



数据挖掘与商务分析

第4讲 关联规则分析

主讲教师：肖升生



课程导入：关联分析实例

商品已成功加入购物车!



数据挖掘导论 完整版 Introduction to Data Mining(图灵出品)

颜色: 数据挖掘导论完... / 数量: 1

[查看商品详情](#)

[去购物车结算 >](#)

购买了该商品的用户还购买了



数据挖掘导论 完整版

¥49.00

[加入购物车](#)



鸟哥的Linux私房菜 基础学习篇 第四版

¥96.40

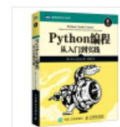
[加入购物车](#)



包邮 数据挖掘: 概念与技术 (原书第3版) |3683062

¥51.40

[加入购物车](#)



Python编程 从入门到实践(图灵出品)

¥72.70

[加入购物车](#)



数据挖掘 华章图书 计算机科学丛书

¥51.40

[加入购物车](#)



数学之美 第三版

¥34.50

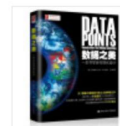
[加入购物车](#)



数据挖掘 概念与技术 (原书第3版)

¥55.30

[加入购物车](#)

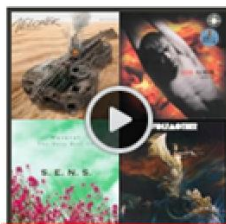


数据之美: 一本书学会可视化设计

¥69.00

[加入购物车](#)

音乐



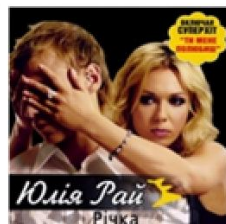
基于你的 个人音乐库

今日推荐歌单



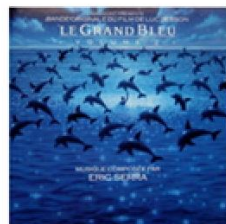
基于歌曲 Set Fire To ...

Piano Tribute to Ad...
Piano Tribute Players



基于歌曲 Pічка

Pічка
Юлія Pай



基于歌曲 The Diva D...

Le Grand Bleu volu...
Eric Serra



基于歌曲 Claude De...

Poulenc: Le bal ma...
Pascal Rogé

商品



讲授提纲

- 01** 关联规则分析基本概念
- 02** 频繁项集生成
- 03** 关联规则生成
- 04** 关联规则分析拓展
- 05** 商务案例分析



讲授提纲

01 关联规则分析基本概念

02 频繁项集生成

03 关联规则生成

04 关联规则分析拓展

05 商务案例分析



基本概念介绍

- 关联规则挖掘 (Association Rule Mining): 从海量的数据中挖掘隐藏在数据间相互联系的一种方法
- 数据挖掘中最为活跃的分支之一
- 用规则来表述发现的联系:

数据挖掘导论->{数学之美、Python 编程}

- 关联关系具备的两个特点:

频繁性 & 规则性



关联规则挖掘与商务决策

- 物品摆放和商场布局
- 交叉销售和促销
- 用户分类及个性化服务
- 医生决策支持助手
- ...



基本概念

■ 购物篮数据

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

■ 交易(Transaction): 一次购买的物品记录

■ 项集(Item): 购买物品的集合

- 1-项集 {Milk},
- 2-项集 {Bread, Milk},
- k-项集 {...}

■ 规则(Rule): $A \rightarrow B$

- $A \cap B \in \Phi$



基本概念

■ 支持度计数 ($\sigma()$)

- 项集发生的频次
- $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$

■ 支持度 (Support)

- 项集发生的频率
- $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$

■ 频繁项集 (Frequent Itemset)

- 超过某一个阈值(i.e., min-support)的项集
- 如果 $\text{min-support} = 0.3$, 则 $\{\text{Milk, Bread, Diaper}\}$ 为频繁项集

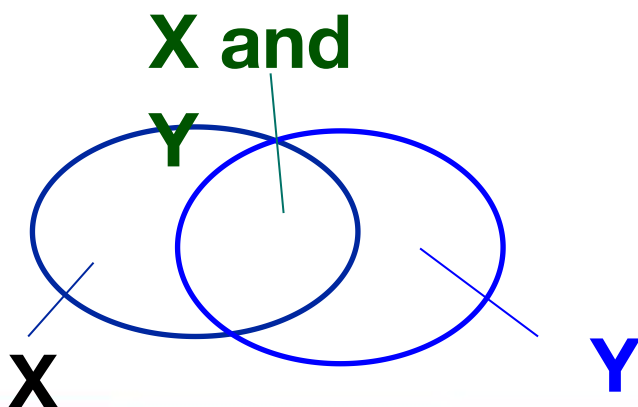
<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke



基本概念

- 规则: $X \rightarrow Y$
- 规则的支持度(Support)
 $s(X \rightarrow Y) = \sigma(\{X, Y\})$
- 规则的置信度(Confidence)

$$C(X \rightarrow Y) = \frac{\sigma(\{X, Y\})}{\sigma(X)}$$



<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

例子:

$\{\text{Milk, Diaper}\} \rightarrow \text{Beer}$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$



基本概念

■ 感兴趣的规则

- 发生频繁
- 可信度高

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

■ 关联规则挖掘

- 找出所有类似 $X \rightarrow Y$ 的规则
- 支持度 $\geq \text{min-support}$ (发生频繁)
- 置信度 $\geq \text{min-confidence}$ (可信度高)



关联规则挖掘

■ 蛮力法

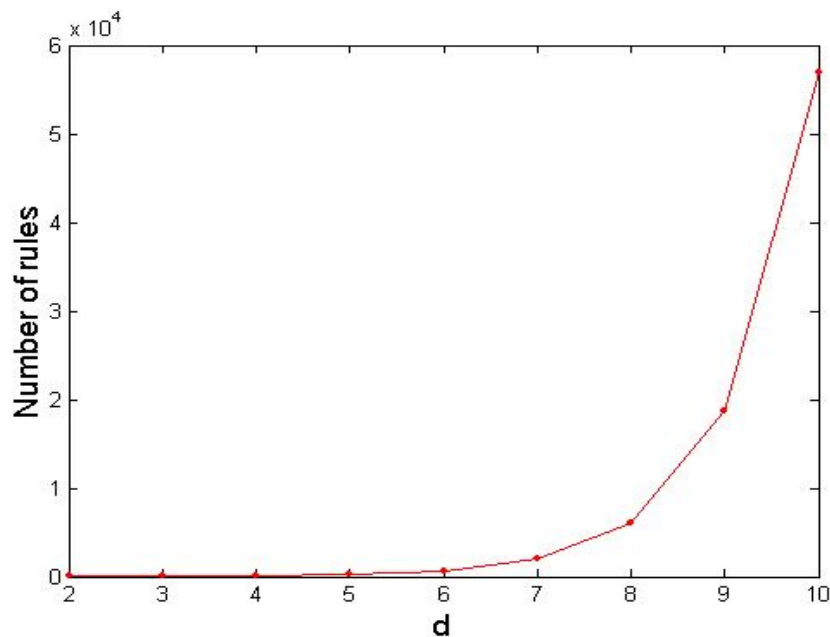
- 列出所有规则: $X \rightarrow Y$
- 计算支持度/置信度
- 筛选规则

例如: 给定 d 个物品

$$R = \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right]$$
$$= 3^d - 2^{d+1} + 1$$

如果 $d=6$, 候选规则数: 602

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke





关联规则挖掘

■ 常用的关联规则挖掘算法

- Apriori
- FP-growth
- ...

