关联规则挖掘作业报告

数据集：作业一中的cbg\_patterns.csv数据集

**一.预处理操作**

由于apriori算法只能处理二值型数据，所以要将感兴趣的标称属性与数值属性进行离散化。

标称属性使用'属性=取值'的编码方式进行离散化。为每一种种类都进行唯一的编码，从而进行离散化。

数值属性使用'属性=百分比'的编码方式进行离散化，使用四分位数将原属性切分为四部分，使用0-0.25,0.25-0.5,0.5-0.75,0.75-1.0共4个离散化属性来替代原属性。

实验中只选取了原始数据中的'related\_same\_day\_brand', 'related\_same\_month\_brand','top\_brands','popularity\_by\_hour','popularity\_by\_day','raw\_visit\_count','raw\_visitor\_count', 'distance\_from\_home' 8种属性作为预处理的输入。最后将预处理结果保存到文件中。

**二.关联规则挖掘过程**

执行apriori算法，设置最低支持度为1000，最低置信度为0.8，得到如下的关联规则挖掘结果。





