Research on Application of Improved Association Rules Mining Algorithm in Personalized Recommendation

Jie Gao\textsuperscript{1,*}

\textsuperscript{1}Jiangsu Vocational Institute of Commerce, Nanjing, Jiangsu, China, 211168

*Corresponding author e-mail:jie222234@jvic.edu.cn

Abstract. Traditional association rule mining does not consider the importance of each item, so the actual process lacks certain pertinence. Based on the New-Apriori algorithm and the Fp-growth algorithm idea, this paper proposes an improved association rule algorithm based on Fp-tree, Constructs the general process of personalized recommendation of association rules. And uses the Web log file to use the frequency of web pages being selected by users as the weight value, and realizes the algorithm of the personalized recommendation system. The experimental results show that the algorithm has high accuracy and effectiveness.

Keywords: Improved association rules, New-Apriori algorithm, Weighted support, Weighted frequent set, Personalized recommendation

1. Introduction

Association rule mining was originally to discover the connection between different commodities in the transaction database. Currently, association rules have not only been successfully applied in the business field, but also have good applications and development in other fields. It is an important part of the KDD (Knowledge Discovery in Database) research. Direction\textsuperscript{[1-2]}. Traditional association rules believe that the importance of each data item is equal, but in reality the importance of each item is not the same, and each item has its own weight, so it is called a weighted association rule. Weighted association rules can be divided into vertical weighted association rules and horizontally weighted association rules\textsuperscript{[3-4]}. Vertically weighted association rules also give different weights to the same item in time sequence, and horizontally weighted association rules are based on the importance of the project to decision-making. The degree assigns different weights\textsuperscript{[5-6]}. In addition, data mining technology can be used to analyze and dig the data set composed of website Daken's usage data (user visits) and other related data, so as to obtain valuable knowledge of the patterns of website visits. Web log mining is mainly The goal is to extract patterns of confusion and interest from the access records of the Web. Because every WWW server keeps access logs, it records information about the access and interaction of thousands of users. Therefore, these data are analyzed. It can help understand user behavior, improve site structure, and provide users with personalized services.

In this paper, based on the weighted support of the New-Apriori algorithm, combined with the idea of the Fp-growth algorithm, a weighted association rule algorithm based on the Fp-tree is proposed, and
the general process of personalized recommendation of association rules is given. Using Web log files
The frequency with which the web page is selected by the user is used as the weight value, which
implements the algorithm of the personalized recommendation system.

2. Improve association rule mining
This article gives the following conceptual explanation on this basis.

Let \( I = \{i_1, i_2, \cdots, i_m\} \) be the full set of items, and each item has a weight corresponding to it. Their
weights are \( \{h_{i_1}, h_{i_2}, \cdots, h_{i_m}\} \) \((h_i \in [0,1])\), the minimum weighted support is \( w_{\text{minsup}} \), and the minimum
credibility is \( w_{\text{minconf}} \).

Definition 1: Set item sets \( \{X_1, X_2, \cdots, X_k\} \) and \( X \subseteq I \), then the weighted support \( w_{\text{sup}} \) of item
set \( X \) is shown in formula (1):

\[
 w_{\text{sup}}(X) = \max \{h_{i_1}, h_{i_2}, \cdots, h_{i_k}\} \times \text{sup}(X) \quad (1)
\]

Among them, \( \text{sup}(X) \) is the traditional support count of \( X \), and \( \max \{h_{i_1}, h_{i_2}, \cdots, h_{i_k}\} \) is called the
weight of \( X \). If \( w_{\text{sup}}(X) \geq w_{\text{minsup}} \), \( X \) is frequently weighted.

Definition 2: For item sets \( X, Y, X \subseteq I \), \( X \cap Y = \emptyset \), if there is \( w_{\text{sup}}(X \cup Y) \geq w_{\text{minsup}} \), and \( w_{\text{conf}}(X \Rightarrow Y) \geq w_{\text{minconf}} \), then \( X \Rightarrow Y \) is a rule. The weighted credibility is defined as:

\[
 w_{\text{conf}}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}
\]

Property 1: For item set \( X \), \( w_{\text{sup}}(X) \leq \text{sup}(X) \).

