Journal of

Software Engineering

ISSN 1819-4311

Academic
Journals Inc.

www.academicjournals.com

OPEN ACCESS Journal of Software Engineering

ISSN 1819-4311
DOI: 10.3923/jse.2017.217.223

Research Article
Application of Association Rule Mining Algorithm in Logistics Information System Design

Jiaxin Wang

Department of Basic Education, Beijing Polytechnic, 100176 Beijing, China

Abstract

Background: Currently, the rapid development of computers and the internet technology has intensified social informatization and business intelligentization. Under the circumstance, the modern logistics industry is more dependent on economic globalization and integration, consequently, it is necessary for the logistics industry to take full advantage of information technology to enhance market competitiveness and increase profits for enterprises. **Materials and Methods:** The traditional Apriori algorithm can generate lots of candidate item sets in association rule mining and the data are needed to be read for repeated times. Therefore, the traditional algorithm is very inefficient and a large number of invalid item sets are produced, which can result in the waste of storage and time. To deal with these problems, this study introduced a kind of mining algorithm of weighted association rules based on partition technology. It was loaded to structured query language server 2012 for data processing. The data set was processed with mining while parameters were varied. Moreover, the algorithm was applied in logistics information system for further experimental verification. The association rule mining algorithm (provided by SQL server) was applied for the mining of the same data set under the same conditions and contrast experimental verification was performed in logistics information system. **Results:** When the minimum support was low, the execution time of the algorithm proposed in this study was shorter than that of the association rule mining algorithm provided by SQL server and its performance was better. **Conclusion:** It proves that the proposed algorithm is superior, it is beneficial to the development of the logistics information system.

Key words: Data mining, association rule mining, logistics information system, weighted, WPAR algorithm

Received: August 04, 2016 Accepted: November 28, 2016 Published: March 15, 2017

Citation: Jiaxin Wang, 2017. Application of association rule mining algorithm in logistics information system design. J. Software Eng., 11: 217-223.

Corresponding Author: Jiaxin Wang, Department of Basic Education, Beijing Polytechnic, 100176 Beijing, China

Copyright: © 2017 Jiaxin Wang. This is an open access article distributed under the terms of the creative commons attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Competing Interest: The author has declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Data mining, also referred to as knowledge discovery from databases is a process of extracting valuable knowledge from a large amount of random data[1]. In the face of massive data and small amount of information, association rule, an important research branch of data mining is in the ascendant as an advanced and intelligent data processing and analysis technique[2].

Based on interval concept lattice, Qodmanan *et al.*[3] proposed an association rule mining model and proved the validity of the mining algorithm. Mao *et al.*[4] put forward a multilevel association rule mining algorithm and generalized association rule mining algorithm which had better mining efficiency and expansibility than the current similar algorithms.

The flourishing of the logistics industry makes it particularly important to develop logistics information technology. However, the existing data mining algorithms can not fully meet the requirements of high-efficiency processing of big data. Accordingly, based on the previous study achievements regarding data mining technology, this study applies the association rule mining algorithm, which is the WPAR algorithm, to design the logistics information system and develop the logistics information technology. Such algorithm could partition databases and allow each portion to operate in the internal storage. Moreover, only 2 times of database scanning were needed, which could reduce the operation time and thus increase the speed. In production of association rules, methods that were based on profits weighting were adopted, which catered to the requirements of enterprises. In addition, the superiority of such algorithm was verified through experiments.

MATERIALS AND METHODS

Data mining and data warehousing: The reason why data mining is in great demand is that enterprises need to search and obtain important information which is related with market development trend, suppliers, enterprise competitors, clients and demands from a mass of data[5]. The data is massive while useful information is usually inexplicit, therefore, it is necessary for enterprises to be armed with data mining technology.

Business data are usually original data which are not sorted or extracted. In order to use the original data effectively, data warehouse needs to be designed. Data warehouse contains the processed data which is obtained from database, information can be extracted by analyzing and mining the processed data to support decision making.

Association rule mining: There is an extensive application of association rule mining in exposing the potential links, rules or patterns between things. Its earliest application is aimed at discovering the sale relationship among commodities in supermarket transaction database[6]. The association rules are in X→Y form, in which the X and Y refer to one or several properties in data mining and are also called items. The theoretical basis of the association rule mining algorithm in this study was data mining theory algorithm.

