关联分析:基本概念和算法

第六章讲义

Tan,Steinbach,Kumar 的幻灯片由 Michael Hahsler 改编

在课程网站上查找随附的 R 代码。

话题

·定义

·挖掘频繁项集（APRIORI） ·简明项集表示 · 查找频繁项集的替代方法 ·关联规则生成

·支持分布 ·模式评估

关联规则挖掘

·给定一组交易,找出可以预测交易的规则
基于交易中其他项目的发生,一个项目的发生

市场一篮子交易

时间项目	
1	面包、牛奶
2	面包、尿布、啤酒、鸡蛋
3	牛奶、尿布、啤酒、可乐
4	面包、牛奶、尿布、啤酒
5	面包、牛奶、尿布、可乐

关联规则示例

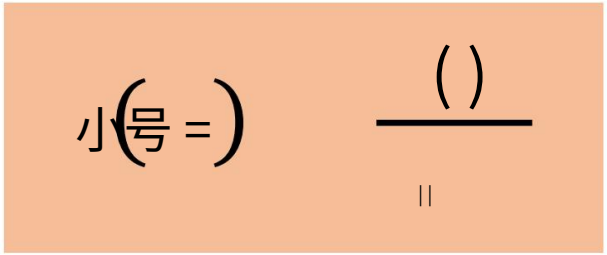
{ } → { },
{ , } → , },
{ , } { → { },

暗示意味着同时发生,而不是因果关系!

定义:频繁项集

- 项目集
  - 一个或多个项目的集合
    - 示例:{牛奶、面包、尿布}
  - k 项集
    - 包含k个项目的项目集
- 支持计数( )
  - 项集的出现频率
  - 例如 ({牛奶、面包、尿布}) = 2支持
-
  - 包含项集的事务的分数
  - 例如 s({牛奶、面包、尿布})
    = ({牛奶、面包、尿布}) / |T| = 2/5
- 频繁项集
  - 支持度大于或等于minsup阈值的项集

时间项目	
1	面包、牛奶
2	面包、尿布、啤酒、鸡蛋
3	牛奶、尿布、啤酒、可乐
4	面包、牛奶、尿布、啤酒
5	面包、牛奶、尿布、可乐



定义:关联规则

- 关联规则
  - 形式的隐含表达
    X → Y,其中 X 和 Y 是项集
  - 示例:
    { , } → { }

- 规则评估指标
  - 支持
    - 交易中包含的部分
      X 和 Y
  - 信心 (c)
    - 测量 Y 中项目的频率
      出现在交易中
      包含 X

时间	项目
1	面包、牛奶
2	面包、尿布、啤酒、鸡蛋
3	牛奶、尿布、啤酒、可乐
4	面包、牛奶、尿布、啤酒
5	面包、牛奶、尿布、可乐

例子:

{ , } → { }

$$= \frac{(\{\ ,\ ,\ \})}{|\ |} = \frac{2}{5} = 0.4$$

$$c = \frac{(\{\ ,\ ,\ \})}{(\{\ ,\ \})} = \frac{2}{3} = 0.67$$

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{s(X \cup Y)}{s(X)}$$

话题

·定义

·挖掘频繁项集（APRIORI） ·简明项集表示 ·

查找频繁项集的替代方法 ·关联规则生成

·支持分布 ·模式评估

关联规则挖掘任务

·给定一组交易T,关联规则的目标
挖掘是找到所有具有- support ≥
minsup阈值- confidence ≥
minconf阈值的规则

·蛮力方法： -列出所有可能
的关联规则-计算每个规则的支持度和置
信度-修剪未通过minsup和minconf阈值的规则 计算上禁止！

矿业协会规则

时间项目	
1	面包、牛奶
2	面包、尿布、啤酒、鸡蛋
3	牛奶、尿布、啤酒、可乐
4	面包、牛奶、尿布、啤酒
5	面包、牛奶、尿布、可乐

规则示例：

{牛奶、尿布} → {啤酒} (s=0.4, c=0.67)
{牛奶、啤酒} → {尿布} (s=0.4, c=1.0)
{尿布、啤酒} → {牛奶} (s=0.4, c=0.67)
{啤酒} → {牛奶、尿布} (s=0.4, c=0.67)
{尿布} → {牛奶、啤酒} (s=0.4, c=0.5)
{牛奶} → {尿布、啤酒} (s=0.4, c=0.5)

观察：

- 上述所有规则都是同一项目集的二元分区：
  {牛奶、尿布、啤酒}
- 源自相同项集的规则具有相同的支持度，但可以具有不同的置信度
- 因此，我们可以将支持度和置信度要求解耦