Proof: Because of \( 0 \leq h_i \leq 1 \), there is \( w_{\text{sup}}(X) = \max \{h_{i_1}, h_{i_2}, \cdots, h_{i_k}\} \times \text{sup}(X) \leq \text{sup}(X) \).

Proposition 1: If the item set \( X \) is frequently weighted, then \( \forall Y \subseteq X \) has \( \text{sup}(Y) \geq w_{\text{minsup}} \).

Proof: Because \( X \) is frequently weighted, from property 1, \( \text{sup}(X) \geq w_{\text{minsup}} \). And \( \forall Y \subseteq X \), \( \text{sup}(Y) \geq \text{sup}(X) \). So \( \text{sup}(Y) \geq w_{\text{minsup}} \).

Corollary 1: If the item set \( Y \) has \( \text{sup}(Y) < w_{\text{minsup}} \), then any superset of \( Y \) cannot be weighted
frequent.

Property 2: In horizontally weighted association rule mining, the two basic properties in the Apriori
algorithm are not valid.

3. Improved association rule mining algorithm
After data cleaning, user identification, and session identification, web logs can divide user access
records into individual sessions. The data form consists of multiple pages, such as session
\( S_j = ([i_1, \text{time}_{i_1}],[i_2, \text{time}_{i_2}],[i_n, \text{time}_{i_n}]) \), where \( i_n \) represents the nth page, and \( \text{time}_{i_n} \) is the nth page.
Time spent on the page.

Given a page set \( I = \{i_1, i_2, \cdots, i_n\} \), a weight \( w(i_j) \) is assigned to each page \( i_j \), which is used to
indicate that the importance of page access is different. The weight calculation formula is as follows:

\[
 w(i_j) = ATime(i_j) / \sum ATime(i_j) \quad (2)
\]

Among them: \( ATime(i_j) \) is the average access time of page \( i_j \); \( \sum ATime(i_j) \) is the sum of
the average access time of all pages; \( 0 \leq w(i_j) \leq 1 \), \( j = \{1,2,\cdots,n\} \). When the data item has a weight,
the item set to which the data item belongs also has a corresponding weight. The weight of the item set
\( X, Y \subseteq I, X \) \( W(X) \) is defined as

\[
 W(X) = \text{MAX}(w(x)) \quad (3)
\]

Obviously \( 0 \leq W(X) \leq 1 \). The weighted support of item set \( X \) \( W \text{ Sup}(X) \) is defined as
Sup \( (X) = W(X) \times \text{Sup}(X) \) \tag{4}

Weighted Confidence (W Confidence): Under the premise that the item set \( X \) appears, the probability of the item set \( Y \) appears:

\[
W \text{ Confidence}(X \Rightarrow Y) = W \text{ Sup}(X \cup Y)/W \text{ Sup}(X)
\] \tag{5}

Here \( \text{Sup}(X) \) is the support number of the support degree item in the Boolean association rule, which is recorded as \( \text{sup\_count} \), the minimum support number input by the user is \( \text{minsup\_count} \), and the minimum weighted support threshold value is expressed as \( \text{wminsupt} \). If \( W \text{ Sup}(X) \geq \text{wminsupt} \), then \( X \) is called the weighted frequent itemset. When the weighted frequent item \( X \) contains \( K \) items, it is called the weighted frequent \( K \)-item set, and \( Lk \) is the set of all weighted frequent \( K \)-items. All weighted frequent items The set of itemsets is denoted as \( L \).

