北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

Management in Big Data Era

马宝君 博士 讲师

经济管理学院
电子商务中心
2014年11月24日

12月15日上课时间调整

2014 年 12 月				December		
星期一	星期二	星期三	星期四	星期五	星期六	星期日
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15 18:30-20:20 此次课程不上	16	17	18	19 改到周五 18:30-20:20	20	21
22	23	24	25	26	27	28
29	30	31				

课前分享1：中国人平均每人每天摸手机150次

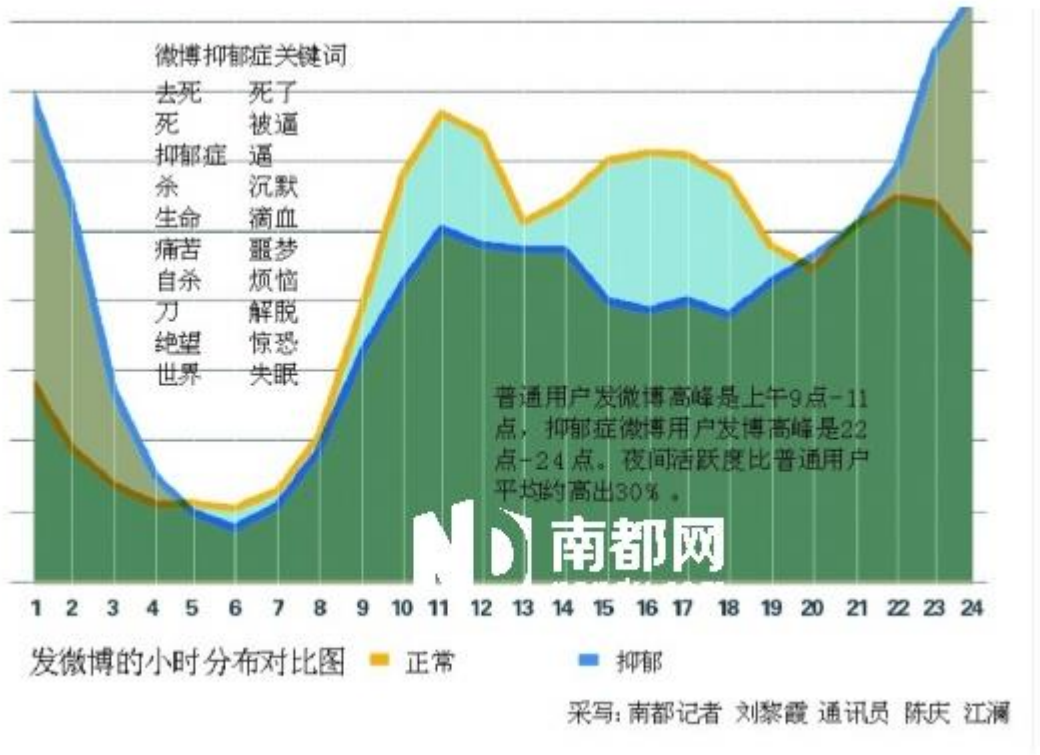
随时随地刷微信，上厕所也要带手机，手机一没电就像世界末日……近日，清华大学新闻与传播学院沈阳教授称，调查显示，每个中国人平均每天摸手机150次！ 除去睡着的8小时，差不多6分钟就要看一次…

新东方创始人俞敏洪听后深有同感，因为“他每天摸手机的次数比这还要多”。而且，“我女儿用手机打字比我在电脑上打字快，现在来新东方报名的学生三分之二是看了微信来报名的，比看了报纸广告来报名的多多了。”

来源：http://bbs.jurong.cn/thread-1157261-1-1.html

课前分享2：大数据挖出微博抑郁患者

- 在微博上经常抱怨“去死”，或许会被甄别为抑郁患者
- 近期，一项“利用社交媒体数据挖掘识别抑郁倾向人群”的研究成果在网络引发热议，来自哈尔滨工业大学的研究人员称，通过构建抑郁倾向识别模型，实验室在新浪微博近亿用户中识别出几百名重度抑郁症患者，研究结果经医学机构确认准确度可达83%。相关人士表示，这项研究结果或成为抑郁症临床诊断之外的新兴诊断方法。



课前分享2：大数据挖出微博抑郁患者

- 建构预测模型在新浪过亿用户中扫描
  - 如何识别抑郁群体？研究团队首先是挑选新浪微博用户中被确认为抑郁症的人群作为样本，通过计算机强大的计算能力分析样本数据，从这些数据获取出规律后建构预测模型。有了数据模型，计算机就可以用这一模型扫描新浪微博上过亿用户了。
  - “计算机算法会包括自然语言处理、时间序列、机器学习等，比如失眠在抑郁症患者中比例非常高，会成为语言处理的关键词，机器还会对关键词出现的频率和时间段打分。”
- 约200用户被人工判定为抑郁患者
  - 计算机最终统计的数据比他们想象中要更为丰富：存在抑郁倾向的微博用户与普通用户发博时间有明显差异，这部分人群发博高峰在23点，其夜间活跃度比普通用户平均约高出30%。该群体微博关键词为：死、抑郁症、生命、痛苦、自杀。有60%为女性，40%为男性，女性比例比男性略高。
  - “有很多数据很值得关注的，比如有些表现出抑郁症倾向的用户除了喜欢用小号来表达痛苦情绪，还有群落聚集趋势，他们会同时关注很多其他同类人群，有的甚至会习惯每天到已经自杀的用户微博上评论‘今天你还好吗？’，这听起来有点瘆人。