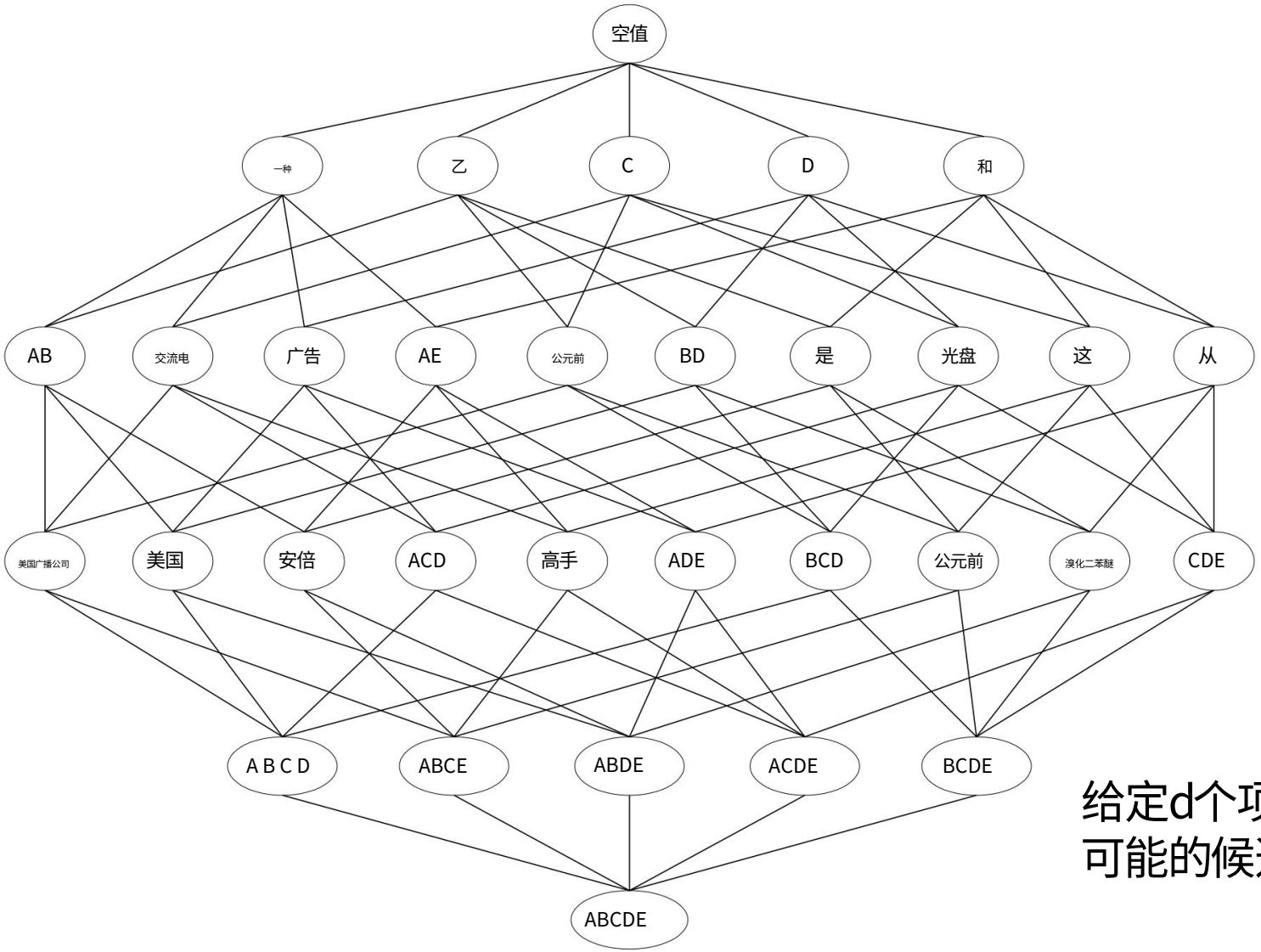
矿业协会规则

- 两步法： 1.频繁项集生成
  - –生成所有支持 minsup 的项集

  2.规则生成
  - –从每个频繁项集生成高置信度规则,其中每个规则是频繁项集的二分法

- 频繁的项集生成仍然是计算的昂贵的

频繁项集生成



给定d个项目,有2d个可能的候选项目集

减少候选人人数

先验原则：

- 如果一个项集是频繁的，那么它的所有子集也必须是频繁的
  频繁

由于支持度量的以下性质，Apriori 原则成立：

$$\forall \quad , \quad :(\subseteq \Rightarrow \not\ni ( ) \quad (\quad)$$

- 项目集的支持永远不会超过其子集的支持
- 这被称为支持的反单调特性

说明先验原理

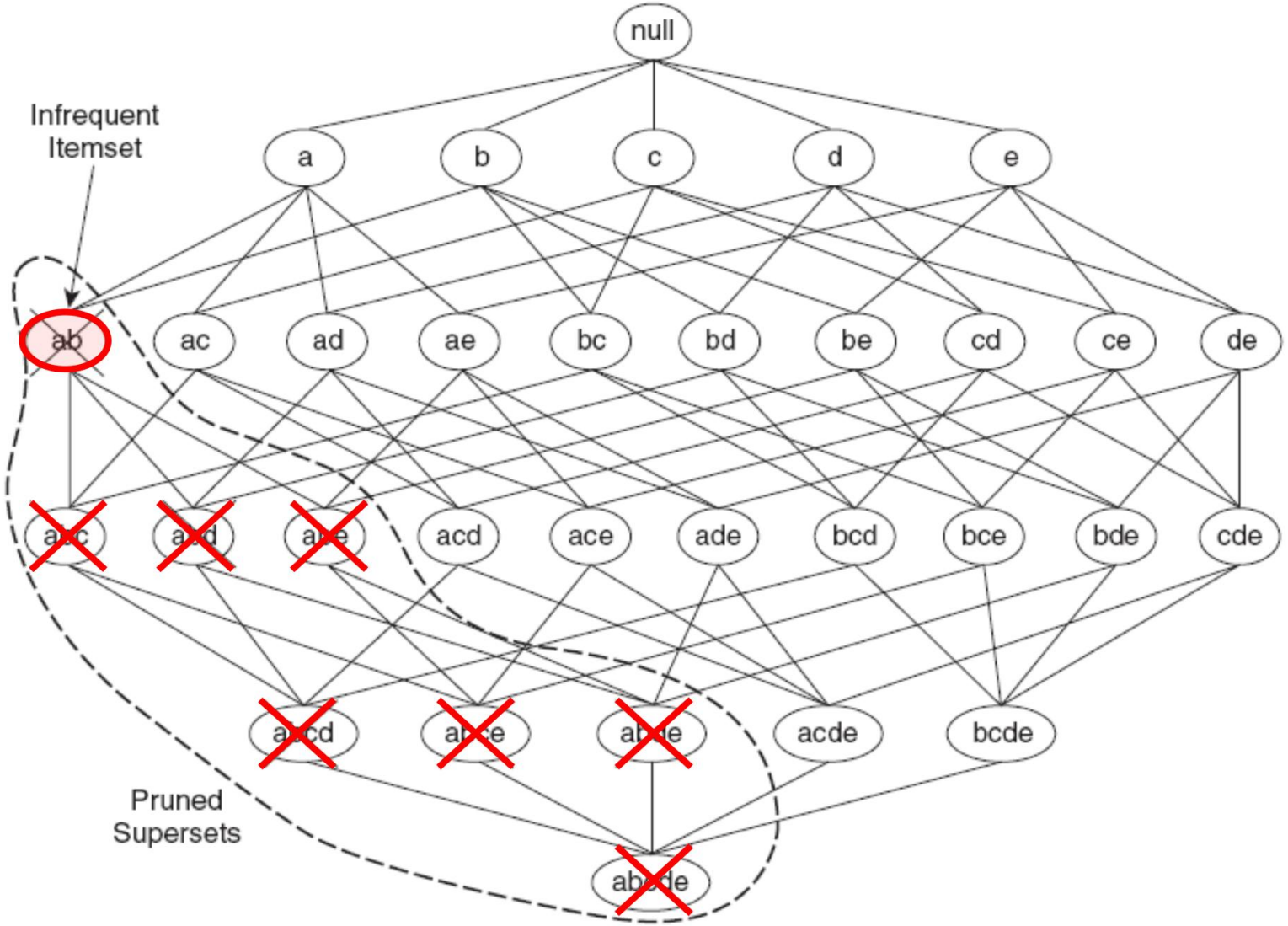


Figure 6.4. An illustration of support-based pruning. If $\{a, b\}$ is infrequent, then all supersets of $\{a, b\}$ are infrequent.

说明先验原理

项目（1-项目集）

物品	数数
面包	4
可乐	2
牛奶	4
啤酒	3
尿布	4
蛋	1

对（2项集）

物品集	数数
{面包,牛奶}	3
{面包,啤酒}	2
{面包,尿布}	3
{牛奶,啤酒}	2
{牛奶,尿布}	3
{啤酒,尿布}	3

（无需生成
涉及可口可乐的候选人
或鸡蛋）

最低支持 = 3

三胞胎（3项集）

物品集	数数
{面包,牛奶,尿布}	3

如果考虑每个子集，
6C1 + 6C2 + 6C3 = 41
使用基于支持的修剪，
6 + 6 + 1 = 13

先验算法

- 方法：
  - 让 k=1
  - 生成长度为 1 的频繁项集
  - 重复直到没有新的频繁项集被识别
    - 从长度为k的频繁项集生成长度为(k+1)的候选项集
    - 修剪包含不频繁的长度为 k 的子集的候选项目集
    - 通过扫描数据库统计每个候选者的支持度
    - 剔除不常出现的候选人，只留下经常出现的候选人

影响复杂性的因素

- 最低支持阈值的选择
  - 降低支持阈值可能导致更多的频繁项集
  - 这可能会增加候选人的数量和频繁的最大长度项集
- 数据集的维度（项目数）
  - 需要更多空间来存储每个项目的支持计数
  - 如果频繁项的数量也增加，计算和 I/O 成本也可能增加
- 数据库大小
  - 由于 Apriori 进行多次传递，算法的运行时间可能会增加交易数量
- 平均交易宽度
  - 交易宽度随着更密集的数据集而增加
  - 这可能会增加频繁项集的最大长度和哈希树的遍历（事务中的子集数量随其宽度增加）

