上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

数据挖掘与商务分析

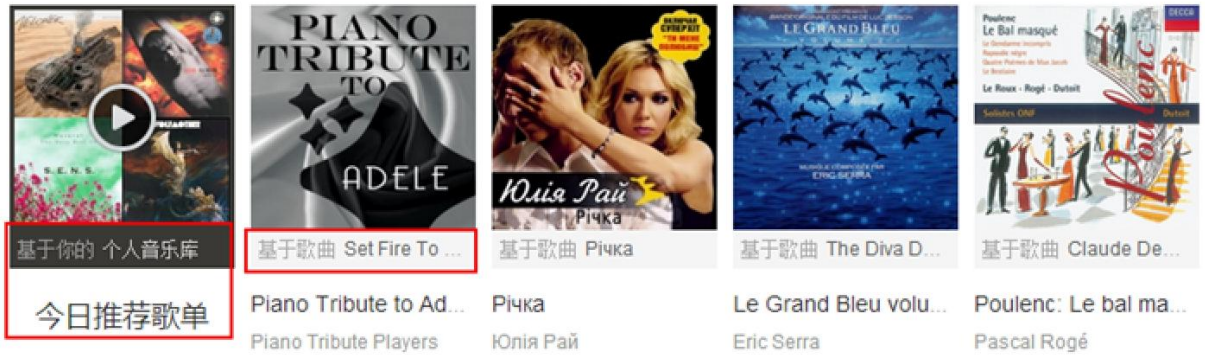
关联规则分析

主讲教师：肖升生
xiao.shengsheng@shufe.edu.cn

课程导入：关联分析实例

商品

音乐



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

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

- 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->B
  - $A \cap B \in \Phi$

基本概念

- **支持度计数 (σ())**
  - 项集发生的频次
  - σ({Milk, Bread,Diaper}) = 2
- **支持度 (Support)**
  - 项集发生的频率
  - s({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min-support = 0.3, 则{Milk, Bread, Diaper}为频繁项集

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

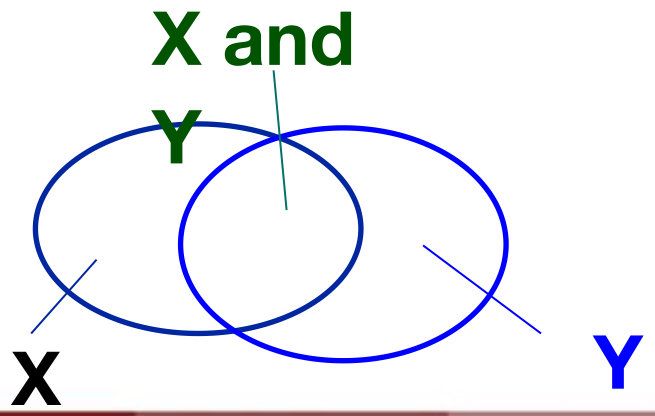
基本概念

- 规则： X -> Y
- 规则的支持度(Support)

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

- 规则的置信度(Confidence)

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



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

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

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

- **关联规则挖掘**
  - 找出所有类似 X->Y的规则
  - 支持度>= min-support (发生频繁)
  - 置信度>= min-confidence (可信度高)

关联规则挖掘

- **蛮力法**
  - 列出所有规则： X-> Y
  - 计算支持度/置信度
  - 筛选规则

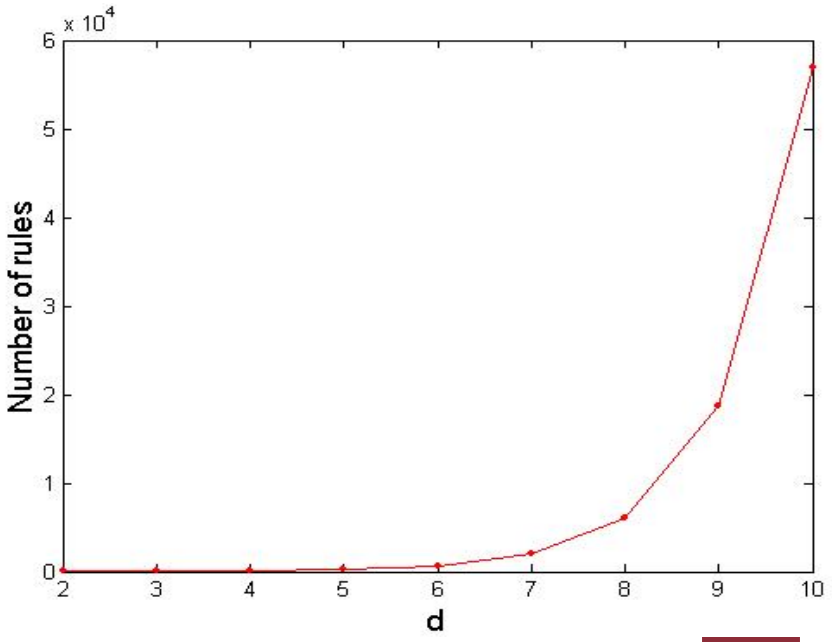
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

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



关联规则挖掘

- **常用的关联规则挖掘算法**
  - Apriori
  - FP-growth
  - …
- **整体解决思路**
  - 寻找频繁的项集
  - 产生规则