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

■ 整体解决思路

- 寻找频繁的项集
- 产生规则



讲授提纲

01 关联规则分析基本概念

02 频繁项集生成

03 关联规则生成

04 关联规则分析拓展

05 商务案例分析



频繁项集的生成

■ 给定购物篮数据集和 min-support

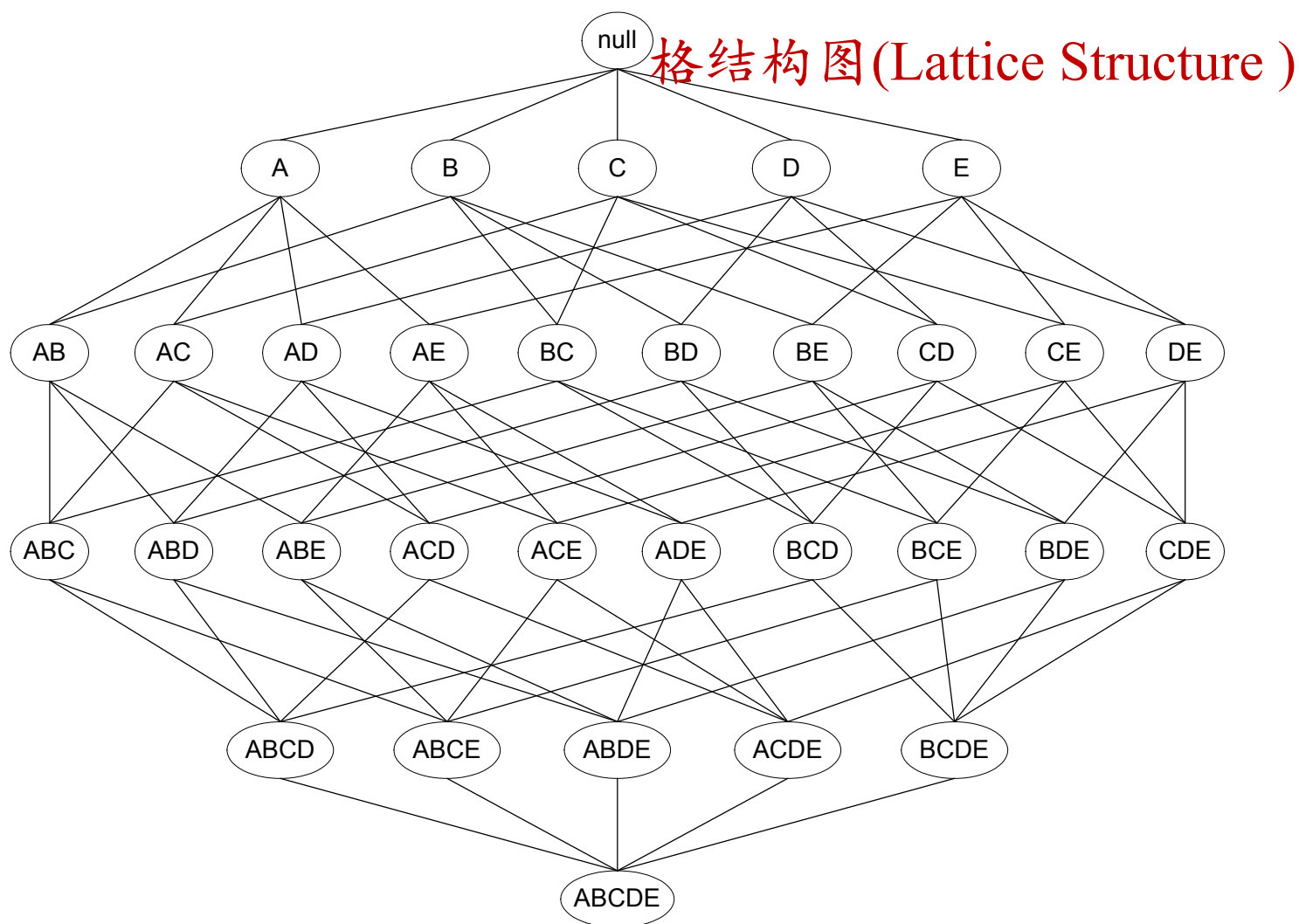
<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

■ 找出所有的频繁项集 (支持度 $\geq$ min-support)

