Improvement of Apriori algorithm based on matrix compression

Jigang Zheng[1, a*] and Jingmei Zhang[2, b]

1Department of Mathmatic, Baoshan College,Baoshan,Yunnan,678000,China.

2Library of Baoshan College, Baoshan, Yunnan, 678000, China.

a6913641@qq.com,b279619568@qq.com

Keywords: Data mining; Time; Improvement; Relational matrix; Association rule

Abstract. The traditional association rule mining algorithm for Apriori time cost, the lack of Apriori algorithm, based on the theory of relational algebra, relation matrix and related operations by given search association rules from the frequent itemsets mining algorithm based on relation algebra theory. Using the relation matrix to scan the database only once, in order to reduce the running time of the algorithm, frequent itemsets mining, finally the simulation results comparing the two execution time of the algorithm, the effect of sample data and the minimum support degree on the performance of the algorithm is discussed. The simulation results show that the improved algorithm is efficient and reduces the running time of mining frequent itemsets.

Introduction

Apriori algorithm is a classical algorithm for discovering association rules in data mining. In 1993, Agrawal, Imielinski and Swami put forward the concept of association rule mining. In 1994, Agrawal and Srikant proposed Apriori algorithm [1],used to find interesting association rules or relationships among data items in a given data set.

Table 1 is an example of a supermarket shopping basket [2],each row in the table corresponds to a transaction that contains a unique identifier and a set of goods purchased by the customer.

Table 1 Shopping basket

TID	Commodity collection
1	{Bread, Milk}
2	{Bread, Diaper, Beer, Egg}
3	{Milk, Diaper, Beer, Cola, Salt}
4	{Bread, Diaper, Beer}
5	{Bread, Milk, Diaper, Salt}

Using Apriori algorithm, the following rules:

$$\{\text{Diaper}\}\rightarrow\{\text{Beer}\}$$

$$[\text{support}=2\%,\ \text{confidence}=40\%]$$

The rule's support rating support=2% shows that 2% of consumers buy diapers and beer at the same time, and the reliability of confidence=40% means that more than 40% of customers who buy diapers also buy beer. Through the discovery of association rules, it is helpful for the decision maker to design the catalogue, find out the new cross marketing opportunities or make other relevant business decisions.

Basic Concept

Association Rules

Association rule is "if...... Then......" In order to get useful rules, we need two important information related to the rules: support - the probability that the rules appear, and the probability that the rules are correct. The degree of support is a measure of the importance of association rules, which shows how much of this association rule is representative in all transactions. Credibility is a measure of the accuracy of association rules, although some of the association rules are highly reliable, but the support is very low, indicating that the association rules are very small, so it is not important.

Definition1 Set $I = \{I_1, I_2, \cdots, I_m\}$ is a collection of data items, D transaction is a collection of all[5], A transaction T has a unique identifier TID . If items, transaction support items claimed T set A , also known as T transaction that contains the item set A .

Definition2 Association rules are shaped like $A \Rightarrow B$ type of implication, among them $A \subset I$,$B \subset I$, and $A \cap B = \Phi$.

$A \Rightarrow B$ supports the association rule is defined as: $\sup port(A \Rightarrow B) = \dfrac{\sup port(A \cup B)}{N} \times 100\%$,credibility is defined as: $confidence(A \Rightarrow B) = \dfrac{\sup port(A \cup B)}{\sup port(A)} \times 100\%$.

Definition 3 Support and confidence required to be greater than the threshold set by the user (ie, minimum support threshold and minimum confidence threshold), that:

$\sup port(A \Rightarrow B) \ge \min\_\sup$, $confidence(A \Rightarrow B) \ge \min\_conf$.

Apriori Algorithm

```
In: Database D and $\min\_\sup$ ,[3]
Out: Database D itemsets $L$ ,
Algorithm:
$L_1$= Looking frequent two sets($D$);
For $k$=2;$L_{k-1} \neq \Phi$; $k$++
  { $C_k$=apriori_gen($L_{k-1}$);
      For each transaction $t \in D$
        { $C_t = subset(C_k, t)$;
            For each candidate $c \in C_t$
              $c.count++$;}
            $L_k$=($c \in C_k \mid c.count \ge \min\_\sup$)}
        Return $L$={All $L_k$}.
```

The apriori_ gen is a key step in Apriori algorithm, according to the L_{k-1} for L_k ,need to do two things: pruning and connection. The connection step: to produce C_k, by connecting the pruning step: if a candidate k set ($k-1$) a subset of frequent itemsets in ($k-1$),then the candidate set is not frequent, so as to remove from C_k [4].

Apriori _ gen is described as follows:

```
Apriori _ gen($L_{k-1}$ :frequent(k-1)-itemsets)
    For each itemsets $l_1 \in L_{k-1}$
        For each itemsets $l_2 \in L_{k-1}$
If $(l_1[1]=l_2[1]) \wedge \ldots \wedge (l_1[k-2]=l_2[k-2]) \wedge (l_1[k-1]<l_2[k-1])$
Then
{$c = l_1\ l_2$;
            If has_infrequent_subset$(c, L_{k-1})$
                Then delete;
            Else add $c$ to $C_k$;
```

}

Return C_k.

The Bottleneck of Apriori Algorithm

In search of 1 sets,2 sets......,k sets, each mining a layer of L_k,to scan the transaction database D again[4],k second scan, get k sets, due to D in a short period of time with little change or no change, do is repeat scanning. When the transaction database D is large, the overhead of Apriori algorithm is relatively large, which is to reduce the I/O overhead. The Apriori algorithm is improved in this paper.

Improvement of Apriori algorithm

Ideas

In view of the deficiency of Apriori algorithm, based on relational algebra theory, the relationship matrix and correlation operation are obtained by Optimization Relation Association Rule, this algorithm only needs to scan the database once, and overcomes the shortcoming of the Apriori algorithm which needs to scan the database for many times.