Apriori algorithm: Using Apriori algorithm, frequent item sets were expanded constantly, when larger frequent item sets were generated, every 2 of them were connected based on a certain rule, generating candidate item sets, then by reading the database, with counting statistics on the item sets, if the count satisfied the threshold, it was an frequent item set, otherwise, it was deleted from the candidate item sets. The pseudo codes are as follows:

```
C 1 = {candidate 1-item sets};
L1 = {c∈C1|c.count≥minsupport};
For (k = 2, Lk-1 ≠ Φ, k++) // until no larger item sets can be generated
Ck = sc_candidate (Lk-1); // generating candidate item sets with k elements
for all transactions t∈D // transaction and processing
Ct = count support (Ck, t); // candidate item sets included in transaction t
for all candidates c∈Ct
c.count = c.count+1; // item sets count
Lk = {c∈Ck|c.count≥minSupport};
resultset = resultset∪Lk;
```

The first step of Apriori algorithm was to scan the database and count the frequency of each item in the database, the items that exceeded the minimum support were frequent item sets, based on which frequent 1-item sets were obtained (also known as large item sets). The set of frequent 1-item sets was denoted by $Large_1$. The set of 2 candidate 2-item sets from frequent 1-item sets was denoted by $Candidate_2$ (followed by procedures). According to the database statistics count through scanning of the items in item sets of $Candidate_2$, frequent 2-item sets were obtained, which was also known as pruning, namely, to remove the items that did not belong to the frequent item sets[7]. If a database 1-item set included 103 items, there were about 10^6 candidate 2-item sets. Generally, the number of item sets recorded in enterprise-class database is even greater.

Weighted association rule mining algorithm based on the partitioning technique: In database mining using the Apriori algorithm, lots of candidate item sets are generated. Thus, there are multiple times of database reading. The efficiency of Apriori algorithm can be extremely low when it comes to a

large quantity of data. Besides, in generation of frequency item sets, all items in the candidate item sets should be connected. Thus, invalid middle item sets can be generated, resulting in an extra waste of time and space[8]. Based on these, this study put forward an improved weighted association rule mining algorithm WPAR algorithm based on the partitioning technology.

Data structure and definition: Before data mining, some signs and structure needed to be defined. Transaction format was defined as <Tid, i_1, i_2, ⋯, i_n>, the items were stored in lexicographic order, transaction Tid was monotonically increasing. Furthermore, it was assumed that data were stored in secondary storage and the exact size of the database was known in advance.

Local support count of an item set is referred to as support count in a fragment that contains the item set[9]. A local frequent item set refers to a large or frequent item set in a subblock. Local candidate item sets are candidate item sets in a subblock but not necessarily in the whole database. Corresponding, global support count refers to support count in the entire database, global frequent item sets refer to the item sets whose support count sum in all blocks satisfies the threshold, global candidate item sets refer to the candidate item sets with regard to the whole data set .

The algorithm in this study adopted a data structure in vertical congruent relationship, which could avoid frequency item sets produced by the matched pattern, each item or item set was given a Tidlist to store all transaction record ID that contained the item or item set. Besides, the horizontal data structure was not adopted in this study[10]. Figure 1 and 2, respectively show the horizontal and vertical data structure. Horizontal corresponding data structure refers to the collection of each transaction ID and the items it contains, while vertical corresponding structure refers to the collection of each item and all the transaction IDs that contain the items.

TID	Itemset
01	ABE
02	BCDE
03	ACDE

Fig. 1: Horizontal data structure

Item	A	B	C	D	E
Tidset	01	01	02	02	01
	02	02	03	03	02
					03

Fig. 2: Vertical data structure

When the database was first scanned, Tidlist table of each item or item set was set up. Transaction IDs that contained the items or item sets were added to Tidlist in ascending order. By solving the common item amount of Tidlist of two items or item sets, it could be determined whether the two items or item sets could constitute a larger frequent item set.

