

# 第15讲 炒股不看股盘看微博 ——关联规则挖掘与大数据思维

---

李玉华

华中科技大学计算机学院

[ldcliyuhua@hust.edu.cn](mailto:ldcliyuhua@hust.edu.cn)

# 炒股不看股盘，看微博，为什么？

2

假如炒股，你看什么--股盘？微博？



学习完本讲  
后再回答吧？

搜索结果

- 综合
- 实时
- 用户
- 文章
- 视频
- 图片
- 话题
- 高级搜索



天道酬勤小鲨鱼

39秒前 来自 新浪微博 weibo.com

京山轻机这主力洗盘真牛逼，先跌破均线给人一种下杀的假象，反抽不过均线震荡震荡，给你一种反抽不过均线出货的假象，震荡后直接一根线封板，不给你加仓入手机会。#微博股票##今日看盘#京山轻机 sz000821



转发

评论

赞



满仓喝酒吃药

40秒前 来自 股票超话

贵州茅台，这次突破2000元，很可能2000元下方，就会成为历史！我可以毫不掩饰的说，我最喜欢的股票就是贵州茅台，贵州茅台让我感觉到有安全感。这种安全感，别的股票都给不了。股票贵州茅台 sh600519 五粮液 sz000858 泸州老窖 sz000568 #白酒 #片仔癀 sh600436 展开

转发

评论

赞



听芹凌翠儿

47秒前 来自 微博 weibo.com

股票 云天化 sh600096 今年的10万吨磷酸铁预期利润终究占比不到5%，主线还是化肥，不要说没有给新能源预期，磷今年仅次于煤，其他大宗都跌成shi了，即使磷矿能支撑住，合成氨和硫磺也撑不住原油暴跌带来的预期指引，必须等原油企稳



# 怎样运用数据库—数据挖掘

3

## 什么是数据挖掘

数据挖掘，又称为数据库中知识发现，它是一个从大量数据中抽取挖掘出未知的、有价值的模式或规律等知识的复杂过程。简单地讲就是从大量数据中挖掘或提取出知识。

- 概要归纳
- **关联规则挖掘**
- 分类与预测
- 聚类分析
- 异类分析
- 演化分析

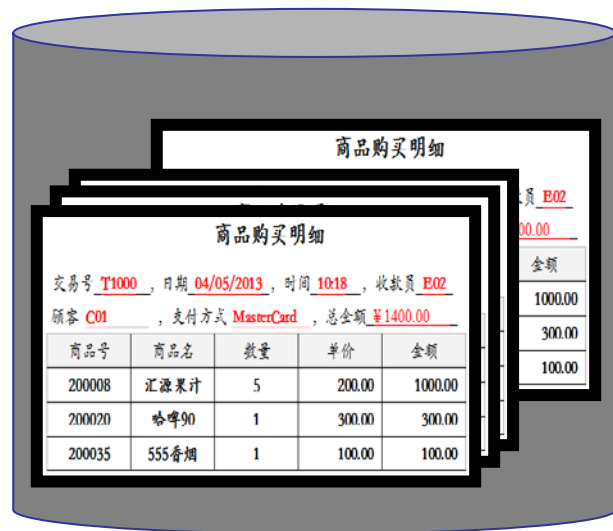
怎样挖掘呢？

啤酒与尿布  
的故事...

# 数据挖掘示例之背景

4

## 超市数据库



客户购买习惯  
商品组合方式及策略

...

营销策略  
价格策略  
货源组织



数据对超市经营有无帮助呢?

# 数据挖掘示例之背景

5

## 超市数据库

### 超市数据库

#### 商品购买明细

交易号 T1000 , 日期 04/05/2013 , 时间 10:18 , 收款员 E02

顾客 C01 , 支付方式 MasterCard , 总金额 ¥1400.00

商品号	商品名	数量	单价	金额
200008	汇源果汁	5	200.00	1000.00
200020	哈啤90	1	300.00	300.00
200035	555香烟	1	100.00	100.00

#### 商品购买单

交易号	日期	时间
T1000	04/05/2013	10:18
T1001	04/05/2013	11:10
...	...	...
T1101	04/06/2013	09:10

