大数据分析实验
——关系挖掘

目录

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；
- 3、PCY算法解决问题。

二 任务描述

二 任务描述

◆ 任务一

- 以Groceries.csv作为输入文件，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。
- 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；

二 任务描述

◆任务二

- 在Apriori算法的基础上，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。
- 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。
- 输出PCY或PCY变式算法中的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0.005，关联规则的最小置信度为0.5。

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 基础知识

理论回顾 —— 购物篮模型

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Groceries.csv

	items
1	{citrus fruit,semi-finished bread,margarine,ready soups}
2	{tropical fruit,yogurt,coffee}
3	{whole milk}
4	{pip fruit,yogurt,cream cheese ,meat spreads}
5	{other vegetables,whole milk,condensed milk,long life bakery product}
6	{whole milk,butter,yogurt,rice,abrasive cleaner}
7	{rolls/buns}
8	{other vegetables,UHT-milk,rolls/buns,bottled beer,liquor (appetizer)}
9	{pot plants}
10	{whole milk,cereals}
11	{tropical fruit,other vegetables,white bread,bottled water,chocolate}
12	{citrus fruit,tropical fruit,whole milk,butter,curd,yogurt,flour,bottled water,dishes}
13	{beef}
14	{frankfurter,rolls/buns,soda}
15	{chicken,tropical fruit}

项：每一个商品，如Bread

项集：一些商品的集合，如{Coke, Milk}，含k个项的集合称为k阶项集

支持度：项集在所有购物篮中出现次数或频率：

s(itemset) = count(itemset) / len(market-basket)

支持度达到某个阈值的项集称为频繁项集，实验需要得到**1~3阶频繁项集**

关联规则：I->j，I是一个项集，j是一个项

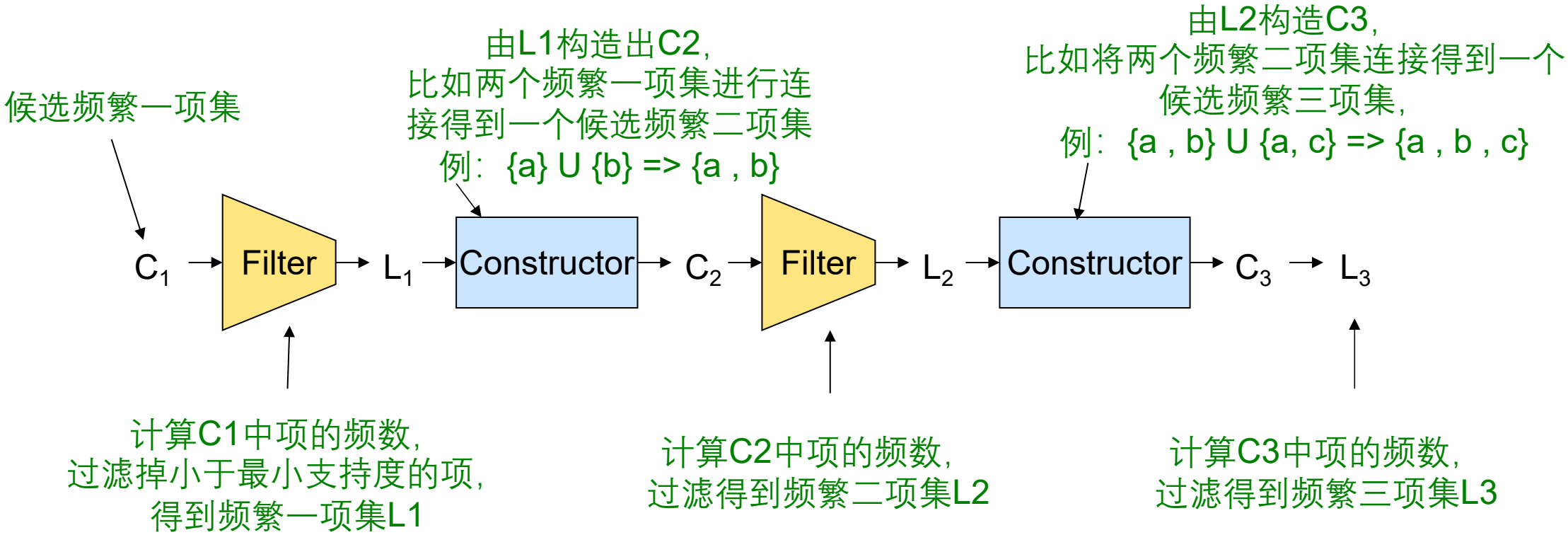
置信度：某个关联规则的可信程度。Rule: J-{j} -> j. J是一个频繁集，j是J中的一个项

conf(rule) = s(J) / s(J-{j})