The WPAR algorithm was used to solve the local frequent k-item set of a partition P_i. With regard to the items A and B in frequent (k-1)-item set, if the common item amount of A.Tidlist and B.Tidlist was greater than the minimum support count, the common items of A and B were saved, denoted as Tidlist of {A, B}. If the minimum support was not satisfied, the information was not saved[11,12]. When local operation in the subblock was completed, if the maximal local frequent item set was k-item set, Tidlist of the non-frequent i-item sets was deleted to release memory space.

WPAR algorithm: Compared with the traditional Apriori algorithm, the weighted association rule algorithm based on partitioning in this study has following innovation points. First of all, databases are partitioned and each part can operate in the internal storage, which can reduce the times of input/output (I/O) operations when the database is too big to be stored. Thus, it is time-saving. Secondly, only 2 times of database scanning are required, thus it saves the operation compared with the repeated scanning of Apriori algorithm. Thirdly, starting from the interest of users, merchandises that have different profits are given profit weighting processing when association rules are generated. Thus, it can be avoided that item sets that have few counts and big profits are ignored. The implementation of this algorithm included three stages.

The first scanning of database was aimed at weighting processing of the transaction records based on profits, the total profit of each item (within a certain period of time) was counted, profit range was determined according to the profits of all items, weight values of the items were determined according to the profit range, eventually, weighting table of each item was set up based on the profits.

By re-scanning, the database was randomly divided into n blocks. This stage included an n cycle, in the ith cycle, only P_i (partition i) was considered, according to the weight of each item and the data obtained from normalization processing, local frequent item sets of all length (namely, frequent 1-item set, frequent 2-item set, ⋯, frequent k-item set in partition P_i) were generated.

Then, merging operation was performed, counting statistics for the local frequent item sets of the same length of

all the blocks was implemented with regard to the whole data set and global frequent item sets were generated.

In the third stage, of the generated local frequent item sets, Tidlist tables of the same local frequent item sets were merged, the number of transactions in Tidlist was the total count of the item sets in all the n subblocks, so that global frequent item sets were generated. Then, the frequent item sets generated association rules.

The WPAR algorithm only required scanning twice on the database. Initially, each item or item set was set up with a counter. The record contained the items or transaction amount of item sets, namely, the length of Tidlist. When global frequent item sets were solved, intersection set of all local frequent item sets belonged to each subblock, therefore, these item sets were global frequent itemsets[13-15].

Correctness verification: The global frequent item sets, which were the subsets of all local frequent item sets were obtained according to local frequent item sets in each subblock. What needed to be proved was that all potential global frequent item sets were local frequent item sets in at least one subblock. Since, C^g (the set of candidate item sets) was the set of local frequent item sets in all subblocks, global frequent item sets were solved from the set, suppose L was the set of actual global frequent item sets in the database, then $L \subseteq C^g$.

Assume that item 1 in L was not contained in C^g, namely, $1 \notin C^g \wedge 1 \in L$, which meant an actual global frequent item set was not contained in local frequent item sets of any subblock. Suppose w_i (1) was the Tidlist size of item set 1 in subblock i, w_i was the count of transaction record in subblock i. Since $1 \notin C^g$ (item set 1 was not the local frequent item set of any block with regard to i = 1, 2, 3, 4, ···, n), there is:

$$w^1(1)/w^1<\text{minSupport} \wedge w^2(1)/w^2<\text{minSupport} \wedge \ldots \wedge w^n(1)/w^n<\text{minSupport}$$

or:

$$w^1(1)<\text{minSupport}*w^1(1)<\text{minSupport}*w^2 \wedge \ldots \wedge w^n(1)<\text{minSupport}*w^n$$

which was equal to $w^1(1)+w^2(1)+\ldots+w^n(1)<\text{minSupport}*(w^1+w^2+\ldots+w^n)$.

$w^1(1)+w^2(1)+\ldots+w^n(1)$ was the total count of item set 1 and $w^1+w^2+\ldots+w^n = n$ was the count of all transaction records in the database, then, S^g (1)<minSupport, which was contradictory with the case that 1 was a global frequent item set, therefore, in this algorithm, the generation of global frequent item sets was correct.

