关联规则

麦国炫

2017/10/25

目录

关联规则的简介

“尿布与啤酒”的故事。

- 这是一个来自沃尔玛超市的真实案例，美国的沃尔玛超市对一年多的原始交易数据进行了详细的分析，得到一个意外发现：与尿布一起被购买最多的商品竟然是啤酒。

- 借助于数据仓库和关联规则，商家发现了这个隐藏在背后的事实：美国的妇女们经常会嘱咐她们的丈夫下班以后要为孩子买尿布，而30%～40%的丈夫在买完尿布之后又要顺便购买自己爱喝的啤酒。有了这个发现后，超市调整了货架的设置，把尿布和啤酒摆放在一起销售，从而大大增加了销售额。

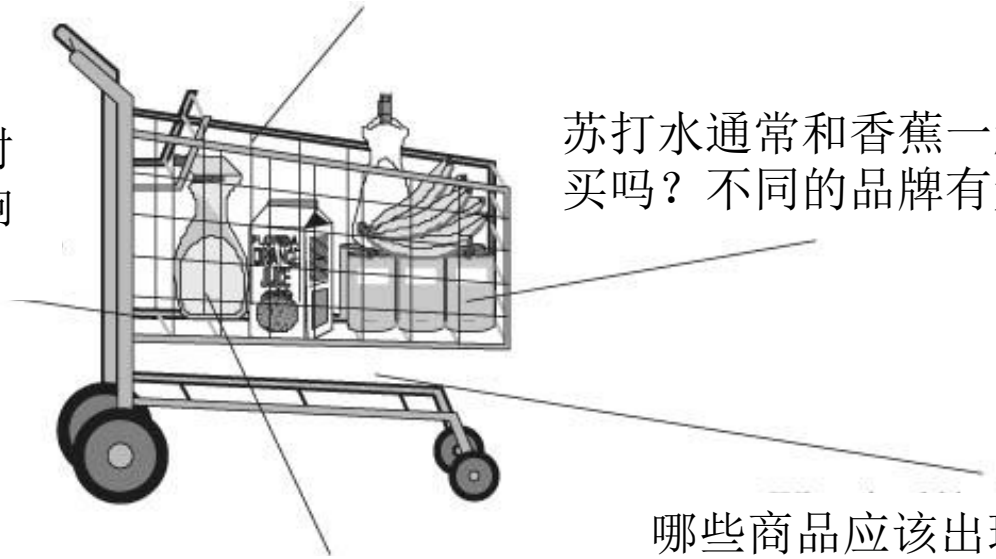
关联规则的简介

- 关联规则（Association Rules）反映一个事物与其他事物之间的相互依存性和关联性。如果两个或者多个事物之间存在一定的关联关系，那么，其中一个事物就能够通过其他事物预测到。首先被Agrawal, Imielinski 和Swami在1993年的SIGMOD会议上提出.
- 关联规则挖掘是数据挖掘中最活跃的研究方法之一。典型的关联规则发现问题是对超市中的购物篮数据（Market Basket）进行分析。通过发现顾客放入购物篮中的不同商品之间的关系来分析顾客的购买习惯。

关联规则

关联规则分析也称成为购物篮分析，最早是为了发现超市销售数据库中不同的商品之间的关联关系。找到顾客经常同时购买的商品，进而合理摆放货架，方便顾客选取

这个购物篮中有橙汁、香蕉、洗洁精和玻璃清洁剂，以及6罐苏打水



商品的摆放位置对顾客的购买有影响吗？

苏打水通常和香蕉一起被购买吗？不同的品牌有无区别

哪些商品应该出现而没有出现？

当同时购买洗洁精和橙汁时，玻璃清洁剂也常常同时够买吗？

关联规则的实际意义

- 70%购买了牛奶的顾客将倾向于同时购买面包；
- 某网店或某宝买了商品时，会向顾客推荐相关商品；
- 在保险业务方面，如果出现了不常见的索赔要求组合，则可能为欺诈，需要作进一步的调查；
- 在医疗方面，可以找出可能的治疗组合；
- 在银行方面，对顾客进行分析，可以推荐感兴趣的服务等等。