Definition4 $T=\{t_1,t_2,\cdots t_m\}$ is transaction data collection,$I=\{i_1,i_2,\cdots i_n\}$ is a collection of data items, the relational matrix R is defined as

$$R=[r_{kj}]=\begin{cases}1 & (\text{Item}\,i_j\text{ appears in the item }t_k)\\ 0 & (\text{Item}\,i_j\text{ dit not appears in the item }t_k),\end{cases}$$

Among them $k=1,2,\cdots,m$, $j=1,2,\cdots,n$, R is the T to I matrix,

That is:

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n}\\ r_{21} & r_{22} & \cdots & r_{2n}\\ \vdots & \vdots & \cdots & \vdots\\ \vdots & \vdots & \cdots & \vdots\\ r_{m1} & r_{m2} & \cdots & r_{mn}\end{bmatrix},$$

The value of r_{ij} is 0 or 1,indicates that the i transaction data contains or does not contain j data item.

Data item $I=\{Bread, Milk, Diaper, Beer, Egg, Cola, Salt\}$,table 1 of the shopping basket transaction is converted to the corresponding relationship matrix R

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 1 & 1 & 1 & 0 & 0\\ 0 & 1 & 1 & 1 & 0 & 1 & 1\\ 1 & 0 & 1 & 1 & 0 & 0 & 0\\ 1 & 1 & 1 & 0 & 0 & 0 & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\end{bmatrix}.$$

Definition5 According to relational matrix R,the support for the j data item as the 1 set is $\frac{\sum_{i=1}^{m} r_{ij}}{m}\times 100\%$.

Algorithm Description

In: Database D and $\min\_\sup$,

Out: Database D itemsets L,

Algorithm:

```
$\{a_1\}$=find_frequent_1-itemsets( $R$ );
For $i$=2;$i<n$;$i$++
 {    For Each frequent $i-1$ itemsets $c_{i-1}=\{c_{i-11},c_{i-12},\cdots,c_{i-1n}\}$
            For $h$=1;$h$<=$n$;$h$++
                For $j$=1;$h$<=$m$;$j$++
                    Calculation $r_{jc_{ii-1}}$ and $r_{jh}$;
                    If $\{c_{ji-1},h\}$ support degree>=min_sup
                        Then $\{c_{i1},c_{i2},\cdots,c_{ii-1},h\}$ is a $i$ set
        For each $i$ item set $c_j=\{c_{j1},c_{j2},\cdots,c_{ji}\}$
            For $k$=1;$k$<=$m$;$k$++
                If $i$ item set $c_j$ support degree>=min_sup
                    Then out $i$ item set $c_j$
}.
```

Simulation Experiment

Experimental environment: Intel Core I3 CPU 3.1GHz,memory 3G,hard disk 320G;Microsoft Windows 7 system; algorithm with Java7.0, SQL Server 2008 implementation. Experimental data sets from “KDDCUP.data_10_percent” data sets of KDDCUP99 subsets, Under the minimum support of 20%, 40% and 50%, respectively, under different data samples, the experimental data are shown in table 2.

Table 2 The running time of the algorithm under different support and sample size(Company: s)

Algorithm	Apriori			Improvement of Apriori algorithm		
Data quantity Support degree	s=20%	s=40%	s=50%	s=20%	s=40%	s=50%
12×5	0.496	0.511	0.603	0.078	0.063	0.047
100×6	73.797	69.249	68.171	0.109	0.062	0.015
600×8	71.719	71.515	74.812	0.125	0.094	0.046
1000×8	73.187	72.043	78.469	0.344	0.109	0.016
1000×16	80.438	84.015	91.562	7.312	0.063	0.011

It can be seen that the number of samples gradually increased, the superiority of the improved algorithm began to reflect.

Summary

This paper based Apriori algorithm for mining association rules in the analysis on the present association rule mining algorithm based on relation algebra theory, in order to reduce the running time of the algorithm, frequent itemsets mining, finally the simulation results comparing the two execution time of the algorithm, the effect of sample data and the minimum support degree on the performance of the algorithm is discussed. However, in practical applications, uncertain data mining and uncertain data mining[5], Association Rules Mining for spatial data[6]and network intrusion detection[7],It's very meaningful, and that's what we're going to do next.

Reference:

[1] Agrawal R, Srikant R. Fast Algorithm for Mining Association Rules. In Proceeding 1994 International conference Very Large Data Base(VLDB’94). Santiago, Chile, Sept, 1994: 487-499.

[2] Pang-Ning Tan, Michael Steinbach. Introduction to data mining [M].Beijing: People's Posts and Telecommunications Press, 2006:201-204.

[3] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques [M].Beijing: Machinery

Industry Press, 2007: 151-154.

[4] LI Xue-bin, ZHU Yan-qin, LUO Xi-zhao. Research and Improvement for Apriori Algorithm of Association Rule Mining [J].Computer Knowledge and Technology, 2009, 5(19):5084-5085.

[5] Lu Ye, Wang Lizhen, Zhang Xiaofeng. Mining Frequent Co-location Patterns from Uncertain Data [J].Journal of Frontiers of Computer Science and Technology, 2009, 3(6):656-664.

[6] Li Deren, Wang ShuLiang, Shi Wenzhong, Wang Xinzhou. On Spatial Data Mining and Knowledge Discovery [J]. Geomatics and Information Science of Wuhan University, 2001, 26(6):491-499.

[7] Chen Hongquan, Huo Zhikai. Association Rules Based Network Intrusion Detection Method [J]. Journal of University of Electronic Science and Technology of China, 2009, 38(S):94-96.