Chi-square test: Based on the idea of chi-square distribution, consensus of observing frequency and expected frequency was assumed and the deviation between them was calculated. A was defined as the observing frequency of an item, E was defined as expected frequency, the difference between A and E was called deviation:

$$X^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i-E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(A_i-np_i)^2}{np_i}$$

Chi-square test was generally used to measure the correlation between 2 items or patterns, however, it could not distinguish whether it was negative or positive correlation and certain premise conditions were required. Suppose item A and item B were independent of each other, then P (A, B) = P (A)*P (B), namely, support (A, B) = support (A)*support (B). If A and B were associated, there was deviation between P (A, B) and P (A)*P (B).

Application of association rule mining algorithm and experimental verification: Association rule mining was applied in logistics information system to seek potential rules between item sets or attributes which might be significant to the management and operation of enterprises. In general, it is applied to cross-selling, customer relationship management and analysis of lost customers.

To verify performance of the improved algorithm, this study applied Microsoft SQL Server and Microsoft Visual Studio to perform the improved algorithm. Meanwhile, the association rule mining algorithm provided by Microsoft SQL server was also performed. The two kinds of algorithms were applied to mining of the same data set so that the results were compared and analyzed.

Data set in this experiment was obtained from chess data set, which was usually used as testing data set to test the performance of association rule mining algorithm. The other data sets that could be applied to testing data set included mushroom, connect, Pumsb, Pumsb*, T10I4D100k, T25I1OD10k and T25I20D100, which were commonly used in normal data sets, provided by UCIrvine Machine Learning Database Repository and IBM Almaden. The data set was used to simulate the sales transaction record set of an enterprise. The data set had 4037 transaction records and 78 items, the average length of each transaction was 35 items.

RESULTS

Association rule mining algorithm is mainly concerned with exponential order[16], namely, when the value of minimum

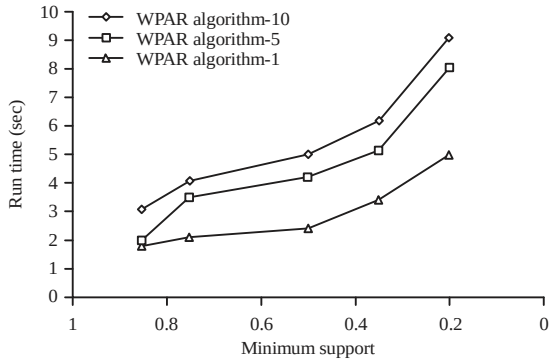

Fig. 3: Comparison of run time when the amounts of subblocks were 1, 5 and 10

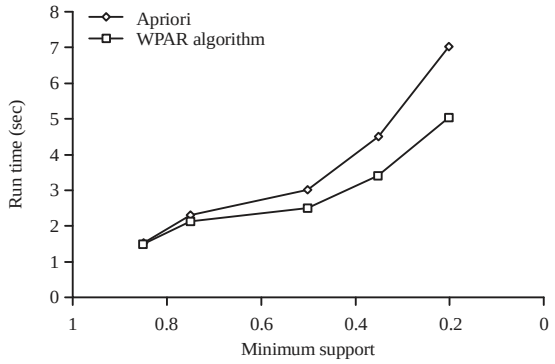

Fig. 4: Comparison of run time with different minimum support

support decreased, execution time of the algorithm would increase exponentially. The increase of frequent item sets resulted in the increase of execution time.

Comparison of run time with varying subblock amount: The requirement for the scale of partition was that each subblock and its operation could be implemented in internal storage, at each stage, partition could be read only once. After partition, the merging procedure also consumed extra space and time, in this study, the amount of subblocks was associated with data set size and memory size. Execution time of the algorithm varied with the amount of subblocks. On condition that other parameters were constant, the amounts of subblocks were 1, 5 and 10, respectively. Figure 3 shows the run time of the algorithm with different amounts of subblocks. The WPAR algorithm-1, WPAR algorithm-5 and WPAR algorithm-10 respectively show the operation conditions of algorithms when the amounts of subblocks were 1, 5 and 10 (the minimum support was the same).

The data set could be operated in internal storage since it was not large. After partition, execution time was longer than that without partition. After partition, merging operation was required and the local frequent item sets were not necessarily global frequent item sets[17], a lot of local frequent item sets needed to be processed in the merging phase, which required time consumption.