来源：http://www.d1net.com/bigdata/news/316953.html

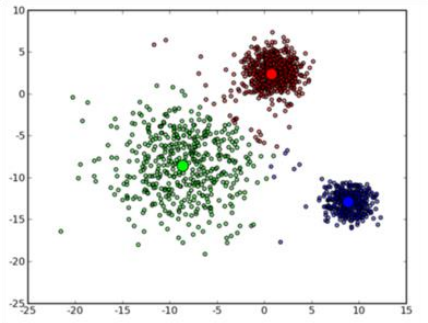
课程回顾：深度业务分析——原方法

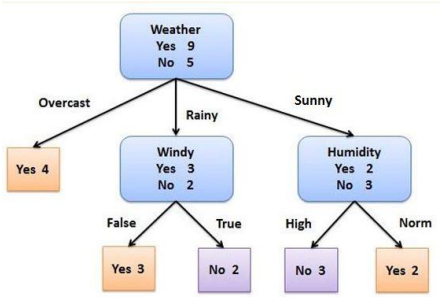
- 聚类（Clustering）
- 分类（Classification）
- 关联（Association）
- 模式（Pattern）
- ……

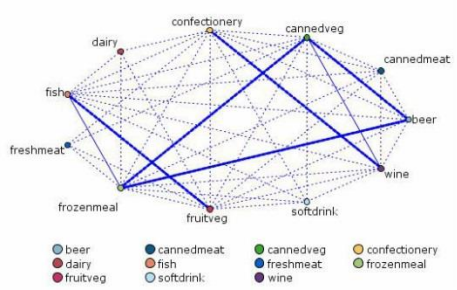
类别

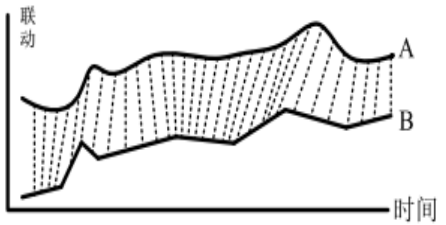
联系

轨迹







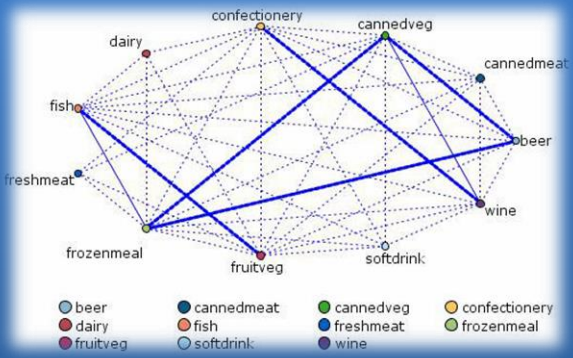


上次课程内容回顾

- 数据挖掘/商务智能分析方法 参考资料
- 划分式聚类方法的代表：K-means
- 层次聚类方法（Hierarchical Clustering）
- 聚类分析小案例：客户细分

关联分析方法

Association Analysis



关联分析（Association Analysis）

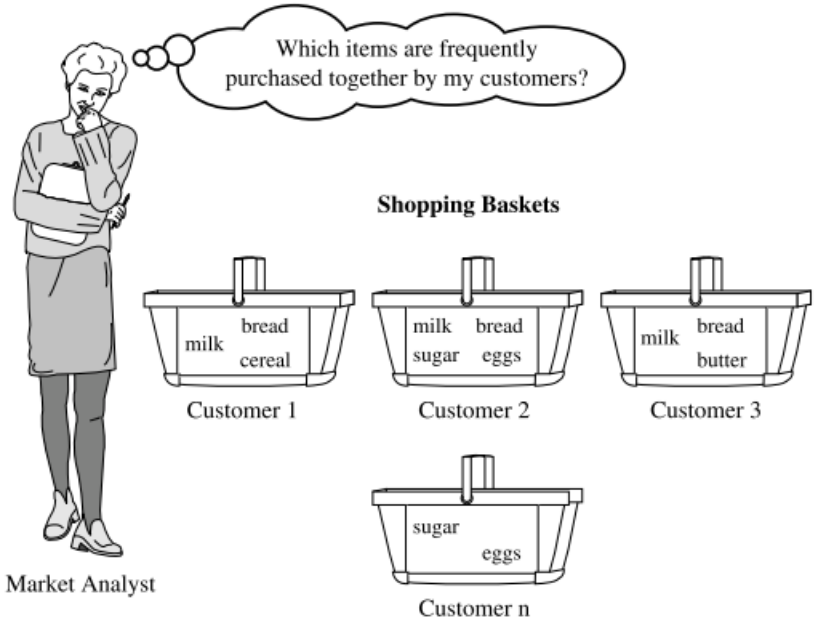
- 基本概念
- 关联规则挖掘的有效方法
- 不同种类的关联规则
- 关联规则的应用
- 总结

