读书报告

09118119 黄一凡

# 自己提出的问题

#### 1.Fig.2.8的第五行中 是否为了避免重复讨论， 以及它们的大小是按什么标准比较的？（可能是字典序？）

经过讨论后，该问题并没有得出确定的结论。进一步思考后，我认为这里的顺序是按事先在itemset中依照MIS值升序排列。 如果按照字典序进行，那么就不能保证生成的candidate中item的顺序是完全依照MIS升序排列，对后面的算法实现造成困难。 这里的作用只要是为了确定生成后的顺序以及避免对于同样的和重复讨论，改善算法效率。

#### 2.为什么只需记录，就可以得到所有？

Candidata set的生成总体上看是一个递归的过程，在算法实现的过程中从逐渐生成到。 举例来说，我们不妨设任一，为的任意子集，那么是在生成时计算并记录的。 因此，只需记录，便可以得到所有

# 别人提出的问题

#### 1.在MS-Apriori算法中，为什么第一步要生成和两个集合

因为在MS-Apriori算法中，我们对每一个item引入了MIS值，而frequent itemsets需要满足的minsup是由其item的最小 MIS值决定的。因此，对于单元素的itemsets来说，可能它本身并不满足frequent的条件，但是如果再加入MIS值较小的item后， 很有可能又会满足frequent的条件，所以需要多生成一个集合，来储存这种单元素itemsets。

#### 2.在MS-Apriori算法中，时的condidate-generation算法为什么和其他不同，是否可以合并？

不可以，最本质的区别是两者引入的参数不同。时的condidate-generation算法需要， 而时的condidate-generation算法需要，具体原因可见上一个问题。 其次，在时，算法中另外引入了prunning step，而这一步是时不需要的。

# 读书计划

#### 本周所读：

2.3-2.5

#### 下周计划：

3.1-3.3

# 读书摘要

下面是我读书时做的一些笔记整理：

## 2.3 Data Formats for Association Rule Mining

Association rule mining can also be performed on relational tables, which needs table data sets converted to transcation data sets.

* If attribute takes categorical values: simply change each value to an attribute-value pair
* If attribute takes numerical values: We need to first discretize its value range into intervals, and treat each interval as a categorical value

We can also convert a transaction data set to a table data set.

## 2.4 Mining with Multiple Minimum Supports

In real life applications, some items appear very frequent while some others rarely appear. If we use only a single minsup, then will encounter two problems:

Problem 1: If the minsup is set too high, we will not find rules that involve infrequent items or rare items in the data.

Problem 2: In order to find rules that involve both frequent and rare items, we have to set the minsup very low. However, this may cause combinatorial explosion and make mining impossible because those frequent items will be associated with one another in all possible ways.

The solution to problem 1 is to allow the user to specipy multiple minimum supports. Thus, different itemsets need to satisfy different minimum supports depending on what items are in the itemsets.

The solution to problem 2 is to set a constraint, which prevent itemsets containing both very frequent and very rare items being generated.

### 2.4.1 Extended Models

To allow multiple minimum supports, the original model needs to be extended. - Let be the minimum item support(MIS) value of item .

The minimum support of a rule is the lowest MIS value among the items in the rule.

In this extended model, downward closure property no longer holds. In 2.4.2, we will present an algorithm to solve the problem of the lost of downward closure property

Till now, this extended model has solve the problem 1. When it comes to problem 2, we introduce the support difference constraint:

Let be the actual support of item in the data - For each itemset , the support difference constraint is:

is the user specified maximum difference

This constraint will not affect downward closure property

### 2.4.2 Mining Algorithm

We can call this new algorithm: MS-Apriori

The key operation in the new algorithm is the sorting of the items in I in ascending order of their MIS values. This order is fixed and used in all subsequent operations of the algorithm.