An Improved Recognition Method of Weighted Rules and Its Application in Recommendation Algorithm

Lei Shi^{a,*}, Shuqing Li^b, Yong Zhang^c

College of Information Engineering, Nanjing University of Finance & Economics, Nanjing, China ^a1176466796@qq.com, ^bleeshuqing@163.com, ^c1341035456@qq.com *Corresponding author

Abstract: The method proposed in this paper mainly uses the difference of the importance of items in the database to improve the problem of mining a large number of redundant and useless rules by traditional association rules mining algorithm. The method has made great improvement in three aspects of association rule design, which are effective length recognition of recommendation rule, weighted association rule mining combined with frequency analysis and time constraint application. The weighted temporal association rules mined by the effective frequency length weighted association rules mining algorithm can improve the accuracy of recommendation, and the accuracy of recommendation prediction is increased from 62% to 69%. The purpose of this algorithm is to mine hidden high value rules, and optimize the algorithm considering the time complexity of algorithm execution while ensuring the accuracy of the algorithm.

Keywords: Recommender System, Data Mining, Association Rules, Frequency Length, Sliding Window

1. Introduction

As an intelligent and personalized information service system, recommendation system describes the user's long-term information with the help of user modeling technology, and realizes targeted personalized information customization through certain intelligent recommendation strategies according to the user model^[1]. At present, the main recommendation methods are text-based content recommendation, collaborative filtering recommendation and knowledge-based recommendation, among which the most important one is association rule recommendation^[2]. Association rules can find hidden relationships between items, and make recommendations more simple, accurate and effective^[3]. In addition, association rules contain many kinds of patterns, among which frequent patterns and sequential patterns are widely used. Sequence pattern is to find out all sequences whose support degree is greater than the minimum support degree threshold in a given sequence data set. The biggest difference between sequence pattern and frequent pattern is that time information is added to the sequence pattern, and the data in the project is no longer in disorder state^[4].

The close relationship between association rules and recommendation system is obvious, and recommendation based on association rules is a simple and effective recommendation model^[5]. Sandwig's research shows that the recommendation based on association rules is more robust than the collaborative filtering model, that is, it is not vulnerable to the negative impact of trust attack^[6]. Recommendation based on association rules is a simple and effective recommendation method, especially in the fields of marketing, retail and e-commerce, for example, Tencent uses association rules to recommend its products and manage its business in real time^[7].

Association rule is a main branch of data mining research field, and weighted association rule is an important topic of association rule mining. The main task of the algorithm is to solve the problem of data association and pattern mining^[8]. Moreover, most researchers only focus on the improvement of weighted association rules mining algorithm, but ignore two urgent problems faced by traditional association rules mining algorithm. The first is that the Apriori algorithm only considers the frequency of items in the database, but ignores the difference of the importance of each item in the database. The second is how to deal with the weight of items and weighting criteria to make the association rules more suitable for practical needs. To solve the above two problems, we can set weights and allow weights to be associated with each item in the transaction to reflect the interest of each item in the transaction, and develop a new recommendation algorithm based on weighted association rule mining method to extend the traditional association rule mining algorithm^[9]. The method of introducing weights is only an improvement of

traditional association rule mining. To make a breakthrough, we need to innovate the rule mining algorithm itself. Therefore, this paper proposes an algorithm for mining weighted association rules based on effective length of frequency, which can mine non-redundant rare weighted association rules and improve the accuracy of recommendation.

