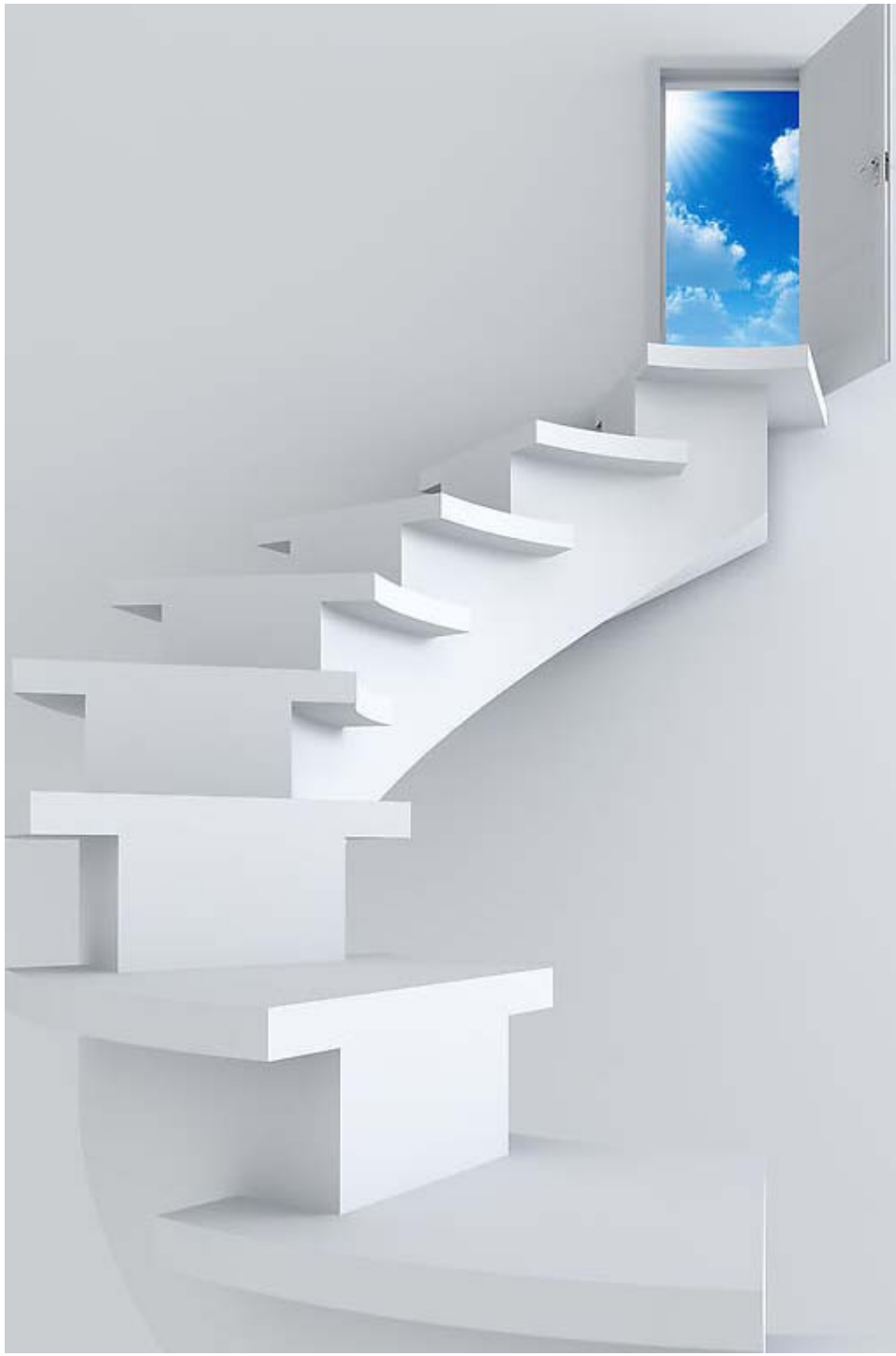

商务智能
Business Intelligence

关联分析

朱卫平 博士
计算机学院
武汉大学

“啤酒与尿布”