#### 商品购买单明细

交易号	商品号	商品名	数量	单价	金额
T1000	200008	汇源果汁	5	200.00	1000.00
T1000	200020	哈啤90	1	300.00	300.00
T1000	200035	555香烟	1	100.00	100.00
T11001	200020	哈啤90	2	300.00	600.00
T11001	200009	巧克力	2	300.00	600.00
...	...	...	...	...	...
T1101	200008	汇源果汁	1	200.00	200.00
T1101	200020	哈啤90	1	300.00	300.00

三味的烘焙小屋			
销售单号:XS130113020013			
顾客名称:普通顾客			
商品名称	折后价	数量	金额
欧登宝淡奶油			
5900002	12.00	2	24.00
雪花烘培纸杯 高温杯/马芬杯/耐高			
温纸杯 10个			
5900226	1.50	1	1.50
sweet烘培纸杯 高温杯/马芬杯/耐高			
温纸杯 1.5元/10个			
5900061	1.50	1	1.50
裱花嘴 曲奇 嘴			
5900410	3.20	1	3.20
水玉波点/雪花/马芬杯/蛋糕杯 小			
号 3元/10个(折)			
5900073	3.00	1	3.00
中号格子烘培纸杯 高温杯/马芬杯/			
耐高温纸杯 1.5元/10个			
5900093	1.50	1	1.50
邮费			
5900177	1.00	5	5.00
消费7项,折后合计:39.7元			
原价合计:39.7元,为您节省:0.0元			
信用卡付款:¥39.70			
收银员:可可			
销售时间:2013-01-13 15:31:55			

# 关联规则挖掘的基本概念

6

## 什么是关联规则

### 数据挖掘之关联规则挖掘

#### 商品的关联规则

商品购买明细

交易号 T1000 , 日期 04/05/2013 , 时间 10:18 , 收款员 E02  
顾客 C01 , 支付方式 MasterCard , 总金额 ¥1400.00

商品号	商品名	数量	单价	金额
200008	汇源果汁	5	200.00	1000.00
200020	哈啤90	1	300.00	300.00
200035	555香烟	1	100.00	100.00

“由尿布的购买，能够推断出啤酒的购买”

“尿布”  $\Rightarrow$  “啤酒” [支持度=2%，置信度=60%]

支持度2%意味着所分析事务的2%同时购买尿布和啤酒

置信度60%意味着购买尿布的顾客60%也购买啤酒。

这样的规则  
能否由机器  
挖掘出来呢？

是否相信这  
条规则呢？

# 关联规则挖掘的基本概念

7

## 基础概念

### 1. 项、项集与事务

设  $P = \{p_1, p_2, \dots, p_m\}$  是所有**项(Item)**的集合。 $D$ 是数据库中所有事务的集合，其中每个**事务  $T$  (Transaction)**是项的集合，是  $P$  的子集，即  $T \subset P$ ;

每一个事务有一个关键字属性，称作**交易号**或**事务号**以区分数据库中的每一个事务。设  $A$  是一个**项集(ItemSet)**，事务  $T$  **包含**  $A$  当且仅当  $A \subseteq T$ 。

### 2. 关联规则

关联规则是形如  $A \Rightarrow B$  的蕴涵式，即命题A(如“项集A的购买”)蕴涵着命题B(“如项集B的购买”)，或者说由命题A能够推出命题B，其中  $A \subseteq P$ ， $B \subseteq P$ ，并且  $A \cap B = \emptyset$ 。

P是超市所有商品的集合

T是某一顾客一次购买的商品的集合

商品购买明细

交易号 T1000，日期 04/05/2013，时间 10:18，收款员 E02  
顾客 C01，支付方式 MasterCard，总金额 ¥1400.00

商品号	商品名	数量	单价	金额
200008	汇源果汁	5	200.00	1000.00
200020	哈啤90	1	300.00	300.00
200035	555香烟	1	100.00	100.00