- {Bread}, {Milk}, ...
- {Bread, Milk}, {Bread, Diaper}, ...
-



频繁项集的生成

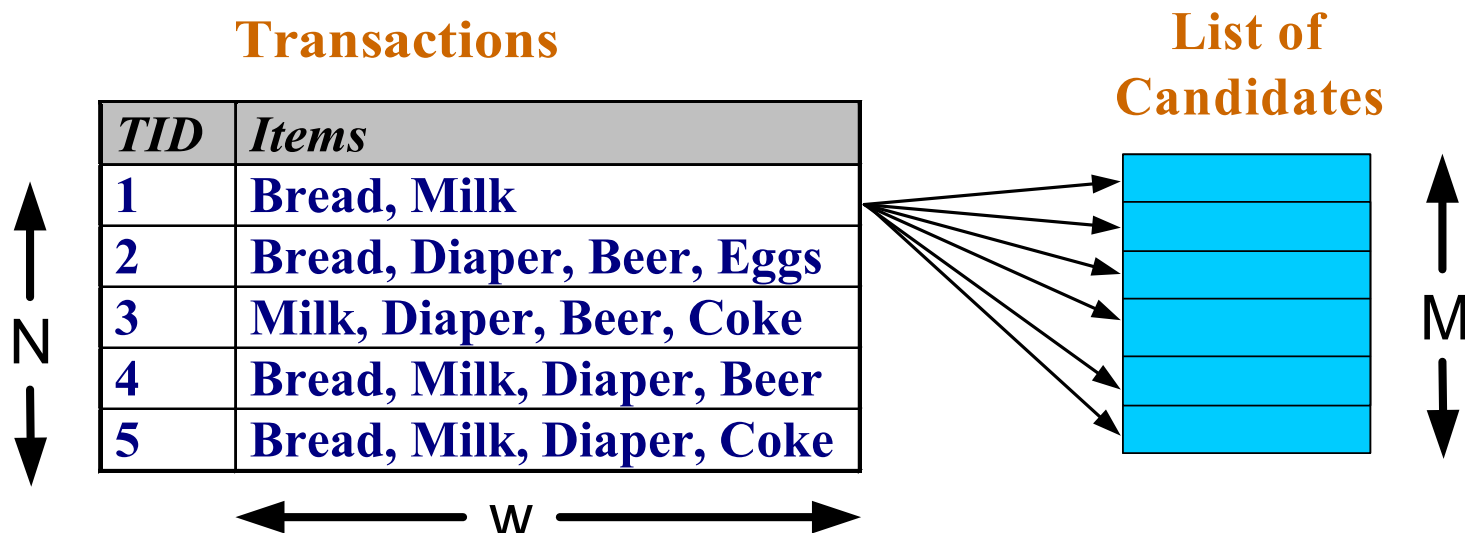


d items, $2^d - 1$ 候选项集



频繁项集的生成

■ 频繁项集的寻找和比对过程



■ 复杂度: $O(NMW)$

- $M=2^d-1$
- 复杂度惊人



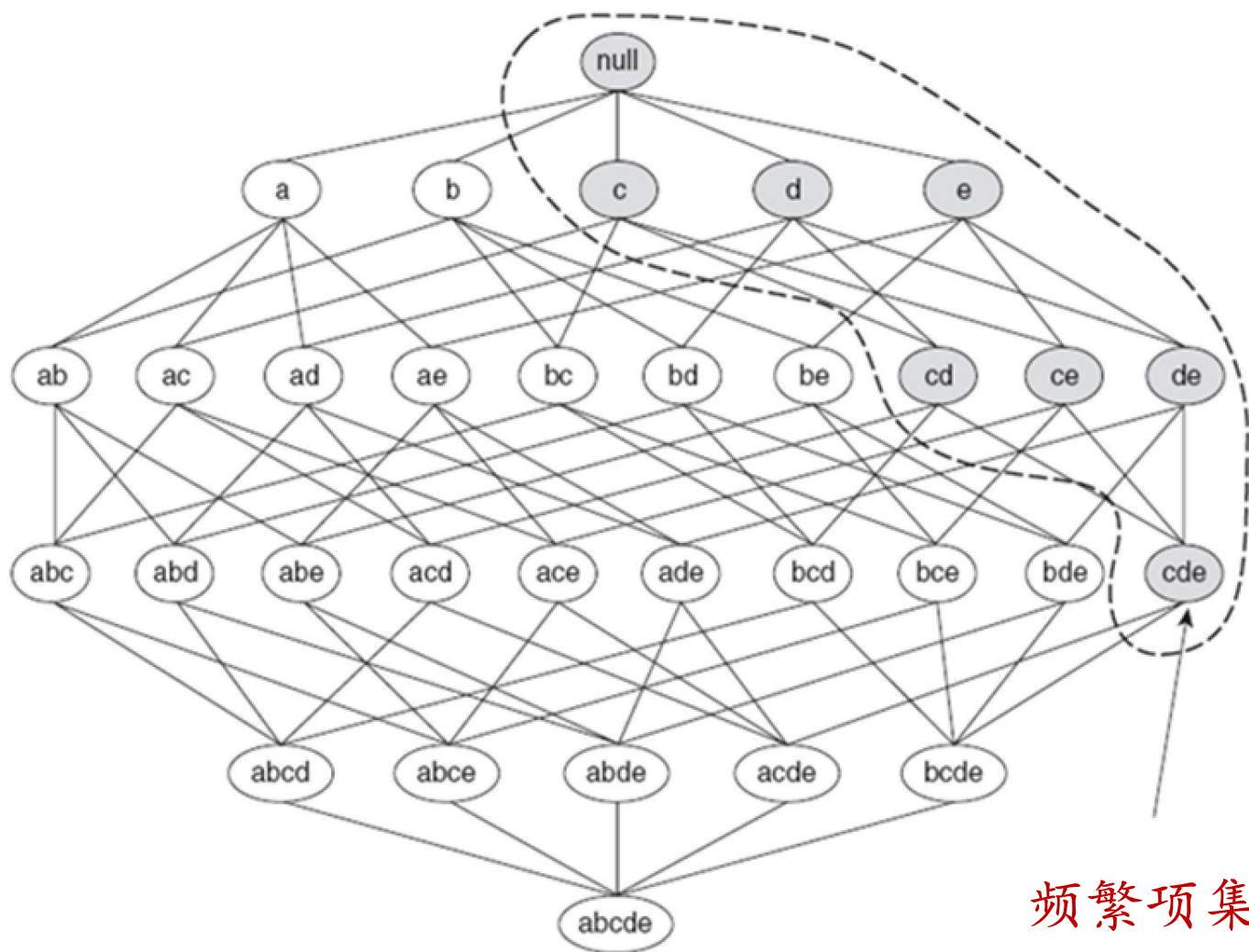
Apriori 算法与先验原理

- Apriori: 数据挖掘领域经典的十大算法之一
- 先验原理: 如果一个项集是频繁项集, 则它的所有子集也一定是频繁的
- 推理: 如果一个项集是非频繁项集, 则它的所有超集也一定是非频繁的