话题

·定义

·挖掘频繁项集（APRIORI） ·简明项集表示 ·

查找频繁项集的替代方法 ·关联规则生成

·支持分布 ·模式评估

最大频繁项集

一个项目集是最大频繁的,如果它的直接超集都不是频繁的

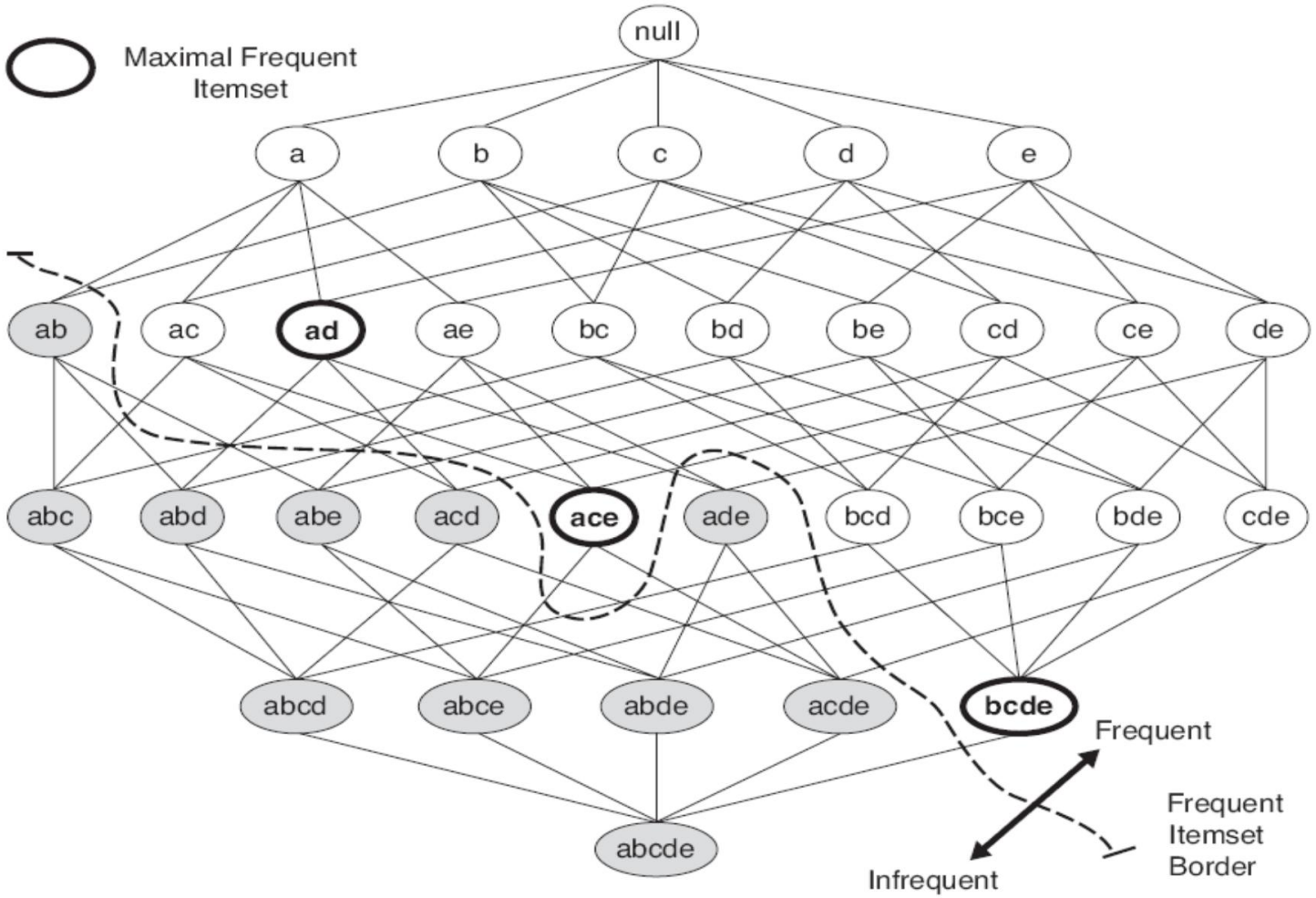


Figure 6.16. Maximal frequent itemset.

封闭项目集

- 如果项目集的直接超集都没有与项目集相同的支持，则项目集是封闭的（只能有较小的支持 -> 参见 APRIORI 原则）

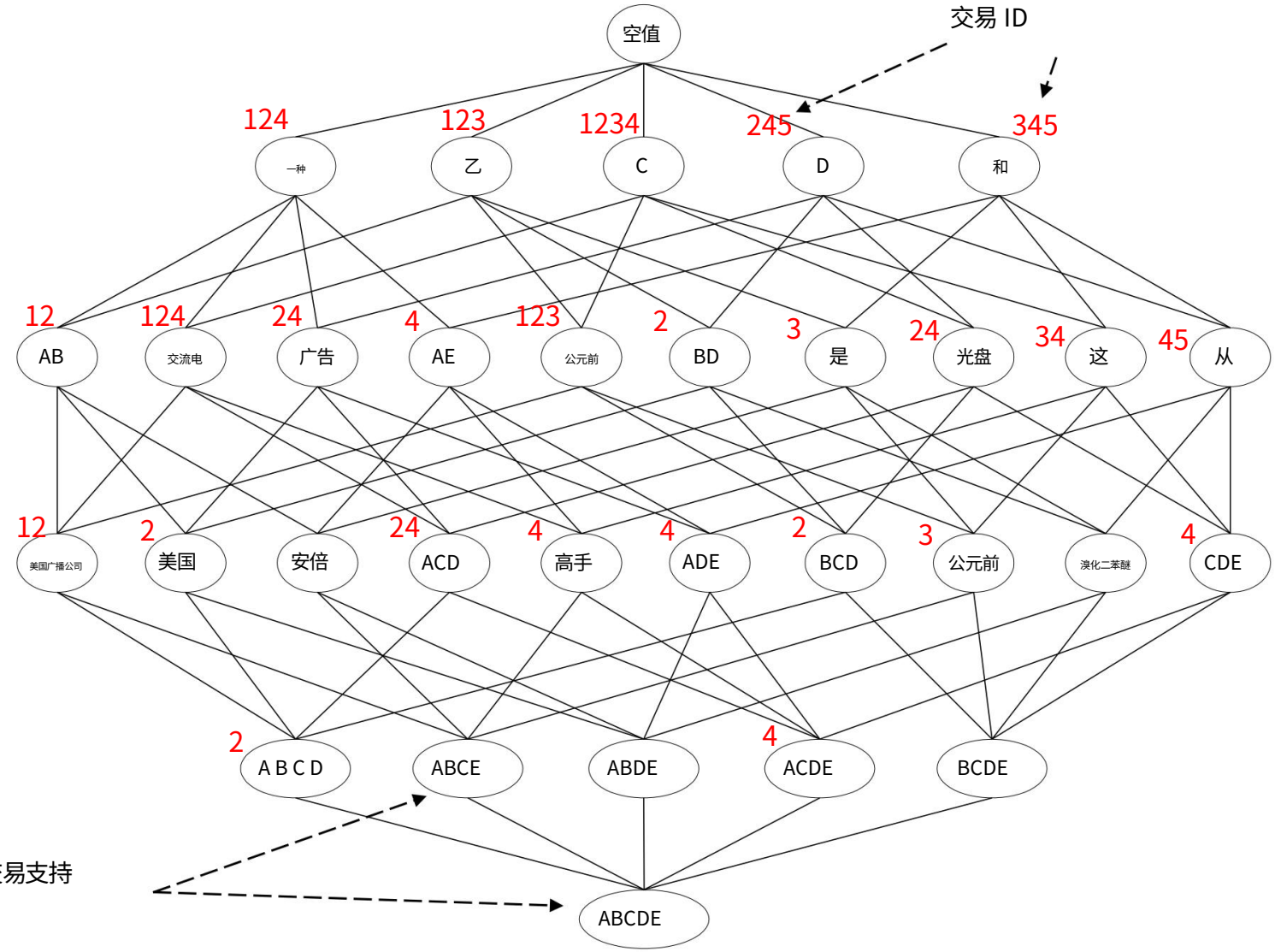
时间	项目
1 2	{A,B}
3 4	{B,C,D}
5	{A B C D}
	{A,B,D}
	{A B C D}