关联分析（Association Analysis）

- **基本概念**
- 关联规则挖掘的有效方法
- 不同种类的关联规则
- 关联规则的应用
- 总结

关联分析的基本概念

- 关联(association)是商务智能领域关注的重要知识类型，关联分析用于分析对象之间的关联性和相关性。
- 关联知识具有多种形式，如关联规则、数据依赖、模式关联等。关联规则是商务智能领域中最基本的一种关联知识形式。



What Is Association Rule Mining?

- **关联规则的挖掘：**
  - 从交易数据库、关系数据库以及其他的数据集中发现项或对象的频繁模式（frequent patterns）、关联规则（ association rules）的过程
  - 例如：buys(x, “diapers”) →(⇒) buys(x, “beers”) [0.5%, 60%]
  - Rao, Srikumar S. “Diaper-beer Syndrome,” Forbes, April 6, 1998. pp. 128–130.

TID	Items Bought
1	beer,nuts,diaper
2	beer,diaper
3	beer,bread
4	nuts,cheese,butter

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes

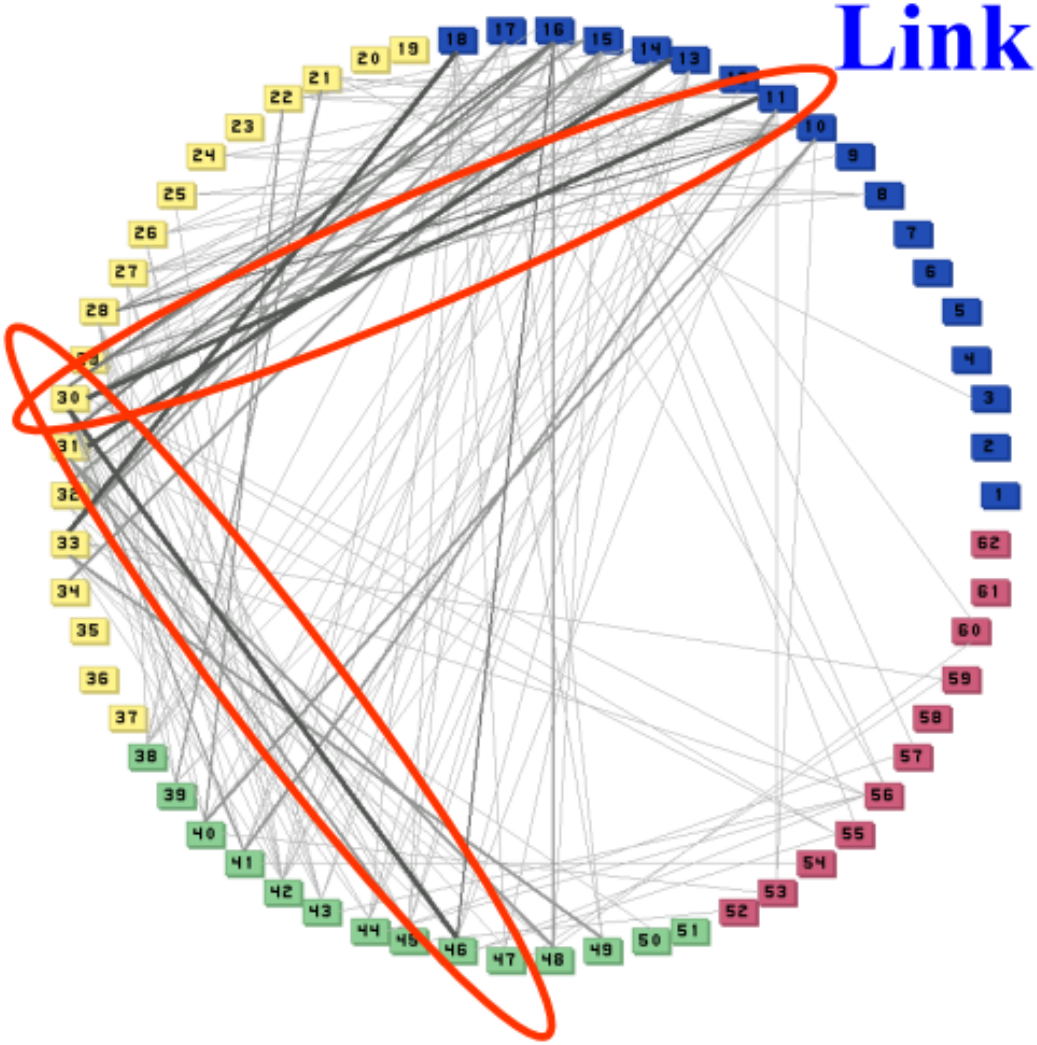
Enterprise Miner - [Detailed Results - Links]

