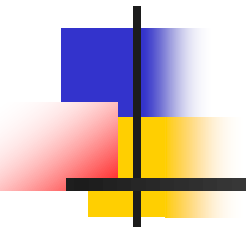


关联规则挖掘

杜逆索

三、关联规则挖掘

1.基本概念

关联规则挖掘用来发现大量数据中项集之间有趣的关联联系。如果两项或多项属性之间存在关联，那么其中一项的属性就可以依据其他属性值进行预测。

关联规则挖掘问题两个子问题：

第一步是找出事务数据库中所有大于等于用户指定的最小支持度的数据项集；

第二步是利用频繁项集生成所需要的关联规则，根据用户设定的最小置信度进行取舍，最后得到强关联规则。

三、关联规则挖掘

1.基本概念

识别或发现所有频繁项目集是关联规则发现算法的核心，关联规则的基本描述如下:

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），用符号表示，项的集合称为项集。

2.事务

设$I=\{i_1,i_2,\cdots,i_k\}$是由数据库中所有项目构成的集合，事务数据库$T=\{t_1,t_2,\cdots,t_n\}$是由一系列具有唯一标识的事务组成。每一个事务$t_i(i=1,2,\cdots,n)$包含的项集I都是的子集。

三、关联规则挖掘

1.基本概念

识别或发现所有频繁项目集是关联规则发现算法的核心，关联规则的基本描述如下:

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

关联规则是形如$X \Rightarrow Y$的蕴含式，其中X，Y分别是I的真子集，并且$X \cap Y = \phi$其中X称为规则的前提，Y称为规则的结果。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。

三、关联规则挖掘

1.基本概念

识别或发现所有频繁项目集是关联规则发现算法的核心，关联规则的基本描述如下：

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（$X \Rightarrow Y$），即

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

三、关联规则挖掘

1.基本概念

识别或发现所有频繁项目集是关联规则发现算法的核心，关联规则的基本描述如下：

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（$X \Rightarrow Y$），置信度反映了包含X的事务中出现Y的条件概率。

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

三、关联规则挖掘

1.基本概念

识别或发现所有频繁项目集是关联规则发现算法的核心，关联规则的基本描述如下：

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。

8.强关联规则

support（$X \Rightarrow Y$）≥ min_ sup且confidence($X \Rightarrow Y$) ≥ min_ conf，称关联规则 $X \Rightarrow Y$为强关联规则，否则称 $X \Rightarrow Y$为弱关联规则。通常所说的关联规则一般是指强关联规则。

三、关联规则挖掘

1.基本概念

识别或发现所有频繁项目集是关联规则发现算法的核心，关联规则的基本描述如下:

9.频繁项集

设$U \subseteq I$，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

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

式中:$\|\square\|$ 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，称为频繁项目集或大项目集。

三、关联规则挖掘

1.基本概念

识别或发现所有频繁项目集是关联规则发现算法的核心，关联规则的基本描述如下:

10.项目集空间理论

理论的核心为：频繁项目集的子集仍是频繁项目集，非频繁项目集的超集是非频繁项目集。

三、关联规则挖掘

2.关联规则挖掘算法-Apriori算法

（1）Apriori算法基本思想

Apriori算法基本思想是通过对数据库的多次扫描来计算项集的支持度，发现所有的频繁项集从而生成关联规则。

三、关联规则挖掘

2.关联规则挖掘算法-Apriori算法

（2）Apriori算法产生频繁项集的过程

产生频繁项集的过程主要分为连接步和剪枝步两步。

三、关联规则挖掘

2.关联规则挖掘算法-Apriori算法

（3）Apriori算法的主要步骤

（1）扫描全部数据，产生候选1-项集的集合C1；

（2）根据最小支持度，由候选1-项集的集合C1产生频繁1-项集的集合L1；

（3）对k>1，重复执行步骤（4）、（5）、（6）；

（4）由Lk执行连接和剪枝操作，产生候选（k+l）-项集的集合Ck+1；

（5）根据最小支持度，由候选（k+l）-项集的集合Ck+1，产生频繁（k+1）-项集的集合Lk+1；

（6）若L≠Φ，则k=k+1，跳往步骤（4）；否则，跳往步骤（7）；

（7）根据最小置信度，由频繁项集产生强关联规则，结束。

三、关联规则挖掘

2.关联规则挖掘算法-Apriori算法

（4）Apriori算法描述

输入：数据库D，最小支持度阀值min_ sup

输出：D中的频繁集L

（1）Begin

（2）L1=1-频繁项集；

（3）for（k=2；Lk-1≠Φ；k++）do begin

（4）Ck=Apriori_ gen（Lk-1）；{调用函数Apriori_ gen（Lk-1）通过频繁（k-1）-项集产生候选k-项集}

（5）for所有数据集tD do begin {扫描D用于计数}

三、关联规则挖掘

2.关联规则挖掘算法-Apriori算法

（4）Apriori算法描述

输入：数据库D，最小支持度阀值min_ sup

输出：D中的频繁集L

（6）Ct=subset（Ck，t）；{用subset找出该事务中是候选的所有子集}

（7）for所有候选集cCt do

（8）c. count++;