Select a transaction database. There are 10 transactions in the database. Table 1 is the weight table of the corresponding items. Take \( \text{wminsupt}=0.3 \) and \( \text{wminconf}=0.4 \).

| Item | \( I_1 \) | \( I_2 \) | \( I_3 \) | \( I_4 \) | \( I_5 \) |
|------|------|------|------|------|------|
| Weights | 0.3 | 0.4 | 0.7 | 0.6 | 1 |

According to definition 1, \( \text{wsup}(I_1)=0.3 \times 0.6=0.18 \), \( \text{wsup}(I_2)=0.32 \), \( \text{wsup}(I_3)=0.42 \), \( \text{wsup}(I_4)=0.18 \), \( \text{wsup}(I_5)=0.3 \). Because of \( \text{wsup}(I_1) \) and \( \text{wsup}(I_4) \) is smaller than \( \text{wminsupt} \), so the Fp-tree created.

The steps for mining weighted frequent sets are as follows: First consider \( I_5 \). Since \( \text{wsup}((I_3,I_5))=0.1 \), the constrained Fp subtree of \( I_5 \) does not include \( I_3 \). Constrained Fp-subtree. Mining this subtree can obtain the weighted frequent set \( \{I_3,I_5\}, \{I_2,I_5\} \), the weighted support is 0.3. For \( I_3 \), since \( \text{wsup}((I_2,I_3))=0.28 \), the constraint Fp-subtree of \( I_3 \) is an empty tree. So the weighted frequent set \( \{I_3\} \), and the weighted support is 0.42. Since \( I_2 \) is the last item in the header table, the weighted frequent set \( \{I_2\} \) can be obtained, and the weighted support is 0.32.

The generated weighted frequent set \( \{\{I_3\},\{I_2,I_3\},\{I_3\},\{I_2\}\} \), according to the calculation of definition 2 reliability, the generated rules are \( \{\{w\text{conf}(I_3 \Rightarrow I_5)=1\}, \{w\text{conf}(I_2 \Rightarrow I_5)=0.375\}\} \), since \( \text{wminconf}=0.4 \), the generated rule is \( \{I_5 \Rightarrow I_2\} \).

4. Personalized recommendation of association rules

With the rapid development of the Internet, in order to attract more Web users, various sites have launched rich and personalized services, so personalized technology has become a research hotspot. Association rules, as one of the technical methods of Web mining, are The personalized recommendation system plays an important role in obtaining user access patterns. A personalized recommendation system based on association rules generally includes two parts: an offline part and an online part. The offline part mainly uses various association rule mining algorithms to establish an association rule recommendation model. Because the algorithm execution is relatively time-consuming, it can be performed in an offline cycle. The online part provides users with real-time recommendation services based on the established association rule recommendation model and the user's browsing or purchasing behavior. At present, most of the association rule recommendations are still based on support. Reliability mode, the algorithm steps are as follows:

1. Create transaction records for each user based on the historical data of all items visited by each user in the transaction database to construct a transaction database.
2. Use various association rule mining algorithms to conduct association rule mining on the constructed transaction database, and obtain all association rules that meet the minimum support threshold \( \text{minsup} \) and the minimum credibility threshold \( \text{minconf} \), which are recorded as the association rule set \( R \).
3. For each current user \( u \), set a candidate recommendation set \( P_u \), and initialize the candidate recommendation set \( P_u \) to empty.
4. For each current user \( u \), search the association rule set \( R \) to find all the association rule sets \( R_1 \)

\[
W \text{ Sup}(X) = W(X) \times \text{Sup}(X) \tag{4}
\]

Weighted Confidence (W Confidence): Under the premise that the item set \( X \) appears, the probability of the item set \( Y \) appears:

\[
W \text{ Confidence}(X \Rightarrow Y) = W \text{ Sup}(X \cup Y)/W \text{ Sup}(X) \tag{5}
\]
supported by the user, that is, all the items on the left of the association rule appear in the current access data and historical access records of user u.

(5) Add all the items on the right of the association rule set R1 to the candidate recommendation set Pu.

(6) Delete items that the user has visited from the candidate recommendation set Pu.

