

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级： CS1908**

**学 号： U201915160**

**姓 名： 陈千鹤**

**指导教师： 杨驰**

**报告日期： 2021年12月20日**

**计算机科学与技术学院**

**目录**

[实验三 关系挖掘实验 1](#_Toc90230186)

[**3.1 实验目的** 1](#_Toc90230187)

[**3.2 实验内容** 1](#_Toc90230188)

[**3.3 实验过程** 1](#_Toc90230189)

[3.3.1 编程思路 2](#_Toc90230190)

[3.3.2 遇到的问题及解决方式 2](#_Toc90230191)

[3.3.3 实验测试与结果分析 2](#_Toc90230192)

[**3.4 实验总结** 3](#_Toc90230193)

# 实验三 关系挖掘实验

## **3.1 实验目的**

1、加深对Apriori算法的理解,进一步认识Apriori算法的实现；

2、分析Apriori算法的缺点，使用pcy等变式对Apriori算法进行优化。

## **3.2 实验内容**

**必做：**

1. 实验内容

编程实现Apriori算法，要求使用给定的数据文件进行实验，获得频繁项集以及关联规则。

2. 实验要求

以Groceries.csv作为输入文件。

输出1~3阶频繁项集与关联规则，各个频繁项的支持度，各个规则的置信度，各阶频繁项集的数量以及关联规则的总数。

固定参数以方便检查，频繁项集的最小支持度为0.005，关联规则的最小置信度为0.5。

**加分项：**

1. 实验内容

在Apriori算法的基础上，要求使用pcy或pcy的几种变式multiHash、multiStage等算法对二阶频繁项集的计算阶段进行优化。

2. 实验要求

以Groceries.csv作为输入文件。

输出1~4阶频繁项集与关联规则，各个频繁项的支持度，各个规则的置信度，各阶频繁项集的数量以及关联规则的总数。

输出pcy或pcy变式算法中的vector的值，以bit位的形式输出。

 参数不变，频繁项集的最小支持度为0.005，关联规则的最小置信度为0.5。

## **3.3 实验过程**

1. 读物文件，预处理文件：用pandas读取csv文件，并且处理分割每行的数据。将数据映射为连续数字离散化。

2. 计算一阶频繁项集：统计每件物品的出现次数，计算支持度，并去除支持度低于阈值的，得到一阶频繁项集。

3. 计算二阶频繁项集：首先使用一阶频繁项集生成二阶备选集，然后使用PCY算法，构造频繁桶，生成位图，得到二阶频繁项集。

4. 计算三阶频繁项集：使用二阶频繁项集生成三阶备选集，计算支持度，去除支持度低于阈值的，得到三阶频繁项集。

5. 计算四阶频繁项集：使用三阶频繁项集生成四阶备选集，计算支持度，去除支持度低于阈值的，得到四阶频繁项集。

6. 生成关联关系。

7. 输出结果。

### 3.3.1 编程思路

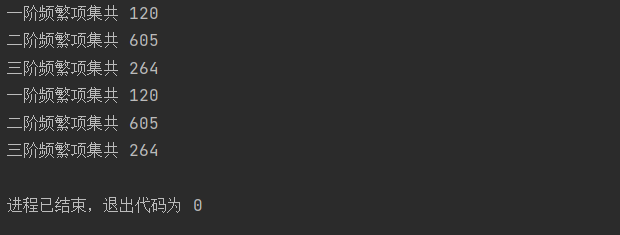
### 3.3.2 遇到的问题及解决方式

1. 一开始以为每次生成频繁项集都需要考虑置信度导致二阶频繁项集就成了0

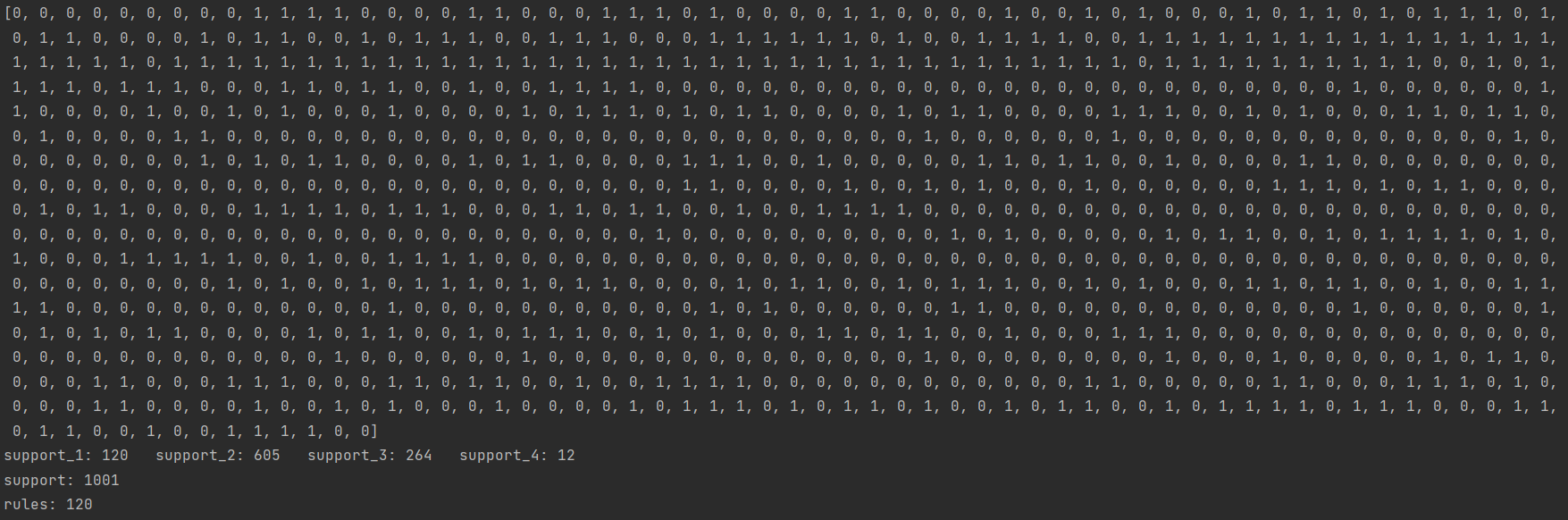
解决方法：研究了Apriori算法，发现不需要将置信度也作为筛选条件。

### 3.3.3 实验测试与结果分析

普通方法生成结果见下图。



PCY方法生成结果如下图所示。



## **3.4 实验总结**

1. 学习Apriori算法,进一步实现了Apriori算法；

2. 了解Apriori算法的缺点，使用pcy变式对Apriori算法进行了优化。