物品集	支持
{一种}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{广告}	3
{公元前}	3
{B,D}	4
{光盘}	3

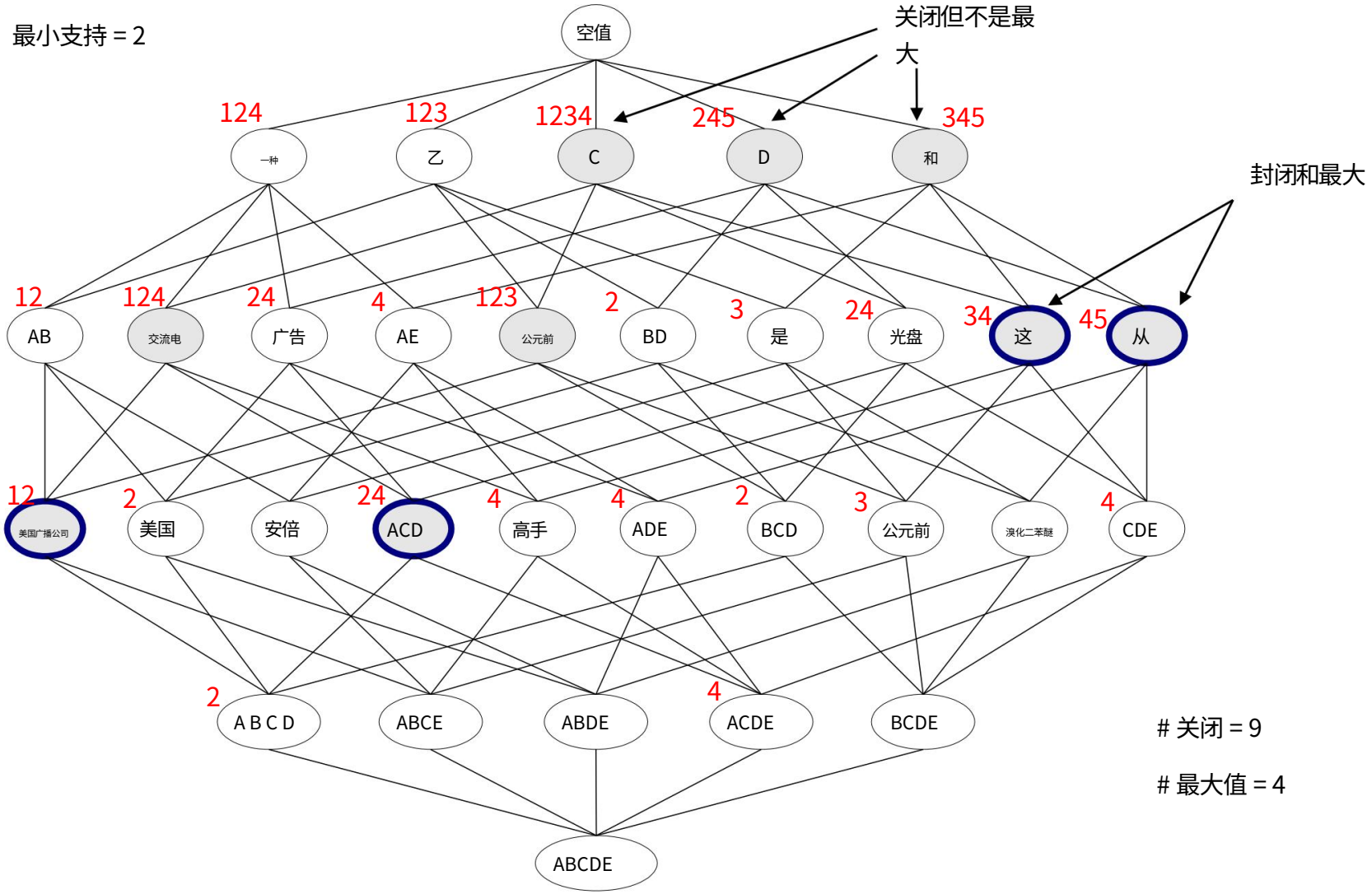
物品集	支持
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A B C D}	2

最大与封闭项集

时间项目	
1	美国广播公司
2ABCD	
公元前 3 年	
4 ACDE	
5	从



最大与封闭频繁项集



最大与封闭项集

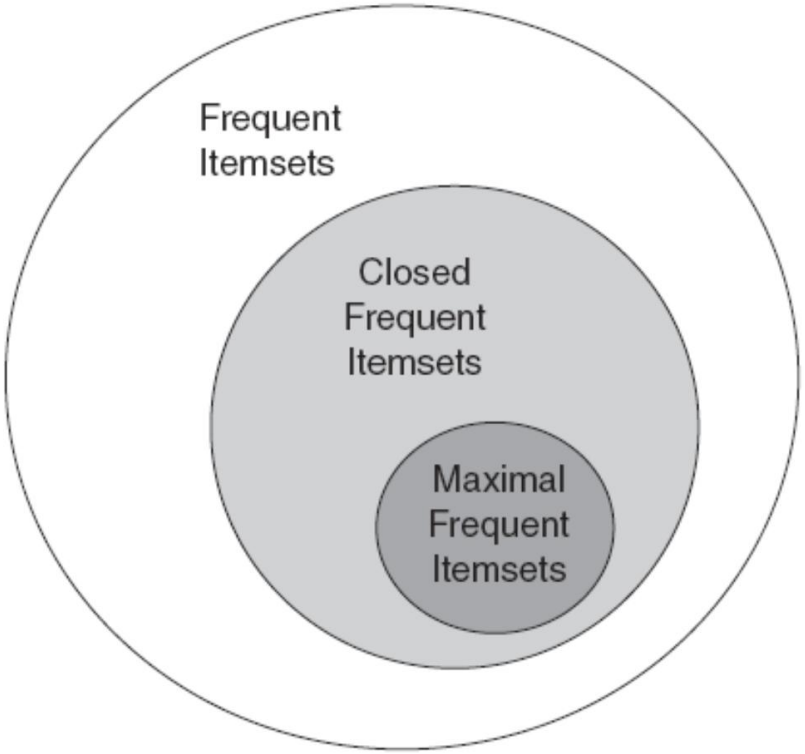


Figure 6.18. Relationships among frequent, maximal frequent, and closed frequent itemsets.