(7) Sort all candidates in the candidate recommendation set Pu according to the credibility of the association rule set R1. If an item appears in multiple association rules, the association rule with the highest credibility is selected as the sorting criterion.

(8) Select the top N items with the highest credibility from the candidate recommendation set Pu as the recommendation result and return it to the current user u.

This recommendation system establishes the association between items based on the selection habits of all users, without considering the importance of the user's selection of items, and lacks a certain degree of pertinence in the actual process, making the recommendation accuracy rate not high. Consider in step (2) The importance of each item is generated based on the weighted association rule set, so that the final recommendation result is more accurate.

After data preprocessing of the log file, unnecessary rows and columns are deleted, and then classified by users. A user's transaction database is formed, Userid represents the user, and Itemset represents all items browsed by the user. Then all items are calculated according to the transaction database The weight of the project forms the project weight table. Suppose the number of users who selected the item $I_j$ is $C_j$, then $W_j = C_j / m$, $m$ is the number of users.

In the experiment, 40 users were selected, and each user visited at least 15 pages. 30 users were used as the training set and 10 user data were used as the test set for testing. The traditional association rule recommendation algorithm (mining algorithm Implemented with Fp-growth), New-Apriori algorithm, and improved association rule algorithm based on Fp-tree. At the same time, the execution time of improved association rule based on Fp-tree and New-Apriori in generating weighted frequent itemsets was carried out. Comparison: Compared with traditional association rule recommendation algorithms in terms of accuracy, accuracy refers to the ratio of the number of accurate predictions to the total number of predictions. Because the Fp-tree method only needs to scan the DB in the process of generating weighted frequent sets Two times, and the created Fp-tree has greatly compressed the transaction database, which can reduce the time required to calculate the support for the candidate set. New-Apriori needs to scan the database multiple times and generate a large number of candidate item sets. There is a large difference in execution time, and the comparison result is shown in Figure 1. Because the project weight is considered, the accuracy of the recommendation is also improved, as shown in Figure 2.

![Figure 1. Comparison of execution time for generating weighted frequent itemsets](image-url)
5. Conclusions
This paper proposes an improved association rule mining algorithm based on Fp-tree, and implements the algorithm according to the personalized recommendation step of association rules. The association rule mining and analysis of the pre-processed data of Web log records are performed to generate weighted frequent itemsets. And then generate association rules on this basis, which has certain reference value for website managers to understand user browsing behavior, improve site structure, and provide customers with personalized and characteristic services.

Acknowledgments
The school level project of Jiangsu Vocational Institute of Commerce “Research on Accurate Marketing classification of e-commerce based on classification algorithm” (JSJM19023).

References
[1] Tayou J. Identification of subsets of actionable genetic alterations in KRAS-mutant lung cancers using association rule mining[J]. Cellular Oncology, 2018, 41(4):395-408.
[2] Sizheng Z , Daniel T , Natasha M , et al. The prevalence of depression in axial spondyloarthritis and its association with disease activity: a systematic review and meta-analysis[J]. Arthritis Research & Therapy, 2018, 20(1):140-148.
[3] Altaf W , Shabbaz M , Guergachi A . Applications of association rule mining in health informatics: a survey[J]. Artificial Intelligence Review, 2017, 47(3):313-340.
[4] Lovell-Smith J W , Feistel R , Harvey A H , et al. Metrological challenges for measurements of key climatological observables. Part 4: atmospheric relative humidity[J]. Metrologia, 2016, 53(1):40-49.
[5] Miura, Masatomo. Therapeutic drug monitoring of imatinib, nilotinib, and dasatinib for patients with chronic myeloid leukemia.[J]. Biological & Pharmaceutical Bulletin, 2015, 38(5):645-654.
[6] Calèes, P, Boursier J , Lebigot J , et al. Liver fibrosis diagnosis by blood test and elastography in chronic hepatitis C: agreement or combination?[J]. Alimentary Pharmacology & Therapeutics, 2017, 45(7):991-1003.