# 关联规则挖掘的基本概念

8

## 基础概念

### 3. 支持度与置信度

$Support(A \Rightarrow B) = P(A \cup B)$  = 包含A和B的事务数  $\div$  D中事务总数。

$confidence(A \Rightarrow B) = P(B/A)$  = 包含A和B的事务数  $\div$  包含A的事务数。

支持度反映一条规则的实用性，置信度反映规则的“值得信赖性”的程度

### 4. 强规则

同时满足最小支持度阈值( $min\_s$ )和最小置信度阈值( $min\_c$ )的规则称作强规则。

### 5. k-项集与k-频繁项集

项的集合称为**项集**，包含 $k$ 个项的项集称为**k-项集**。

{面包, 果酱} --- 2-项集  
{面包, 果酱, 奶油} --- 3-项集

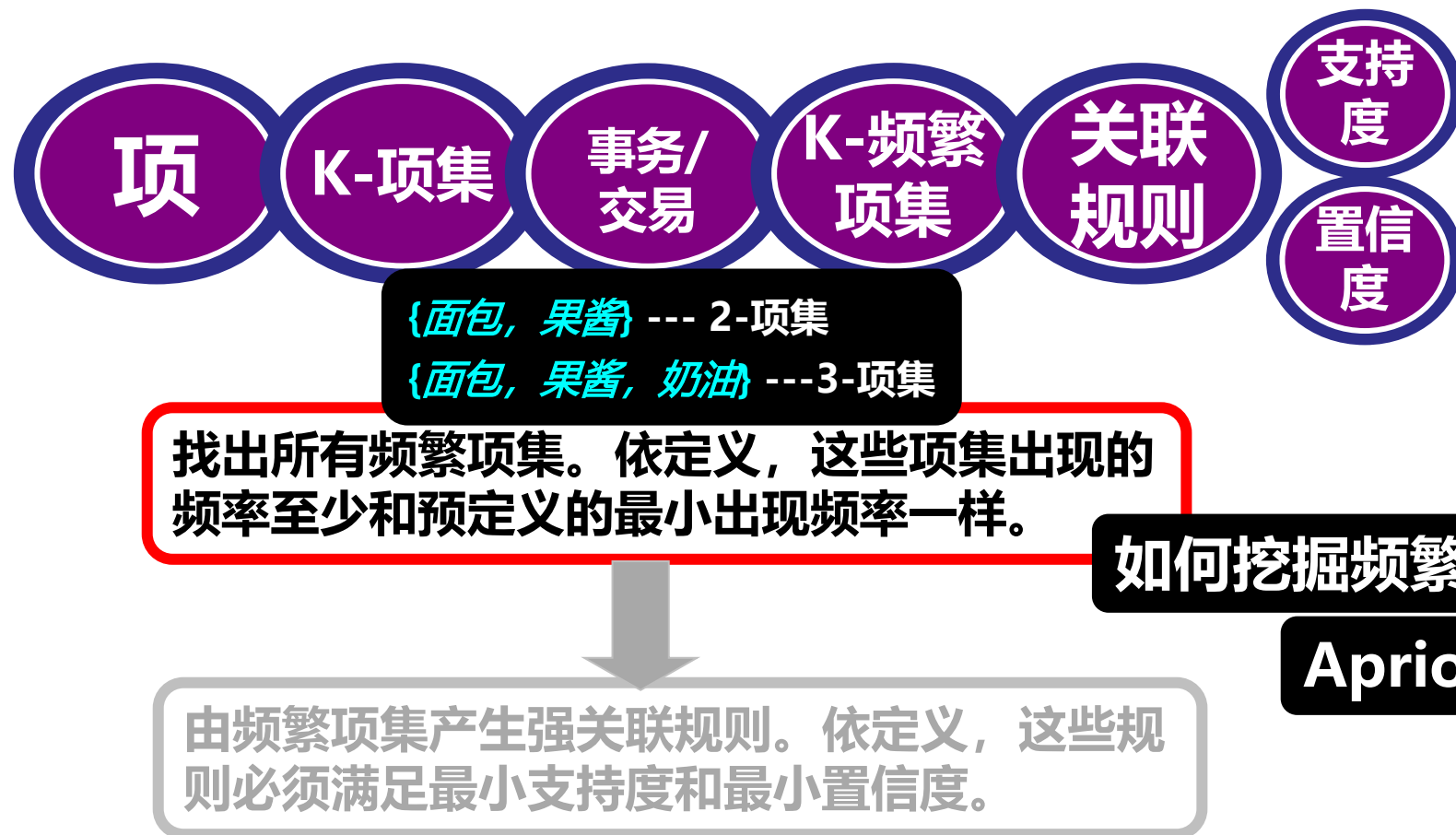
项集的出现频率是包含项集的事务数，简称为项集的**频率**、**支持计数**或**计数**。如果项集的出现频率大于或等于 $min\_s$ 与D中事务总数的乘积，则项集满足最小支持度 $min\_s$ 。如果项集满足最小支持度，则称它为**频繁项集**。频繁 $k$ -项集的集合通常记作 $L_k$ 。



