# 一、实验目的

1、了解关联规则的基本思想，理解频繁项集挖掘的一般过程和基本原理；

2、巩固Apriori算法的算法思想，能够进行关联规则分析。

# 二、实验原理

1、频繁项集

设I={i1, i2,…,im}是项的集合，D是数据库事务的集合，其中每个事务T是一个非空项集，使得T包含于I。每一个事务具有唯一的事务标识符TID。设A是一个项集，事务T包含A，当且仅当A包含于T。如果项集A中包含k个项，则称其为k项集。项集的出现频度是包含项集的事务数，简称为项集的频度、支持度计数或计数。如果项集I的相对支持度满足预定义的最小支持度阈值，则I是频繁项集。

2、关联规则、关联规则的挖掘

关联规则是形如A=>B的逻辑蕴含式，其中A包含于I，B包含于I，A不等于空集，B不等于空集，并且A交B等于空集。如果事务数据库D中有s%的事务包含A并B，则称关联规则X=>Y的支持度为s%，实际上，支持度是一个概率值。若项集A的支持度记为support(A)，规则的置信度为support(A并B)／support (A)。这是一个条件概率P(B|A)。也就是：

support(A=>B)=P(A并B)

confidence(A=>B)=P(B|A)=support(A并B)／support(A)

同时满足最小支持度阈值（min\_sup）和最小置信度阈值（min\_conf）的规则称为强规则。

为方便计算，用0%~100%之间的值，而不是0.0~1.0之间的值来表示支持度和置信度。

SouthEast

该式表明，规则A=>B的置信度容易从A并B的支持度计数推出。也就是说，一旦得到A、B和A并B的支持度计数，则导出对应的关联规则A=>B和B=>A，并检查它们是否是强规则是直截了当的。因此，挖掘关联规则的问题可以归结为挖掘频繁项集。

一般而言，关联规则的挖掘是一个两步的过程：

（1）找出所有的频繁项集：根据定义，这些项集的每一个频繁出现的次数至少与预定义的最小支持度计数min\_sup一样。

（2）有频繁项集产生强关联规则：根据定义，这些规则必须满足最小支持度和最小置信度。

由于第二步的开销远低于第一步，因此挖掘关联规则的总体性能由第一步决定。

3、Apriori算法

Apriori算法将发现关联规则的过程分为两个步骤：

* + 通过迭代，检索出事务数据库中的所有频繁项集，即支持度不低于用户设定的阈值的项集；

（1）连接步：为找出Lk，通过将Lk-1与自身连接产生候选k项集的集合。该候选项集的集合记为Ck。

 （2）剪枝步：为了压缩Ck，可以使用先验性质。任何非频繁的（k-1）项集都不是频繁k项集的子集。注意，由于Apriori算法使用逐层搜索技术，给定候选k项集后，只需检查它们的（k-1）项所有子集是否频繁。

* + 利用频繁项集构造出满足用户最小信任度的规则。

# 三、实验内容：

数据库有10个购物篮事务，要求利用Apriori算法求出由项目集I={ I1，I2，I3，I4，I5，I6}组成的事务表中的频繁项集和关联规则。其中，最小支持度设定为3，最小置信度设定为70%。（项目集I中个项目对应的信息为：I1-牛奶，I2-啤酒，I3-尿布，I4-面包，I5-黄油，I6-饼干）

|  |  |
| --- | --- |
| TID | 商品 |
| T100 | 牛奶，啤酒，尿布 |
| T200 | 牛奶，面包，黄油 |
| T300 | 牛奶，尿布，饼干 |
| T400 | 面包，黄油，饼干 |
| T500 | 啤酒，尿布，饼干 |
| T600 | 牛奶，尿布，面包，黄油 |
| T700 | 尿布，面包，黄油 |
| T800 | 啤酒，尿布 |
| T900 | 牛奶，尿布，面包，黄油 |
| T1000 | 啤酒，饼干 |

# 四、实验步骤

（代码）

# 五、实验结果

# 六、心得体会

# 分组任务

请同学们自由组合分组，准备合作完成期末大作业，要求：

（1）每个小组不得超过4人，自设一名组长组织小组的分工合作；

（2）请大家尽快完成分组，并将小组名单下周二报给我，谢谢。