DATA MINING TECHNOLOGY BASED ON ASSOCIATION RULES ALGORITHM

Guihong Zhang[1], Caiming Liu[2], Men Tao[3]
[1,2,3] School of Computer Science
Leshan Normal University, LeShan, 614000, China
e-mail: 3093399546@qq.com

Abstract - The current data mining technology has the mismatch problem of mining keywords and used words, which leads to low mining precision and low efficiency. In this paper, a new data mining algorithm based on association rule algorithm is proposed. Firstly, the weight of the keywords in the data mining process is calculated by using the TF-IDF function. Based on the analysis of the association rules algorithm and the vector space model, the information retrieval model is constructed. On the basis of this, the association rule algorithm including the initial mining items is used to establish the rule database, and several words with the highest degree of relevance to the mining words are selected as the extension words, which is combined with the initial mining into a new mining. Finally, the clustering analysis of new mining results are carried out by K-means clustering algorithm, and according to the descending order, the relevance is ordered, to output mining results. The experimental results show that the algorithm is accurate and efficient.

Keywords: Association Rule Algorithm; Data Mining; Key Words.

1. Introduction

At present, China is in the information age, with the gradual increase of network users, large amount of information and the information which is easy to lose become the challenges in the process of mining data [1, 2]. As the Internet is used on a large scale, more and more people use search engines to perform mining operations. However, many search engines are mining through keywords, it often mines a lot of unwanted information, resulting in reduced mining rate [3, 4]. So, how to improve the recall ratio and precision ratio of data mining, is the current faced key issue [5-8].

In the mining of data, the optimized space-time indexing algorithm is usually used to make the mining effect better. The algorithm uses the multidimensional indexing technique to transform the index to the spatiotemporal dimension, and realize the establishment and improvement of the space-time index [9].

At present, R-Tree index and its expansion are the most commonly used space-time indexing technology. In the literature [10], the R-Tree and the index structure of the inverted table are used to index the same spatiotemporal attribute and time dimension to achieve the purpose of mining. In the literature [11], the MAH_TRP indexing algorithm is used to realize the mining, to improve the spatiotemporal index structure and the updating algorithm, it is effective to deal with the problems such as the spatiotemporal index reconstruction and the poor mining effect caused by frequent updating.

In the literature [12], the R tree optimization algorithm is used to achieve data mining. The coverage area and the blank part of the node are reduced by spatial clustering, and the efficiency of data mining is improved. Compared with the traditional R-tree algorithm, the performance of this algorithm is improved. However, in the case of large amount of data and increasing, it often consumes a lot of resource cost to establish and maintain the index. In the process of data mining, it will be due to search for complex index information too much, causing mining performance decline. In the literature [13], the dynamic algorithm is used for mining data.

The algorithm uses the method of constantly correcting the size of the HASH bucket to suppress the HASH barrel overflow and uses the HASH function to divide the relationship into several small HASH buckets. The size of all the buckets need to less than the available storage capacity of the corresponding processing nodes so as to implement data mining. The processing speed of this algorithm is fast, but the utilization rate of resources is not satisfactory.

A new data mining algorithm based on association rule algorithm is proposed to solve the shortcomings of the above algorithm.
The experimental results show that the algorithm is accurate and efficient.

2. Key Technologies

This section first analyzes the weight calculation process, the association rule algorithm and the information retrieval model, and provides the basis for the subsequent processing.

- ***Weight calculation***

The weight calculation method is more, this section uses TF-IDF function to calculate the weight of the keyword in the data mining process [14], the formula is as follows:

$$w_{i,j} = TF \times IDF = TF \times \lg\left[D/DF(W)\right] \quad (1)$$

Where TF is used to describe the word frequency of the keyword W; D is used to describe the amount of mining; $DF(W)$ is used to describe the mining number of occurrence of the keyword W.

- ***Association rules algorithm***

Association rules algorithm is widely used in different fields, and is the most perfect and most commonly used data mining algorithm. In the association rule algorithm, D is used to describe the set of all datas, and the set of data items is described by X, Y, then the association rule algorithm can be described as $X \Rightarrow Y$ [15]. Where X represents the predictor of the association rule algorithm, Y represents the consequent of the association rule algorithm, and $\Rightarrow$ represents the associated operation. Through the support degree s and the credibility c, the rules in the transaction set D are limited. The credibility is the embodiment of the rule strength, support is the rule frequency of the embodiment. The transaction containing the project set X in the transaction set D is called the number of support for the project set X and is described by σ_x.