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

讲授提纲

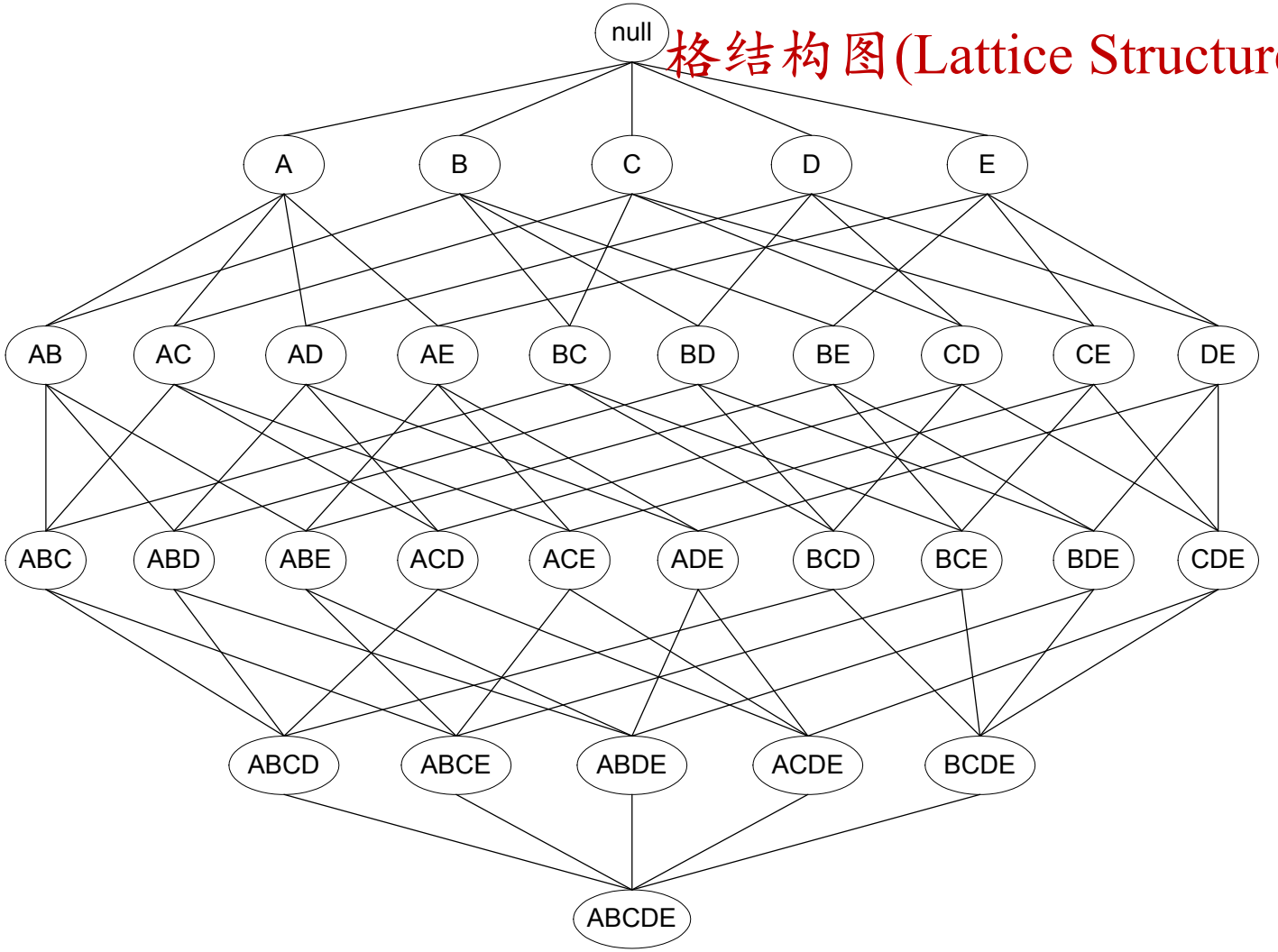
频繁项集的生成

- 给定购物篮数据集和 min-support

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

- 找出所有的频繁项集（支持度>=min-support）
  - {Bread}, {Milk}, …
  - {Bread, Milk}, {Bread, Diaper}, …
  - …..

频繁项集的生成

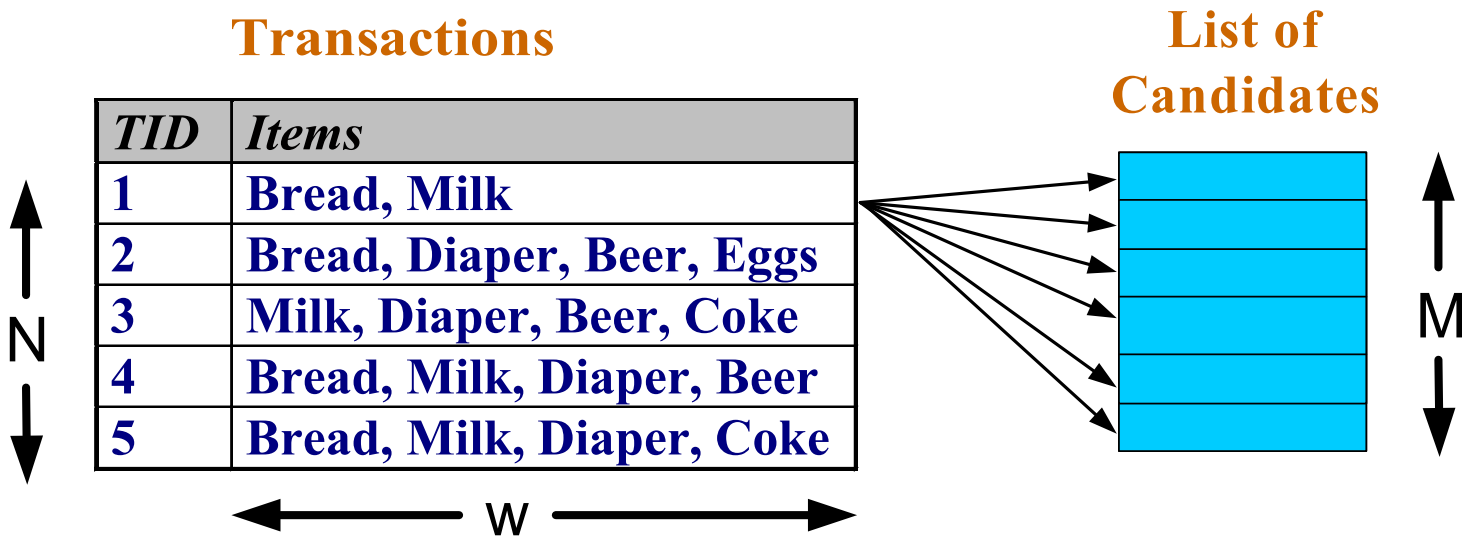


格结构图(Lattice Structure)