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$



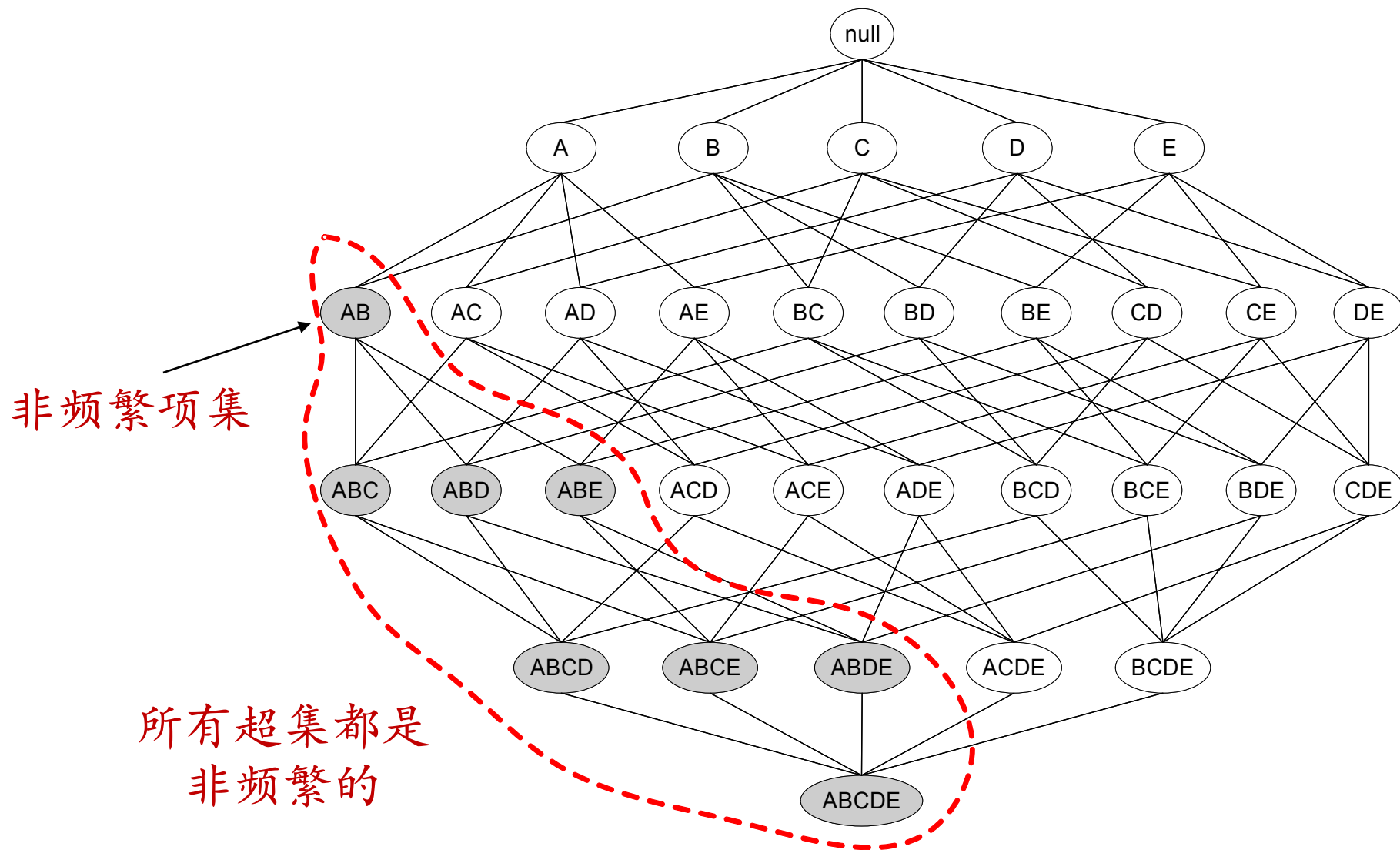
先验原理与频繁项集生成



频繁项集



先验原理与频繁项集生成





Apriori 算法示例

Minimum Support = 3

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)



Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3



Item set	Count
{Bread,Milk,Diaper}	3

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



Apriori 算法过程

■ 基于Apriori 算法的频繁项集生成:

- 列出候选1-项集 (C_1)
- 根据支持度筛选频繁1-项集(F_1)
- 生成候选2-项集 (C_2)
- 根据支持度筛选频繁2-项集(F_2)
- ...
- 生成候选k-项集 (C_k)
- 根据支持度筛选频繁k-项集(F_k)
- ...

连接步

剪枝步



Apriori 算法：连接步

■ 连接步任务

根据已有的频繁 k -项集(L_k), 生成候选的 C_{k+1} 项集

■ 连接策略

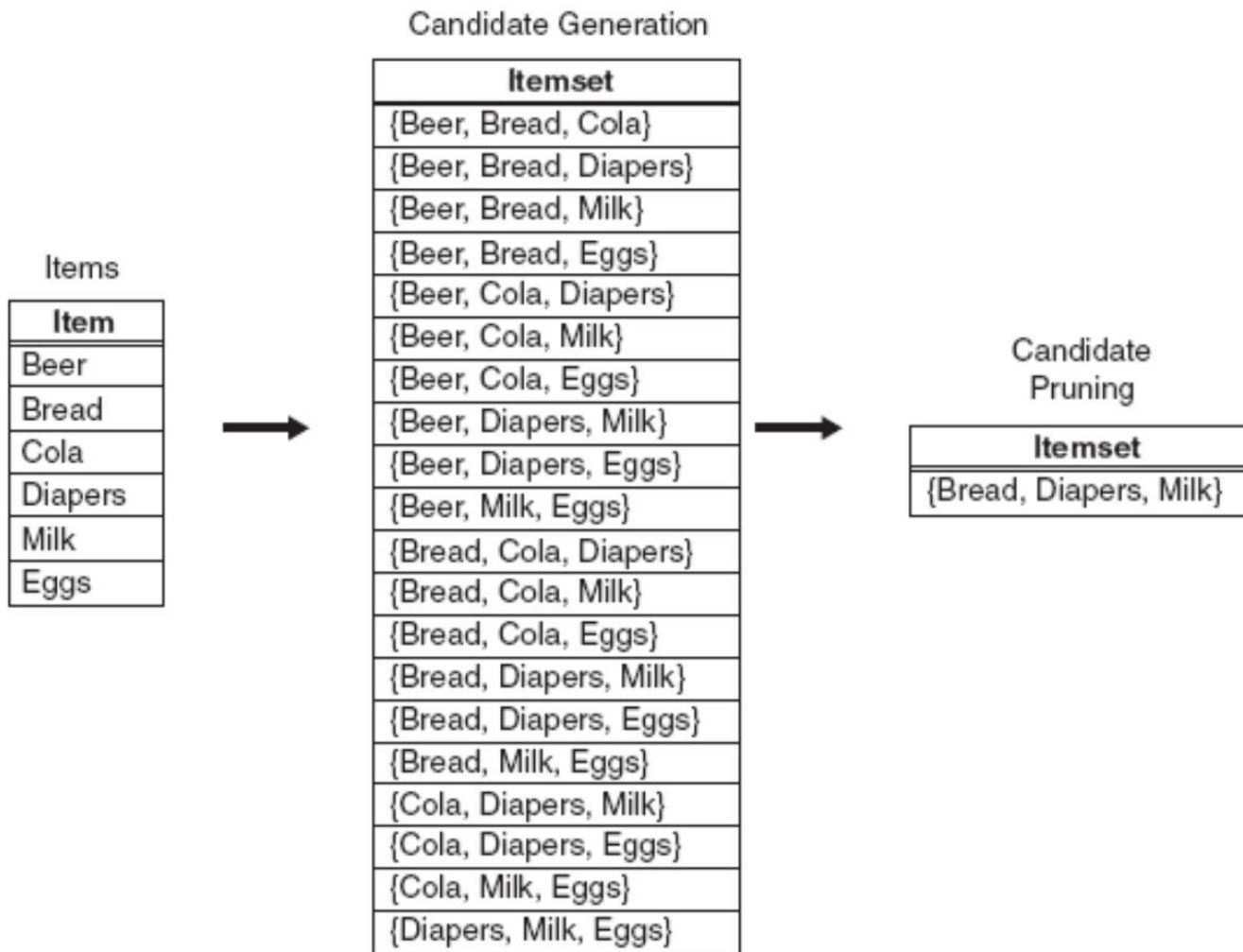
- 策略1: 蛮力法
- 策略2: $L_k \times L_1 \rightarrow C_{k+1}$
- 策略3: $L_k \times L_K \rightarrow C_{k+1}$

■ 连接要求

- 高效: 避免产生不必要的候选项集;
- 完备: 所有频繁的 $k+1$ 项集不能遗漏;