实验中要求**筛选**出置信度不低于最小置信度的规则

三 Apriori基础知识

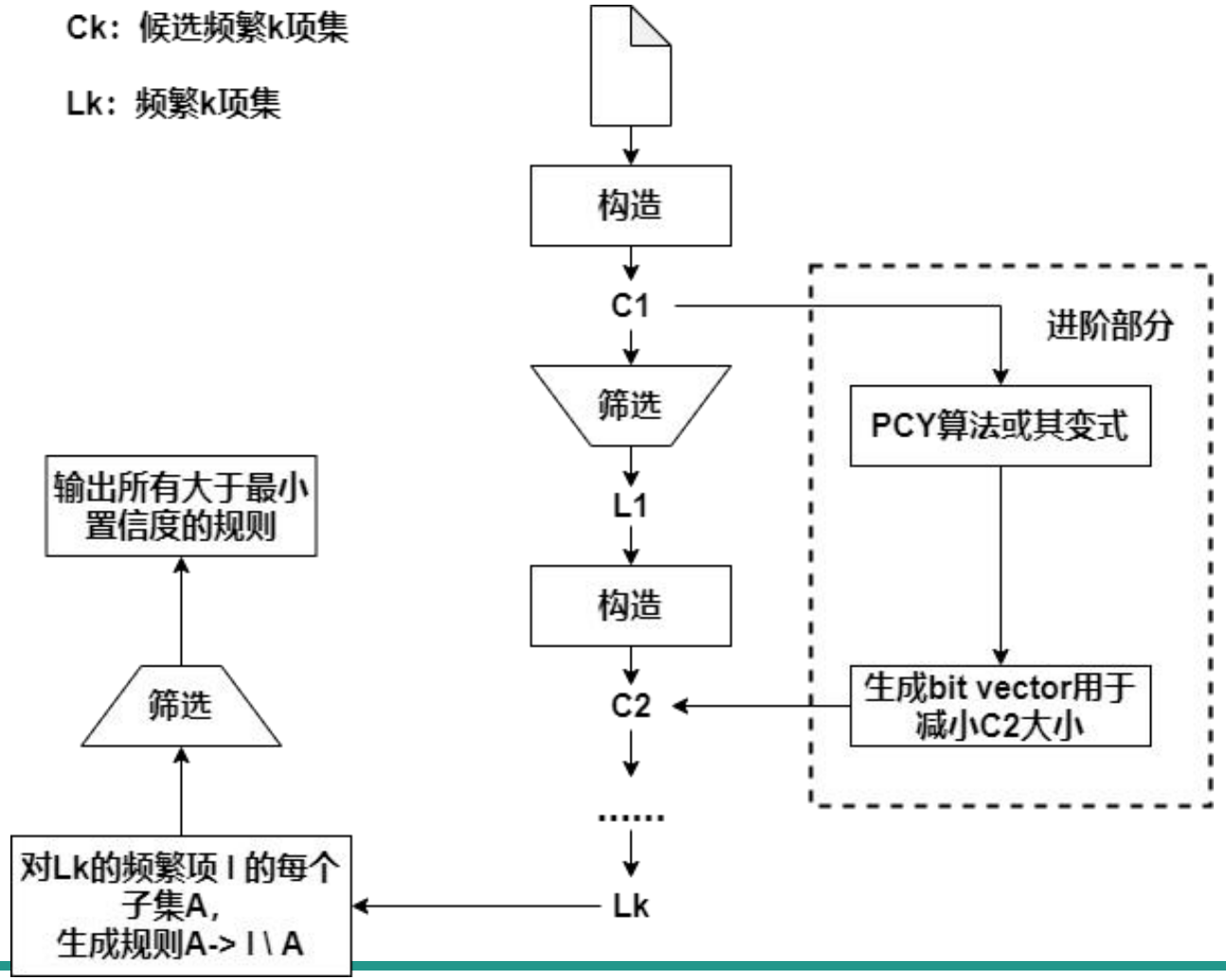
基本思路：**频繁项集的所有子集也一定是频繁项集。**



注意，由L_k构造出C_{k+1}的方法有很多，其目的只是减小后续筛选阶段的工作量。但必须保证L_{k+1}是C_{k+1}的子集。

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

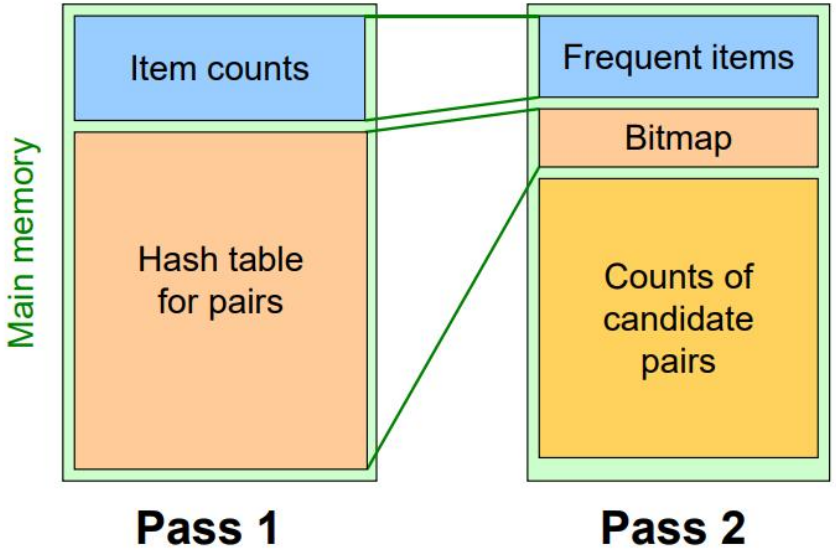
- 对于一个频繁项I，I的每个子集 A，生成一个规则 $A \rightarrow I \setminus A$
- 筛选出所有置信度大于最小置信度的规则
- 一个规则的置信度计算公式：$confidence(A \rightarrow I \setminus A) = support(I) / support(A)$

三 算法流程

◆ PCY算法

```
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1~3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。