话题

·定义

·挖掘频繁项集（APRIORI） ·简明项集表示 ·

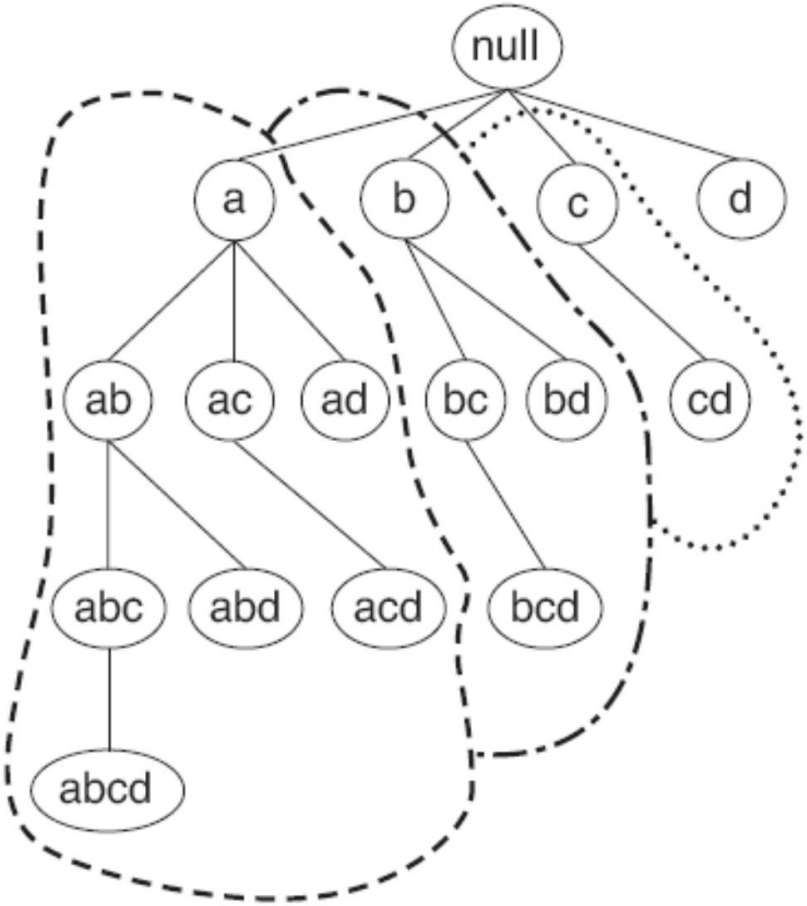
查找频繁项集的替代方法 ·关联规则生成

·支持分布 ·模式评估

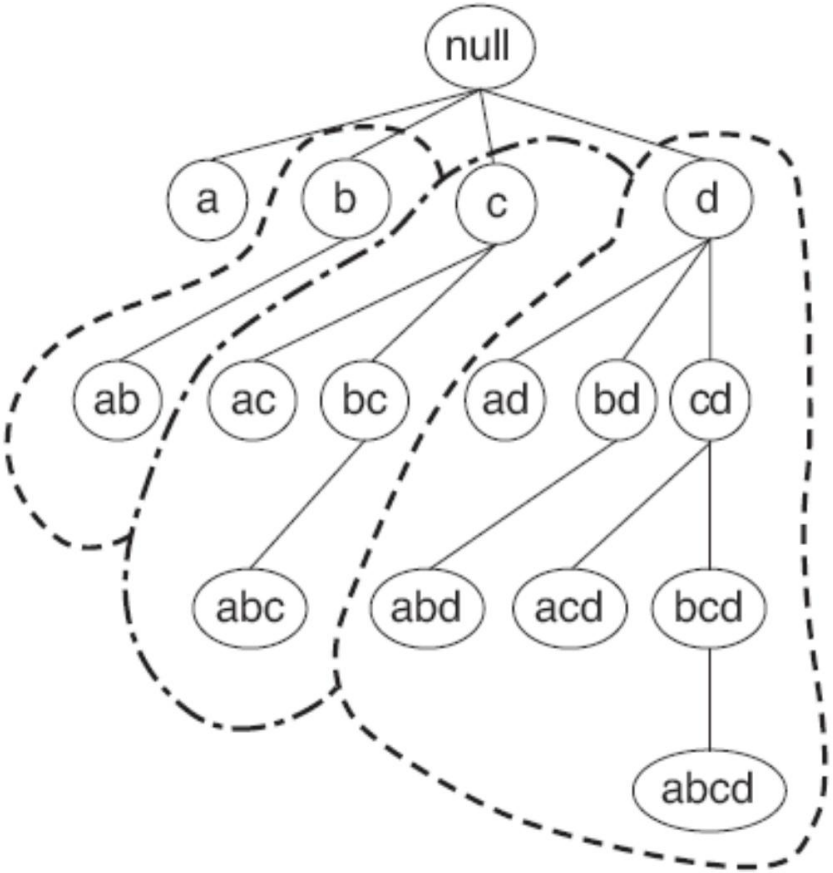
频繁项集生成的替代方法

·项集格的遍历

-等效类



(a) Prefix tree.



(b) Suffix tree.

频繁项集生成的替代方法

- 数据库的表示:水平与垂直数据布局

Horizontal
Data Layout

TID	Items
1	a,b,e
2	b,c,d
3	c,e
4	a,c,d
5	a,b,c,d
6	a,e
7	a,b
8	a,b,c
9	a,c,d
10	b

Vertical Data Layout

a	b	c	d	e
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

Figure 6.23. Horizontal and vertical data format.

替代算法

- FP-增长
  - 使用数据库的压缩表示,使用 FP树
  - 一旦构建了 FP-tree,它使用递归 分治法挖掘频繁项集
- ECLAT
  - 存储交易 ID 列表（垂直数据布局）。
  - 执行快速 tid-list 交集（按位异或)来计数 项集频率

话题

·定义

·挖掘频繁项集（APRIORI） ·简明项集表示 ·

查找频繁项集的替代方法 ·关联规则生成

·支持分布 ·模式评估

规则生成

·给定一个频繁项集L,找出所有非空子集
X=f　　L 和 Y=L – f 使得 X → Y 满足最小置信度要求

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

-如果 {A,B,C,D} 是频繁项集,候选规则：

ABC→D,	美国→C,	ACD →B,	BCD →A,
A →BCD,	B→ACD,	C →ABD,	D →ABC
AB →CD,	交流→BD,	公元→公元前,	公元前→公元,
BD →交流,	CD →AB,		