The support degree s of the project set X, that is, $\sup port(X)$ can be obtained by the following formula: $\sup port(X) = \sigma_x / |D|$ (2)

The credibility c $confidence(X \Rightarrow Y)$ of the project set X can be described as follows:

$$confidence(X \Rightarrow Y) = \sup port(X \cup Y) / \sup port(X) \quad (3)$$

The basic idea of data mining using association rule algorithm is as follows: firstly, the frequency set of the transaction set D without exceeding the minimum support degree is filtered, and then all the rules that do not exceed the minimum confidence from the frequency set is filtered to obtain the association rule algorithm [16].

This section considers all databases as transaction databases, treats the data in the database as a set of items, and calculates the correlation between them by obtaining the association rule algorithm among the keywords in the database.

- ***Construction of information data retrieval model***

The vector space model (VSM), developed by Saltond et al., describes any piece of text in a set of terms $(t_1, t_2, \cdots t_j, \cdots t_n)$ [17], assigning it corresponding weight w_j through the criticality of each term t_j, and at the same time it is converted to n-dimensional coordinate system, $w_1, w_2, \cdots w_n$ is the corresponding coordinate value.

Assuming that the user mining vector is described by $(w_{q,1}, w_{q,2}, \cdots w_{q,n})$, the mined vector is described by $d_i = (w_{i,1}, w_{i,2}, \cdots w_{i,n})$, then the computational expression of the information retrieval model is:

$$Sim(q, d_i) = \cos(q, d_i) = \frac{\sum_{k=1}^{n} w_{ik} \times w_{qk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \sum_{k=1}^{n} w_{qk}}} \quad (4)$$

3. Data Mining Technology Based on Association Rules Algorithm

The association rules algorithm is used to mine the first N document data pair of the initial mining result. First, the association rule algorithm containing the initial mining item is selected to establish the rule database, and then the K words with the largest relevance are selected and mined, to combine the initial mining and form a new mining. Finally, the new mining results are clustered by K-means clustering algorithm. The correlation is ranked in descending order to output the mining result.

Through the association rule algorithm to achieve the data mining needs two processes [18], the first process is as follows: for the first N pieces of the initial mining results, the association rule algorithm is collected by local feedback, and the rule base is established to select K words with the closest degree of mining words, as the expansion words combining with the initial mining to form a new mining, and then circulate the above process, the detailed steps are:

(1) The initial mining vector $q_{initial} = (w_{q,1}, w_{q,2}, \cdots, w_{q,n})$ and all vectors $d_i = (w_{i,1}, w_{i,2}, \cdots, w_{i,n})$ are constructed;

(2) According to the formula (4), the initial mining is completed, to access to the relevance of $Sim_{initial-doc}$

of each text and the initial mining, and obtain the set arranged from the large to the small according to the close degree of relevance;

(3) The pre- N texts in the set are picked out, and the words are mined by using association rules algorithm [19]:

a. Through the association rule algorithm, the above N piece of data collection is completed, and the frequent item sets including the initial mining items are selected, to establish the rule base [20].

b. The extended words in the rule base are picked out, to find its weight, the K words with the maximum weight in the operation results are as the mining expansion word, and combining the initial mining to form a new mining. The new mining vector q_{new} is calculated.

(4) According to formula (4) the new mining is performed, to access to the correlation degree of $Sim_{new-doc}$ of each data, and according to the close degree of correlation, the text set arranged from large to small can be obtained [21].

The second process is as follows: complete the clustering analysis of the new text set, get the correlation degree of each data, and re-arranged [22-24], the implementation steps are:

(1) The mining results obtained by the new mining are clustered to solve the corresponding center vector c_k for each cluster. The detailed process is:

a. K vectors are randomly selected, as a starting center vector.

b. each mining vector and the interval between it and K class-center vectors are solved, and the minimized interval is included in this class.

c. all the mining results are clustered, and each type of center vector is solved once again.

d. step b and step c are running cycling until all cluster center vectors remain unchanged;

(2) The correlation between the new mining vector and each cluster can be described by the $\cos ine$ - angle between the new mining vector q_{new} and each cluster center vector c_k, which is calculated as:

$$Sim_{new-class} = \cos\left(q_{new}, c_k\right) = \frac{q_{new} \square c_k}{\left|q_{new}\right| \times \left|c_k\right|} \quad (5)$$

Where c_k is used to describe the central vector of each cluster.

