第4章 关联规则

本章目标

- 基本的概念；
- 了解Apriori算法基本原理、关键技术和应用。

4.1 关联规则概述

- 关联分析（Association Analysis）
  - 关联分析是发现事物之间关联关系（Associations）的分析过程。
  - 典型应用——就是购物篮分析（Market Basket Analysis）。
- 购物篮分析
  - 确定顾客在一次购物中可能一起购买的商品，发现其购物篮中不同商品之间的联系，分析顾客的购买习惯，从而发现购买行为之间的关联。
  - 显式关联：比如顾客购买了手机，很可能还会买手机套和膜。
  - 隐式关联

交易号	产品
T01	啤酒
T01	尿布
T02	啤酒
T02	尿布
T03	尿布

- 关联关系以一组特殊的规则形式出现——关联规则（Association Rules）

4.1 关联规则概述

基本概念

- 1.项与项集
- 数据库中不可分割的最小单位信息称为项（或项目），项的集合称为项集。

- 2.事务
- 设 $T=\{t_1,t_2,\cdots,t_n\}$ 是由数据库中所有项目构成的集合，事务数据库 $t_i(i=1,2,\cdots,n)$ 是由一系列具有唯一标识的事务组成。每一个事务包含的项集都是I的子集。

4.1 关联规则概述

- 3.项集的频数（支持度计数）
- 包括项集的事务数称为项集的频数（支持度计数）。

- 4.关联规则
- 一般表现为蕴涵式规则形式：$X \Rightarrow Y$
- **其中——**
  - X和Y分别称为关联规则的前提或先导条件（Antecedent）和结果或后继（Consequent）。
  - 关联规则反映X中的项目出现时，Y中的项目也跟着出现的规律。

4.1 关联规则概述

- 5.关联规则的支持度（support）
- 关联规则的支持度是交易集中同时包含X和Y的交易数与所有交易数之比，它反映了X和Y中所含的项在事务集中同时出现的频率，记为support（ ），即 $X \Rightarrow Y$

$$\sup port(X \Rightarrow Y) = \sup port(X \cup Y) = P(XY)$$

4.1 关联规则概述

- 5.关联规则的支持度（support）
  - 解释一：比如选秀比赛，那个支持和这个有点类似，那么多人（资料库），其中有多少人是选择（支持）你的，那个就是支持度；
  - 解释二：在100个人去超市买东西的，其中买苹果的有9个人，那就是说苹果在这里的支持度是 9，9/100；
  - 解释三：P(X)，意思是事件X出现的概率；
  - 解释四：关联规则当中是有绝对支持度（个数）和相对支持度（百分比）之分的。

4.1 关联规则概述

- 6.关联规则的置信度（confidence）
- 关联规则的置信度是交易集中同时包含X和Y的交易数与包含X的交易数之比，记为confidence（ ），置信度反映了包含X的事务中出现Y的条件概率。$X \Rightarrow Y$

$$\text{confidence}\ (X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} = P(Y|X)$$

4.1 关联规则概述

- 6.关联规则的置信度（confidence）
  - 在历史数据中，已经买了某某（例如：A、B）的支持度和经过挖掘的某规则（例如：A=>B）中A的支持度的比 例，也就是说买了A和B的人和已经买了 A的人的比例，这就是对“A推荐B”的置信度（A=>B的置信度）

4.1 关联规则概述

- 7.最小支持度与最小置信度
- 通常用户为了达到一定的要求，需要指定规则必须满足的支持度和置信度阈限值，此两个值称为最小支持度阈值(min_ sup)和最小置信度阈值(min_ conf)。其中，min_ sup描述了关联规则的最低重要程度，min_ conf规定了关联规则必须满足的最低可靠性。

4.1 关联规则概述

- 8.强关联规则
- 同时满足最小支持度阈值和最小置信度阈值的关联规则称为强关联规则。

- 经过关联规则分析后，针对某些人推销（根据某规则）比盲目推销（一般来说是整个数据）的比率，这个比率越高越好，这个规则称为强规则。

4.1 关联规则概述

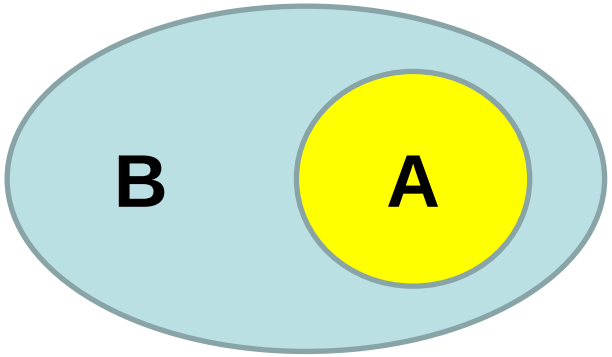
- 9.频繁项集
- 设 ，项目集U在数据集T上的支持度是包含U的事务在T中所占的百分比，即

$$\text{support}（U）=\frac{\left\|\{t \in T \mid U \subseteq t\}\right\|}{\|T\|}$$

- 式中： 表示集合中元素数目。对项目集I，在事务数据库T中所有满足用户指定的最小支持度的项目集，即不小于min_sup的I的非空子集，称为频繁项目集。
- 识别或发现所有频繁项集是关联规则发现算法的核心。

4.1 关联规则概述

- 10.项目集空间理论
- 阿戈沃（Agrawal）等人建立了用于事务数据库挖掘的项目集空间理论。
- 理论的核心是：频繁项目集的子集仍是频繁项目集，非频繁项目集的超集为非频繁项目集。



【例2.4】

根据顾客实际购买行为数据（表2.3，值为1表示购买了该种商品；值为0表示未购买该种商品），分析顾客在网络购物中购买图书、运动鞋、耳机、DVD和果汁五种商品时，是否存在购买行为上的关联。

表2.3 网络购物交易记录表

表2.3 网络购物交易记录表

序号	Book	Sneaker	Earphone	DVD	Juice
1	1	1	1	1	1
2	1	1	1	1	0
3	0	1	1	0	0
4	0	1	0	1	1
5	0	0	1	1	0
6	1	0	1	1	0
7	1	0	1	1	1
8	0	1	0	1	1
9	0	0	1	1	1
10	1	0	0	0	1

置信度（Confidence）和支持度（Support）

- 若得到4条关联关系

（1）如果顾客购买了Sneaker（运动鞋），那么他们也会购买Earphone（耳机）。

（2）如果顾客购买了Book（图书），那么他们也会购买Juice（果汁）。

（3）如果顾客购买了Book（图书）和DVD，那么他们也会购买Earphone（耳机）。

（4）如果顾客购买了Book（图书）、Sneaker（运动鞋）和Earphone（耳机），那么他们也会购买DVD。

- 使用置信度度量每个关联规则在前提条件下结果发生的可能性。

关联关系（1）的置信度为：3/5 = 60%。

- 使用支持度度量包含了关联关系中出现的属性值的交易占所有交易的百分比。

关联关系（1）的支持度为：3/10 = 30%

- 关联分析过程中设置置信度和支持度的阈值，当得到的关联关系达到置信度和支持度的阈值时，这样的关联关系被认为是有趣的，而保留下来应用到实际问题中。