An Improved Apriori Algorithm for Association Analysis

Xi ZHU, Qing YANG* and Fei-yang DU

Department of Computer Science and Technology Central China Normal University
*Corresponding author

Keywords: Association analysis, NewApriori, Weka.

Abstract. In recent years, association analysis has been gradually used in school teaching. In this paper, we proposed an improved NewApriori algorithm based on the Apriori algorithm. The New Apriori algorithm has combined the transaction data compression with the pruning of candidate item sets by using open source data mining tools weka. Three different transaction data of students scores are separately used on the experiment of NewApriori algorithm, the experiment result shows that the NewApriori algorithm is better than the original Apriori algorithm in time efficiency.

Introduction

The association analysis has been gradually used in school teaching in recent years. Association analysis refers to the process of finding strong association rules from large data sets. Some common algorithms are widely used in association analysis, such as Apriori [1] algorithm and FP growth[5] algorithm. Rakesh Agrawal advocated Apriori algorithm for the first time in 1994 in Fast Algorithms for Mining Association Rules [1]. In recent years, various kinds of improvement methods of Apriori algorithm have been brought up. These methods are mainly focused on the following categories:

1) The compression of the transaction database;
2) Reducing the scanning times of the transaction database;
3) Combining the correlation analysis algorithm with other algorithms, such as genetic algorithm [6], clustering algorithm [8] and so on;
4) Transforming the transaction database into other kinds of data structure, such as a two-dimensional matrix [11], linkedlist [13] and so on;
5) Apply MapReduce to Apriori algorithm [7].

In this paper, The NewApriori algorithm is designed based on the improvement of the classic Apriori algorithm, it has compressed the transaction data and prunned the unnecessary itemsets before the generation of candidate itemsets. The experiments of NewApriori algorithm and the Apriori algorithm are separately carried on based on three different transaction data of students scores.

Apriori Algorithm

Apriori is a classic algorithm to mine association rules. It aims at mining the frequent item sets and association rules by traversing transaction data with a breadth-first search strategy.

However, the apriori algorithm has two main limitations: First, the increase of the time spent on I/O will slow down the speed of the Apriori algorithm. Second, large number of redundant rules will be generated if the transaction data is big or the min sup is low, which will reduce the time efficiency of the apriori algorithm.

The Improved NewApriori Algorithm

Here is the basic idea of the NewApriori algorithm based on the improvement of apriori algorithm:

1) The program scans transaction data D and generate the candidate 1-itemset \( \mathcal{C}_1 \), then choose the frequent 1-itemset \( \mathcal{L}_1 \) from \( \mathcal{C}_1 \) that has reached the value of min sup.
2) The candidate 2-itemset is generated by the self-joins of \( L_1 \). Similarly, the candidate item sets of length \( k \) \( C_k \) is generated by the self-joins of \( L_{k-1} \) through iterative method. Delete the item-set with the minimum support of the itemsets in \( L_{k-1} \) before \( C_k \) is generated by the self-join of \( L_{k-1} \), and then generate \( C_k \) by the self-join of the rest itemsets in \( L_{k-1} \), in this way some redundancy candidates itemsets of low support value can be deleted.

3) The process of the self-joins of \( L_k \) : During the process of generating \( C_k \) by \( L_{k-1} \), ensure that the the same former \( k \) items was found between two separate item-set from \( L_k \), and then combine the different latter items.

4) After the \( C_k \) is generated, update the support value of every item-set in \( C_k \) by scanning the transaction data \( D \). The process of data sampling is to scan the records in transaction data and select the records randomly from transaction data according to a certain proportion. In this way, the time of scanning the sample transaction data was reduced. For each instance in \( D \), if \( C_k \) contains the subset of the instance, plus one to the old counting value of candidates in \( C_k \).

5) The infrequent candidate itemsets need to be pruned after being generated. For a frequent itemset, all its subsets are also frequent and thus for an infrequent itemset, all its super-sets must also be infrequent [9]. In this way, the infrequent itemsets can be pruned.

6) Mining association rules: Output the rules, which has reached the userspecified minimum confidence value.

The pseudo code for the NewApriori algorithm is given below:

**Algorithm 1: NewApriori:**

```plaintext
k=1;
for all instances \( \in DB \) do
NewDB=get_instances_randomly(DB);//random sampling of transaction data.
findLargeItemSets(NewDB);//frequent 1-itemset \( L_1 \).
end for
while( \( L_{k-1} \neq \emptyset \) ) do
\( C_k = \text{NewApriori\_gen}( L_{k-1} ) \);//generate candidate item sets
k++
end while
for all itemset \( \in C_k \) do
if(\( \in C_k \) ) then
\( \text{c.count}++; //adding to the support count
end if
\( L_k = \{ \ c \in C_k | \text{c.count}>N*\text{min\_sup} \};
end for
```

**Algorithm 2: NewApriori\_gen:**

```plaintext
for all itemset \( \in L_{k-1} \) do
\( \text{add c to } C_k; \)
\( \text{c = get\_item\_minsup}( C_k ); //get the item-set with minimum support in } C_k.
\end{for
\( \text{delete } c \text{ from } C_k; \)
end for
for all itemset \( c_1 \in C_k, c_2 \in C_k \) do
if( \( |c_1[1] = c_2[1]\) \( \cap (c_1[2] = c_2[2]) \cap \ldots (c_1[m] = c_2[m]) \rangle \),m<\( k \)
\( \text{c= c_1*c_2; //the self-joins of } c_1, c_2.
\end{for
\( \text{add c to } C_k; \)
return \( C_k; \)
end if
end for
```
Data Processing

In this paper, the experimental data is based on the students computer scores of 2014 grade undergraduates in Central China normal university (CCNU). The elective computer courses are as follows: Database Technology and Application, Multimedia Technology and Application, Advanced algorithm language.