关联规则的实际意义

- 案例：沃尔玛每20秒就售出一个芭比娃娃，而购买芭比娃娃的顾客中有60%的可能性购买棒棒糖。
- 讨论：沃尔玛可以如何利用这一信息呢？

关联规则的实际意义

- 1、将芭比娃娃和棒棒糖放在一起，促进两者的销售量
- 2、将芭比娃娃放在玩具区，而将棒棒糖摆放在远离玩具区的另外一个销售区，在两个销售区间的通道上摆放一些特别的儿童商品，如促销类商品、高利润商品、沃尔玛自有品牌商品等。消费者可能在超市里逗留更长时间，购买更多商品。
- 3、降低芭比娃娃价格，适当增加棒棒糖的价格，为超市带来更多利润。
- 例如，假设每个娃娃的利润为1$，每只棒棒糖的利润是0.75$。当前每天售出10000个芭比娃娃，关联地每天售出6000个棒棒糖(注意是关联地，棒棒糖实际的销售量要大于6000)。
- 在促销芭比娃娃的活动中，将每个芭比娃娃的利润降低到0.95$，而将棒棒糖每只的利润增加到0.85$促销后每天能销售11 000个芭比娃娃，相关联售出的棒棒糖为每天6 600只。促销前的利润为14 500$，促销后为16 060$

关联规则的实际意义

- 4、如果同时购买芭比娃娃、棒棒糖和沃尔玛希望推广的另外一种商品，消费者可以获得折扣。
- 5、不用同时做芭比娃娃和棒棒糖的广告，只要对芭比娃娃进行宣传，棒棒糖的销售量自然会有所增加，这样做可以节省广告费用
- 6、将棒棒糖生产成芭比娃娃形状

目录

常用的关联规则算法

常用关联算法如下表：

算法名称	算法描述
Apriori	关联规则最常用也是最经典的挖掘频繁项集的算法，其核心思想是通过连接产生候选项及其支持度然后通过剪枝生成频繁项集。
FP-Tree	针对Apriori算法的固有的多次扫面事务数据集的缺陷，提出的不产生候选频繁项集的方法。Apriori和FP-Tree都是寻找频繁项集的算法。
Eclat算法	Eclat算法是一种深度优先算法，采用垂直数据表示形式，在概念格理论的基础上利用基于前缀的等价关系将搜索空间划分为较小的子空间。
灰色关联法	分析和确定各因素之间的影响程度或是若干个子因素（子序列）对主因素（母序列）的贡献度而进行的一种分析方法。

- 本节重点详细介绍Apriori算法。

Apriori算法介绍

- 以超市销售数据为例，提取关联规则的最大困难在于当存在很多商品时，可能的商品的组合（规则的前项与后项）的数目会达到一种令人望而却步的程度。
- 各种关联规则的算法是从不同的方面来减小可能的搜索空间的大小以及减小扫描数据的次数。
- Apriori算法是最经典的挖掘频繁项集的算法，第一次实现了在大数据集上可行的关联规则提取，它的关键是先验原理。
- 如果一个项集是频繁的，则它的所有子集一定也是频繁的

关联规则的基本概念

事务和项集

关联规则的分析对象是事务(Transaction)。

事务可以理解为一种商业行为，含义极为宽泛。

- 超市顾客的购买行为是一种事务
- 网页用户的页面浏览行为是一种事务
- 一份保险公司的人寿保单也是一种事务

项集也即购物篮，在实际应用中有多种不同的理解

- 将一个用户的一次网页浏览记录看作购物篮
- 将一个用户的所有网页浏览记录看作购物篮

Apriori算法介绍

- 令I={i1,i2,...,id}是购物篮数据中所有项的集合，而T={t1,t2,...,tN}是所有事务的集合。
- 每个事务ti包含的项集都是I的子集。
- 项集：在关联分析中，包含0个或多个项的集合称作项集

{牛奶, 面包, 尿布}

- k-项集：一个包含k个项的项集
- 事务的宽度定义为事务中出现的项的个数

Apriori算法介绍

关联规则和频繁项集

项集的重要性质是它的支持度计数(　)

- 一个项集出现的个数
- ({牛奶，面包，尿布}) = 2

关联规则的强度可以用它的支持度(Support)和置信度(Confidence)度量

规则评价方法

- 支持度 (s)：包含X和Y的事务的几率
- 置信度 (c)：包含X的事务中Y出现的几率（X出现的前提下Y的条件概率）

支持度和置信度的例子

$$\{牛奶, 尿布\} \Rightarrow 啤酒$$

$$s = \frac{\sigma(牛奶, 尿布, 啤酒)}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(牛奶, 尿布, 啤酒)}{\sigma(牛奶, 尿布)} = \frac{2}{3} = 0.67$$