# 关联规则挖掘的基本概念

9

怎样进行关联规则挖掘-基本思想



# 频繁项集挖掘计算过程

10

对问题域数据进行抽象

## 商品购买明细数据库

交易号	一次交易中购买的商品列表	交易号	一次交易中购买的商品列表
T0000	P1, P2, P3, P5	T0050	P1, P3, P5
T1000	P1, P2, P6, P8	T1500	P2, P4, P8
T2000	P2, P3, P7, P8	T2500	P1, P3, P5
T3000	P1, P2, P6	T3500	P2, P3, P7
T4000	P1, P2, P3, P5, P6, P7	T4500	P1, P2, P6, P8
T5000	P1, P3, P5, P6	T5500	P1, P2, P5, P6
T6000	P2, P3, P6	T6500	P1, P2, P5, P6
T7000	P1, P4, P6	T7500	P1, P2, P4, P6
T8000	P2, P3, P4, P5	T8500	P1, P2, P4, P5, P6
T9000	P3, P4, P5	T9500	P1, P2, P4, P5, P6
总交易次数: 20			

### 商品购买明细

交易号 T1000 , 日期 04/05/2013 , 时间 10:18 , 收款员 E02  
顾客 C01 , 支付方式 MasterCard , 总金额 ¥1400.00

商品号	商品名	数量	单价	金额
200008	汇源果汁	5	200.00	1000.00
200020	哈啤90	1	300.00	300.00
200035	555香烟	1	100.00	100.00

# 频繁项集挖掘计算过程

11

形成候选1-项集，并求出频繁1-项集

候选1项集.

项集	支持度计数
{ P1 }	14
{ P2 }	15
{ P3 }	10
{ P4 }	7
{ P5 }	11
{ P6 }	12
{ P7 }	3
{ P8 }	3



频繁1项集. 支持度计数 $\geq$ 最小支持度计数5  
( $\text{min\_sup}=5/20=25\%$ )

项集	支持度计数
{ P1 }	14
{ P2 }	15
{ P3 }	10
{ P4 }	7
{ P5 }	11
{ P6 }	12

# 频繁项集挖掘计算过程

12

形成候选2-项集，并求出频繁2-项集

频繁1项集

项集	支持度计数
{ P1 }	14
{ P2 }	15
{ P3 }	10
{ P4 }	7
{ P5 }	11
{ P6 }	12

候选2项集.  $C_2 = L_1 \text{ Join } L_1$

项集	支持度计数
{ P1, P2 }	10
{ P1, P3 }	5
{ P1, P4 }	4
{ P1, P5 }	9
{ P1, P6 }	11
{ P2, P3 }	6
{ P2, P4 }	5
{ P2, P5 }	7
{ P2, P6 }	10
{ P3, P4 }	2
{ P3, P5 }	7
{ P3, P6 }	3
{ P4, P5 }	4
{ P4, P6 }	3
{ P5, P6 }	6

频繁2项集. 支持度计数  $\geq$  最小支持度计数5

项集	支持度计数
{ P1, P2 }	10
{ P1, P3 }	5
{ P1, P5 }	9
{ P1, P6 }	11
{ P2, P3 }	6
{ P2, P4 }	5
{ P2, P5 }	7
{ P2, P6 }	10
{ P3, P5 }	7
{ P5, P6 }	6

# 频繁项集挖掘计算过程

13

形成候选3-项集，并剪枝，进一步求出频繁3-项集

频繁2项集

项集	支持度计数
{ P1, P2 }	10
{ P1, P3 }	5
{ P1, P5 }	9
{ P1, P6 }	11
{ P2, P3 }	6
{ P2, P4 }	5
{ P2, P5 }	7
{ P2, P6 }	10
{ P3, P5 }	7
{ P5, P6 }	6

