关联规则挖掘

2120171107

朱佳琪

计算机科学与技术

# 1. 数据源

本次作业中，对下面一个数据集进行关联规则挖掘。

* 数据集1: San Francisco Building Permits

# 2. 实验环境

* 语言及环境依赖

语言： python

依赖的包：pandas、numpy

# 3. 实验过程

## 3.1对数据集进行处理，转换成适合关联规则挖掘的形式

* 在数据集San Francisco Building Permits上我们选取'Permit Type', 'Permit Type Definition', 'Plansets', 'TIDF Compliance', 'Existing Construction Type', 'Proposed Construction Type'6个属性进行关联规则挖掘，以减轻Apriori算法的运行负担。
* 将属性名和属性值进行拼接，形成新的属性值。例如：Existing\_Construction\_Type\_5.0', 'Permit\_Type\_Definition\_otc\_alterations\_permit'。
* 去除原数据集中缺失数据的元素，从而得到不等长的新属性，以获得更加真实的关联规则挖掘。

## 3.2找出频繁项集

* 使用Apriori算法，获取1到K项的频繁项集。调用ass\_mining.py中generate\_L(data\_set, k, min\_support)函数，其中k=3,将结果按支持度降序排列，存入results/ frequent\_items.txt文件中，部分频繁项集展示如下：

|  |  |
| --- | --- |
| frequent\_item\_set | support |
| frozenset(['Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.899 |
| frozenset(['Permit\_Type\_8']) | 0.899 |
| frozenset(['Permit\_Type\_8', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.899 |
| frozenset(['Proposed\_Construction\_Type\_5.0']) | 0.575 |
| frozenset(['Existing\_Construction\_Type\_5.0']) | 0.570 |
| frozenset(['Existing\_Construction\_Type\_5.0', 'Proposed\_Construction\_Type\_5.0']) | 0.563 |
| frozenset(['Proposed\_Construction\_Type\_5.0', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.514 |
| frozenset(['Proposed\_Construction\_Type\_5.0', 'Permit\_Type\_8']) | 0.514 |
| frozenset(['Proposed\_Construction\_Type\_5.0', 'Permit\_Type\_8', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.514 |

## 3.3导出关联规则，计算其支持度和置信度

* 使用Apriori算法，可由第二步中得到的频繁项集导出关联规则。调用ass\_mining.py中generate\_big\_rules(L, support\_data, min\_conf)函数，将结果按置信度降序排列，存入results/ rules\_order\_by\_confidence.txt文件中，部分关联规则展示如下：

|  |  |  |  |
| --- | --- | --- | --- |
| rule | support | confidence | lift |
| frozenset(['Permit\_Type\_8']) ==> frozenset(['Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.899 | 1.000 | 1.112 |
| frozenset(['Permit\_Type\_Definition\_otc\_alterations\_permit']) ==> frozenset(['Permit\_Type\_8']) | 0.899 | 1.000 | 1.112 |
| frozenset(['Proposed\_Construction\_Type\_5.0', 'Permit\_Type\_8']) ==> frozenset(['Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.514 | 1.000 | 1.112 |
| frozenset(['Proposed\_Construction\_Type\_5.0', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) ==> frozenset(['Permit\_Type\_8']) | 0.514 | 1.000 | 1.112 |
| frozenset(['Existing\_Construction\_Type\_5.0', 'Permit\_Type\_8']) ==> frozenset(['Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.508 | 1.000 | 1.112 |
| frozenset(['Existing\_Construction\_Type\_5.0', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) ==> frozenset(['Permit\_Type\_8']) | 0.508 | 1.000 | 1.112 |
| frozenset(['Plansets\_2.0', 'Permit\_Type\_8']) ==> frozenset(['Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.397 | 1.000 | 1.112 |
| frozenset(['Plansets\_2.0', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) ==> frozenset(['Permit\_Type\_8']) | 0.397 | 1.000 | 1.112 |

## 3.4使用Lift对规则进行评价

* 关联规则将结果按Lift值降序排列，存入results/ rules\_order\_by\_lift.txt文件中，部分关联规则展示如下：

|  |  |  |  |
| --- | --- | --- | --- |
| rule | support | confidence | lift |
| frozenset(['Plansets\_0.0']) ==> frozenset(['Existing\_Construction\_Type\_5.0', 'Permit\_Type\_8']) | 0.285 | 0.896 | 1.762 |
| frozenset(['Plansets\_0.0']) ==> frozenset(['Existing\_Construction\_Type\_5.0', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.285 | 0.896 | 1.762 |
| frozenset(['Plansets\_0.0']) ==> frozenset(['Proposed\_Construction\_Type\_5.0', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) | 0.286 | 0.899 | 1.750 |
| frozenset(['Plansets\_0.0']) ==> frozenset(['Proposed\_Construction\_Type\_5.0', 'Permit\_Type\_8']) | 0.286 | 0.899 | 1.750 |
| frozenset(['Plansets\_0.0', 'Proposed\_Construction\_Type\_5.0']) ==> frozenset(['Existing\_Construction\_Type\_5.0']) | 0.285 | 0.995 | 1.745 |
| frozenset(['Existing\_Construction\_Type\_5.0', 'Permit\_Type\_8']) ==> frozenset(['Proposed\_Construction\_Type\_5.0']) | 0.508 | 0.999 | 1.737 |
| frozenset(['Proposed\_Construction\_Type\_5.0']) ==> frozenset(['Existing\_Construction\_Type\_5.0', 'Permit\_Type\_8']) | 0.508 | 0.883 | 1.737 |
| frozenset(['Existing\_Construction\_Type\_5.0', 'Permit\_Type\_Definition\_otc\_alterations\_permit']) ==> frozenset(['Proposed\_Construction\_Type\_5.0']) | 0.508 | 0.999 | 1.737 |