事物ID	商品项
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

$$Support\ (A \Rightarrow B) = \frac{A, B同时发生的事务个数}{所有事务个数} = \frac{Support\_count\ (A \cap B)}{Total\_count\ (A)}$$

$$Confidence\ (A \Rightarrow B) = P(B \mid A) = \frac{Support\ (A \cap B)}{Support\ (A)} = \frac{Support\_count\ (A \cap B)}{Support\_count\ (A)}$$

Apriori算法介绍

最小支持度和最小置信度

最小支持度是用户或专家定义的衡量支持度的一个阈值，表示项目集在统计意义上的最低重要性；

- 支持度大于最小支持度的项集称为频繁项集，如频繁3项集
- 低支持度的规则不令人感兴趣，所以对顾客很少同时购买的商品进行促销可能并无益处

最小置信度是用户或专家定义的衡量置信度的一个阈值，表示关联规则的最低可靠性。

- 对于给定的规则X→Y，置信度越高，Y包含在X的事物中出现的可能性就越大

同时满足最小支持度阈值和最小置信度阈值的规则称作强规则。

规则的支持度和置信度

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

规则的例子:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

- 注意:
- 上述所有的规则都是关于同一个项集{Milk, Diaper, Beer} 的二分集合的
- 从同一个项集中产生的规则集拥有相同的支持度，但可以有不同的置信度
- 所以需要分别考虑支持度和置信度方面的要求

Apriori算法介绍

将关联规则挖掘任务分解为如下两个主要子任务:

1、频繁项集的产生

- 发现满足最小支持度阈值的所有项集，这些项集称作频繁项集

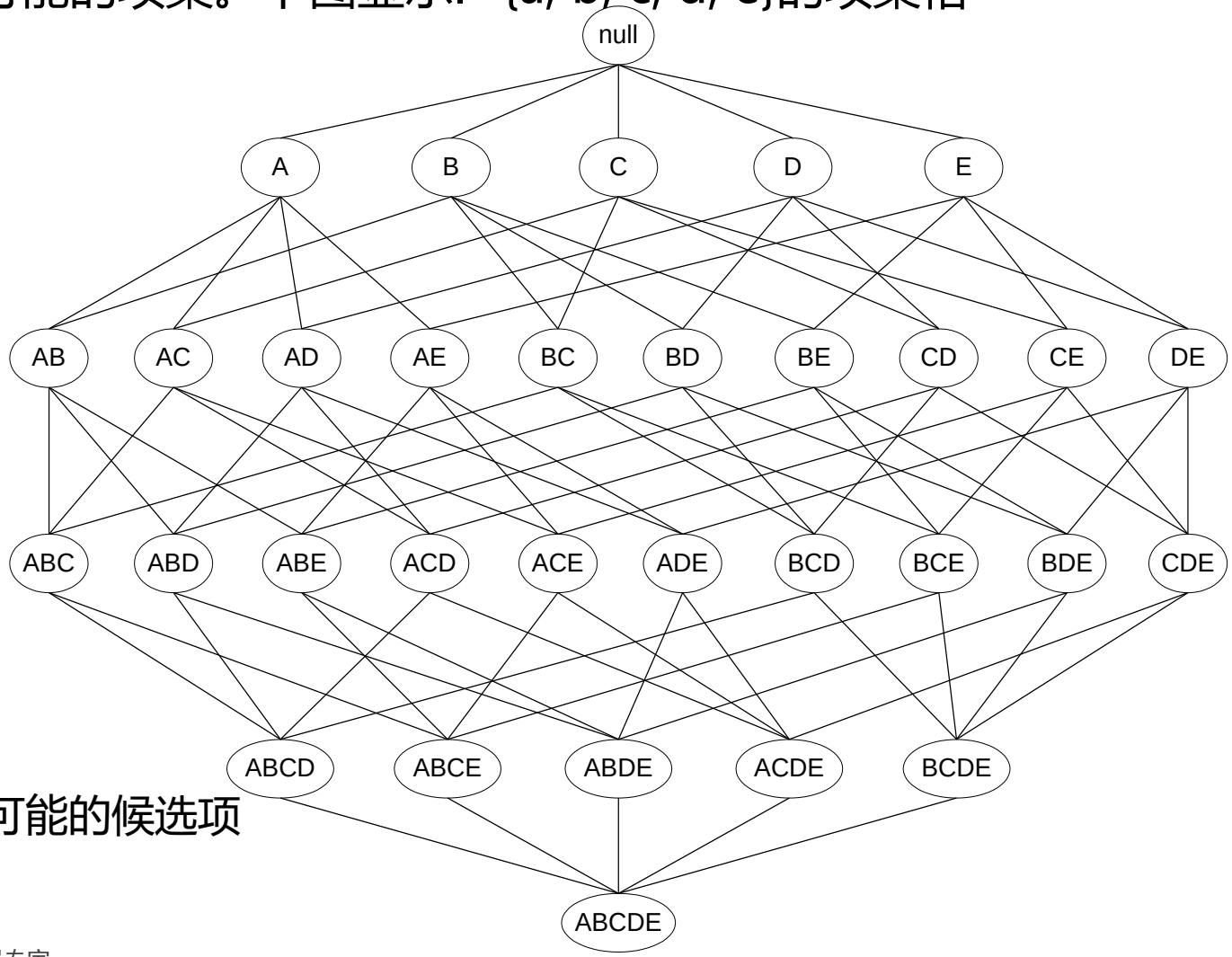
2、规则的产生

- 从上一步发现的频繁项集中提取所有高置信度的规则，这些规则称为强规则

频繁项集的产生是一个计算量巨大的工作

Apriori算法介绍

格结构常用来枚举所有可能的项集。下图显示I={a, b, c, d, e}的项集格



- 如果有d 项,则有 2d个可能的候选项

Apriori算法介绍

先验原理:

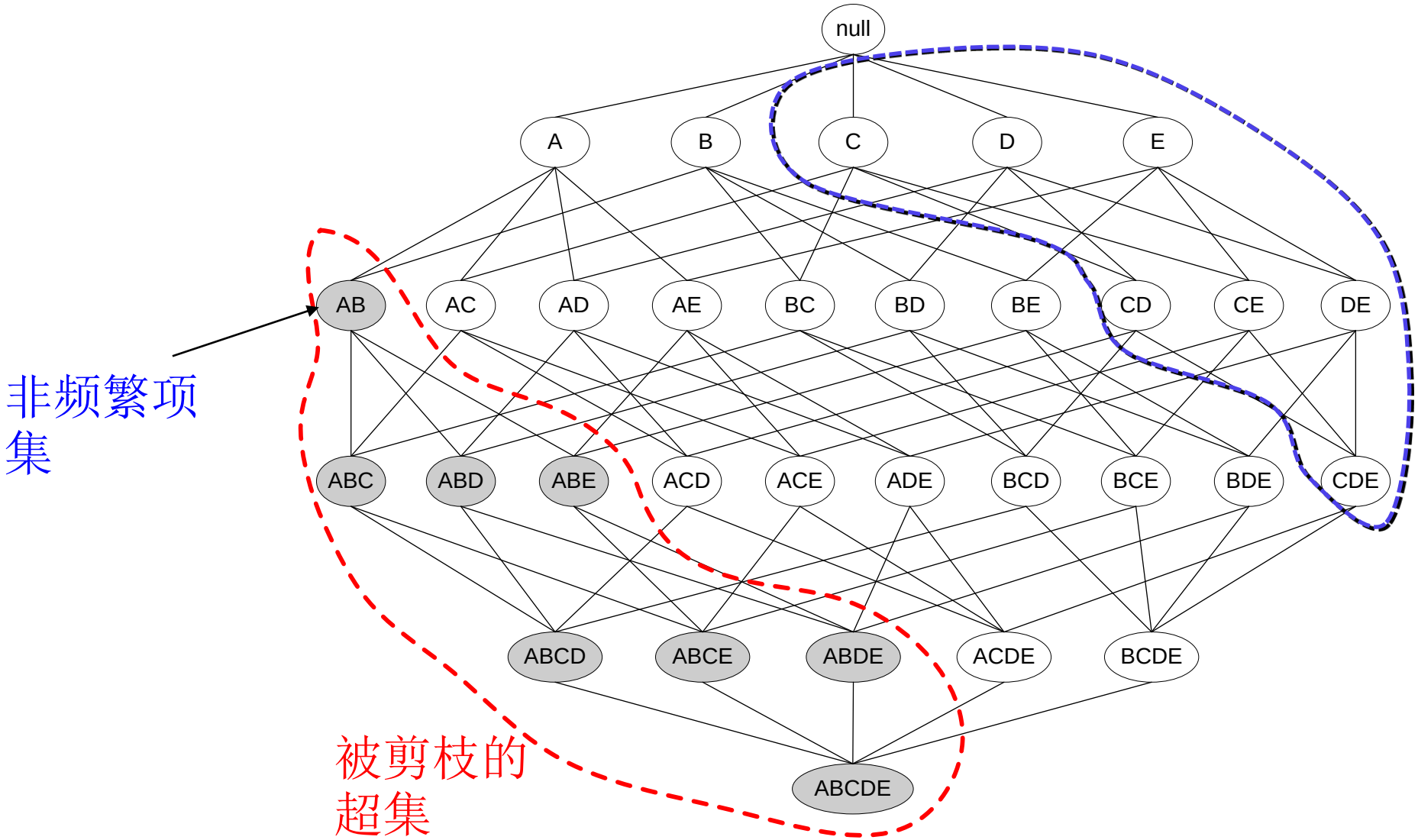
- 如果一个项集是频繁的，则它的所有子集一定也是频繁的

Apriori 算法根据如下的支持度度量的特性:

- 一个项集的支持度决不会超过它的子集的支持度，称作支持度度量的反单调性
- 如{c,d,e}是频繁的，则其所有的子集一定是频繁的
- 若项集{a,b}是非频繁的，则其所有超集也一定是非频繁的

一旦发现{a,b}是非频繁的，则整个包含{a,b}超集的子图可以被立即剪枝，这种基于支持度度量修剪指数搜索空间的策略称为基于支持度的剪枝

先验原理


null
A
B
C
D
E
AB
AC
AD
AE
BC
BD
BE
CD
CE
DE
ABC
ABD
ABE
ACD
ACE
ADE
BCD
BCE
BDE
CDE
ABCD
ABCE
ABDE
ACDE
BCDE
ABCDE
非频繁项集
被剪枝的超集

Apriori算法介绍

- Apriori算法是第一个关联规则挖掘算法，它开创性地使用基于支持度的剪枝技术，系统地控制候选项集指数增长。
- 以购物篮事务数据为例说明该算法。假定支持度阈值为60%，相当于最小支持数为3。
- 最小支持数=支持度 x 事务总数

频繁1项集

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

- 最初每个项都被看作候选1-项集
- 当支持数大于或等于最小支持数时，为频繁项集
- 对其支持度计数之后，候选项集{可乐}和{鸡蛋}被丢弃，因为它们出现的事务少于3个

Minimum Support = 3

Apriori算法

- 首先通过单遍扫描数据集，确定每个项的支持度。得到频繁-1项集的集合F1
- 使用上一次迭代发现的频繁(k-1)-项集，产生候选k-项集
- 为了对候选项集的支持度计数，算法需要再次扫描一遍数据集。
- 计算候选项的支持度计数后，删去支持度计数小于minsup的所有候选项集。
- 当没有新的频繁项集产生时，算法结束。