2. Literature Review

There are many methods to improve the efficiency of mining frequent itemsets and to mine strong association rules by weighted method. However, the number of association rules in traditional model mining is usually very large, and it is difficult to mine non-redundant association rules^[10]. R. forsati et al.^[11] proposed a new recommendation algorithm based on weighted association rule mining method to expand the traditional association rule mining algorithm, and used the time and frequency spent by each user on each page to allocate the quantitative weight of the page, rather than the traditional binary weight. Unil Yun^[12] proposed a weighted sequential pattern mining algorithm, which applies weights to deep mining sequential patterns. Kuo Cheng Yin et al.^[13] proposed a frequent pattern based on the interval of time series association rule mining algorithm, taking into account the necessity of time characteristics in data mining. Khan et al.^[8] proposed the Weighted Utility Association Rule Mining (WUARM) framework, which can handle project weights and utilities in a mixed manner. This framework can be integrated into the mining process, which is different from most utility and weighted association rule mining algorithms, overcoming the challenges of using weights and utilities together, especially the invalidity of the downward closure property. Zhai et al.^[10] proposed an efficient method for mining weighted association rules from weighted transaction databases, using the unresolvable matrix to quickly find all nodes of the lattice. At the same time, a Frequent Weighted Closed Itemsets Lattice (FWCIL) incremental algorithm was proposed. Ouyang^[12] proposed an algorithm for mining rare weighted association rules on data streams using sliding windows, and used sliding windows to mine weighted association rules in online real-time data streams. Li Chengjun et al.^[13] proposed a new method to calculate the support and confidence of weighted association rules mining. This algorithm reflects the actual importance of each item by weighting, and maintains the downward closure of Apriori algorithm.

Although many scholars have have made good achievements in mining weighted association rules, but they neglect the consideration of long transaction sequence with temporal constraints when mining weighted association rules. Restricted by traditional mining methods, mining methods lack consideration and analysis of the data itself in the database, such as periodicity, generalization and multi-level data. At this time, different mining methods and appropriate weighting methods are needed. In this paper, an algorithm for mining weighted association rules based on effective length of frequency is proposed.

3. Relevant research work

3.1. Summary of Association Rules and Weighted Association Rules

Set $I = \{i_1, i_2, \dots, i_n\}$ to be a collection of items, define data D to be a collection of database transactions, each transaction T to be a collection of items, $T \subseteq I$, and each transaction has an identifier $T_i D$. Assuming that X is an itemset, transaction T contains X if and only if $X \subseteq T$, association rules are implications similar to $X \rightarrow Y$, X is the precursor of rules, Y is the successor of rules, in which $X \subset I, Y \subset I, X \cap I = \emptyset$. Sup is the percentage of $X \to Y$ in transaction D, and the confidence Conf is the percentage of transactions containing X and Y in transaction set D.

MinSup can be set as the minimum support, and if $Sup \ge MinSup$, the itemset is called frequent itemset. Confidence Conf is the percentage of transactions D that contain both X and Y in transaction set D. Then the support degree of rule $X \to Y$ is $P(X \cup Y)$ and the confidence degree is P(X|Y).

Set $I = \{i_1, i_2, \dots, i_n\}$ to be a collection of items, then there are n items in transaction set D of database, and each item is assigned to a corresponding weight. Then their weights can be expressed as $W = \{w_1, w_2, \dots, w_n\}$ respectively, in which $0 \le wi \le 1, i = \{1, 2, \dots, n\}$. The minimum weighted support threshold WMinSup and the minimum weighted confidence threshold WMinConf are specified for pruning weighted association rules.

The essence of association rules mining is mining frequent sets. In the traditional Apriori algorithm, if $\{XY\}$ and $\{YZ\}$ are not frequent, then $\{XYZ\}$ and $\{YZW\}$ are not frequent, we can draw a conclusion that the superset of the infrequent set is also infrequent. During the execution of the algorithm,

the efficiency of the algorithm is improved by pruning the non-frequent sets.But this conclusion is not valid in the mining of weighted association rules. In other words, any subset of weighted frequent itemsets may not be weighted frequent itemsets; any subset of non-weighted frequent itemsets may be weighted frequent sets^[14]. Therefore, the strategy of this paper is to mine association rules, and then prove that some rules are weighted frequently.

3.2. Frequency Effective Length Method

In order to solve the problem of mining a large number of redundant low-value rules by traditional mining algorithms, this paper proposes an algorithm of mining weighted association rules based on effective frequency length. On the basis of studying the value information of item frequency in database, the method of effective frequency length and the method of setting weights are established. Through the combination of these two methods, the mining of rare weighted association rules is carried out.

3.2.1. Sliding Window Technology and Time Series Association Rules

In this paper, one of the basic elements of mining weighted association rules algorithm is that the algorithm should be carried out in sequential data with temporal constraints. Because the real world is developing constantly, so most transaction information has the time characteristic. Because of the huge amount of time series data and the long time span, the data can be divided into "new" and "old". In order to make good use of this time series data, the process of mining rules is put into a sliding window. The mechanism of sliding window can be divided into two stages: initial stage and sliding stage. In the initial stage, the new data is continuously moved into the sliding window. When the window length is reached, the data will stop moving in.In the sliding stage, the classical sliding window adopts the method of moving coverage, with the sliding of the window, "old" data moves out, and "new" data moves in .The purpose is to mine the most valuable temporal association rules in the current sliding window period.