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



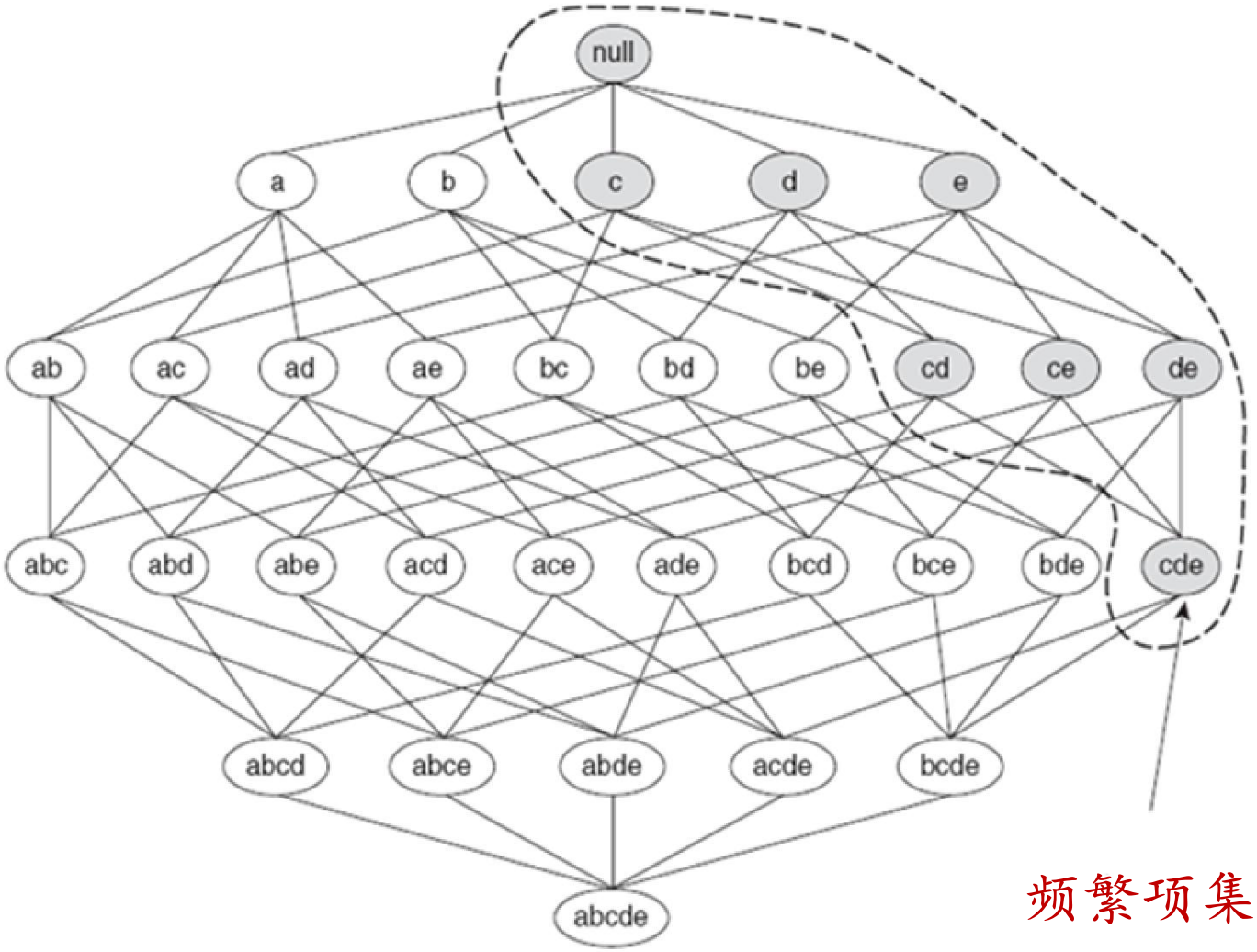
- 复杂度： O(NMW)
  - $M=2^d-1$
  - 复杂度惊人

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**

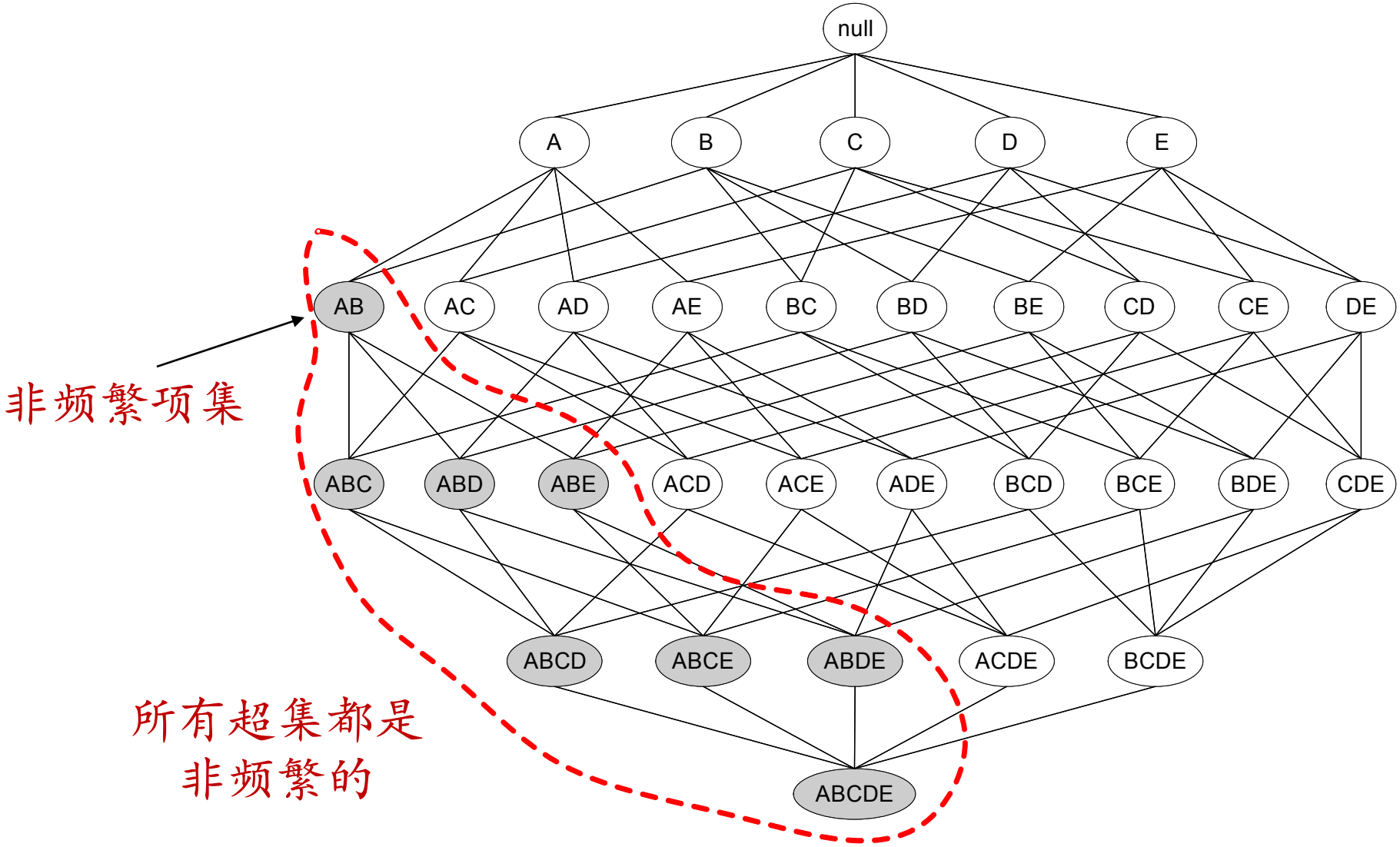
$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

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

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

Items (1-itemsets)

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

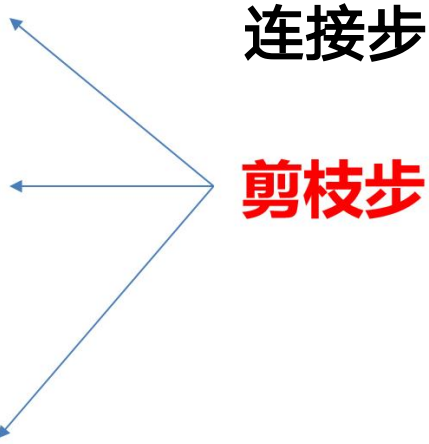
Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Itemset	Count
{Bread,Milk,Diaper}	3

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)
  - …



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 X L_1-> C_{k+1}
- 策略3： L_k X L_K-> 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