(3) By calculating, the final correlation $Sim_{final-doc}$ of all text and mining is obtained, the formula is:

$$Sim_{final-doc} = \alpha \times Sim_{new-class} + (1-\alpha) \times Sim_{new-doc} \quad (6)$$

Where α is used to describe the adjustment factor.

(4) The final relevance is the arranged from large to small, and the output is the result of mining.

4. Experimental Results Analysis

- **Selection of test sets**

In this experiment, 10 authoritative databases are selected as the test samples, to orderly mine in the above database and then 10 mining $\left(q_1, q_2, \cdots, q_{10}\right)$ are established, and the small capacity database and large capacity database are established by two rounds collection. Firstly, in each database, a keyword is used to mine, to get 800 × 10 data, which will be defined as a small capacity database $D1$. And then in each database 1300×10 data are collected, which will be defined as large-capacity database $D2$.

The results of the word segmentation in each database are sorted from large to small, the first 100 characteristic words with high frequency are as a mining database.

The details of test set are described in Table 1.

Table 1. Details of test set

Database number	Data volume of D1 database /pcs	Data volume of D2 database /pcs
Database 1	766	1332
Database 2	811	1294
Database 3	712	1421
Database 4	814	1312
Database 5	798	1265
Database 6	605	1364
Database 7	764	1289
Database 8	836	1318
Database 9	637	1158
Database 10	751	1220
Total	7494	12973

- ***The mining results of using the proposed algorithm***

In this paper, the mining experiments are carried out on the small-capacity database D1 and the large-capacity database D2 respectively to verify the effectiveness of the proposed algorithm. In this section, the mining time of two different capacity databases by using the proposed algorithm in case of the support degree of 30%, 20%, 15%, 10% and 5%.

The results are described in Table 2.

As shown in Table 2, we can see that the total number and running time of the association rules are higher than those of the small-capacity database when using the proposed algorithm for mining the data of large-capacity database, and the running time of the algorithm increases with the support degree gradually.

Table 2. The mining results of using the proposed algorithm

Support degree/%	Small Capacity Database D1		Large Capacity Database D2	
	The total number of association rules	Operation time/s	The total number of association rules	Operation time/s
30	38	37	56	39
20	104	45	165	47
15	201	49	236	50
10	258	54	295	56
5	399	62	342	58

- **Mining efficiency test**

In order to further verify the performance of the algorithm in mining efficiency, the R-Tree algorithm and the dynamic algorithm are compared, and the mining efficiency is evaluated by mining time, mining speed and acceleration ratio.

The mining time is calculated as follows:

$$T_{comm} \approx o*d*\log\ N + o*\log\ N \tag{7}$$

Where o represents the time required to mine the unit number; d represents the data dimension; N represents the total amount of mining.

The mining algorithm is based on the following formula:

$$S = T_{comm} / \left(T_{comm} + o*\left(d+1\right)\right) \tag{8}$$

Mining speed calculation formula of the mining algorithm is as follows:

$$v = \frac{N}{T_{comm}} \tag{9}$$

The proposed algorithm, R-Tree algorithm and dynamic algorithm are used to mine the small-capacity database D1 and the large-capacity database D2 respectively.

The mining time, mining acceleration ratio and mining speed in the mining process are statistically compared. The results are described in Table 3.

Table 3. Comparison of mining efficiency of three algorithms

algorithm	Small Capacity Database D1			Large Capacity Database D2		
	Mining Time/s	Mining acceleration ratio/n·ms^{-2}	Mining speed/n·ms^{-1}	Mining Time/s	Mining acceleration ratio/n·ms^{-2}	Mining speed/n·ms^{-1}
The proposed algorithm	41.5	14.2	5.92	50.3	16.7	5.37
R-Tree algorithm	49.6	10.68	4.37	55.6	12.93	4.92
dynamic algorithm	53.8	8.35	4.25	61.5	10.56	3.88

As shown in Table 3, it can be seen that the mining time of the proposed algorithm is always lower than that of R-Tree algorithm and dynamic algorithm, whether it is for small-capacity database D1 or large-capacity database.

The mining acceleration ratio and mining speed are always higher than that of R-Tree algorithm and dynamic algorithm, indicating the overall mining efficiency of this algorithm is higher.

- **Mining accuracy test**

Data mining is through mining algorithms to find more and more information which are relevant to mining. So in this section, on the basis of the test data in the above section, for the different databases, some irrelevant data are collected, to constitute the test data set.