For example, by analyzing supermarket data sets in traditional association rule mining, association rules such as "turkey—pumpkin pie (support = 0.0001, confidence = 0.05)" can be obtained, which means that 0.01% of transactions contain Turkey and pumpkin pie, and 5% of all transactions containing Turkey contain pumpkin pie. The above rules cannot be regarded as a prominent association rule, because their support and confidence are too low. If a promotional campaign is launched in the summer based on the discovered rules, such as a 10% discount on Turkey and pumpkin pie" will not be sold in large quantities until a few weeks before the Thanksgiving (winter) holiday. From this example, it can be seen that item sets are frequent in specific time intervals. Therefore, it is very important to find the time interval of frequent occurrence of these patterns effectively.

3.2.2. Frequency Effective Length Method

From the above example of "turkey and pumpkin pie", it can be seen that "turkey" and "pumpkin pie" are frequent in specific time intervals, that is to say, "turkey" and "pumpkin pie" frequent time length (cycle) is a year. Therefore, there is a basis for guessing that there is a periodic relationship between the data in the time series. Although it can be guessed that the time series data in the database has periodicity, the periodic objects are unknown. Therefore, by observing and experimenting the time series data in the database, it is found that there are types of items with different frequencies in the transaction data, that is, the type items with higher frequencies tend to be more important than other types of items. Thus, we can count the type with the highest frequency in each transaction, and then calculate the frequency effective period of this type.

If the recommended prediction range is only temporal sequence data, the next node must have defects, because there may be multiple successors of association rules, it is obviously unreasonable to select only one of them to de prediction. Therefore, it is necessary to explore which nodes of sequential data have the highest accuracy. The effective length of frequency is the sum of the effective period of frequency and the length of the predicted node. That is to say, the effective length of frequency will be the length of sliding window. Finally, association rules can be mined within the effective length of frequency, which can discover high-value rules. Therefore, the new method is called the effective length of frequency method.

3.3. Mining Weighted Association Rules with Frequency Effective Length

3.3.1. Algorithmic description

Generally, the classical mining of weighted association rules is divided into two steps^[15]: mining the initial association rules and selecting the rules that meet the requirements by weighting calculation. The basic steps of the improved weighted association rule algorithm are consistent with the classical algorithm^[16].

Algorithm1: Frequency Effective Length Method for Mining Unweighted Association Rules

Input: database dataD; Output: Unweighted Association Rules

Begin

1) for i = 1 to *D* do // Traversing database data *D*

- for j = 1 to sliding windows(*sw*) do // Traversing data in sliding windows
- 2) Generate rules by new method
- 3) Filter by their own sequence

// Filtering Irrelevant Rules with Successive Sequences of Mining Rules

4) End

5) End

In algorithm 1, the association rules of time series are first mined in the effective length of frequency (sliding window), and then some useless rules are filtered out through the sequence data of itself.

Algorithm2: Mining Weighted Association Rules in Sliding Window

Input: Unweighted Association Rules; Minimum Weighted Support; Minimum Weighted Confidence; Database DataD; Rating Data *Rate*

Output: Weighted association rules satisfying minimum weighted support and minimum weighted confidence

Begin

3) 4)

1) for each $R' \in R$ do //Traversing through each rule in the rule base

- 2) for s = 1 to D, do D //Traversal Time Series Data and Scoring Data
 - for x = 1 to U do //Traversing through each sequence of time series
 - Calculate Frequency and Rating by using eq1 and eq2
- 5) End
- 6) End
- 7) Calculate the weight of transaction and the sum weight of transaction by using eq3
- 8) If R' in sw then //If the rule is in the sliding window, proceed to the next step

9) Calculate the *wsup* and *wconf*//Calculating Weighted Support and Weighted Confidence10) End

- 11) If $R'.wSup \ge WMinSup$ and $R'.wConf \ge WMinConf//Comparing Minimum Weighted Support and Minimum Weighted Confidence$
- 12) Add R' to queue //Add rules that meet the requirements to the list
- 13) End
- 14) End

In algorithm 2, the weighted support and the weighted confidence of association rules are calculated, and the association rules higher than the minimum weighted support and the minimum weighted confidence are selected.