候选3项集.  $C_3 = L_2 \text{ Join } L_2$

项集	
{ P1, P2, P3 }	
{ P1, P2, P5 }	
{ P1, P2, P6 }	
{ P1, P3, P5 }	
{ P1, P3, P6 }	被剪掉, 因 {P3,P6}
{ P1, P5, P6 }	
{ P2, P3, P4 }	被剪掉, 因 {P3,P4}
{ P2, P3, P5 }	
{ P2, P3, P6 }	被剪掉, 因 {P3,P6}
{ P2, P4, P5 }	被剪掉, 因 {P4,P5}
{ P2, P4, P6 }	被剪掉, 因 {P4,P6}
{ P2, P5, P6 }	
{ P3, P5, P6 }	被剪掉, 因 {P3,P6}

候选3项集的支持度计数

项集	支持度计数
{ P1, P2, P3 }	2
{ P1, P2, P5 }	6
{ P1, P2, P6 }	8
{ P1, P3, P5 }	4
{ P1, P5, P6 }	6
{ P2, P3, P5 }	3
{ P2, P5, P6 }	5

频繁3项集

项集	支持度计数
{ P1, P2, P5 }	6
{ P1, P2, P6 }	8
{ P1, P5, P6 }	6
{ P2, P5, P6 }	5

# 频繁项集挖掘计算过程

14

迭代地求出最终结果-频繁项集

频繁3项集

项集	支持度计数
{ P1, P2, P5 }	6
{ P1, P2, P6 }	8
{ P1, P5, P6 }	6
{ P2, P5, P6 }	5



候选 4 项集---频繁 4 项  
集支持度计数  $\geq 5$

项集	支持度计数
{ P1, P2, P5, P6 }	5

频繁项集全集 = 频繁1项集  
 $\cup$  频繁2项集  $\cup$  频繁3项集  
 $\cup$  频繁4项集

项集	支持度计数
{ P1 }	14
{ P2 }	15
{ P3 }	10
{ P4 }	7
{ P5 }	11
{ P6 }	12
{ P1, P2 }	10
{ P1, P3 }	5
{ P1, P5 }	9
{ P1, P6 }	11
{ P2, P3 }	6
{ P2, P4 }	5
{ P2, P5 }	7
{ P2, P6 }	10
{ P3, P5 }	7
{ P5, P6 }	5
{ P1, P2, P5 }	6
{ P1, P2, P6 }	8
{ P1, P5, P6 }	6
{ P2, P5, P6 }	5
{ P1, P2, P5, P6 }	5

# 关联规则挖掘的基本概念

15

怎样进行关联规则挖掘-基本思想



{面包, 果酱} --- 2-项集  
{面包, 果酱, 奶油} --- 3-项集

找出所有频繁项集。依定义，这些项集出现的频率至少和预定义的最小出现频率一样。

由频繁项集产生强关联规则。依定义，这些规则必须满足最小支持度和最小置信度。

关联规则生成

项集	支持度计数
{ P1 }	14
{ P2 }	15
{ P3 }	10
{ P4 }	7
{ P5 }	11
{ P6 }	12
{ P1, P2 }	10
{ P1, P3 }	5
{ P1, P5 }	9
{ P1, P6 }	11
{ P2, P3 }	6
{ P2, P4 }	5
{ P2, P5 }	7
{ P2, P6 }	10
{ P3, P5 }	7
{ P5, P6 }	5
{ P1, P2, P5 }	6
{ P1, P2, P6 }	8
{ P1, P5, P6 }	6
{ P2, P5, P6 }	5
{ P1, P2, P5, P6 }	5



# 关联规则产生过程

16

## 潜在关联规则的生成并计算

关联  
规则

{P1,P2,P5,P6}可以产生的潜在规则 $A \Rightarrow B$ , 其中  
 $A \cup B = \{P1,P2,P5,P6\}$ ,  $A \cap B = \emptyset$ .