The algorithm is tested if it can effectively get more relevant results in the same time period.

In ensuring the recall ratio and the precision ratio at the same time, the mining algorithm needs to have higher performance.

The test data set is described in Table 4.

Table 4. Testing Data Set

Name	Number of relevant mining/pcs	Number of irrelevant mining/pcs	The proportion of the number of relevant mining to total/%
Database 1	499	2482	16.74
Database 2	514	2411	17.57
Database 3	402	2426	14.21
Database 4	536	2344	18.61
Database 5	529	1995	20.96
Database 6	511	2334	17.96
Database 7	339	2716	11.10
Database 8	471	2127	18.13
Database 9	552	2753	16.70
Database 10	339	1051	24.39
Total	4692	22639	100

The efficiency of the mining algorithm is measured by the recall ratio and the precision ratio.

The calculation formulas are as follows:

$$\text{Recall ratio}=\frac{|R_A|}{|R|} \tag{10}$$

$$\text{Precision ratio}=\frac{|R_A|}{|A|} \tag{11}$$

Where $|R|$ is used to describe the number in the relevant R; $|R_A|$ represents that the test set is related to the initial mining q; $|A|$ is used to describe the number of set A.

In the process of mining the database, the recall ratio and the precision ratio is generally regarded as two contradictory indicators. In some cases, the precision ratio will decrease with the increase in the recall ratio, that is, in order to achieve the information recall, it will produce some useless information to a large extent.

And in the case of that the recall rate is gradually increased, the mining speed will be greatly reduced.

Therefore, in the precision testing process of the mining algorithm, it is necessary to test the recall ratio of the algorithm for various recall ratios.

In order to facilitate the analysis, the precision ratio under the level of recall ratio is made treatment in this section, the formula is described as follows:

$$\bar{P}(r)=\sum_{i=1}^{Nq}\frac{P_i(r)}{Nq} \tag{12}$$

Where $P(r)$ is used to describe the average precision ratio when the recall ratio is r; Nq is used to describe the total amount of mining; $P_i(r)$ is used to describe the precision ratio of the i-th mining when the recall ratio is r.

In this section, the test data in Table 4 are used to compare the accuracy of the proposed algorithm, R-tree algorithm and dynamic algorithm at the level of each recall rate. After the treatment of the precision ratio by the above equation, the obtained results are described in Table 5.

Table 5. Comparison results of the average of precision ratio by three algorithms at different recall ratio levels

Recall ratio/%	The proposed algorithm/%	R-tree algorithm/%	Dynamic algorithm/%
10	83.2	80.7	81.1
20	77.8	62.2	65.3
30	69.3	50.4	52.6
40	53.8	32.6	41.5
50	51.5	27.8	36.2
60	48.9	25.1	31.1
70	39.2	23.2	25.7
80	31.2	17.9	22.9

From Table 5, we can see that in the same recall rate to ensure that the state, the precision ratio of the proposed algorithm is significantly higher than that of the R-tree algorithm and the dynamic algorithm.

At the same time, with the gradually increasing of the recall ratio, the proposed algorithm can effectively control the noise data caused by mining, so that the precision rate is kept at a very high level, indicating that the algorithm has the highest accuracy.

In this paper, a new data mining algorithm based on association rule algorithm is proposed. Firstly, the weight of the keywords in the data mining process is calculated by using the TF-IDF function.

Based on the analysis of the association rules algorithm and the vector space model, the information retrieval model is constructed. On the basis of this, the association rule algorithm including the initial mining items is used to establish the rule database, and several words with the highest degree of relevance to the mining words are selected as the extension words, which is combined with the initial mining into a new mining.

Finally, the clustering analysis of new mining results are carried out by K-means clustering algorithm, and according to the descending order, the relevance is ordered, to output mining results. The experimental results show that the algorithm is accurate and efficient.

5. Conclusions

In this paper, a new data mining technology based on association rule algorithm is proposed. The weight calculation process, association rule algorithm and information retrieval model are analyzed.

The K-means clustering algorithm is used to make clustering analysis of the new mining results, and according to the descending order, the correlation is ordered, to output the mining results.

The experimental results show that the algorithm is accurate and efficient.

Acknowledgements

A Project Supported by Scientific Reserch Fund of Sichuan Proincial Education Department - Application of data mining algorithm in intelligent service - (No. LYC15-16);

A Project Supported by Scientific Reserch Fund of Sichuan Proincial Education Department - Research on network public opinion supervision and hotspot discovery based on topic crawler technology (No. 15ZB0258).