Data Processing of Students Scores

The method of processing the student's raw scores data are as follows: A student's raw score is the original test scores of this student. The raw score can't objectively reflect the students' real level because the raw score value is not equivalent. That means the score evaluation is not limited to the students raw score of the course. Therefore, the raw scores are transformed to 5 kinds of levels. These levels are as follows:

E(0<score<60), D(60<score<70), C(70<score<80), B(80<score<90), A(90<score<100). A represents highest scores level and E represents the lowest scores level. The students' scores are transformed to corresponding knowledge level and capability level of the course, the level of the transformed score can be A, B, C, D, E.

Here are the three processed transaction data: data of Database Technology and Application (Database data for short), data of Multimedia Technology and Application (Multimedia data for short), data of Advanced algorithm language (Advanced data for short). The attributes number of the three different transactions is 6. The number of three different transactions is separately 339 for Database data, 181 for advanced data, 578 for Multimedia data.

Course Evaluation Standards

According to The computer teaching syllabus for college computer teaching, the evaluation criteria of the mastery degree of students' current skill levels in computer course is divided into two sections by course content, the knowledge section and the capability section.

Experiment and Results

In this paper, three experiments is designed to compare the performance of the two algorithms based on an open source data mining tools called weka[3,10]. We designed the NewApriori algorithm and added it to the weka. The experiments compare the two algorithms on consuming time trends and the consuming time reducing rate based on three transaction data, they are the Database data, the Multimedia data and the Advanced data. The Figure 1, Figure 2 and Figure 3 below are illustrated according to the minimum support value (min sup) as abscissa and the running time of the two algorithms as coordinate. The minimum support value ranges from 0.1 to 0.4.

![Figure 1. Database Technology and Application Data.](image1)

![Figure 2. Multimedia Technology and Application Data.](image2)

![Figure 3. Advanced algorithm language Data.](image3)

According to the three experimental data above, we analyse the consuming time reducing rate of the NewApriori algorithm based on different transaction data. As is shown in Table 2 below, the second and third columns show the average consuming time in milliseconds of the two algorithms, and the last column shows the time reducing rate between the two algorithms.
Table 2 shows that the NewApriori algorithm reduce the time consuming most by 51.37% from the apriori algorithm in the Database data which has the least number of transactions and by 34.49% in the Advanced data.

Table 1. Database technology and application data.

<table>
<thead>
<tr>
<th>ID</th>
<th>institute</th>
<th>sex</th>
<th>Database Knowledge level</th>
<th>Database ability level</th>
<th>Basic computer skills level</th>
<th>Basic computer knowledge level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>education</td>
<td>male</td>
<td>E</td>
<td>A</td>
<td>C</td>
<td>B</td>
</tr>
<tr>
<td>2</td>
<td>education</td>
<td>female</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>education</td>
<td>female</td>
<td>D</td>
<td>E</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>education</td>
<td>female</td>
<td>C</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
</tbody>
</table>

Table 2. The Consuming time reducing rate of newapriori Algorithm and Apriori Algorithm according to different transaction data.

<table>
<thead>
<tr>
<th>Transaction data</th>
<th>apriori algorithm</th>
<th>newapriori algorithm</th>
<th>Consuming time reducing rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database data</td>
<td>56</td>
<td>29.25</td>
<td>47.77%</td>
</tr>
<tr>
<td>Advanced data</td>
<td>47.25</td>
<td>31</td>
<td>34.49%</td>
</tr>
<tr>
<td>Multimedia data</td>
<td>63.75</td>
<td>31.75</td>
<td>51.37%</td>
</tr>
</tbody>
</table>

Summary

In this paper, we proposed an improved NewApriori algorithm based on the Apriori algorithm. We applied NewApriori algorithm and Apriori algorithm to the open source data mining tools weka based on three different transaction data of college students scores. The experiment result shows that the proposed NewApriori algorithm is superior to the original Apriori algorithm with a better time efficiency.

However, the NewApriori algorithm has been less focused on teaching promotion. Futhermore, we can apply our further research to several aspects, such as the recommendation of students course selection, the personalized course selection and personalized learning recommendation on our further studies.

Acknowledgement

We thank the Central China Normal University for providing us with the research facilities and academic environments. The research in this paper is supported by the project of the 2016thSpecific funding for education science research by self-determined research funds of CCNU from the Colleges' basic research and operation of MOE NO.ccnu16JYKX036 and the project of teaching and research funds of CCNU , the project number is 201649.

References