Comparison of run time with varying minimum support: Minimum support refers to the minimum probability of item sets in transactions. When the minimum support is different, the amount and size of frequent item sets were different, thus different association rules were generated. In this study, the minimum support was varied (at 0.75, 0.5 and 0.35) to analyze the influence of minimum support on the algorithm. Figure 4 shows the time to execute algorithm with different minimum support while other parameters were unchanged.

It can be concluded from Fig. 4 that as the minimum support decreased, the execution time of both WPAR algorithm and association rule mining algorithm (provided by SQL server) increased when there was only one subblock. Since, the decrease of minimum support resulted in the increase of local frequent item sets, the execution time also increased. Moreover, when the minimum support was lower, the execution time of the WPAR algorithm was shorter than that of the association rule mining algorithm, with relatively better performance. In the execution process of WPAR algorithm, data structure and weighting processing were needed. When the minimum support was relatively larger, fewer candidate item sets and frequent item sets were generated from the mining process of the two algorithms. There were few differences in the time to generate frequent item sets, with WPAR algorithm, Tidlist structure of 1-item sets needed to be built and weighting normalized processing was implemented, therefore, when the minimum support was large, the performance of association rule mining algorithm (provided by SQL server) was better than that of the WPAR algorithm.

Because the data set was not big, the performance advantage of parts reducing I/O operations was not reflected under the condition that the improved WPAR algorithm had shorter execution time compared with that of the association rule mining algorithm provided by SQL server. It only avoided the pattern matching and thus improved the time performance when candidate or frequency item sets were generated[18]. However, due to the limitations of data sets and system environment, the improvement of the algorithm should be further studied.

DISCUSSION

Based on the latest technology related to data mining, this study summarized the process and functions of data mining and put forward an improved association rules mining algorithm which was loaded to SQL server 2008. According to this, Dai[19] analyzed the development and application of the association rule mining algorithm to various fields in recent years. Moreover, they applied the algorithm to the logistics industry, which improved marketing services of the logistics, optimized logistics paths as well as enhanced logistic performance[19]. It is found that compared with the traditional Apriori algorithm, the improved algorithm had following advantages. Firstly, the segmentation processing of databases during data processing reduced times of I/O operation. According to comparisons of the minimum support and performances, the study algorithm was better than the traditional one. Secondly, times of database scanning were reduced during the processing of logistics information, thus unnecessary operations were avoided and the algorithm speed was improved. Thirdly, based on the interest of users, a field was added to measure the weight, thus the association rules were determined. Therefore, useful information could be acquired, while item sets that had small counts and big profits could not be neglected.

Combining with the research results of other researches, the association rule mining algorithm could accelerate the processing speed of logistic services as well as optimize the design of the logistic information system. However, in consideration of the influences of hardware and software environments and the complexity of practical logistic data, the algorithm in this study still had some deficiencies. Firstly, the plug-in algorithms had deficiencies in prediction, though it could extend functions as required. Secondly, the validity of association rule mining should be further improved. Thirdly, mined rules should be classified for selection.

CONCLUSION

This study was focused on the research of association rule mining algorithm. By comparing WPAR algorithm and association rule mining algorithm (provided by SQL server), this study performed experimental verification of the algorithms in logistics information system. It proved that WPAR algorithm was superior to some degree. However, due to the limitations in experimental conditions and human factors, there are some defects in the analysis, which will be improved in future studies.

SIGNIFICANCE STATEMENTS

This study advanced the development of data mining algorithm.

This study achieved the proper application of WPAR algorithm and the successful development of the logistics information system.

Data mining technology was applied to logistics information system, which accelerated the development of business logistics.

ACKNOWLEDGMENT

This study funded by the National Science Foundation of China, the study was supported by the Science and Technology Innovation Team Project of Zhejiang Province and the Research Group of Logistics Information System Design.