File Globals Options Window Help

OK

Threshold:

Link analysis



Minimum = 334
Maximum = 1979

1 : ITEM1 = ARTICHOK
2 : ITEM1 = AVOCADO
3 : ITEM1 = BAGUETTE
4 : ITEM1 = BOURBON
5 : ITEM1 = CHICKEN
6 : ITEM1 = COKE
7 : ITEM1 = CORNED_B
8
9
10
11
12
13
14 : ITEM1 = PEPPERS
15 : ITEM1 = SARDINES
16 : ITEM1 = SODA
17 : ITEM1 = STEAK
18 : ITEM1 = TURKEY
19 : ITEM2 = APPLES
20 : ITEM2 = ARTICHOK
21 : ITEM2 = AVOCADO
22 : ITEM2 = BAGUETTE
23 : ITEM2 = BORDEAUX
24 : ITEM2 = BOURBON
25 : ITEM2 = CHICKEN
26 : ITEM2 = COKE
27
28
29
30
31
32
33 : ITEM2 = OLIVES
34 : ITEM2 = PEPPERS
35 : ITEM2 = SARDINES
36 : ITEM2 = SODA
37 : ITEM2 = STEAK
38 : ITEM3 = APPLES
39 : ITEM3 = ARTICHOK
40 : ITEM3 = AVOCADO
41 : ITEM3 = BAGUETTE
42 : ITEM3 = BOURBON
43 : ITEM3 = CHICKEN
44
45
46
47
48
50 : ITEM3 = ICE_CREA
51 : ITEM3 = OLIVES
52 : ITEM4 = APPLES
53 : ITEM4 = ARTICHOK
54 : ITEM4 = AVOCADO
55 : ITEM4 = BAGUETTE
56 : ITEM4 = BOURBON
57 : ITEM4 = CHICKEN
58 : ITEM4 = COKE
59 : ITEM4 = CORNED_B
60 : ITEM4 = CRACKER
61 : ITEM4 = HAM
62 : ITEM4 = HEINEKEN

11: Herring

30: Heinekken

46:Crackers

关联分析指导交叉销售

Link Plot 62 of 62 factors 539 defined links Min=334 Max=1979

基本概念：频繁模式

- Transactional database (交易数据库T, n条交易记录)
  - 每个交易：each transaction is a list of items (顾客一次购买的商品)
  - 所有的商品名称项目集合：$I=\{i_1, i_2, \ldots, i_m\}$
  - 项集(Itemset): $X=\{i_{j1}, i_{j2}, \ldots, i_{jp}\}$， $i_{ji} \in I$
  - 每个项集包含的项的个数，称为项集的长度，一个长度为k的项集又称为k项集。

- I={A,B,C,D,E,F}
- 2项集：
  - AB, AC, BC, AD, BE, BF, EF

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

基本概念：支持度（Support）

- 项集X的支持度 sup(X)
  - 一条交易记录包含项集X的概率
  - sup(X) = $|T_X|/|T| = |T_X|/n$
  - 如果sup(X) >= *minsup* (最小支持度)，则称X为频繁项集(frequent itemset)，即X是频繁的

TID	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

- 给定*minsup*=50%
- 所有的频繁项集：

{A: 3/5, B: 3/5, D: 4/5, E: 3/5, AD: 3/5}

基本概念：关联规则（Association Rule, AR）

- 直观意义考虑：$X \rightarrow Y$ 且满足：
  - $X = \{x_1, \ldots, x_k\}$, $Y = \{y_1, \ldots, y_m\}$, $X \cap Y = \phi$
- 合格的关联规则
  - 所有满足最小支持度(*minsup*)和最小置信度(*minconf*)的关联规则
- 阈值（Threshold）
  - Minimum support : *minsup*
  - Minimum confidence : *minconf*
  - $sup(X \rightarrow Y) \geq minsup$
  - $conf(X \rightarrow Y) \geq minconf$

基本概念：支持度 与 置信度

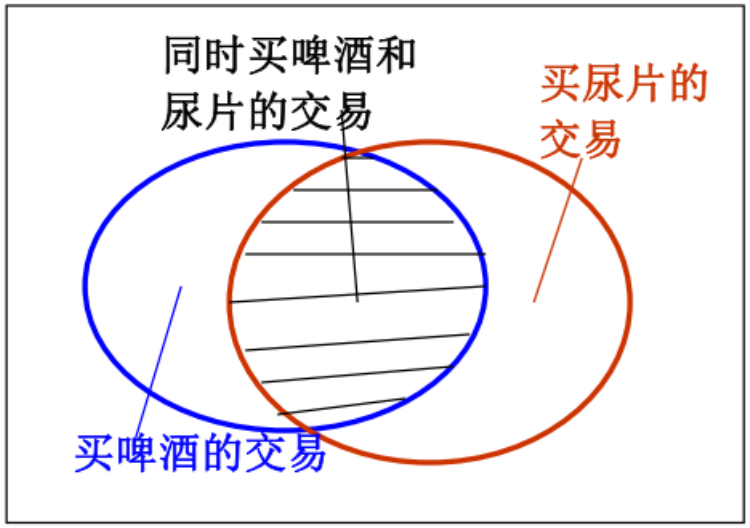
- 关联规则：$X \rightarrow Y$
  - $sup(X \rightarrow Y) = sup(X \cup Y) = |T_{XY}| / n$（同时包含X和Y的记录比例）