References

[1] Wang D W. Android platform of logistics information query and application software design. Electronic Design Engineering, 2016, 24(19): 122-124.

[2] Guo H L, Li J. Aero Engine Gas Path Fault Diagnosis Research Based on Association Rule Mining. Computer Measurement and Control, 2014, 22(2): 379-381.

[3] Kim Y, Yang H, Chung C W. SEDRIS Transmittal Storing and Retrieval System using Relational Databases. Journal of Database Management, 2014, 25(4): 38-65.

[4] Chen K, Wang D D. Database Indexing Algorithm Based on Association Rules Data Structure Rearrangement. Bulletin of Science and Technology, 2015, 31(10): 178-180.

[5] Zhang X, Gong X F. Interaction Mechanism of Dynamic Selection Based on Terminal and Database. Science Technology and Engineering, 2014, 14(28): 256-260.

[6] Kremcheeva, D. and Kremcheev, E. Use of a Quality Management System at the Iron and Steel Enterprise. Journal of Mechanical Engineering Research and Developments, 2018, 41(1): 151-155.

[7] Banerjee, A. and Bej, S. On Extension of Regular Graphs. Journal of Discrete Mathematical Sciences and Cryptography, 2018, 21(1): 13-21.

[8] Ziane, D. and Cherif, M.H. Variational Iteration Transform Method for Fractional Differential Equations. Journal of Interdisciplinary Mathematics, 2018, 21(1): 185-199.

[9] Delhibabu R, Behrend A. A new rational algorithm for view updating in relational databases. Applied Intelligence, 2015, 42(3): 1-15.

[10] Wang B, Wang J, Ning X Q. Algorithm of Mining Association Rules Based on Golden Ratio. Computer Simulation, 2015, 32(8): 302-305.

[11] Zhou L J, Wang X. Research on association rules algorithms based on cloud environment. Computer Engineering and Design, 2014, 35(2): 499-503.

[12] Vainio J, Junkkari M. SQL-based semantics for path expressions over hierarchical data in relational databases. Journal of Information Science, 2014, 40(3): 293-312.

[13] Wang P. The Application of Data Mining Algorithm Based on Association Rules in the Analysis of Football Tactics// International Conference on Robots and Intelligent System. 2016: 418-421.

[14] Martín D, Rosete A, Alcalá-Fdez J, et al. A New Multiobjective Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules. IEEE Transactions on Evolutionary Computation, 2014, 18(1): 54-69.

[15] Cheng M, Xu K, Gong X. Research on audit log association rule mining based on improved Apriori algorithm. Journal of Computer Applications, 2016: 1-7.
[16] Amor G, Rihab C, Moez A. CNOT-based design and query management in quantum relational databases. International Journal of Quantum Information, 2014, 12(04): 507-523.
[17] Liu J, Zhang X. Construction of scientific journal databases from the perspective of "mega-data". Acta Editologica, 2014, 26(1): 59-62.
[18] Jin C Q, Qian W N, Zhou M Q, et al. Benchmarking Data Management Systems: From Traditional Database to Emergent Big Data. Chinese Journal of Computers, 2015, 38(1): 18-34.
[19] Parssian A, Yeoh W, Ee M S. Quality-Based SQL: Specifying Information Quality in Relational Database Queries. Computer, 2015, 48(9): 69-74.
[20] Huang X. An Image Data Query Algorithm Based on Ontology and Singular Value Decomposition. Acta Electronica Sinica, 2014, 42(2): 288-291.
[21] Dokeroglu T, Bayir M A, Cosar A. Robust heuristic algorithms for exploiting the common tasks of relational cloud database queries. Applied Soft Computing, 2015, 30(C): 72-82.
[22] Han X X, Li J Z, Gao H. PAA: An Efficient Approximate Aggregation Algorithm on Massive Data. Journal of Computer Research and Development, 2014, 51(1): 41-53.
[23] Ceci M, Cuzzocrea A, Malerba D. Effectively and efficiently supporting roll-up and drill-down OLAP operations over continuous dimensions via hierarchical clustering. Journal of Intelligent Information Systems, 2015, 44(3): 309-333.
[24] Li B Q. Model Simulation of Big Data Key Characteristics Mining for Peer-to-Peer Networks. Computer Simulations, 2014, 31(11): 294-296.