项集 A	项集 A 支持度 计数(支持度)	项集 B	项集 $A \cup B$ 的支持 度计数(支持度)	置信度 = 项集( $A \cup B$ )的支 持度 / 项集 A 的支持度
{ P1, P2, P5 }	6 (30%)	{ P6 }	5 (25%)	$5/6=83.33\%$
{ P1, P2, P6 }	8 (40%)	{ P5 }	5 (25%)	$5/8=62.50\%$
{ P2, P5, P6 }	5 (25%)	{ P1 }	5 (25%)	$5/5=100.00\%$
{ P1, P5, P6 }	6 (30%)	{ P2 }	5 (25%)	$5/6=83.33\%$
{ P1, P2 }	10 (50%)	{ P5, P6 }	5 (25%)	$5/10=50.00\%$
{ P1, P5 }	9 (45%)	{ P2, P6 }	5 (25%)	$5/9=55.55\%$
{ P1, P6 }	11 (55%)	{ P2, P5 }	5 (25%)	$5/11=45.45\%$
{ P2, P5 }	7 (35%)	{ P1, P6 }	5 (25%)	$5/7=71.42$
{ P2, P6 }	10 (50%)	{ P1, P5 }	5 (25%)	$5/10=50.00\%$
{ P5, P6 }	6 (30%)	{ P1, P2 }	5 (25%)	$5/6=83.33\%$
{ P1 }	14 (70%)	{ P2, P5, P6 }	5 (25%)	$5/14=35.71\%$
{ P2 }	15 (75%)	{ P1, P5, P6 }	5 (25%)	$5/15=33.33\%$
{ P5 }	11 (55%)	{ P1, P2, P6 }	5 (25%)	$5/11=45.45\%$
{ P6 }	12 (60%)	{ P1, P2, P5 }	5 (25%)	$5/12=41.66$

项集	支持度计数
{ P1 }	14
{ P2 }	15
{ P3 }	10
{ P4 }	7
{ P5 }	11
{ P6 }	12
{ P1, P2 }	10
{ P1, P3 }	5
{ P1, P5 }	9
{ P1, P6 }	11
{ P2, P3 }	6
{ P2, P4 }	5
{ P2, P5 }	7
{ P2, P6 }	10
{ P3, P5 }	7
{ P5, P6 }	5
{ P1, P2, P5 }	6
{ P1, P2, P6 }	8
{ P1, P5, P6 }	6
{ P2, P5, P6 }	5
{ P1, P2, P5, P6 }	5

频繁  
项集

# 关联规则产生过程

17

## 潜在关联规则的生成并计算

输出的规则表， $A \cap B = \emptyset$ ，“购买A能够推出购买B”。置信度 $\geq 70\%$ 的规则。

项集 A	项集 A 支持度 计数(支持度)	项集 B	项集 $A \cup B$ 的支持 度计数(支持度)	置信度=项集( $A \cup B$ )的支 持度 $\div$ 项集 A 的支持度
{ P1, P2, P5 }	6 (30%)	{ P6 }	5 (25%)	$5/6=83.33\%$
{ P2, P5, P6 }	5 (25%)	{ P1 }	5 (25%)	$5/5=100.00\%$
{ P1, P5, P6 }	6 (30%)	{ P2 }	5 (25%)	$5/6=83.33\%$
{ P2, P5 }	7 (35%)	{ P1, P6 }	5 (25%)	$5/7=71.42$
{ P5, P6 }	6 (30%)	{ P1, P2 }	5 (25%)	$5/6=83.33\%$

# 关联规则产生过程

## 潜在关联规则的生成并计算

组合形成规则表，频繁3项集能推出哪些频繁项集？  
置信度标记红色为置信度  $\geq 70\%$  的规则。支持度标记蓝色的为满足置信度前提下的支持度  $\geq 40\%$  的规则