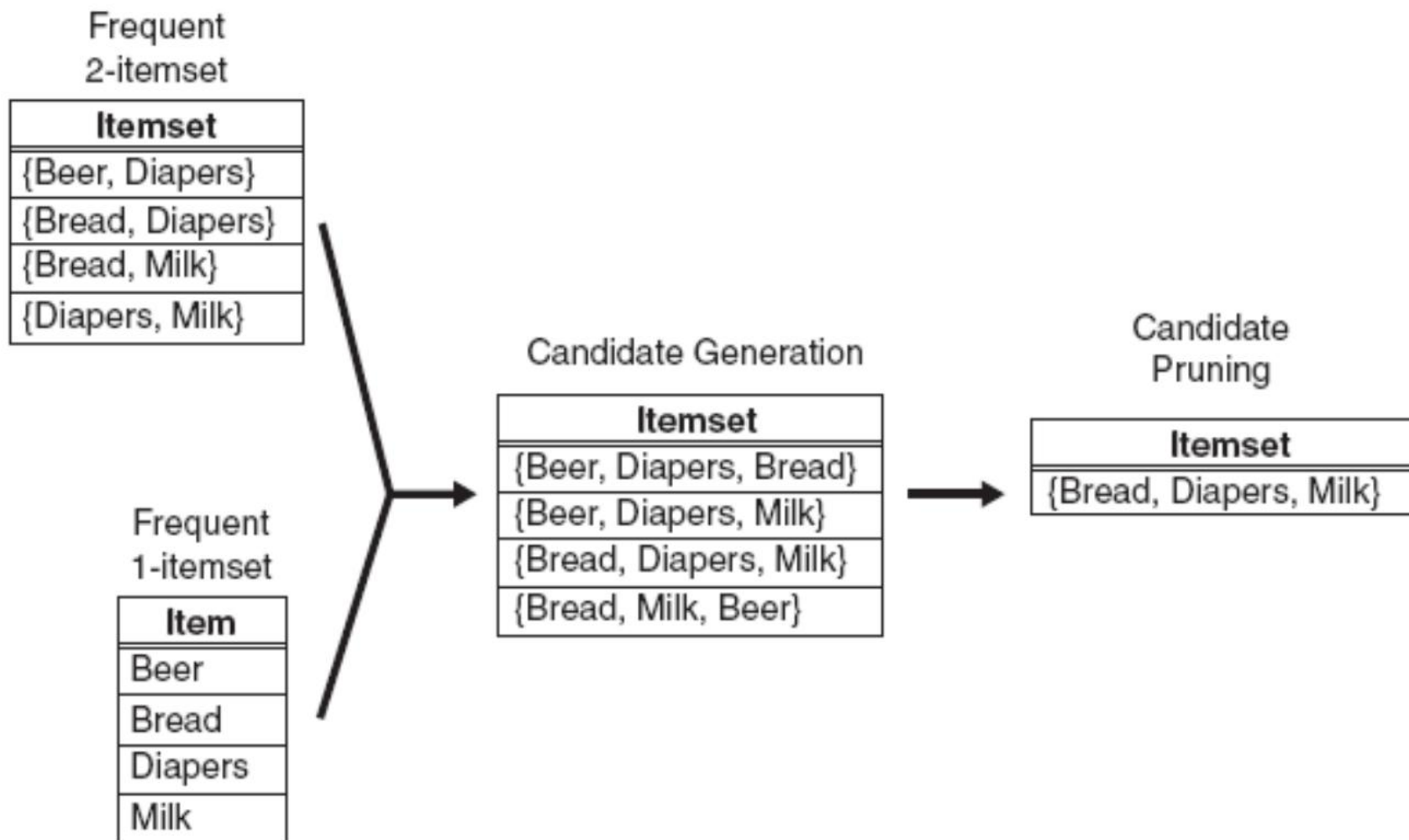


连接步策略1：蛮力法



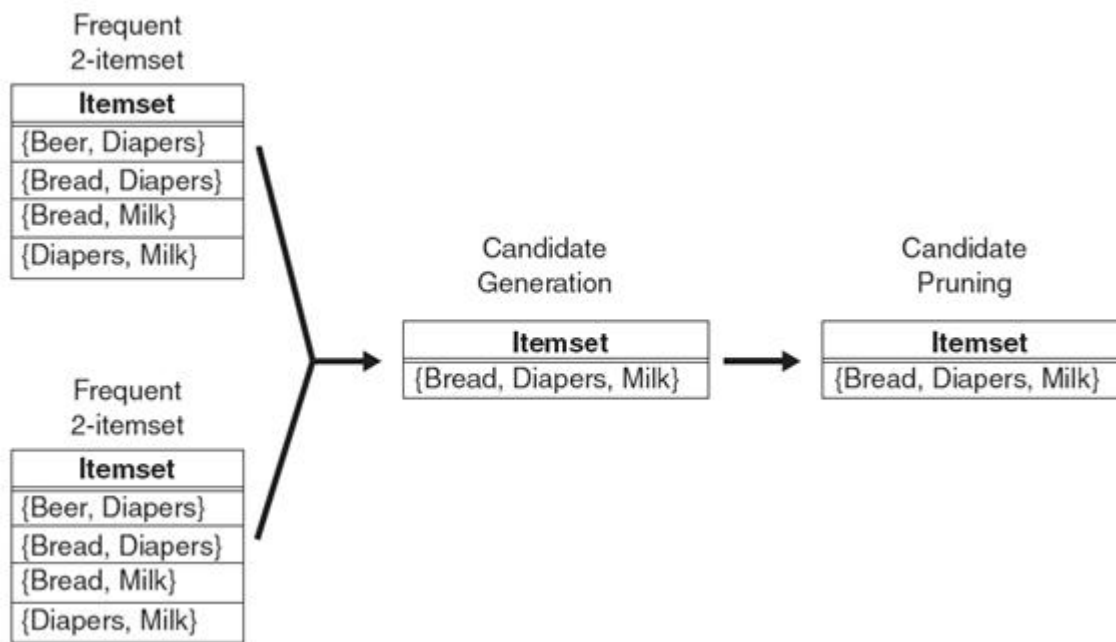


连接步策略2: $L_k \times L_1 \rightarrow C_{k+1}$





连接步策略3: $L_k \times L_k \rightarrow C_{k+1}$



■ Apriori算法中采用的策略: $L_k \times L_k \rightarrow C_{k+1}$



Apriori 算法示例

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Minimum Support = 3

$L_1 = \{\text{Bread, Milk, Beer, Diaper}\}$

$L_2 = \{\{\text{Bread, Milk}\}, \{\text{Bread, Diaper}\}, \{\text{Milk, Diaper}\}, \{\text{Beer, Diaper}\}\}$