Items

Item
Beer
Bread
Cola
Diapers
Milk
Eggs

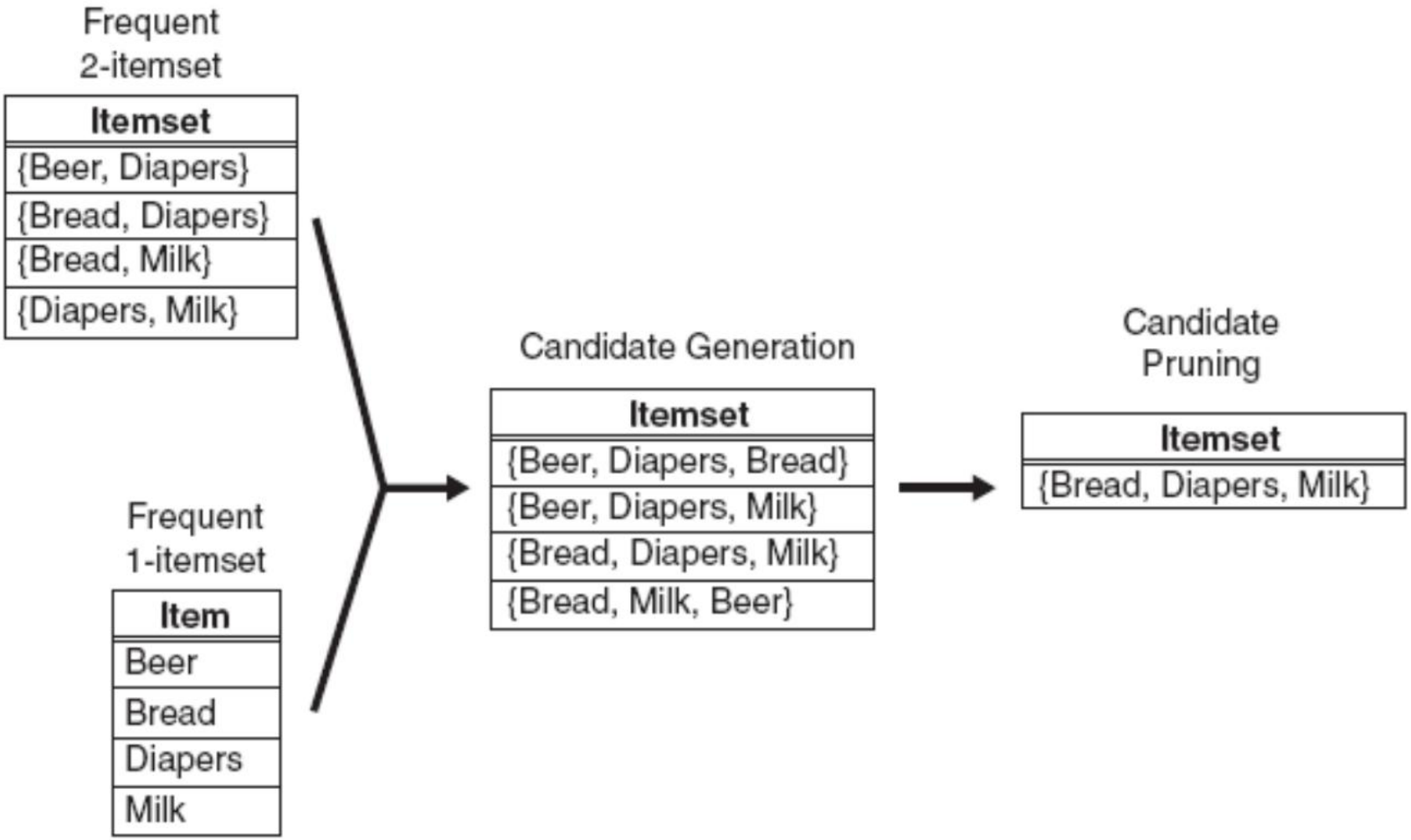
Candidate Generation

Itemset
{Beer, Bread, Cola}
{Beer, Bread, Diapers}
{Beer, Bread, Milk}
{Beer, Bread, Eggs}
{Beer, Cola, Diapers}
{Beer, Cola, Milk}
{Beer, Cola, Eggs}
{Beer, Diapers, Milk}
{Beer, Diapers, Eggs}
{Beer, Milk, Eggs}
{Bread, Cola, Diapers}
{Bread, Cola, Milk}
{Bread, Cola, Eggs}
{Bread, Diapers, Milk}
{Bread, Diapers, Eggs}
{Bread, Milk, Eggs}
{Cola, Diapers, Milk}
{Cola, Diapers, Eggs}
{Cola, Milk, Eggs}
{Diapers, Milk, Eggs}