项集 A	项集 A 支持度 计数(支持度)	项集 B	项集 A $\cup$ B 的支持 度计数(支持度)	置信度=项集(A $\cup$ B)的支持 度÷项集 A 的支持度
{P1, P2, P5}	6 (30%)	{P6}	5 (25%)	5/6=83.33%
{P1, P2, P6}	8 (40%)	{P5}	5 (25%)	5/8=62.50%
{P2, P5, P6}	5 (25%)	{P1}	5 (25%)	5/5=100.00%
{P1, P5, P6}	6 (30%)	{P2}	5 (25%)	5/6=83.33%
{P1, P2}	10 (50%)	{P5, P6}	5 (25%)	5/10=50.00%
{P1, P2}	10 (50%)	{P5}	6 (30%)	6/10=60.00%
{P1, P2}	10 (50%)	{P6}	8 (40%)	8/10=80.00%
{P1, P5}	9 (45%)	{P2, P6}	5 (25%)	5/9=55.55%
{P1, P5}	9 (45%)	{P6}	6 (30%)	6/9=66.66%
{P1, P5}	9 (45%)	{P2}	6 (30%)	6/9=66.66%
{P1, P6}	11 (55%)	{P2, P5}	5 (25%)	5/11=45.45%
{P1, P6}	11 (55%)	{P2}	8 (40%)	8/11=72.72%
{P1, P6}	11 (55%)	{P5}	6 (30%)	6/11=54.54%
{P2, P5}	7 (35%)	{P1, P6}	5 (25%)	5/7=71.42%
{P2, P5}	7 (35%)	{P1}	6 (30%)	6/7=85.71%
{P2, P5}	7 (35%)	{P6}	5 (25%)	5/7=71.42%
{P2, P6}	10 (50%)	{P1, P5}	5 (25%)	5/10=50.00%
{P2, P6}	10 (50%)	{P1}	8 (40%)	8/10=80.00%
{P2, P6}	10 (50%)	{P5}	5 (25%)	5/10=50.00%
{P5, P6}	6 (30%)	{P1, P2}	5 (25%)	5/6=83.33%
{P5, P6}	6 (30%)	{P1}	6 (30%)	6/6=100.00%
{P5, P6}	6 (30%)	{P2}	5 (25%)	5/6=83.33%
{P1}	14 (70%)	{P2, P5, P6}	5 (25%)	5/14=35.71%
{P1}	14 (70%)	{P2, P5}	6 (30%)	6/14=42.85%
{P1}	14 (70%)	{P2, P6}	8 (40%)	8/14=57.14%
{P1}	14 (70%)	{P5, P6}	6 (30%)	6/14=42.85%
{P1}	14 (70%)	{P2}	10 (50%)	10/14=71.42%
{P1}	14 (70%)	{P5}	9 (45%)	9/14=64.28%
{P1}	14 (70%)	{P6}	11 (55%)	11/14=78.57%

A 为 {P2}, {P5}, {P6} 能推出哪些 B。可类同 A 为 {P1} 时那样处理。此处略

# 关联规则产生过程

## 最终输出的关联规则示例

### 最终输出的规则表

项集 A	项集 A 支持度 计数(支持度)	项集 B	项集 $A \cup B$ 的支持 度计数(支持度)	置信度 = 项集 $(A \cup B)$ 的支 持度 ÷ 项集 A 的支持度
{ P1, P2 }	10 (50%)	{ P6 }	8 (40%)	8/10=80.00%
{ P1, P6 }	11 (55%)	{ P2 }	8 (40%)	8/11=72.72%
{ P2, P6 }	10 (50%)	{ P1 }	8 (40%)	8/10=80.00%
{ P1 }	14 (70%)	{ P2 }	10 (50%)	10/14=71.42%
{ P1 }	14 (70%)	{ P6 }	11 (55%)	11/14=78.57%
A 为 {P2}, {P5}, {P6} 能推出哪些 B, 可类同 A 为 {P1} 时那样处理, 此处略.				

“P1, P2”  $\Rightarrow$  “P6” [支持度=40%, 置信度=80%]

“P1, P6”  $\Rightarrow$  “P2” [支持度=40%, 置信度=72.72%]

“P2, P6”  $\Rightarrow$  “P1” [支持度=40%, 置信度=80%]

“P1”  $\Rightarrow$  “P2” [支持度=50%, 置信度=71.42%]

“P1”  $\Rightarrow$  “P6” [支持度=55%, 置信度=78.57%]