- 例如
  - X={A}，Y={C}
  - $sup(A \rightarrow C) = sup(AC) = 0.2$
  - X={A,D}=AD，Y={C}
  - $sup(AD \rightarrow C) = sup(ADC) = 0.2$

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, E
40	B, E, F
50	B, C, D, E, F

基本概念：支持度 与 置信度（续）

- 置信度：Confidence
  - conditional probability that a transaction having X also contains Y
  - 包含X项集的记录中有多大比例（概率）同时也包含Y项集
  - Conf(X → Y)=$|T_{XY}|$ / $|T_X|$=sup(XY) / sup(X)



Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, E
40	B, E, F
50	B, C, D, E, F

$A \rightarrow C$ (20%, 33%)

$AD \rightarrow C$ (20%, 50%)

关联分析（Association Analysis）

- 基本概念
- **关联规则挖掘的有效方法**
- 不同种类的关联规则
- 关联规则的应用
- 总结

To find valid ARs: Big Data Challenge！！

- **场景：假设一个连锁超市共销售3000（m）种商品，每月有60000（n）笔交易。(注: 考虑一个网上超市，m和n可能大很多）**
- **超市高管可能很关心了解客户购买某些商品，还会不会同时也购买其它一些商品？**
- **思考：**
  - 1. 模型Y=f(X）验证（如：Y=a+bX+e），需要验证多少这样的函数关系？
  - 2. 如果 X和 Y为任意商品组合，如： beer & water → diaper, Levis & Swatch → iPod & Lens…，有多少这种 X → Y关系需要验证（X, Y ⊆ I, X, Y ≠ ∅, X ∩ Y= ∅，其中 I是所有商品的集合，即| I | = m）？

TID	Customer	Date	Product
1	Lily	2014. 03.02	A, F
2	David	2014. 03.02	A, D, G
3	Lily	2014. 03.03	B, D
4	Jane	2014. 03.04	A, E, G
5	David	2014. 03.06	B, C, D
6	Jane	2014. 03.06	D
7	David	2014. 03.07	F, E, G, H
8	David	2014. 03.07	F, I
9	Lily	2014. 03.07	H, G
10	Jane	2014. 03.07	I
……	…	…	…

Q1：$Y = f(X), |Y| = 1, |X| = m-1, Y \cap X = \varnothing$

左侧变量选择	右侧变量选择
在m个变量中选择1个 可能情况数：C_m^1	在剩下的$m-1$个变量中选择1个 可能情况数：C_{m-1}^1
	在剩下的$m-1$个变量中选择2个 可能情况数：C_{m-1}^2
	……
	选择所有剩下的$m-1$个变量 可能情况数：C_{m-1}^{m-1}

所有可能的模型数量：

$$C_m^1 \times (C_{m-1}^1 + C_{m-1}^2 + \ldots + C_{m-1}^{m-1}) = C_m^1 \times (\sum_{i=0}^{m-1} C_{m-1}^i - C_{m-1}^0)$$

$$= m \times (2^{m-1} - 1)$$

Q2 : $X, Y \subseteq I; X, Y \neq \Phi$ and $X \bigcap Y = \Phi$

	X	Y
Situation 1	1 item C_m^1	No more than (m-1) items from item set {I-X} $C_{m-1}^1 + \ldots + C_{m-1}^{m-1}$
Situation 2	2 items C_m^2	No more than (m-2) items from item set {I-X} $C_{m-2}^1 + \ldots + C_{m-2}^{m-2}$
.....		
Situation m-1	m-1 items C_m^{m-1}	1 item from item set {I-X} C_1^1

$3^{100} \approx 5.15*10^{47}$

$$N = C_m^1(C_{m-1}^1 + \ldots + C_{m-1}^{m-1}) + C_m^2(C_{m-2}^1 + \ldots + C_{m-2}^{m-2}) + \ldots + C_m^{m-1}C_1^1$$

$$= C_m^1(2^{m-1} - 1) + C_m^2(2^{m-2} - 1) + \ldots + C_m^{m-1}(2^1 - 1)$$

$$= C_m^1 2^{m-1} + C_m^2 2^{m-2} + \ldots + C_m^{m-1}2^1 - (C_m^1 + C_m^2 + \ldots + C_m^{m-1})$$