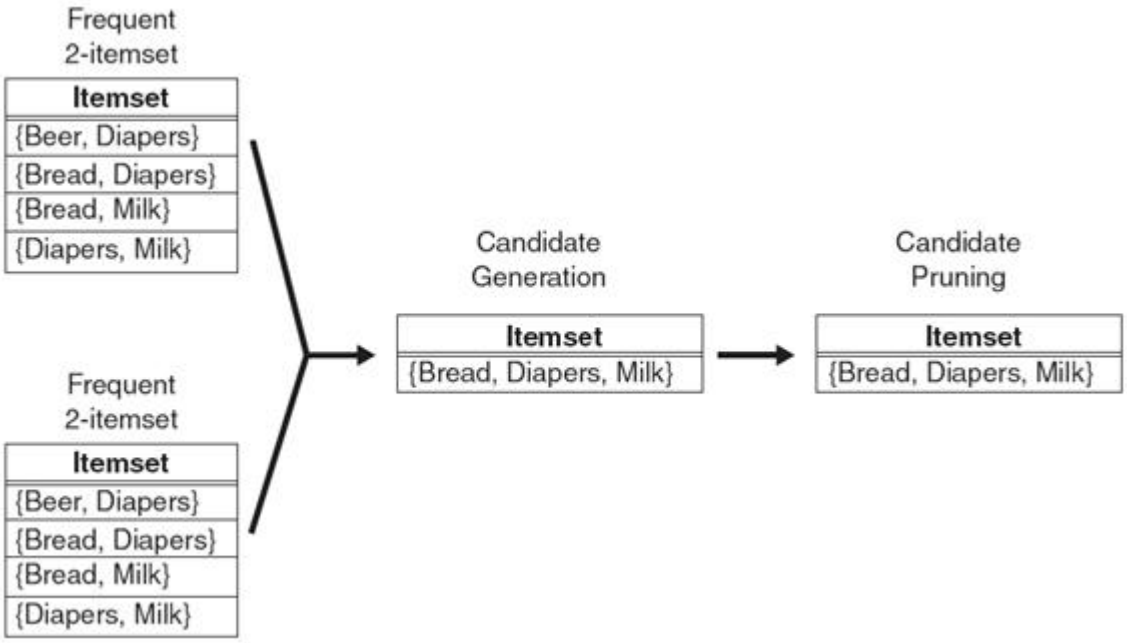
Candidate Pruning

Itemset
{Bread, Diapers, Milk}

连接步策略2： L_k X L_1-> C_{k+1}



连接步策略3： L_k X L_k-> C_{k+1}



- **Apriori算法中采用的策略：** L_k X L_k-> C_{k+1}

Apriori 算法示例

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

Minimum Support = 3

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

Itemset	Count
{Bread,Milk,Diaper}	3

L_1 ={Bread, Milk, Beer, Diaper}

L_2 ={{Bread, Milk}, {Bread, Diaper}, {Milk, Diaper}, {Beer, Diaper}}

L_3 ={Bread, Milk, Diaper}

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

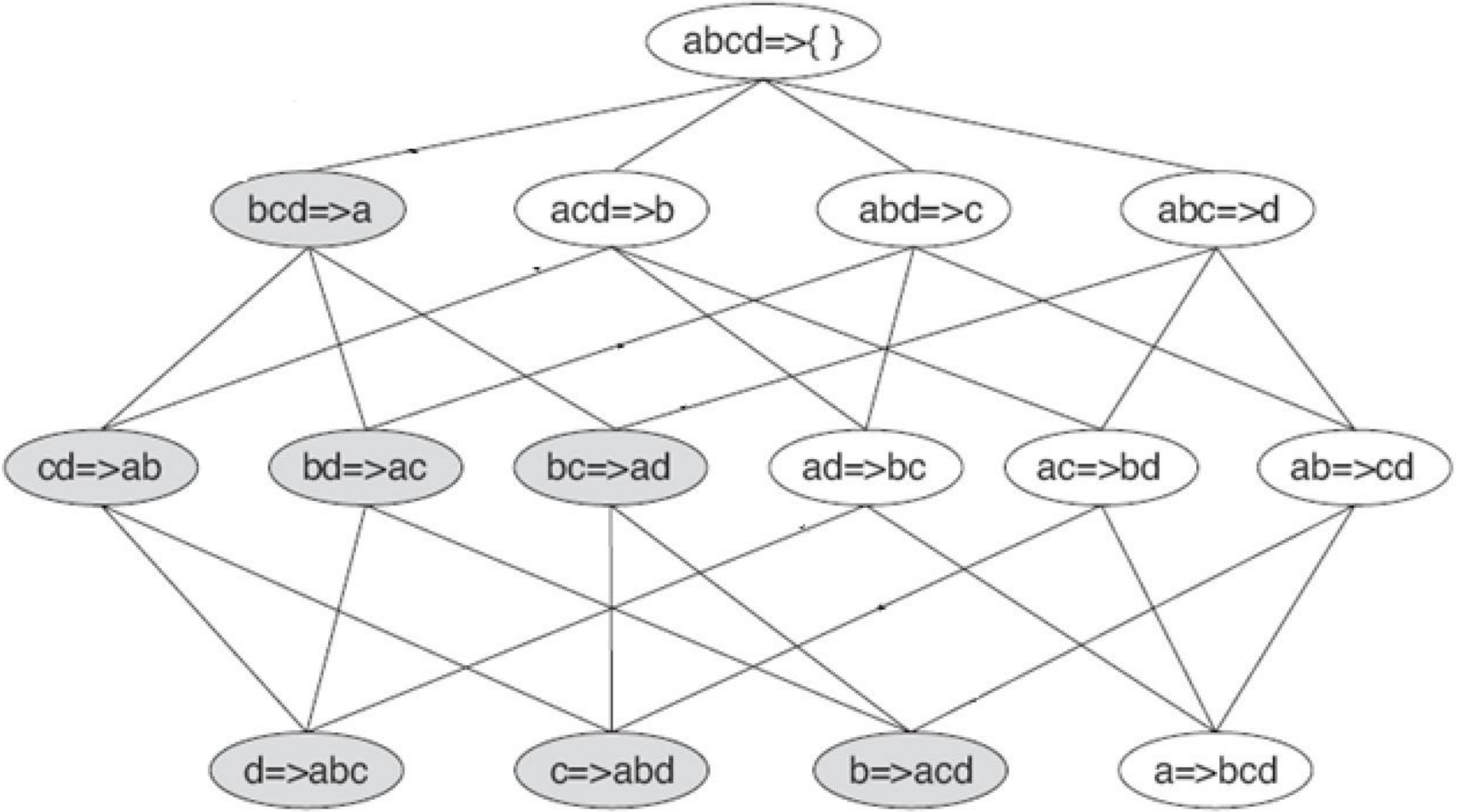
05 商务案例分析

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：**
  - Support ($X \rightarrow L - X$)≥min-support
  - Confidence ($X \rightarrow L - X$) ≥min-confidence
  - 计算量仍然很大

关联规则生成


abcd=>{ }
bcd=>a
acd=>b
abd=>c
abc=>d
cd=>ab
bd=>ac
bc=>ad
ad=>bc
ac=>bd
ab=>cd
d=>abc
c=>abd
b=>acd
a=>bcd