关联规则的实现

数据库中部分点餐数据如下表：

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
…	…	…	…	…

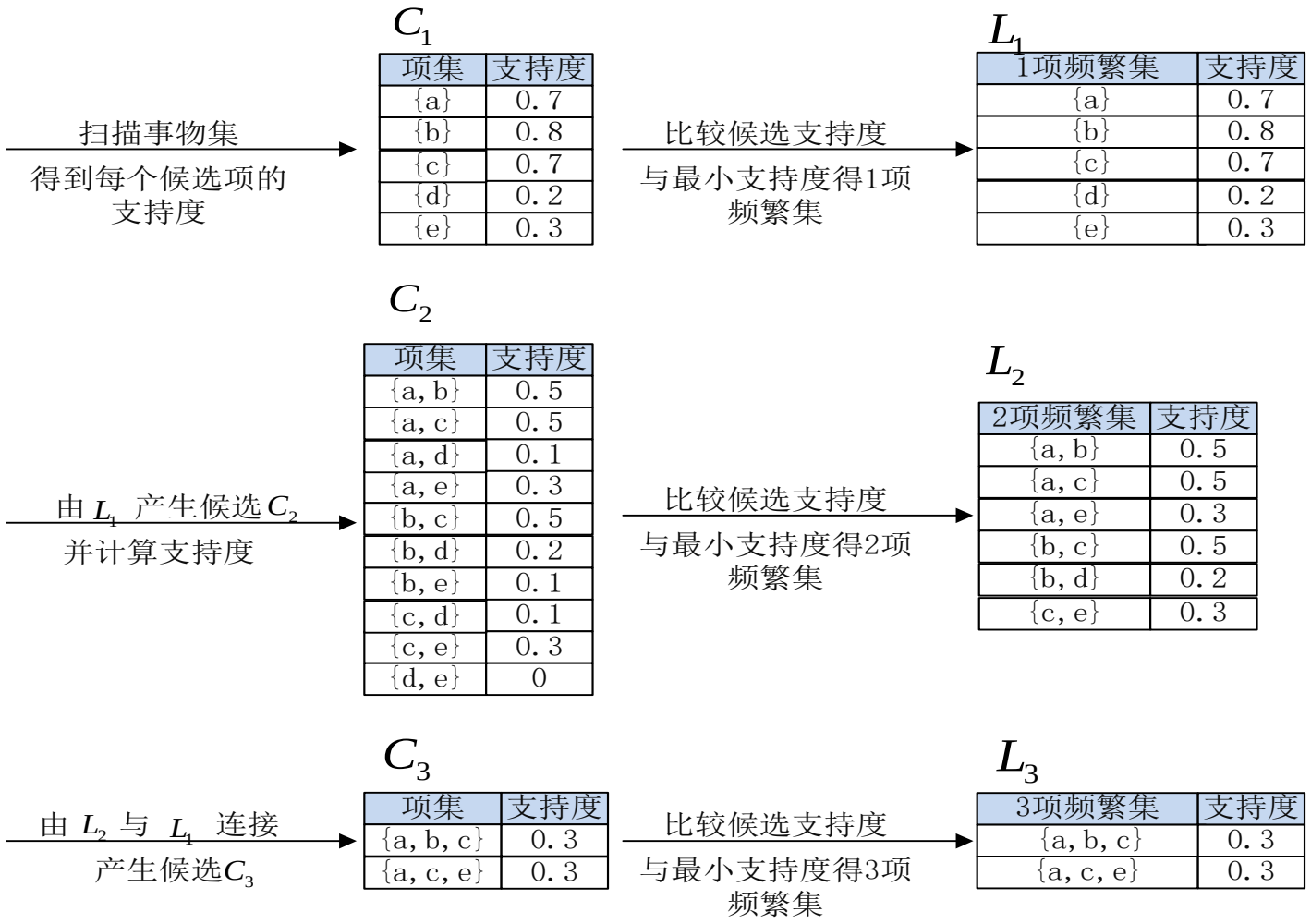
Apriori算法案例

- 首先将事务数据整理成关联规则模型所需的数据结构
- 抽取10个点餐订单作为事务数据集,将菜品{18491，8842，8693，7794，8705}分别简记为{a，b，c，d，e}

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

Apriori算法案例

设支持度为0.2，即支持度计数为2，算法过程如下图：



关联规则运行结果分析

- 转换出的矩阵如下：

	a	b	c	d	e	
0	1.0	0.0	1.0	0.0	1.0	
1	0.0	1.0	0.0	1.0	0.0	
2	0.0	1.0	1.0	0.0	0.0	
3	1.0	1.0	1.0	1.0	0.0	
4	1.0	1.0	0.0	0.0	0.0	
5	0.0	1.0	1.0	0.0	0.0	
6	1.0	1.0	0.0	0.0	0.0	
7	1.0	1.0	1.0	0.0	1.0	
8	1.0	1.0	1.0	0.0	0.0	
9	1.0	0.0	1.0	0.0	1.0	

运行程序的结果如下：

	support	confidence
e---a	0.3	1.000000
e---c	0.3	1.000000
c---e---a	0.3	1.000000
a---e---c	0.3	1.000000
a---b	0.5	0.714286
c---a	0.5	0.714286
a---c	0.5	0.714286
c---b	0.5	0.714286
b---a	0.5	0.625000
b---c	0.5	0.625000
b---c---a	0.3	0.600000
a---c---b	0.3	0.600000
a---b---c	0.3	0.600000
a---c---e	0.3	0.600000

- 对输出结果进行解释：如关联规则“a---b 0.5 0.714286”这条，关联规则a---b的支持度support=0.5，置信度confidence=0.714286。对于餐饮业来说，这条规则意味着客户同时点菜品a和b的概率是50%，点了菜品a，再点菜品b的概率是71.4286%。知道了这些，就可以对顾客进行智能推荐，增加销量同时满足客户需求。

Thank you!

泰迪科技：www.tipdm.com
热线电话：40068-40020