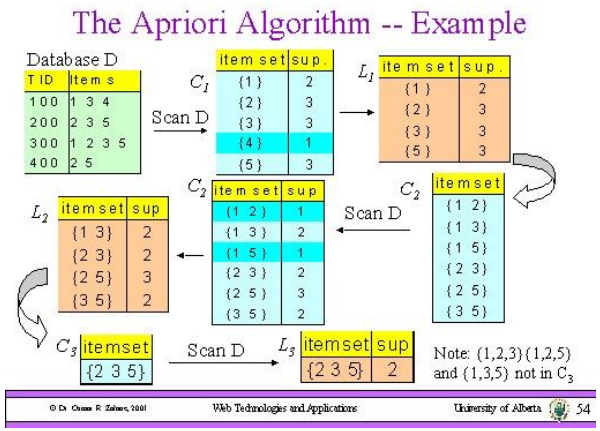
$$= C_m^0 2^m + C_m^1 2^{m-1} + C_m^2 2^{m-2} + \ldots + C_m^{m-1}2^1 + C_m^m 2^0 - 2^m - 1 - (2^m - 2)$$

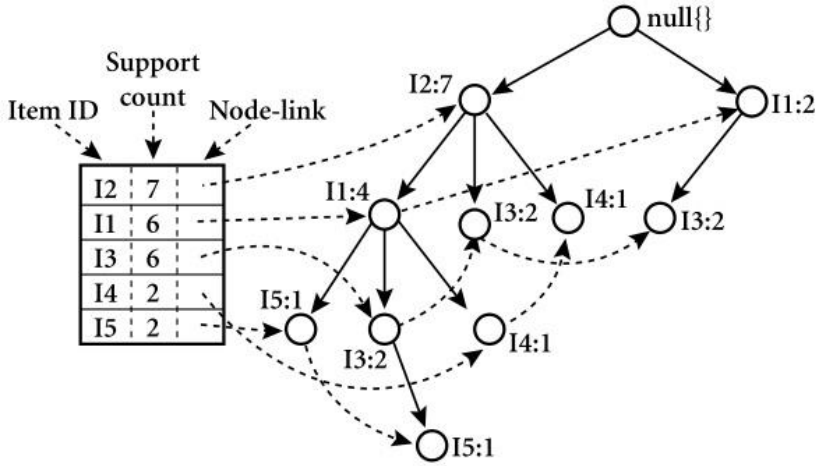
$$= 3^m - 2^{m+1} + 1$$

$$m \times (2^{m-1} - 1)$$

关联规则挖掘的有效方法

- Apriori (Agrawal & Srikant@VLDB'94)
- Freq. pattern growth (FPgrowth—Han, Pei & Yin @SIGMOD'00)
- Vertical data format approach (Charm—Zaki & Hsiao @SDM'02)
- ……





Apriori关联规则挖掘方法

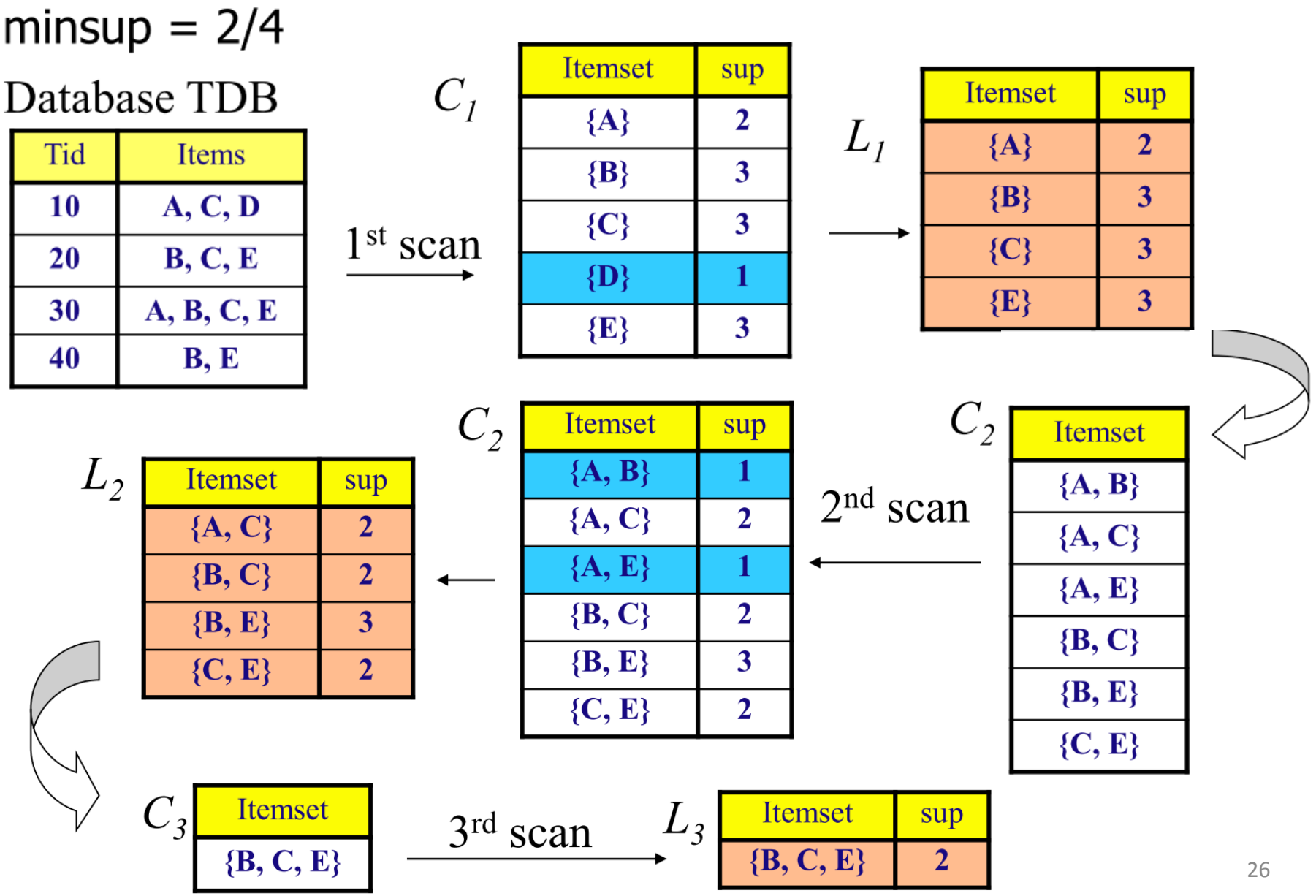
TID	Items Bought
1	beer,nuts,diaper
2	beer,diaper
3	beer,bread
4	nuts,cheese,butter