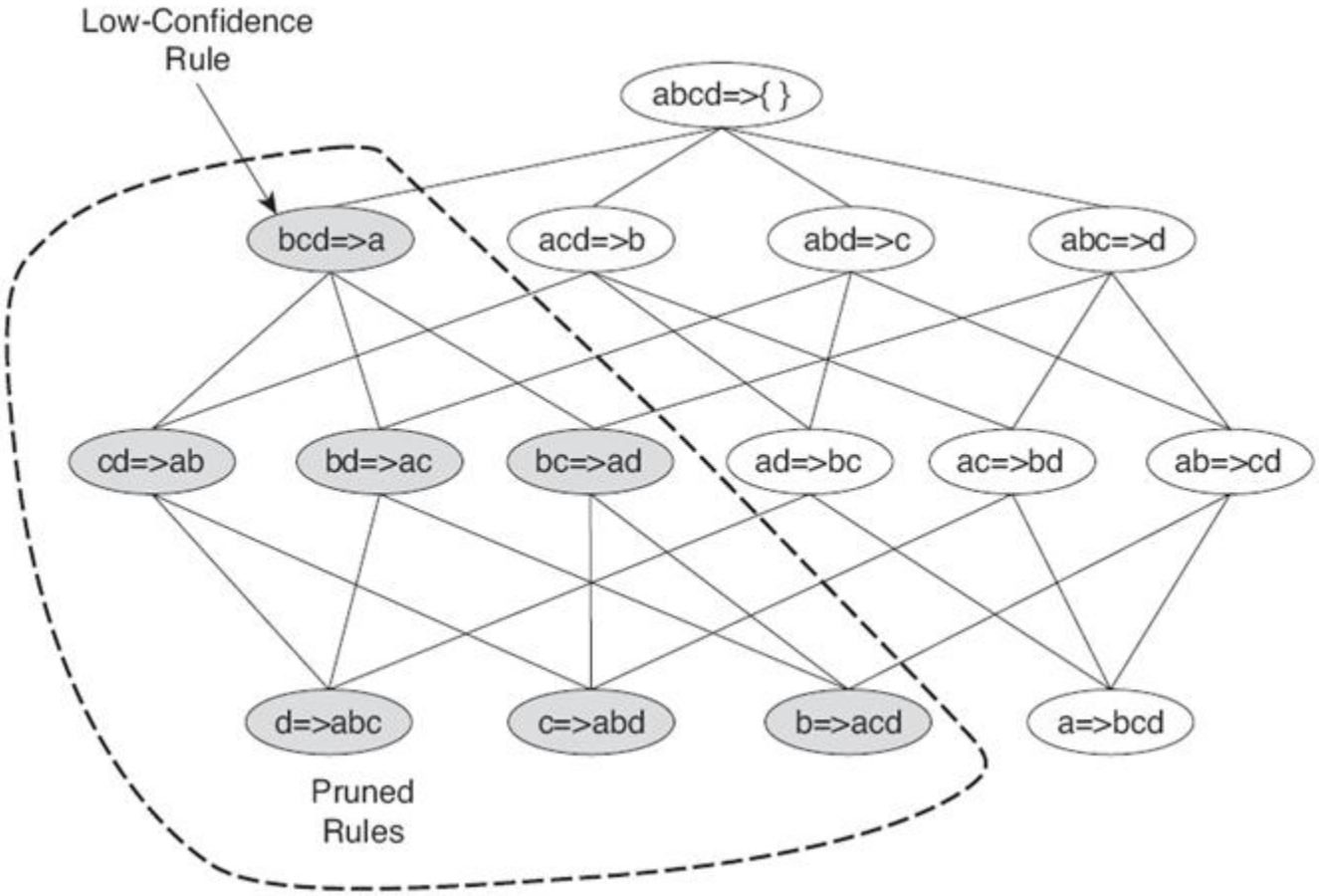
关联规则生成

- **给定基于置信度的剪枝原理(L为频繁项集):**

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

- **证明:**
  - 规则 X->L-X 的置信度为 $\frac{\sigma(\{L\})}{\sigma(X)}$
  - 规则 X'->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


Low-Confidence
Rule
abcd=>{ }
bcd=>a
acd=>b
abd=>c
abc=>d
cd=>ab
bd=>ac
bc=>ad
ad=>bc
ac=>bd
ab=>cd
d=>abc
c=>abd
b=>acd
a=>bcd
Pruned
Rules

关联规则挖掘

- **第一步：寻找频繁项集**
  - Apriori 算法
  - 连接步
  - 剪枝步

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

- **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	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	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(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）→（买健身器=YES）
- 规则2：（买HDTV=NO）→（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

