关联规则过程报告

计算机学院

王钰

2120171076

**数据挖掘大作业二：关联规则挖掘**

**1. 数据源**

从大作业一的两个数据集中任选一个进行分析。

**2. 要求**

1. 对数据集进行处理，转换成适合关联规则挖掘的形式；
2. 找出频繁项集；
3. 导出关联规则，计算其支持度和置信度
4. 对规则进行评价，可使用Lift，也可以使用教材中所提及的其它指标

**3. 提交的内容**

* 对数据集进行处理的源程序
* 关联规则挖掘的源程序
* 挖掘结果及分析
* 挖掘过程的报告

**4.实验过程**

**4.1数据集选取**

Building\_Permits

**4.2 实验环境**

**系统环境：Windows**

**编程语言：python**

**4.3 分析过程**

**（1）数据预处理**

**将数据集进行预处理，转换成适合关联规则挖掘的形式。经过观察，去除数值属性和不具有挖掘价值的标称属性。选取了9个具有挖掘意义的属性，这9个属性分别是：Permit Type(2)、Street Number(7)、Current Status(14)、Structural Notification(20)、Existiong Use(28)、Proposed Use(30)、Existing Construction Type(34)、Proposed Construction Type(36)、Site Permit(38)。**

**（2）频繁规则统计与计算**

**本次数据挖掘使用了Apriori算法，如果要发现强关联规则，就必须先找到频繁集。所谓频繁集，即支持度大于最小支持度的项集。我们把计算出的频繁项集根据项集的元素个数分别按支持度降序排序，结果存储在txt 文件中，部分结果如下图所示，最终结果存储在frequent\_items.txt 中支持度阈值设置为0.01，最小置信度设为0.6。**

**Lift支持度，也叫做提升度，它是可信度与期望可信度的比值，反映了关联规则中的相关性。** 