- **用到的重要“剪枝”性质**
- **任何频繁项集的子集必定是频繁的**
  - if {beer, diaper} is frequent, so is {beer} and {diaper}
- **逆否命题：任何非频繁项集的超集必定是非频繁的**
  - If {beer} is not frequent, {beer, diaper} is not frequent
- **Apriori剪枝规则**
  - **若存在某些项集是不频繁的，则这些项集的任何超集都是不频繁的，因而无须生成和测试。**

Apriori方法：主要步骤

- **1. 发现所有的频繁项集(frequent itemsets)**
  - 支持度>= *minsup*的所有项集
  - 统计每个k项候选集的支持度，找出频繁的k项集：L_k
  - 利用频繁的k项集生成k+1项候选集(Candidate itemset)：C_{k+1}
- **2. 将每个频繁项集生成可能的关联规则**

Apriori生成频繁项集： An Example



Apriori方法：生成关联规则

$$confidence(X \rightarrow Y) = \mathrm{P}(Y \mid X) = \frac{support(XY)}{support(X)}$$

- **For each frequent itemset *l*, generate every non-emptey subset *s*; if *s* satisfies *minconf*:**

$$confidence((l-s) \rightarrow s) = \frac{sup(l)}{sup(l-s)} \geq minconf$$

Output rules: *(l-s)* → *s*

- ***e.g:*** *l=ABCD, s = D , (l-s)= ABC*

confidence(ABC → D)=support(ABCD)/support(ABC)

Apriori生成关联规则： An Example

L_1

itemset	sup.
{A}	2
{B}	3
{C}	3
{E}	3

L_2

itemset	sup
{A C}	2
{B C}	2
{B E}	3
{C E}	2

L_3

itemset	sup
{B C E}	2

- minconf=80%
- For {BCE}: Confidence(BE → C)<80%,
  Confidence(BC → E)>80%
  Confidence(CE → B)>80%
  Confidence(B → CE) <80%
  Confidence(E → BC) <80%
  Confidence(C → BE) <80%

回顾：Apriori方法的主要步骤

- **1. 发现所有的频繁项集(frequent itemsets)**
  - 支持度>= *minsup*的所有项集
  - 统计每个k项候选集的支持度，找出频繁的k项集：L_k
  - 利用频繁的k项集生成k+1项候选集(Candidate itemset)：C_{k+1}
- **2. 将每个频繁项集生成可能的关联规则**

关联分析（Association Analysis）

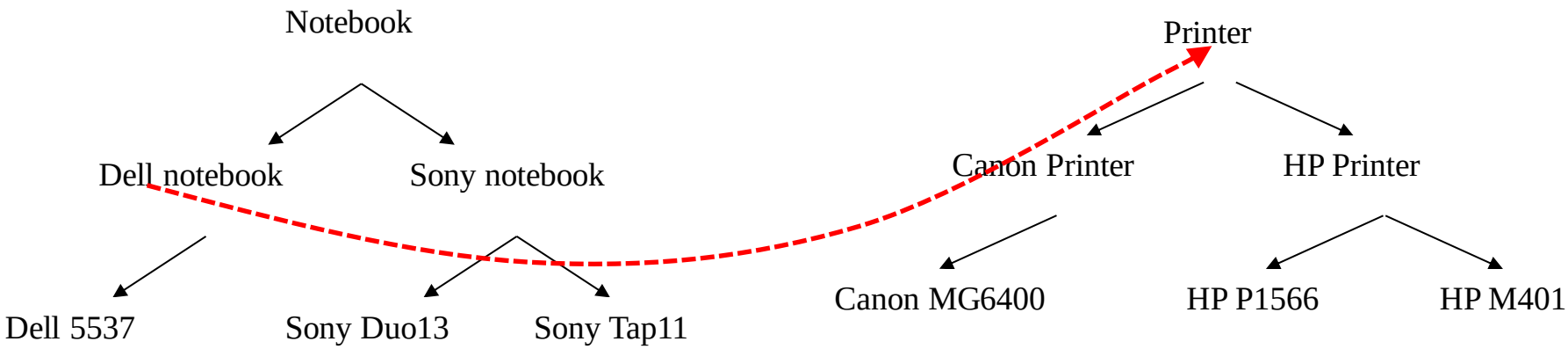
- 基本概念
- 关联规则挖掘的有效方法
- 不同种类的关联规则
- 关联规则的应用
- 总结