After validating the effective length of frequency in sequential data of database, because the effective length of frequency includes the effective period of frequency and the length of predicted nodes, it is strictly arranged as the leader of association rules in the effective period of frequency. The data within the predicted node length are grouped strictly in order to be the successor of association rules. In this way, more association rules with temporal constraints can be mined.

3.3.2. Design of weights

The author puts forward three formulas.

The average frequency of each type(Frequency) refers to "the sum of quantities of each type" divided by "the sum of total quantities of all types", as shown in formula (1).

 $Frequency = \frac{Number \cdot of \cdot every_type}{Number \cdot of \cdot all_type} (1)$

The average score for each type(Rating) refers to "the sum of scores for each type" divided by "the total number of times for that type", as shown in formula (2).

$$Rating = \frac{Number \cdot of \cdot every_rates}{Number \cdot of \cdot every_type}$$
(2)

The final weight(Weight) refers to "the product of the average frequency of each type and the average score of each type for * operation", as shown in formula (3).

$$Weight = \{[Normalize(Frequency)] \times [Normalize(Rating)]\} * (3)$$

Among them, * means to scale the final weight to [0,1]. In order to make the weight meaningful and prevent overflow of [0,1], or infinite approximation to a value in [0,1], the scaling operation is needed witch is commonly used.

3.3.3. Mining Weighted Association Rules in Sliding Window

Since non-weighted association rules have been mined in the effective length of frequency, it is now verified whether the rules are weighted frequently, and then the weighted frequent rules are selected. The data set is shown in Table 1.

Users	Types	Scores
1	A,C,T,W	3, 1, 5, 0.5
2	C,D,T	1,0.5,5
3	A,C,W	3, 1, 1.5
4	A,C,D,W	4, 1, 2, 1
5	C,D,W	1, 2, 1

Table 1: Film Types and Scores.

According to formula (1), the average frequency of type A is equal to $3/17\approx0.18$; Similarly, the average frequencies of C, D, T and W are about 0.29, 0.18, 0.12, 0.24 respectively. According to formula (2), the average score of type A is equal to $(3 + 3 + 4) / 3\approx3.33$; Similarly, the average score of type C, D, T and W is 1, 1.5, 5, 1. According to formula (3), the final weights of type A are $0.18*3.33\approx0.6$; Similarly, the final weights of type C, D, T and W are about 0.3, 0.3, 0.6, 0.2 after normalization. The average weights of each type of item are calculated, and the average weights of user sequence are shown in Table 2.

Users	Types	Computation process	Average Weights of User Sequences
1	A,C,T,W	(0.6+0.3+0.6+0.2)/4	0.43
2	C,D,T	(0.3+0.3+0.6)/3	0.4
3	A,C,W	(0.6+0.3+0.2)/3	0.37
4	A,C,D,W	(0.6+0.3+0.3+0.2)/4	0.35
5	C,D,W	(0.3+0.3+0.2)/3	0.27
sum			1.82

Table 2: Average Weights of User Sequences.

The association rules mined in the effective length of frequency are weighted. The weighted support of association rules is calculated, and the association rules below the minimum weighted support threshold are pruned.Calculate the weighted support of association rules in Table 2, the weighted support of $\langle CT \rangle$ is equal to (0.43+0.4)/1.82=0.46.Similarly, the weighted support of $\langle ACW \rangle$ and $\langle CW \rangle$ is 0.63 and 0.78. If the minimum weighted support is set to 0.6, the rule C \rightarrow T corresponding to $\langle CT \rangle$ will be removed.

In practice, the length of user sequence is very long. IF the relative position of C and T is far away, even if the weighted support of $\langle CT \rangle$ is greater than the threshold of minimum weighted support, it is impossible to blindly determine the applicability of $C \rightarrow T$ rule. Therefore, a sliding window is designed so that the calculation and proof of the average weight of user sequence are carried out in the sliding window. The user sequence is shown in Table 3, where [] represents a sliding window.

