大数据分析实验
——关系挖掘

目录

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

一 任务背景

一 任务背景

- 1、理解Apriori算法思想与流程；
- 2、应用Apriori思想解决问题；

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

二 任务描述

二 任务描述

◆基本任务

- 以实验1得到的频率前1000个title及其相互引用关系作为关系挖掘实验的数据，从实验一中得到的引用关系数据为<<title,<title1,…,titlek>>,…>，将其处理为<<title,title1,…,titlek>,…>作为算法输入。
- 编程实现Apriori算法，使用实验数据进行实验，获得频繁项集以及关联规则。
- 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，固定参数以方便检查，频繁项集的最小支持度为0.15，关联规则的最小置信度为0.3。此处支持度的定义为某个项集出现的频率，也就是包含该项集的数目与总数目的比例（总的购物篮数目为1000）。

二 任务描述

◆进阶任务

- 在Apriori算法的基础上，使用pcy算法对二阶频繁项集的计算阶段进行优化。

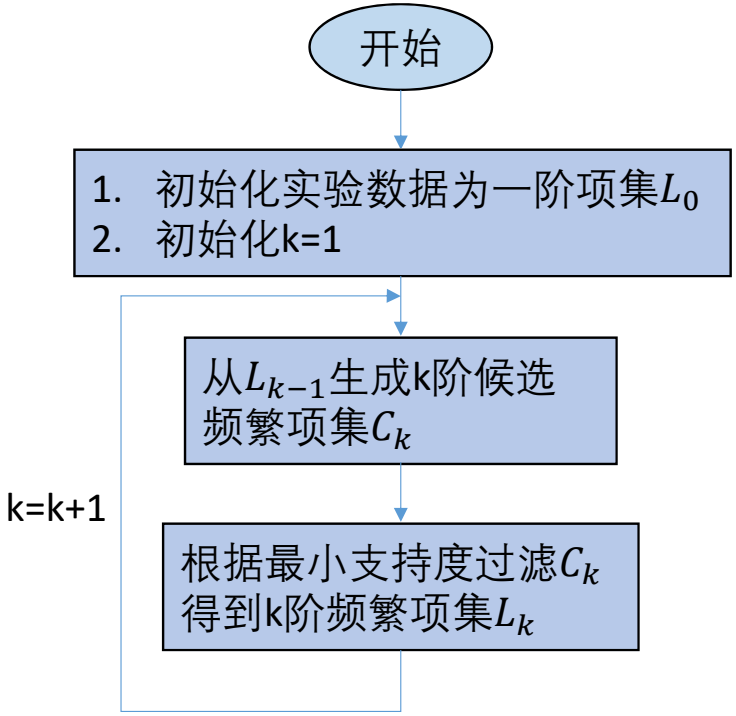
- 除基本任务的输出外，额外输出pcy算法中的vector的值，以bit位的形式输出。固定参数以方便检查，频繁项集的最小支持度为0.15，关联规则的最小置信度为0.3。

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

三 算法流程

三 算法流程

◆ 挖掘频繁项集



三 算法流程

◆ 由频繁项集产生关联规则

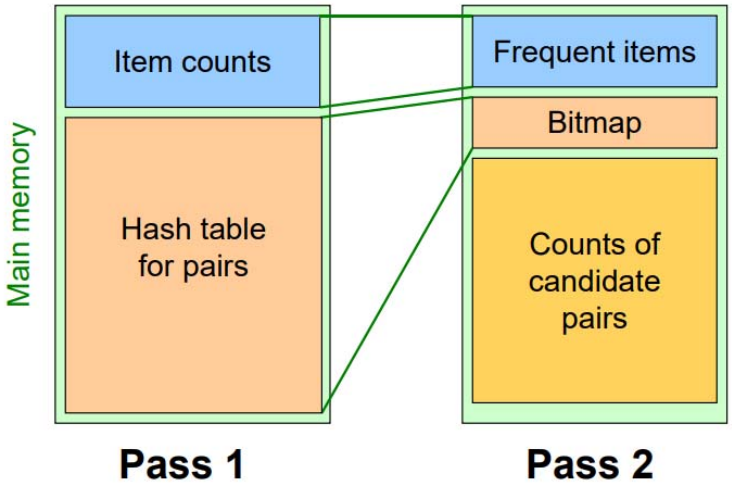
- 对于一个频繁项I，I的每个子集 A，生成一个规则 $A \rightarrow I \setminus A$
- 筛选出所有置信度大于最小置信度的规则
- 一个规则的置信度计算公式：$confidence(A \rightarrow I \setminus A) = support(I) / support(A)$

三 算法流程

◆进阶：PCY

```
FOR (each basket) :
    FOR (each item in the basket) :
        add 1 to item’s count;
    FOR (each pair of items) :
        hash the pair to a bucket;
        add 1 to the count for that bucket;
```

New in PCY



bit vector 的每一位代表一个bucket是否为频繁的，**如果一个bucket中的计数小于最小支持度，那么映射到这个桶的二阶项必然是非频繁的**

e.g. hash(i,j, buckets_len)=(i*j) % buckets_len


華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

四 验收流程

四 验收流程

- 检查1~4阶频繁项集和关联规则。
- 检查频繁项集和关联规则的数量。
- 提问了解编程思路和对Apriori算法和PCY算法的理解。