如果 |L| = k,则有2k – 2 个候选关联规则（忽略 L →　　和　　→ L）

规则生成

·如何有效地从频繁项集中生成规则？

- 一般来说，信心没有反单调
  财产

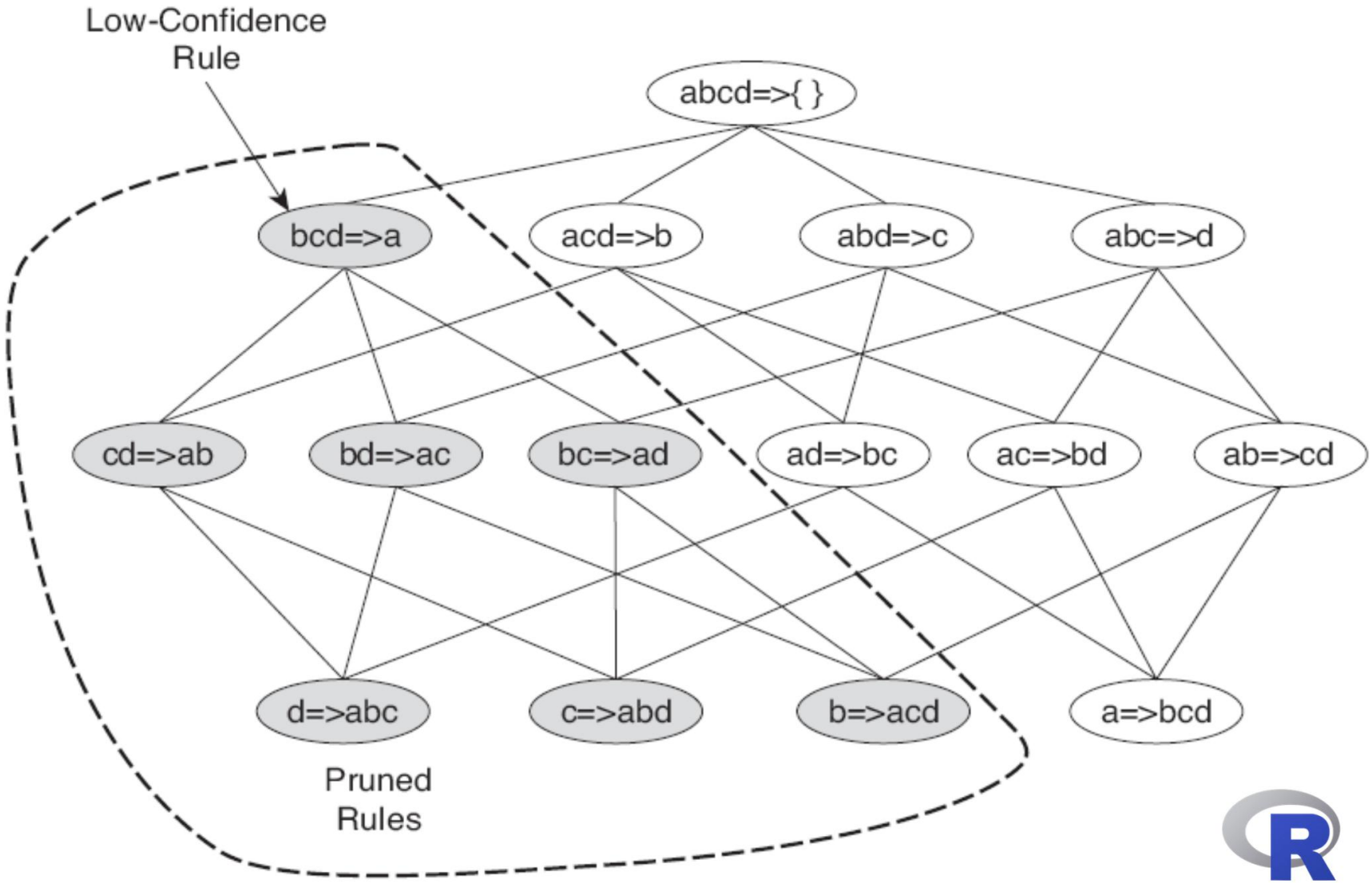
  c(ABC →D) 可以大于或小于 c(AB →D)

- 但是从相同项集生成的规则的置信度
  具有反单调性
- 例如，L = {A,B,C,D}：

$$c(ABC \rightarrow D) \quad c(AB \rightarrow CD) \quad c(A \rightarrow BCD)$$

·置信度是 RHS 上的项目数量的反单调
规则

Apriori 算法的规则生成



话题

·定义

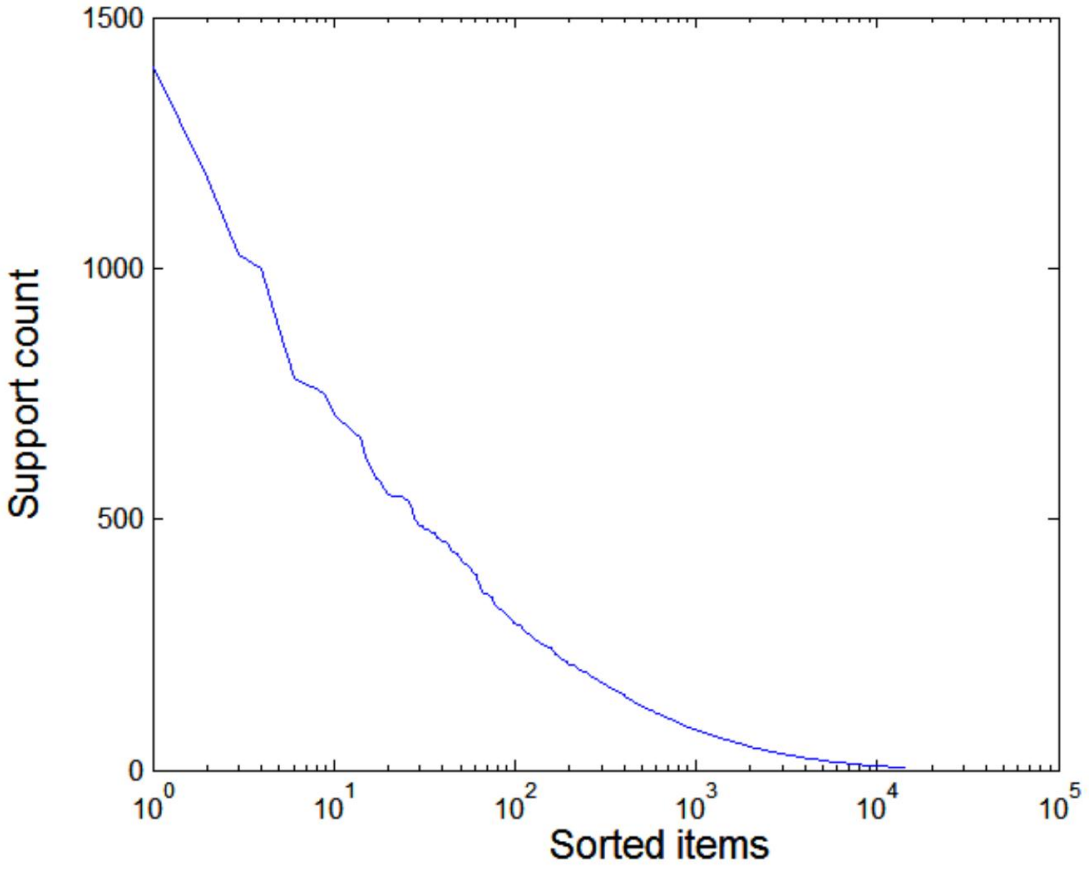
·挖掘频繁项集（APRIORI） ·简明项集表示 ·
查找频繁项集的替代方法 ·关联规则生成

·支持分布 ·模式评估

支持分配的效果

·许多真实数据集的支持分布有偏差

支持零售
数据集的分发



支持分配的效果

·如何设置合适的minsup阈值?