$L_3 = \{\text{Bread, Milk, Diaper}\}$

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1



Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3



Item set	Count
{Bread,Milk,Diaper}	3



讲授提纲

01 关联规则分析基本概念

02 频繁项集生成

03 关联规则生成

04 关联规则分析拓展

05 商务案例分析



关联规则生成

■ 给定频繁项集 L ，生成规则：

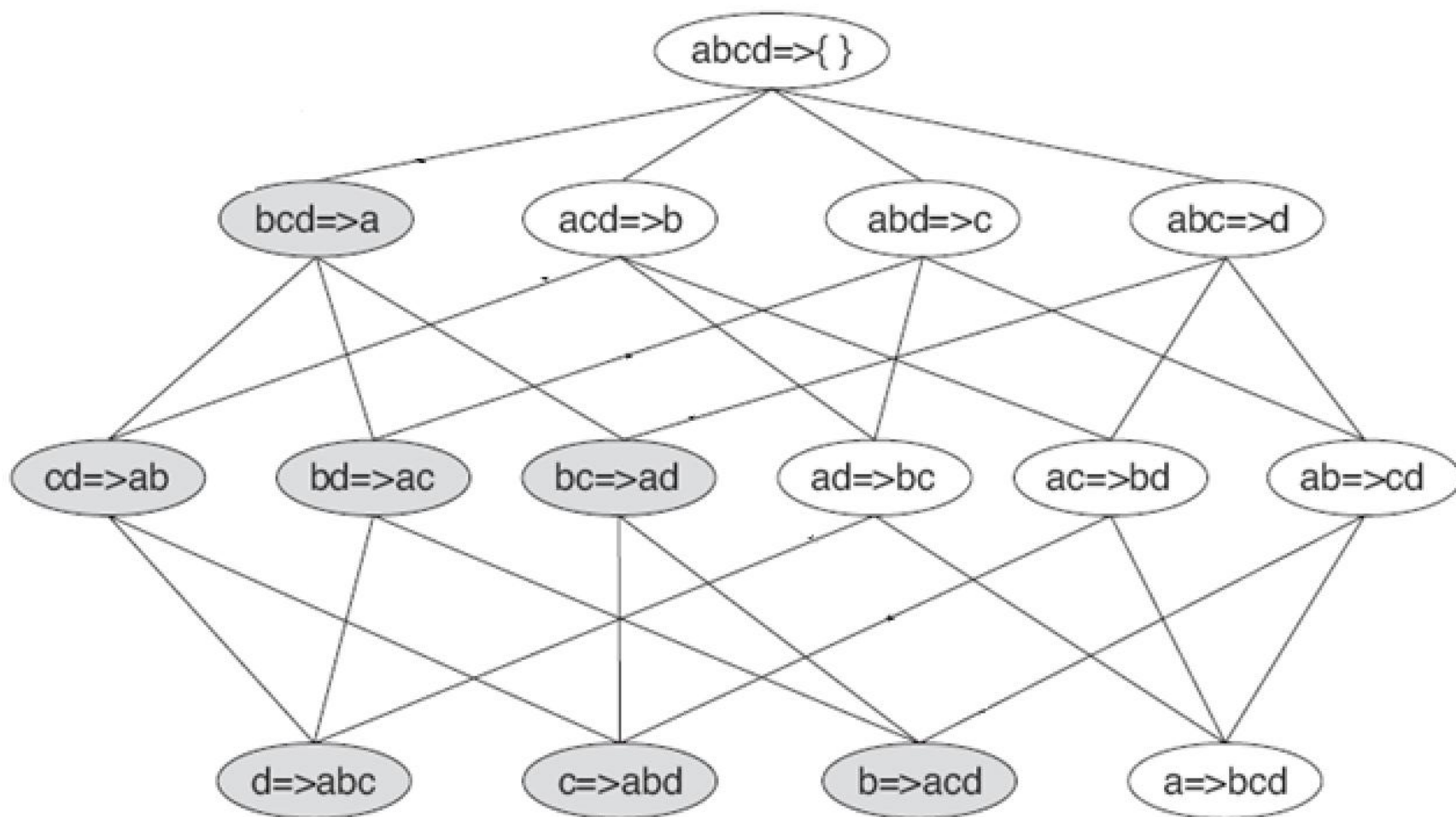
- 找出 L 的所有非空子集 X ($X \subset L$)
- 构建规则： $X \rightarrow L - X$
- 如果 $|L| = k$ ，则候选规则有 $2^k - 2$

■ 感兴趣的规则：

- $\text{Support}(X \rightarrow L - X) \geq \text{min-support}$
- $\text{Confidence}(X \rightarrow L - X) \geq \text{min-confidence}$
- 计算量仍然很大



关联规则生成





关联规则生成

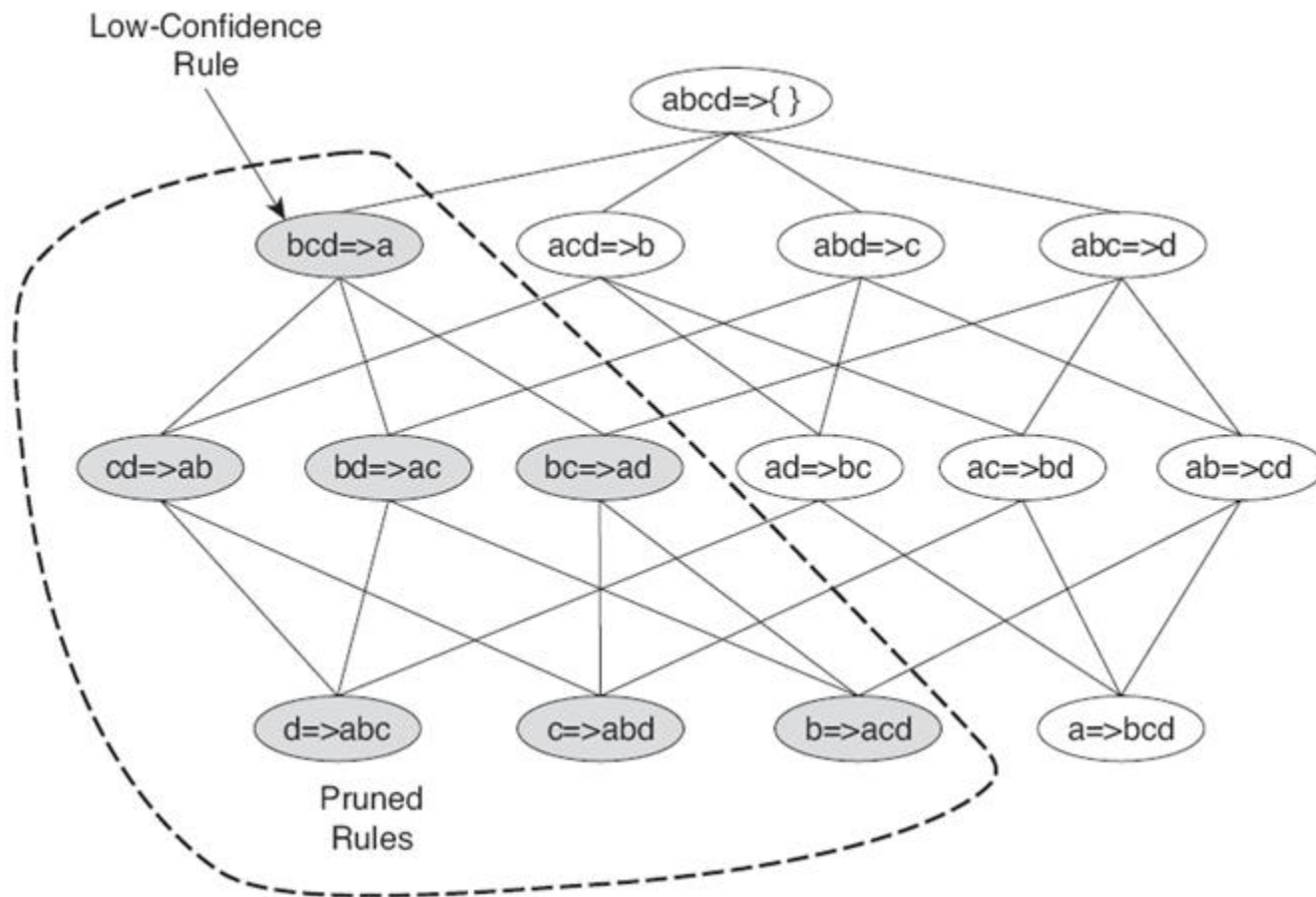
■ 给定基于置信度的剪枝原理(L为频繁项集):