The average weights of sequence in sliding window can be calculated, and the average weights of user sequence in sliding window can be obtained, as shown in Table 4. Suppose there is a rule TWA \rightarrow C, the frequent set of the rule is <TWAC>, and the user sequence of <TWAC> is <1,2>, which is expressed as

 $#1\{<1,2>,<$ TWAC>}. The weighted support (ws)= (0.45+0.51)/2.1=0.46; If we set the minimum weighted support to 0.4, the rule will be retained, and if we set the minimum weighted support to 0.5, the rule will be deleted.

UserS	Types
1	[A,C,T,W,A,W,C,T],W
2	[D,T, T,W,A,W,C,A],W,C
3	[W, T,T,W,A,W,A,T],W,A,C
4	[A,C,D,W,T,A,C,W]
5	[C,D,W]

Table 3: User Sequence and Sliding Window.

Users	Types	Average Weights of User Sequences in Sliding Window
1	[A,C,T,W,A,W,C,T]	0.45
2	[D,T, T,W,A,W,C,A]	0.51
3	[W, T, T,W,A,W,A,T]	0.56
4	[A,C,D,W,T,A,C,W]	0.38
5	[C,D,W]	0.2
sum		2.1

Table 4: Average Weights of User Sequences in Sliding Window.

4. Contrastive analysis of experiments

4.1. Analysis of mining association rules

Apriori algorithm is used to mine non-weighted association rules with temporal constraints in time series, and non-weighted association rules are used for recommendation experiments. The recommendation strategy is to match a segment of a sequence in a sequence of time series once the leader of the rule matches it, and then use the next neighbor node of the sequence as the prediction object to match the successor of the rule. If the result is the same, it is a match.

Since there may be many successors of association rules, it is unreasonable to select only one recommendation. For example, if there are rules (A B, C) and the prediction range of a sequence matched by a pilot is only (B), then the prediction method does not know how to proceed. The classical Apriori algorithm is used to mine temporal constraints association rules in time series, and then expand the prediction range. Through experiments in training set, it is found that the range with the highest prediction hit rate is within five nodes after expanding the prediction range.

For the method in this paper, before Mining Association rules, it is necessary to verify the value of effective period of frequency in training set. For this purpose, time series of different lengths are intercepted, and the error rate corresponding to each length is calculated . It is found that the error rate is lowest when the effective length is 8 nodes. The predicted node length has been proved to be five node lengths. It can be concluded that the frequency effective length (13) of this experiment is the sum of the frequency effective period (8) and the predicted node length (5). According to the new mining method, the weighted association rules are mined in the frequency effective length (sliding window). The rules are formed into intervals according to different weighted support degrees, and the recommended accuracy and coverage of each interval are calculated.

4.2. Contrast experiment

The comparison of accuracy of each method is shown in Figure 1, including the comparison of the highest accuracy of single interval and the comparison of average accuracy of all intervals. The comparison of coverage of each method is shown in Figure 2, including the comparison of the highest coverage of single interval and the comparison of average coverage of all intervals. It can be found that the mining method proposed in this paper has a greater improvement in recommendation accuracy and coverage than before.

Academic Journal of Computing & Information Science

ISSN 2616-5775 Vol. 5, Issue 1: 7-14, DOI: 10.25236/AJCIS.2022.050102





Figure 2: Comparisons of coverage.

"One", "Two", "Three" and "Four" in abscissa represent the traditional Apriori algorithm, Apriori algorithm mining sequence, expanding the prediction range and the method in this paper.

The accuracy of this method and other existing methods are shown in Figure 3. It can be found that under the new weighted association rule mining method, the accuracy of recommendation is as high as 69.0%. Compared with FP-tree based weighted association rule algorithm^[17], improved MINWAL algorithm^[18], evolutionary rule set based recommendation algorithm^[19], and association rule group recommendation algorithm^[20] based on similarity of behavior and score, the proposed algorithm has advantages in accuracy, which shows that the proposed algorithm is effective and feasible.



Figure 3: The accuracy of this method comparing with that of existing methods.