-如果minsup设置得太高,我们可能会错过包含有趣稀有物品（例如,昂贵的产品)的项目集

-如果minsup设置得太低,计算量很大,并且项集的数量非常大

· 使用单一的最小支持阈值可能不是有效的

话题

·定义

·挖掘频繁项集（APRIORI） ·简明项集表示 · 查找频繁项集的替代方法 ·关联规则生成

·支持分布 ·模式评估

模式评估

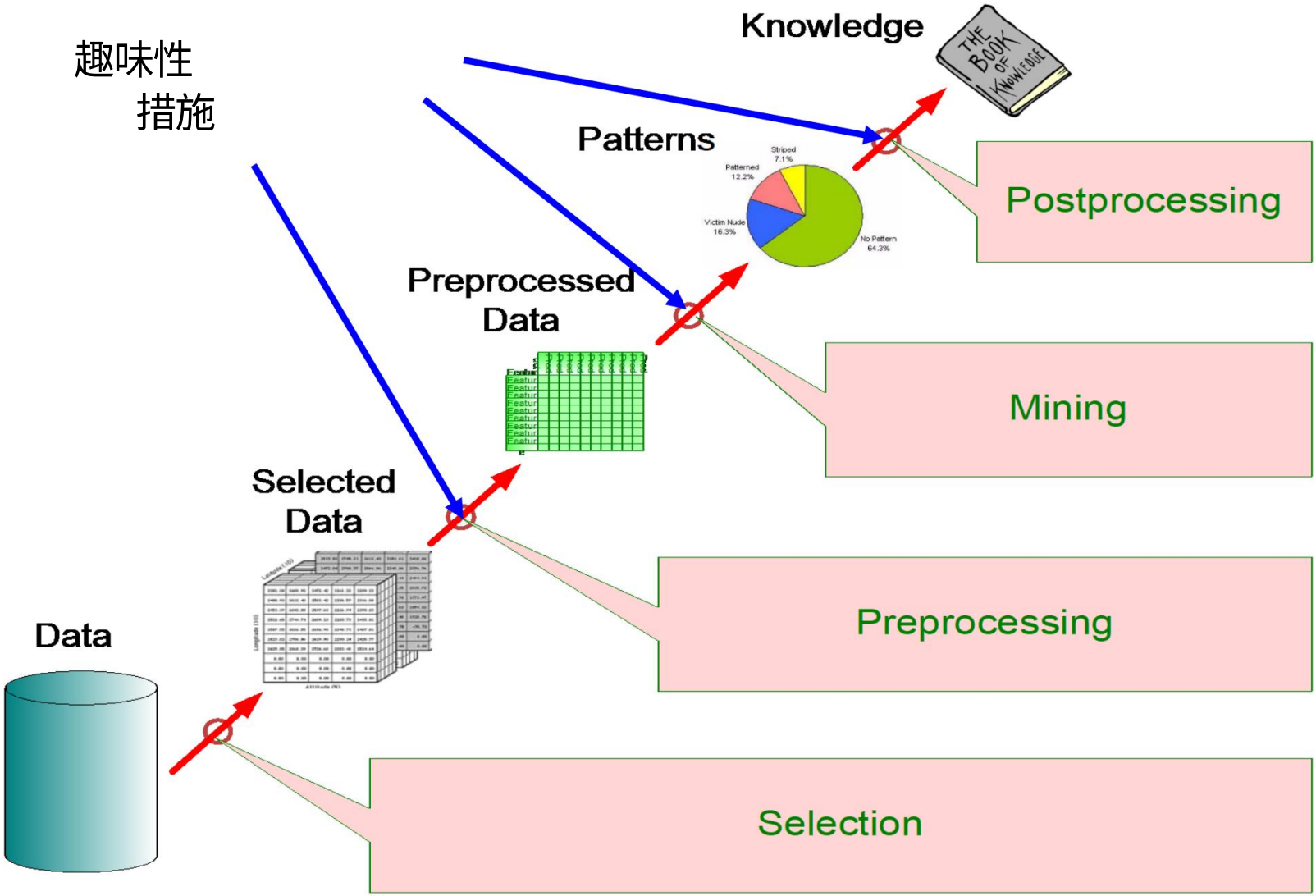
- 关联规则算法往往会产生过多的规则。其中许多是-无趣或

  -冗余

- 兴趣度测量可用于修剪/排列派生模式

·如果{A,B} → {D} 具有相同或更高的置信度,则可以认为规则{A,B,C} → {D} 是冗余的。

兴趣度测量的应用



计算兴趣度测量

·给定规则X → Y,计算规则兴趣度所需的信息可以从列联表中获得

X → Y的列联表

	f11	f10	f1 +
	f01	f00	佛+
	f+1	f+0	|T|

f11:支持 X 和 Y
f10:支持 X 而不是 Y
f01:支持非 X 和 Y
f00:支持非 X 非 Y

错误

用于定义各种度量

例如,支持、信心、提升、基尼,
J-测量等

喂, ({ }) = $\frac{11}{| \; |}$ 估计 (,)

(→=) $\frac{11}{1+}$ 估计 (|)

信心不足

	咖啡	咖啡		
茶	15	5	20
茶	75	5	80
	90	10	100

关联规则:茶→咖啡

支持 = P(咖啡、茶) = 15/100 = 0.15

信心= P（咖啡|茶)= 15/20 = 0.75

但是 P（咖啡)= 90/100 = 0.9

尽管信心很高,但规则具有误导性

P(咖啡|茶) = 75/80 = 0.9375

统计独立性

- 1000 名学生- 600 名学生会游泳
  (S) - 700 名学生会骑车 (B) - 450 名学生会游
  泳和骑车 (S,B)

- P(S,B) = 450/1000 = 0.45 （观察到的关节概率）
- P(S) P(B) = 0.6 0.7 = 0.42 （预计独立）

- P(S,B) = P(S) P(B) => 统计独立性
- P(S,B) > P(S) P(B) => 正相关
- P(S,B) < P(S) P(B) => 负相关

基于统计的措施

·将统计相关性考虑到规则的措施： $X \rightarrow Y$

$$= \quad = \frac{(|)}{()} = \frac{(,)}{(\ )\ ()}$$

$$= \quad (\ ,\ )^{-}\ \ (\ )\ (\ )$$

偏离独立

$$\Phi = \frac{(\ ,\ )^{-}\ \ (\ )\ \ ()}{\sqrt{\ (\ )[1-\ \ (\ )]\ (\ )_{[1-]}\ \ (\ )}}$$

相关性

示例:提升/利息

	咖啡	咖啡	
茶	15	5	20
茶	75	5	80
	90	10	100

关联规则:茶→咖啡

Conf(茶 → 咖啡)= P(咖啡|茶) = P(咖啡,茶)/P(茶) = .15/.2 = 0.75

但是 P(咖啡) = 0.9

提升(茶 → 咖啡) = P(Coffee,Tee)/(P(Coffee)P(Tee)) = .15/(.9 x .2) = 0.8333

注意： Lift < 1,因此咖啡和茶呈负相关

文献中提出了很多措施

有些措施适用于某些应用,但不适用于其他应用

我们应该使用什么标准来确定衡量标准是好是坏?

基于Apriori样式支持的情况如何?
修剪?它如何影响这些措施?

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left. P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\left(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]\right.$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $\left.-P(A)^2-P(\overline{A})^2\right)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$

比较不同的措施

列联表的 10 个示例：

示例f11	f10	f01	f00
E1	8123	83	424 1370
E2	8330	2	622 1046
E3	9481	94	127 298
E4	3954	3080	5 2961
E5	2886	1363	1320 4431
E6	1500	2000	500 6000
E7	4000	2000	1000 3000
E8	4000	2000	2000 2000
E9	1720	7121	5 1154
E10	61	2483	4 7452

