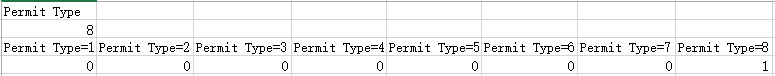
关联规则挖掘作业报告

数据集：作业一中的Building\_Permits.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，如属性'Permit Type'有8种取值，分别是1至8，使用'Permit Type=1'、'Permit Type=2'、'Permit Type=3'、…、'Permit Type=8'共8个离散化的属性来替代'Permit Type'。比如，在原始的'Permit Type'属性列上，有一行取值为8，那么离散的结果就是'Permit Type=8'属性取1，其余新的属性取值为0。



数值属性使用'属性=百分比'的编码方式进行离散化，如属性'Street、Number'，值域为[0,8400]，使用四分位数将原属性切分为四部分，使用'Street、Number=0~0.25'、'Street、Number=0.25~0.5'、'Street、Number=0.5~0.75'、'Street、Number=0.75~1.0'共4个离散化属性来替代'Street、Number'。比如，在原始的' Street、Number '属性列上，有一行取值为在第一四分位数与第二四分位数之间，那么'Street、Number=0.25~0.5'的取值为1，其余新属性的取值为0.

实验中只选取了原始数据中的'Structural、Notification'、'Voluntary、Soft-Story、Retrofit'、'Fire、Only、Permit'、'Site、Permit'、'Permit、Type'、'Current、Status'、'Existing、Construction、Type'、'Street、Number'共8个属性作为预处理的输入。

最后将预处理的结果保存到文件中。

**二.关联规则挖掘过程**

执行apriori算法，设置最低支持度为8000，最低置信度为0.9，得到如下的关联规则挖掘结果。



由于属性'Permit Type'取值为8的频数很大，达到178841，可以看出挖掘出的规则都与'Permit Type=8'有关。