如果规则 $X \rightarrow L-X$ 不满足置信度阈值, 则形如 $X' \rightarrow L-X'$ 的规则一定也不满足上述阈值($X' \subseteq X$)

■ 证明:

- 规则 $X \rightarrow L-X$ 的置信度为 $\frac{\sigma(\{L\})}{\sigma(X)}$
- 规则 $X' \rightarrow L-X'$ 的置信度为 $\frac{\sigma(\{L\})}{\sigma(X')}$
- 由于 $\sigma(X) \leq \sigma(X')$
- 所以: $\frac{\sigma(\{L\})}{\sigma(X)} \geq \frac{\sigma(\{L\})}{\sigma(X')}$

关联规则生成





关联规则挖掘

■ 第一步：寻找频繁项集

- Apriori 算法
- 连接步
- 剪枝步

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

■ 第二步：生成规则：

- 基于频繁项集生成候选规则
- 对候选规则进行剪枝



讲授提纲

01 关联规则分析基本概念

02 频繁项集生成

03 关联规则生成

04 关联规则分析拓展

05 商务案例分析



关联分析的误区1

■ 拥有高置信度度的规则并不一定有意义

	Coffee	<u>Coffee</u>	
Tea	15	5	20
<u>Tea</u>	75	5	80
	90	10	100

规则: Tea $\rightarrow$ Coffee

- Confidence (Tea $\rightarrow$ Coffee) = 0.75
- Confidence ($\overline{\text{Tea}} \rightarrow$ Coffee) = 0.9375
- $P(\text{Coffee}) = 0.9$



关联分析的误区2

■ 虚假关联模式与辛普森悖论

	健身器=YES	健身器=NO	
HDTV=YES	99	81	180
HDTV=NO	54	66	120
	153	147	300

- 规则1: (买HDTV=YES) $\rightarrow$ (买健身器=YES)
- 规则2: (买HDTV=NO) $\rightarrow$ (买健身器=YES)
- Confidence (规则1) = $99/180=0.55$
- Confidence (规则2) = $54/120=0.45$
- 结论: 规则1比规则2更可信!



关联分析的误区2

■ 辛普森悖论

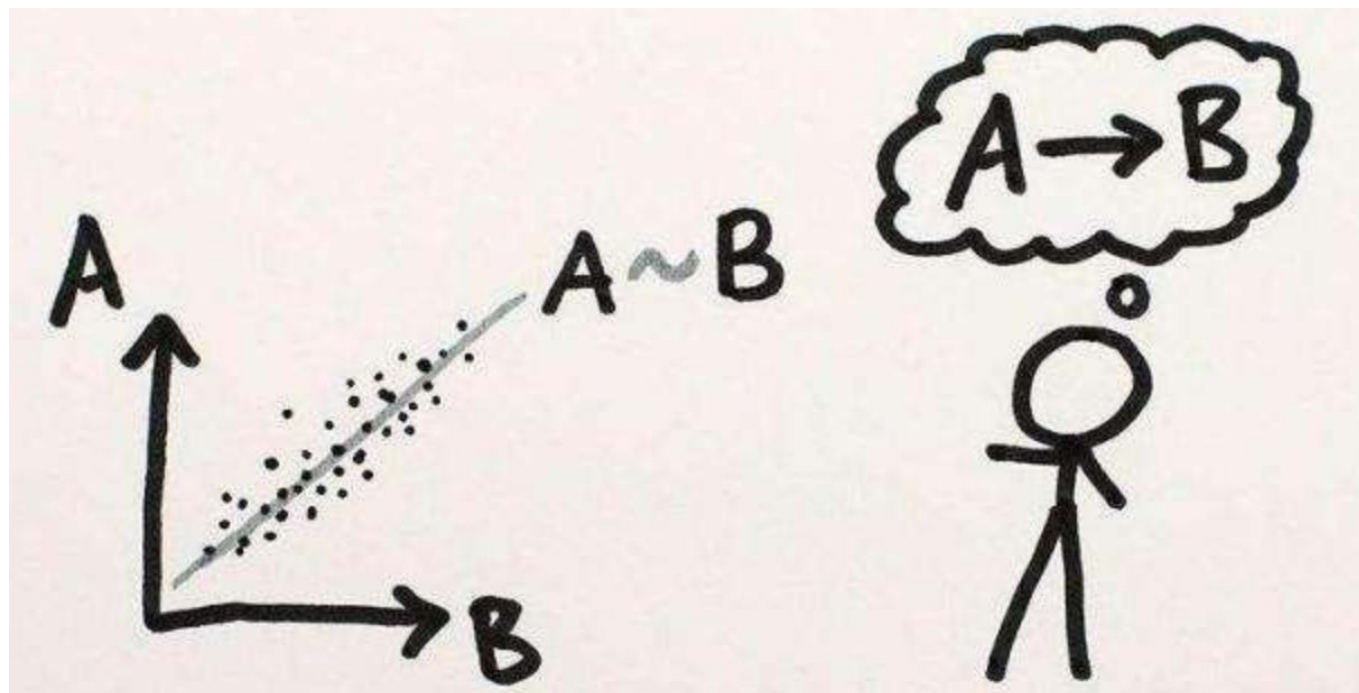
		健身器=YES	健身器=NO	
大学生消费群体	HDTV=YES	1	9	10
	HDTV=NO	4	30	34
在职白领消费群体	HDTV=YES	98	72	170
	HDTV=NO	50	36	86