（9）end;

（10）Lk={cCk |c. countmin_ sup}

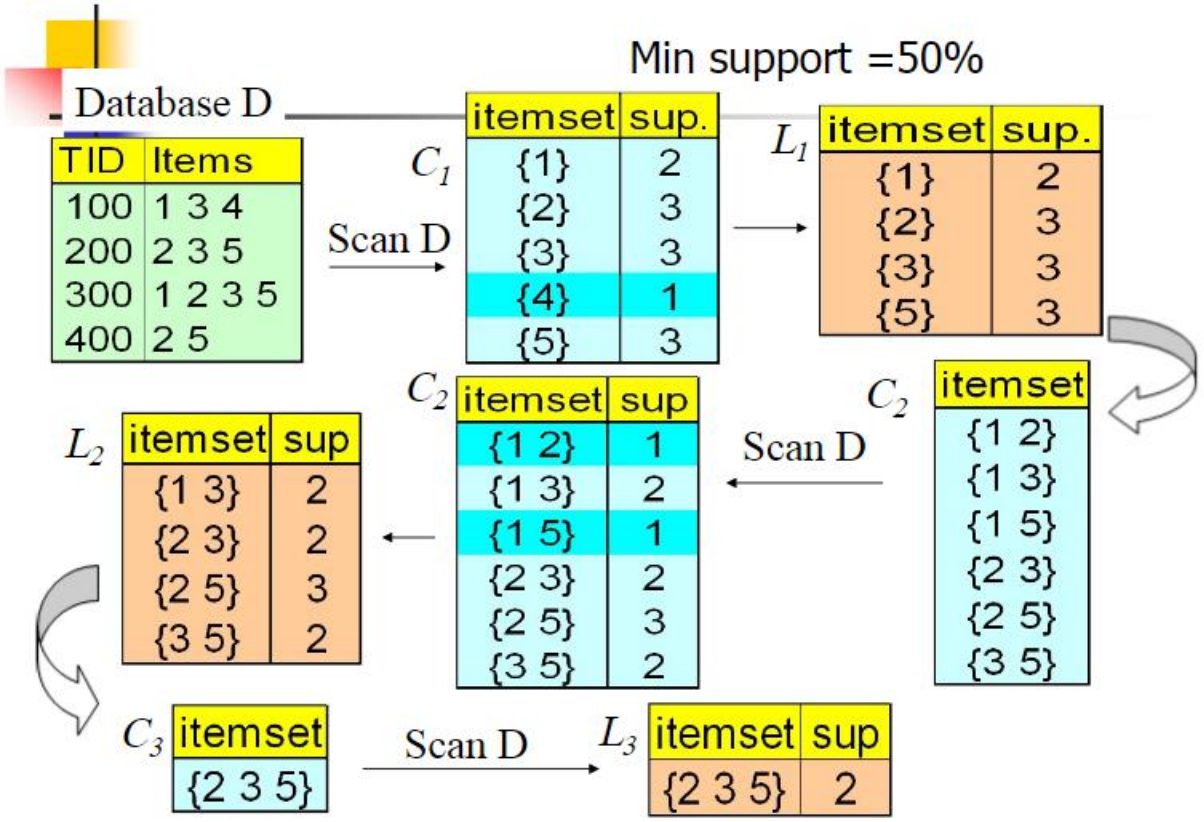
（11）end

（12）end

（13）Return L1∪L2∪Lk…∪Lm {形成频繁项集的集合}

三、关联规则挖掘

2.关联规则挖掘算法-Apriori算法



三、关联规则挖掘

3.Apriori算法特点及应用

3.1 Apriori算法的特点

Apriori算法是应用最广泛的关联规则挖掘算法，它有如下优点：

1） Apriori算法是一个迭代算法。
2）数据采用水平组织方式。
3）采用Apriori优化方法。
4） 适合事务数据库的关联规则挖掘。
5）适合稀疏数据集。

三、关联规则挖掘

3.Apriori算法特点及应用

3.1 Apriori算法的特点

Apriori算法有如下缺点:

1）多次扫描事务数据库,需要很大的I/O负载。

2）可能产生庞大的候选集。

3）在频繁项目集长度变大的情况下,运算时间显著增加。

三、关联规则挖掘

3.Apriori算法特点及应用

3.2Apriori算法的应用

- ✓ 商业
- ✓ 网络安全领域
- ✓ 高校管理
- ✓ 移动通信领域

三、关联规则挖掘

4.小结

详细介绍了关联规则挖掘的基本概念，描述了经典的关联规则挖掘算法-Apriori算法的原理以及发现频繁项目集过程。