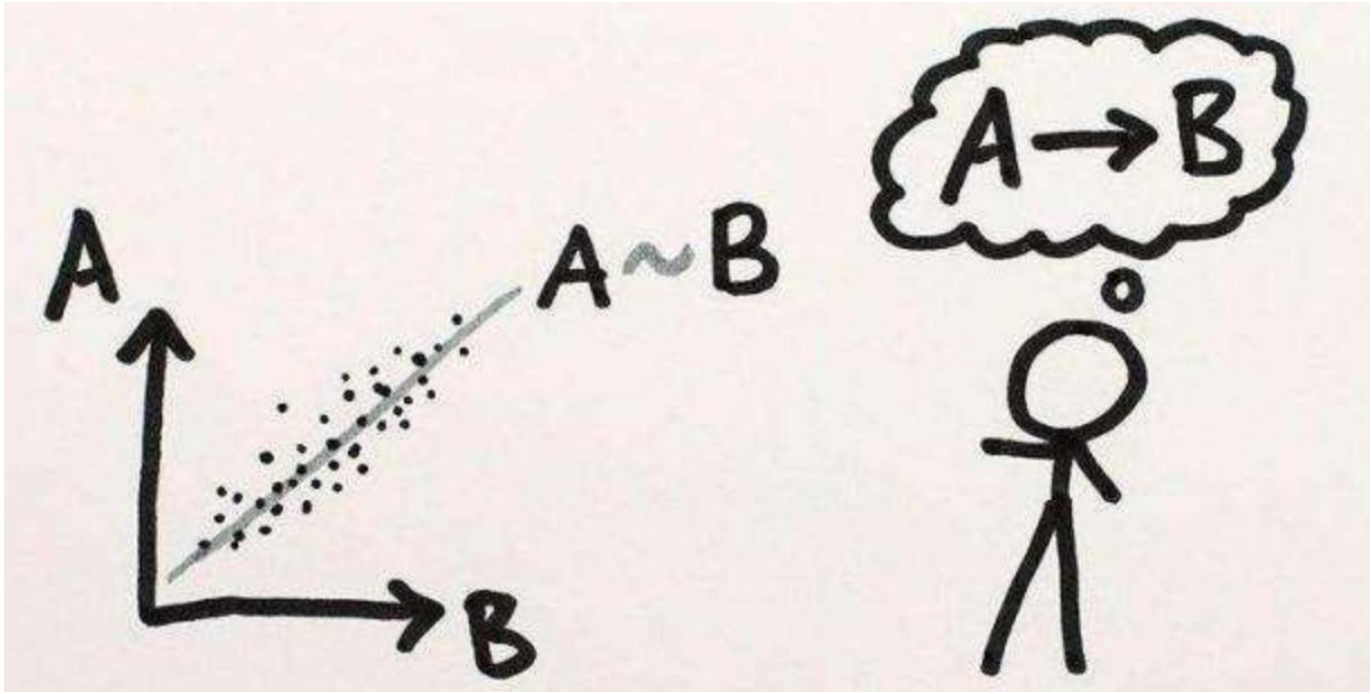
辛普森悖论

		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）→（买健身器=YES）
- 规则2：（买HDTV=NO）→（买健身器=YES）
- 大学生群体：C(规则1)=1/10=0.100　　C(规则2)= 4/34=0.118
- 在职白领：　C(规则1)=98/170=0.576　C(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
…	…	…	…	…

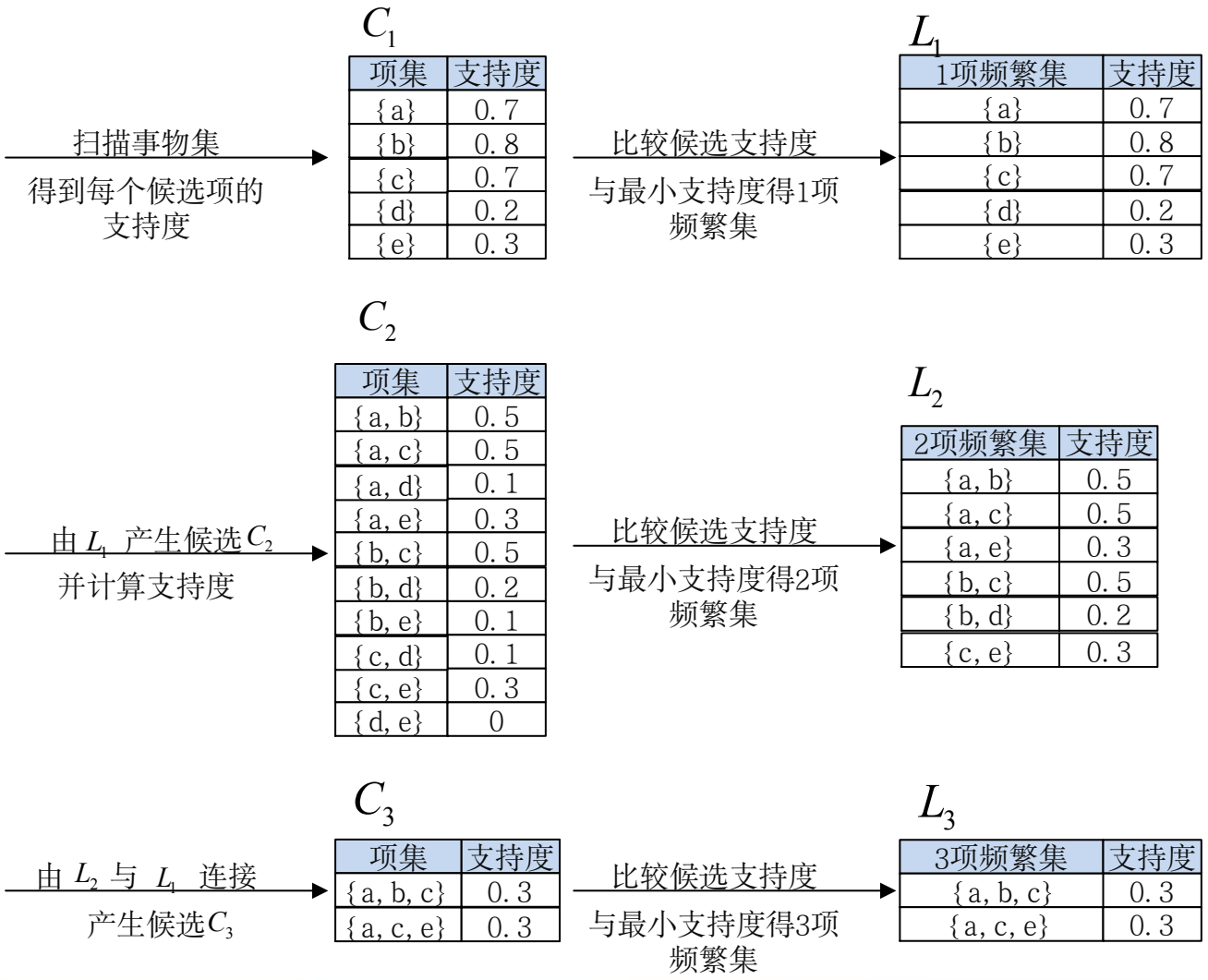
关联规则分析案例

- 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	(50%, 71.4286%)
b -> *a*	(50%, 62.5%)
a -> c	(50%, 71.4286%)
c -> *a*	(50%, 71.4286%)
b -> c	(50%, 62.5%)
c -> b	(50%, 71.4286%)
e -> *a*	(30%, 100%)
e -> c	(30%, 100%)
a,b -> c	(30%, 60%)
a,c -> b	(30%, 60%)
b,c -> *a*	(30%, 60%)
e -> *a*,c	(30%, 100%)
a,c -> e	(30%, 60%)
a,e -> c	(30%, 100%)
c,e -> *a*	(30%, 100%)
d -> b	(20%, 100%)