- 规则1: (买HDTV=YES) $\rightarrow$ (买健身器=YES)
- 规则2: (买HDTV=NO) $\rightarrow$ (买健身器=YES)
- 大学生群体: $C(\text{规则1})=1/10=0.100$ $C(\text{规则2})=4/34=0.118$
- 在职白领: $C(\text{规则1})=98/170=0.576$ $C(\text{规则2})=50/86=0.581$
- 结论: 规则2 比规则1 更可信



关联分析的误区3

- 关联规则： 相关关系 v.s 因果关系





关联分析的延申

■ 产生频繁项集的其他方法

- FP-growth

■ 序列模式挖掘

- 不带时间约束的事件序列挖掘
- 带时间约束事件序列挖掘
- 基因序列挖掘

■ 子图模式挖掘（复杂实体挖掘）

- 蛋白质结构
- 网络拓扑结构



讲授提纲

01 关联规则分析基本概念

02 频繁项集生成

03 关联规则生成

04 关联规则分析拓展

05 商务案例分析



关联规则分析案例

■ 某餐馆的交易点餐数据如下：

序列	时间	订单号	菜品id	菜品名称
1	2014/8/21	101	18491	健康麦香包
2	2014/8/21	101	8693	香煎葱油饼
3	2014/8/21	101	8705	翡翠蒸香茜饺
4	2014/8/21	102	8842	菜心粒咸骨粥
5	2014/8/21	102	7794	养颜红枣糕
6	2014/8/21	103	8842	金丝燕麦包
7	2014/8/21	103	8693	三丝炒河粉
...	...	...	...	...



关联规则分析案例

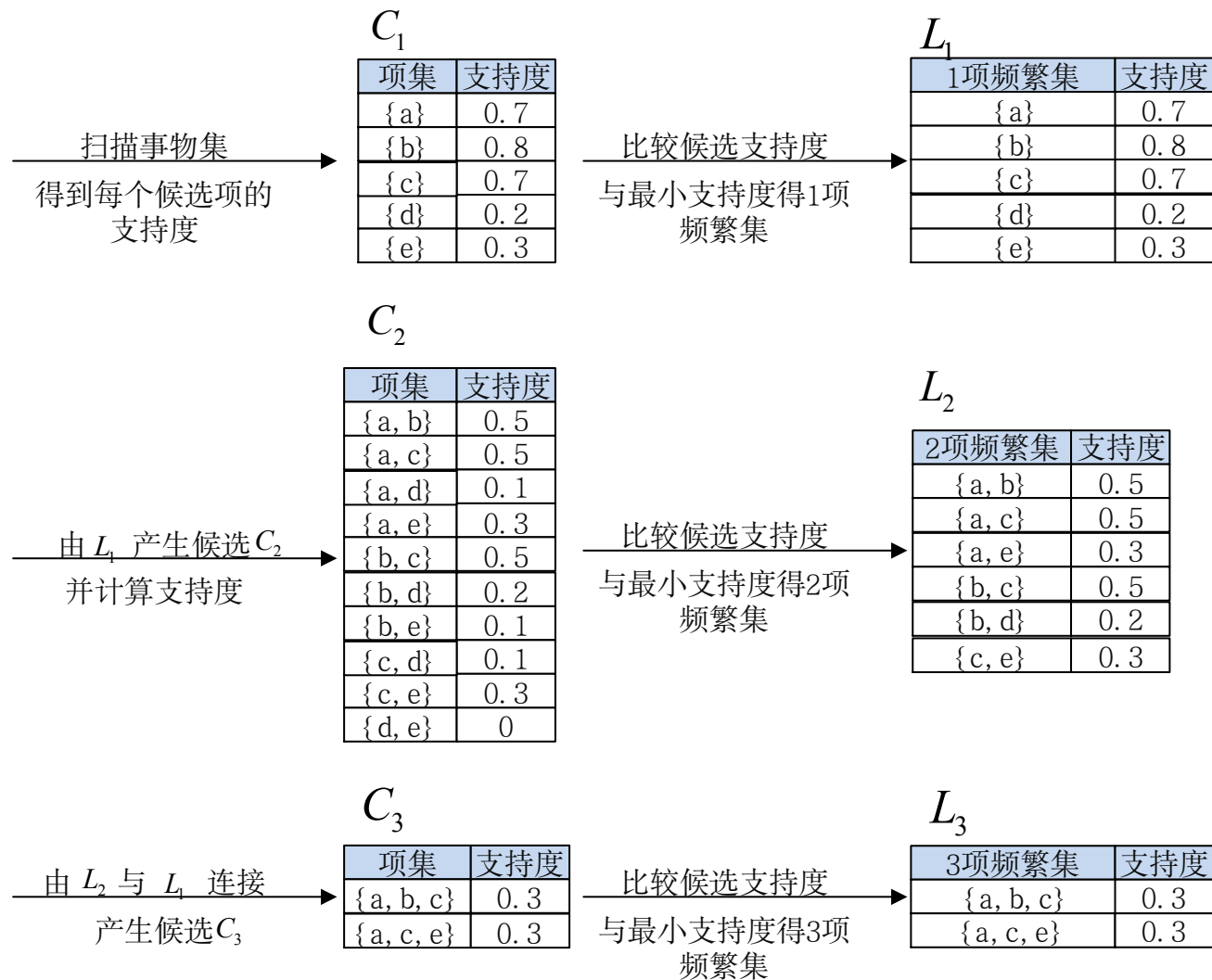
- 首先将上表中的事务数据（一种特殊类型的记录数据）整理成关联规则模型所需的数据结构。

订单号	菜品id	菜品id
1	18491, 8693, 8705	a, c, e
2	8842, 7794	b, d
3	8842, 8693	b, c
4	18491, 8842, 8693, 7794	a, b, c, d
5	18491, 8842	a, b
6	8842, 8693	b, c
7	18491, 8842	a, b
8	18491, 8842, 8693, 8705	a, b, c, e
9	18491, 8842, 8693	a, b, c
10	18491, 8693	a, c, e



关联规则分析案例

■ 设支持度为0.2，即支持度计数为2，算法过程如下图：





关联规则分析案例

■ 由频繁集产生关联规则:

Rule	(Support, Confidence)	
		$a, b \rightarrow c$ (30%, 60%)
$a \rightarrow b$	(50%, 71.4286%)	$a, c \rightarrow b$ (30%, 60%)
$b \rightarrow a$	(50%, 62.5%)	$b, c \rightarrow a$ (30%, 60%)
$a \rightarrow c$	(50%, 71.4286%)	$e \rightarrow a, c$ (30%, 100%)
$c \rightarrow a$	(50%, 71.4286%)	$a, c \rightarrow e$ (30%, 60%)
$b \rightarrow c$	(50%, 62.5%)	$a, e \rightarrow c$ (30%, 100%)
$c \rightarrow b$	(50%, 71.4286%)	$c, e \rightarrow a$ (30%, 100%)
$e \rightarrow a$	(30%, 100%)	$d \rightarrow b$ (20%, 100%)
$e \rightarrow c$	(30%, 100%)	