per group for Apriori algorithm calculation, and the time complexity is $O(t)= 2^n-n-1$. As we all know, traffic events occur in large amounts of data every day, the computing performance is significantly reduced, and the real-time performance is not high, and too little sample data is easy to produce over-fitting problems. In order to solve this problem better, It is necessary to introduce distributed storage and big data technology to solve the problem of low real-time and over-fitting, and adopt full coverage on the amount of data.

Big Data Technology Application

Distributed storage is used in the process of building monitoring system and data acquisition system, which can solve the problem of insufficient storage space and insufficient computing power of application server. From the perspective of system architecture, it is divided into data collection, data warehouse and data application service. And the data visualization four levels, corresponding to the original video library, the basic information database and the police/case event library in the smart traffic business, and the data development dimension corresponds to the theoretical basis of knowledge transfer from knowledge to knowledge.

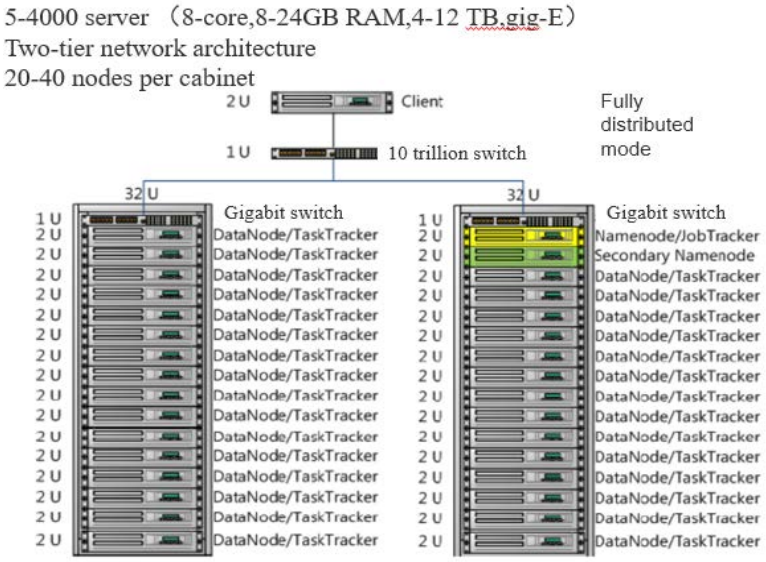

Figure. 1 Environmental deployment

Distributed storage is highly fault-tolerant, allowing data to be automatically saved in multiple copies. Even if the copy is lost, it can be automatically restored. At the same time, it is more suitable for batch processing, using the mobile computing mode instead of the data itself, and the

data location is exposed to the computing framework. It makes the operation faster and more accurate, and is also especially suitable for large data processing data can reach the level of GB, TB, or even PB data, the data file is also the number of files in millions of scale, the node size can handle more than 10K scale. File access form uses streaming file access, one-time write, multiple reads, thus ensuring data consistency, avoiding data inconsistency leading to misjudgment, and most praised is that it can be built on cheap machines, which can be very good Old, through multiple copies to improve reliability, eliminate people's distrust of cheap machines, but also provides a fault-tolerant and recovery mechanism to ensure data consistency. A typical fully distributed model is shown in figure1.

Using distributed computing mode, combined with the traditional Apriori algorithm, real-time feedback of traffic conditions, providing the most accurate first-hand information for the traffic management part, as well as ordinary users, reducing the accident rate, and providing the most reliable reference for people to travel data. Using the Hadoop ecosystem suite, Hive as a technical support, Hive is a data warehousing tool. You can turn raw structured data under Hadoop into a table in Hive, and Hive supports a language HiveQL that is almost identical to SQL. In addition to not supporting updates, indexes, and transactions, almost all other features of SQL can be supported. It can be viewed as a mapper from SQL to Map-Reduce, and provides interfaces such as shell, JDBC/ODBC, Thrift, and Web.

The traffic information is captured in real time and pushed to the end users, which provides a good guarantee for reducing the accident rate and saves social resources. In terms of travel, the information needs to travel to the public, integrate traffic travel service information, and expand the coverage of information services in public transportation, taxi, road traffic, public parking, and road passenger transportation, making public travel more convenient. It can provide comprehensive and multi-level information services, including traffic information, real-time road conditions, bus dynamic information, parking dynamic information, dynamic information services such as water passenger transport, flights and railways, and information interaction services such as travel route planning and rental call.

In terms of management, the use of transportation industry data to support traffic management and decision-making. The use of data mining technology can deeply study the optimization of transportation network, and provide support for industry development trend research, policy formulation and effect evaluation. In addition, the communication with the big data platform of public security, construction management, environmental protection and other related functional departments can improve cross-domain management capabilities.

Conclusion

This paper briefly summarizes the basic principles of Apriori data mining algorithm, and summarizes the effective application of the algorithm in the traffic field. It also expounds that the traditional file storage mode leads to low computational efficiency and easy over-fitting. Therefore, the introduction of big data processing technology, combined with traditional algorithms to improve the data mining process, in order to fully explore the potential of researchers to solve the potential of traffic problems [3]。 In the future research work, the author will further study and explore the principle of the algorithm and its application in the field of transportation.

References

[1] X.J. Zhu. Application of data mining technology in the field of transportation [D].2013.4
[2] W.D. Zhao. Foundation of Business Intelligence [M]. Tsinghua University Press 140-141
[3] X.P. Yang. Summary of the application of three classic algorithms of data mining in the field of transportation[J].Intelligent Processing and Application 208.11.42-44