5. Conclusion

An improved algorithm of mining weighted temporal association rules is proposed in this paper. In order to reflect the importance of the independence of each item, weights are introduced to the item. However, due to the introduction of weights, the subset of frequent item sets is no longer frequent. So rules are first mined, and then rules with the characteristics of weighted frequent sets are proved and selected. Because time characteristic is the basic component of information reflecting the real world, most transaction information has time characteristic. The improved weighted temporal association rules mining algorithm has excellent performance in accuracy and coverage. This paper aims to mine hidden high-value rules and neglect the complexity of algorithm execution time. In the future, the algorithm will be optimized while ensuring the accuracy and taking into account the time complexity of algorithm execution.

Acknowledgements

This work was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province(KYCX20_1348).

References

[1] Jiang, Qingchao, Xu, Shuo, & Chen, Mengjie. (2018). Research on the recommendation algorithm of interdisciplinary fishery scientific data based on geographical location. Fishery modernization, 254(03), 63-67.

[2] Yu, H, F., Hsieh, C, J., & Si, S. (2014). Parallel matrix factorization for recommender systems. *Knowledge and Information Systems*, 41(3), 793-819.

[3] Grahne, G. (2005). Fast Algorithm for Frequent Itemset Mining Using FP-Trees. IEEE Transactions on Knowledge and Data Engineering, 17(10), 1347-1362.

[4] Marcoulides, G, A. (2005). Discovering Knowledge in Data: an Introduction to Data Mining. Publications of the American Statistical Association, 100(472), 1465-1465.

[5] Song, K., Lee, K. (2017). Predictability-based collective class association rule mining. Expert Systems with Applications, 79, 1-7.

[6] Goldberg, D., Nichols, D., & Oki, B, M. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61-70.

[7] Smith, B., Linden, G. (2017). Two Decades of Recommender Systems at Amazon.com. IEEE Internet Computing, 21(3), 12-18.

[8] Khan, M, S., Muyeba, M., & Coenen, F. (2008). A Weighted Utility Framework for Mining Association Rules. Proceedings of the 2nd UKSIM European Symposium on Computer Modeling and Simulation, 87-92.

[9] Forsati, R., Meybodi, M, R. (2010). Effective Page Recommendation Algorithms Based on Distributed Learning Automata and Weighted Association Rules. Expert Systems with Applications, 37(2), 1316-1330.

[10] Zhai, Y., Wang, L., & Wang, N. (2014). Efficient Weighted Association Rule Mining Using Lattice. Proceedings of the 26th Chinese Control and Decision Conference, 4913-4917.

[11] Forsati R, Meybodi M R. (2010). Effective page recommendation algorithms based on distributed learning automata and weighted association rules. Expert Systems with Applications, 37(2):1316-1330. [12] Yun U. (2008). A new framework for detecting weighted sequential patterns in large sequence databases. Knowledge-Based Systems, 21(2):110-122.

[13] Kuo-Cheng Yin, Kim D, Park C, Oh J, et al. (2017). Deep Hybrid Recommender Systems via Exploiting Document Context and Statistics of Items. Information Sciences, 417:72-87.

[14] Ouyang, Weimin. (2016). Mining Weighted Rare Association Rules Using Sliding Window over Data Streams. Proceedings of the 2016 International Conference on Computer Science and Electronic Technology, 116-119.

[15] Li, Chengjun, Yang, Tianqi. (2010). Improved Weighted Association Rules Mining Method. Computer Engineering, 36(7), 55-57.

[16] Ouyang, Weimin, Zheng, Cheng, & Cai, Qingsheng. (2001). Discovery of Weighted Association Rules in Databases. Journal of Software, 12(4), 612-619.

[17] Wang Taowei, Ren, Yibo. (2008). Study on Personalized Recommendation Based on Weighted Association Rule. Computer Applications and Software, 25(8), 242-244.

[18] Wang Bin, Ding, Xiangbin. (2008). A BUC-Based Mining Algorithm for Horizontal Weighted Association Rules. Computer Applications and Software, 25(12), 112-115.

[19] Long Shun, Cai, Tiao, & Lin, Jiaxiong. (2012). A Personalized Recommendation Model Based on Evolving Association Rule Mining. Journal of Jinan University: Natural Science & Medicine Edition, 33(3), 264-267.

[20] Zhang, Jiale, Liang, Jiye, & Pang, Jifang. (2014). Behavior and Score Similarity Based Algorithm for Association Rule Group Recommendation. Computer Science, 41(3), 36-40.