购物篮分析：“尿布与啤酒”

- 沃尔玛发现的规律
  - 一些年轻的父亲下班后经常要到超市去买婴儿尿布，其中有30%～40%的人同时要买一些啤酒。
  - 超市随后调整了货架的摆放，把尿布和啤酒放在一起，明显增加了销售额。
- 同样的，我们可以根据关联规则在商品销售方面做各种促销活动。

目录

- 基本概念
  - 购物篮分析
  - 频繁项集、闭项集和关联规则
- 频繁项集挖掘方法
- 哪些模式是有趣的：模式评估方法

购物篮分析

- 对每种商品都用一个布尔量表示其是否被购买，则购物篮可以用一个布尔向量表示
  - 如 {“beer”，“diaper” } 表示啤酒和尿布被购买的购物篮
- 通过分析购物篮可以得到商品被关联购买的模式，称为关联规则
  - 如 {“diaper”} ⇒ {“beer”}

    表示的是尿布的购买会导致啤酒的购买

关联规则：基本概念

- 给定：
  - 项集：$I=\{I_1, I_2, \ldots, I_m\}$
  - K项集：包含k个项的项集
    - {啤酒,尿布}是2项集，{牛奶,面包,黄油}是3项集
  - 事务集D：事务集合，其中每个事务是项集
    - 每个事务由事务标识符TID标识
    - 比如：TID(2000)={A,B,C}
  - 项集的出现频度是指包含项集的事务数目

D

TID	项集
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

关联规则：基本概念

- 关联规则可表示为如下蕴涵式：

$$A \Rightarrow B[s, c]$$

其中A, B为两个项集并且 A ∩ B = Ø

称规则 $A \Rightarrow B$ 具有支持度s 和置信度c

{“diaper”} ⇒ {“beer”}

[support=2%, confidence=60%]

支持度

置信度

规则存在的普适性

如果购买了尿布后购买啤酒的概率

规则度量：支持度和置信度

$$A \Rightarrow B[s, c]$$

- 支持度s是指事务集D中包含 A∪B的概率

$$\text{support}(A \Rightarrow B) = P(A \cup B)$$

- 置信度c是指D中包含A的同时也包含B的概率

$$\text{confidence}(A \Rightarrow B) = P(B \mid A) = P(A \cup B) / P(A)$$

规则度量：支持度和置信度

$$A \Rightarrow B[s, c]$$

- 最小支持度和最小置信度
- 假设最小支持度为50%，最小置信度为50%，有如下关联规则：

$$\mathrm{A} \Rightarrow \mathrm{C}\ (50\%, 66.6\%)$$

$$\mathrm{C} \Rightarrow \mathrm{A}\ (50\%, 100\%)$$

频繁项集与闭项集

- **频繁项集、闭项集基本概念**
  - 如果项集的出现频度大于（最小支持度×D中的事务总数），则称该项集为频繁项集
  - 项集X在数据集D中是闭的，即不存在真超项集Y使得Y与X在D中具有相同的支持度计数，则项集X是数据集D中的闭项集
  - 闭频繁项集
  - 极大频繁项集：该模式的任何真超模式都是非频繁的

关联规则挖掘过程

- **大型数据库中的关联规则挖掘包含两个过程：**
  - 找出所有频繁项集
    - 大部分的计算都集中在这一步
  - 由频繁项集产生关联规则
    - 找到满足最小支持度和最小置信度的规则

挖掘关联规则实例

- 请从以下事务集中找到关联规则

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

最小支持度 50%
最小置信度 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

- 对规则A => C，其支持度

$$\text{support}(A \Rightarrow C) = P(A \cup C) = 50\%$$

- 置信度

$$\text{confidence}(A \Rightarrow C) = P(C \mid A) = P(A \cup C)/P(A)$$

$$= \text{support}(A \cup C)/\text{support}(A) = 66.6\%$$