# 还能挖掘什么

20

## 还能挖掘什么规则

### 单维度单层次规则

$buys(X, "面包") \Rightarrow buys(X, "果酱")$     X代表顾客

### 多维度多层次规则

$age(X, "30...39") \wedge income(X, "42K...48K") \Rightarrow buys(X, "high\_resolution\_TV")$

# 再看微博数据

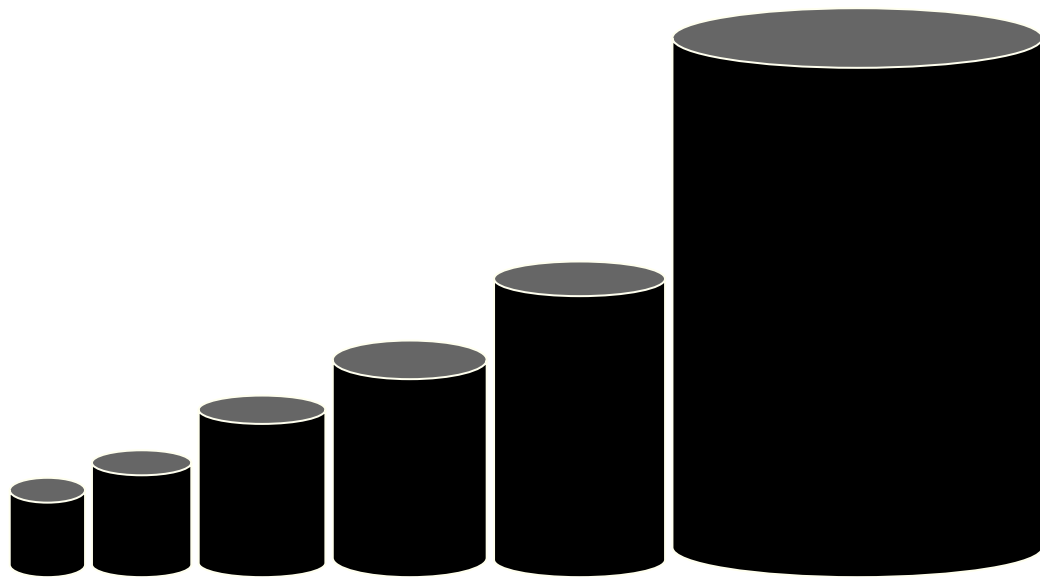
21

## 微博数据挖掘 vs. 超市数据挖掘

	“微博” 挖掘	“超市数据” 挖掘
数据基本组织形式	文本---非结构化数据	“表” ---结构化数据
被挖掘数据D的集合	众多人、众多次：发表的微博	众多人、众多次：购买的商品
事务数据T的涵义	一次发表的“微博” 可以看作是“若干词汇” 的集合	一次购买的商品可以看作是“若干商品” 的集合
项的集合	“词汇” 的集合	“商品” 的集合
频繁项集	频繁使用的“词汇” 集合	频繁购买的“商品” 集合
规则 $A \Rightarrow B$	使用了“词汇A” 也使用了“词汇” B	购买了“商品A” 也购买了“商品B”
规则挖掘的意义	通过分析，可发现“可以组合在一起的关键词汇”，进而进行主题词设置、读者兴趣引导，以提高某主题的关注度、粉丝的聚集度等	通过分析，可发现“可被组合在一起的商品”进而进行位置、政策等的调整，以提高客户的购买兴趣等

关于炒股不看  
股盘看微博，  
你相信吗？

## 小结



只求关系，不求因果

不要相信经验，一切以数据说话

大数据环境下什么不能发生呢？

bit & Byte

1KB(Kilobyte) =  $2^{10}$ 字节

1MB(Megabyte) =  $2^{10}$ KB

1GB(Gigabyte) =  $2^{10}$ MB

1TB(Trillionbyte) =  $2^{10}$ GB =  $2^{20}$ MB

1PB(Petabyte) =  $2^{10}$ TB =  $2^{30}$ MB

1EB(Exabyte) =  $2^{10}$ PB =  $2^{40}$ MB

1ZB(Zettabyte) =  $2^{10}$ EB =  $2^{50}$ MB

1YB(Yottabyte) =  $2^{10}$ ZB =  $2^{60}$ MB

1BB(Brontobyte) =  $2^{10}$ YB =  $2^{70}$ MB