REFERENCES

1. Kacprzyk, J. and S. Zadrozny, 1998. Data Mining via Linguistic Summaries of Databases: An Interactive Approach. In: Methodologies for the Conception, Design and Application of the Soft Computing, Yamakawa, T. and G. Matsumoto (Eds.). World Scientific, Iizuka, Japan, pp: 667-668.
2. Srikant, R. and R. Agrawal, 1997. Mining generalized association rules. Future Generat. Comput. Syst., 13: 161-180.
3. Qodmanan, H., M. Nasiri and B. Minaei-Bidgoli, 2011. Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. Expert Syst. Applic., 38: 288-298.
4. Mao, Y.X., T.B. Chen and B.L. Shi, 2011. Efficient method for mining multiple-level and generalized association rules. J. Software, 22: 2965-2980.
5. Dyer, J.H. and W. Chu, 2011. The determinants of trust in supplier-automaker relationships in the US, Japan and Korea. J. Int. Bus. Stud., 42: 10-27.
6. Xiao, Y., Y. Tian and Q. Zhao, 2012. Discovering frequent itemset with maximum time-window on temporal transaction database using variable neighborhood search. Electron. Notes Discrete Math., 39: 137-144.
7. Fernandez, I., 2015. Relational Database Management Systems. In: Beginning Oracle Database 12c Administration: From Novice to Professional, Fernandez, I. (Ed.). 2nd Edn., Chapter 1, Apress, USA., ISBN: 978-1-4842-0194-7, pp: 3-24.
8. Tanna, P. and Y. Ghodasara, 2014. Using apriori with WEKA for frequent pattern mining. Int. J. Eng. Trends Technol., 12: 127-131.

9. Sathyanarayanan, D. and M. Krishnamurthy, 2015. An efficient association rule based dynamic support count adaptation model for XML databases using XQuery. J. Int. J. Applied Eng. Res., 10: 16129-16147.
10. Gopalakrishnan, S.G., F. Marks Jr., X. Zhang, J.W. Bao, K.S. Yeh and R. Atlas, 2011. The experimental HWRF system: A study on the influence of horizontal resolution on the structure and intensity changes in tropical cyclones using an idealized framework. Monthly Weather Rev., 139: 1762-1784.
11. Chibisov, D.M., 1971. Certain chi-square type tests for continuous distributions. Theory Probab. Applic., 16: 1-22.
12. Baracho, R.A., L. Branquinho, M.B. Almeida and R.R. Souza, 2015. Using ontologies and inference engine in association rule of data mining: Application in sales strategies in medical laboratory diagnostic market. Proceedings of the 19th World Multi-Conference on Systemics, Cybernetics and Informatics, July 12-15, 2015, Orlando, Florida.
13. Kim, Y., 2009. Streaming Association Rule (SAR) mining with a weighted order-dependent representation of Web navigation patterns. Expert Syst. Applic., 36: 7933-7946.
14. Azadnia, A.H., S. Taheri, P. Ghadimi, M.Z. Mat Saman and K.Y. Wong, 2013. Order batching in warehouses by minimizing total tardiness: A hybrid approach of weighted association rule mining and genetic algorithms. Sci. World J., Vol. 2013. 10.1155/2013/246578.
15. Abdullah, Z., T. Herawan and M.M. Deris, 2011. An Alternative Measure for Mining Weighted Least Association Rule and its Framework. In: Software Engineering and Computer Systems, Zain, J.M., W.M. Wan Mohd and E. El-Qawasmeh (Eds.). Springer, New York, ISBN: 978-3-642-22190-3, pp: 480-494.
16. Wanaskar, U., S. Vij and D. Mukhopadhyay, 2013. A hybrid web recommendation system based on the improved association rule mining algorithm. J. Software Eng. Applic., 6: 396-404.
17. Veloso, A., W. Meira Jr., R. Ferreira, D.G. Neto and S. Parthasarathy, 2004. Asynchronous and anticipatory filter-stream based parallel algorithm for frequent itemset mining. Proceedings of the 8th European Conference on Principles of Data Mining and Knowledge Discovery, September 20-24, 2004, Pisa, Italy, pp: 422-433.
18. Sourdis, I. and D. Pnevmatikatos, 2004. Predecoded CAMs for efficient and high-speed nids pattern matching. Proceedings of the 12th Annual Symposiums on Field-Programmable Custom Computing Machines, April 20-23, 2004, IEEE CS Press, pp: 258-267.
19. Dai, X.T., 2014. Analysis of association rule mining algorithm and its application in intelligent logistics. Sci. Technol. Ind., 14: 113-116.