使用各种度量的列联表排名：

支持与信心

#	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

电梯

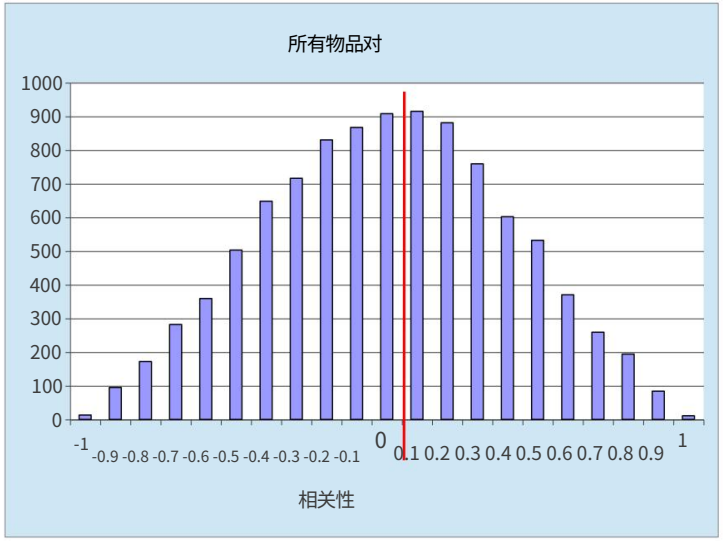
基于支持的修剪

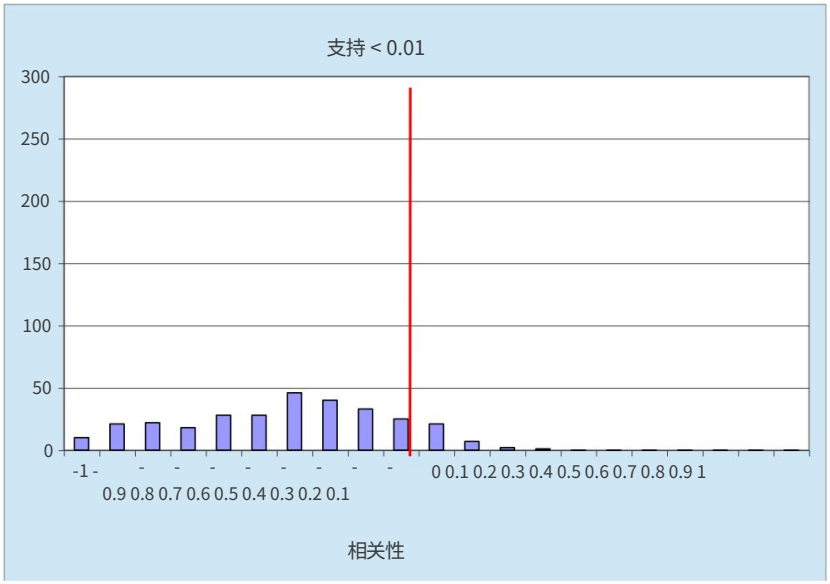
- 大多数关联规则挖掘算法使用支持度量来修剪规则和项集

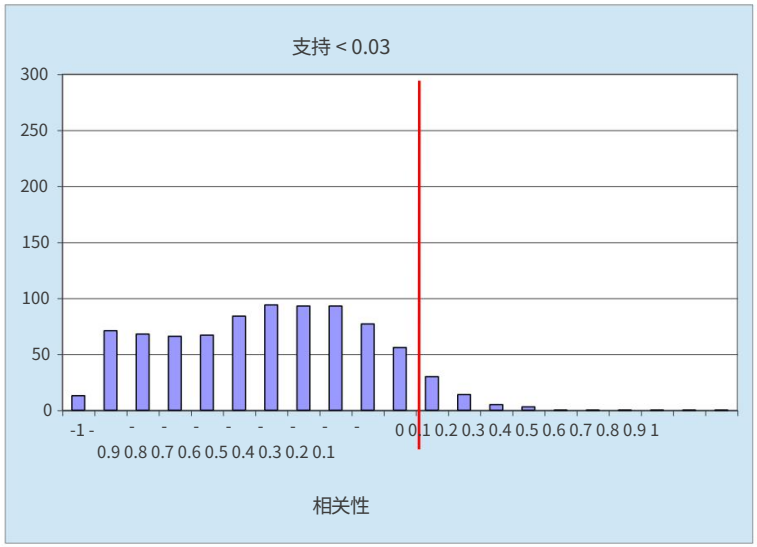
·研究支持修剪对相关性的影响
项集

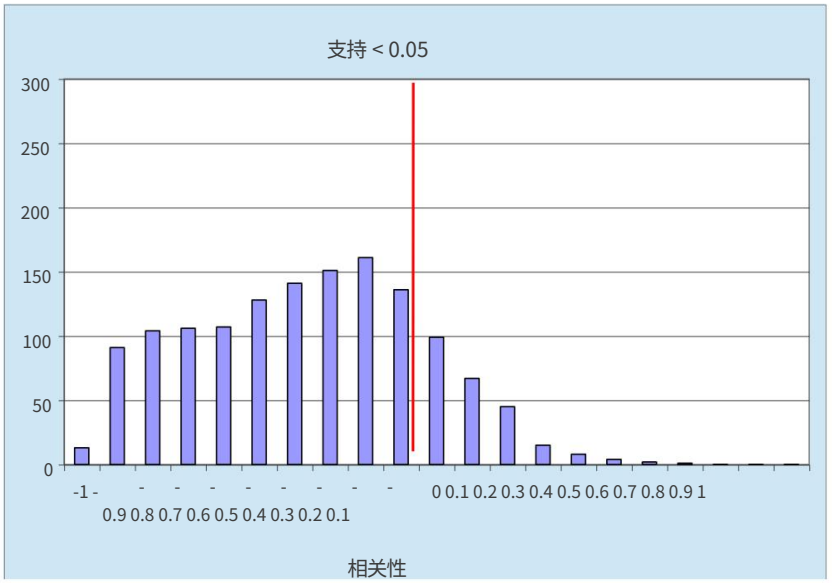
- -生成 10,000 个随机列联表
- -计算每个表的支持和成对相关性
- -应用基于支持的修剪并检查删除的表

基于支持的剪枝效果









基于支持的剪枝消除了大部分负相关项集

主观兴趣度测量

- 客观度量： -基于从
  数据计算的统计数据的排名模式-例如,21 种关联度量（支持、置信度、拉普拉斯、基尼、互信息、Jaccard 等）。

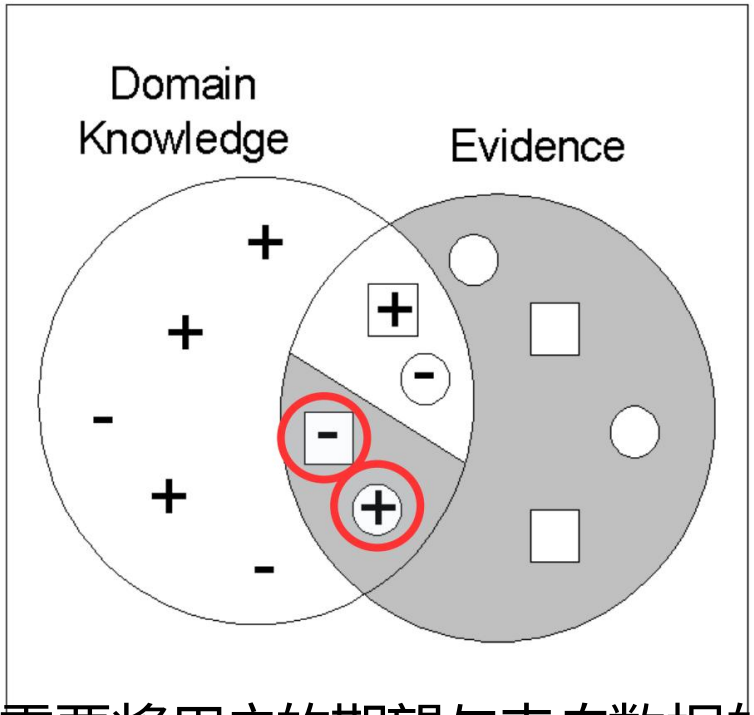
·主观衡量：

-根据用户的解释对模式进行排名

·如果模式与用户的期望相矛盾,那么它在主观上是有趣的
（Silberschatz 和 Tuzhilin） ·如果模式是可操作的,那么它在主观上是有趣的
(银宝&涂之林)

意外带来的趣味

·需要对用户的期望建模（领域知识）



+预计频繁出现的模式

-预计不常见的模式

发现频繁出现的模式

发现不常见的模式

-预期模式

+意想不到的模式

- 需要将用户的期望与来自数据的证据结合起来（即提取的模式）

关联规则申请

·市场篮子分析

营销与零售。例如，频繁项集提供有关“购买此商品的其他客户也购买了 X”的信息

·探索性数据分析

在非常大（=许多事务）、高维（=许多项目）数据中查找相关性

·入侵检测

支持度低但提升度非常高的规则

·构建基于规则的分类器

类关联规则 (CAR)