关联规则的类型

- **单维度关联规则 v.s. 多维度关联规则**
  - beer → diaper [0.2%, 60%]
  - age=30..39, income=medium → buys_PC=yes [1%, 75%]
- **单层次关联规则 v.s. 多层次关联规则**
  - 什么品牌的啤酒和尿布有关联?
- **反向关联规则**
  - play basketball → not eat cereal [20%, 33.3%]
- **数量关联规则**
  - buy 1 apple → buy 2 bananas [5%, 66.6%]
- **时序关联规则**
  - 买电脑的人3天后会买音响 [13%, 51%]

Multiple-Level Association Rules

- 项有概念层次性
- 低层的项通常具有较低的支持度
- 将项抽象到一定高的层次产生的规则更有意义
- 一个超市的库存中至少有10000个项



关联分析（Association Analysis）

- 基本概念
- 关联规则挖掘的有效方法
- 不同种类的关联规则
- **关联规则的应用**
- 总结

关联规则挖掘的应用

- **购物篮分析（Basket data analysis），交叉销售（cross-marketing）等**
  - 家电(Home Electronics) → * (商店需要储备其它什么产品?)
  - 设计商店的内部商品摆放布局：close proximity or at opposite ends of the store
  - 对哪些商品降价销售: computer ->printer, reduce the price of printer

网络日志(Web Log)：点击流(Click stream)

股票分析（Stock Analysis）

Day	stock
day1	A, B, C, D
day2	C, D, E
day3	A, B
day4	A, B, E, F
day5	C, D, E

- Minsup=40%, minconf=60%
- CD→E （40%, 67%)
- C→D （60%, 100%)
- A→B（60%, 100%)

Both A and B up implies E up in the next day with 100% confidence

股票分析（Stock Analysis）

Obs	Hstkcd	exchflg	Dt	Openpr	Highpr	Lowpr	Closepr	Closeprlast
1	600001	1	22JAN1998:00:00:00	8.00	8.49	7.88	7.91	.
2	600001	1	23JAN1998:00:00:00	7.92	8.17	7.91	8.11	7.91
3	600001	1	09FEB1998:00:00:00	8.31	8.35	8.01	8.04	8.11
4	600001	1	10FEB1998:00:00:00	8.01	8.06	7.90	7.94	8.04
5	600001	1	11FEB1998:00:00:00	7.95	8.17	7.89	8.06	7.94
6	600001	1	12FEB1998:00:00:00	8.06	8.06	7.89	7.94	8.06
7	600001	1	13FEB1998:00:00:00	7.91	7.94	7.80	7.82	7.94
8	600001	1	16FEB1998:00:00:00	7.79	7.87	7.66	7.77	7.82
9	600001	1	17FEB1998:00:00:00	7.74	7.89	7.71	7.78	7.77
10	600001	1	18FEB1998:00:00:00	7.86	7.86	7.67	7.72	7.78
11	600001	1	19FEB1998:00:00:00	7.73	7.75	7.58	7.71	7.72

HSTKCD	六位股票代码	600094
EXCHFLG	标识是否为交易日	1
DT	交易日期	2003-11-6
OPENPR	当日开盘价	5.14
HIGHPR	当日最高价	5.15
LOWPR	当日最低价	4.92
CLOSEPR	当日收盘价	4.96
CLOSEPRLAST	前一交易日收盘价	5.14

股票分析（Stock Analysis）

- **Up, down, stable**

$$R = \frac{(closepr - closeprlast)}{closepr}$$

  - **Up: R>α**
  - **Down: R<β**
  - **Stable: others**

Date	HSTKCD	Price change
2010-9-10	601234	up
2010-9-10	601235	down
2010-9-10	601236	up
…		
2010-9-20	601235	down

Date	HSTKCD
2010-9-10	601234, 601236
2010-9-11	601235
…	
2010-9-20	601235

股票分析（Stock Analysis）

- 滑动窗口(Sliding window)

Date	HSTKCD
2010-9-10	601234, 601236
2010-9-11	601235
…	
2010-9-20	601235

Day	stock
day1	A, B, C, D
day2	C, D, E
day3	A, B
day4	A, B, E, F
day5	C, D, E

股票分析：思考题 & 选做作业

- If we want to find rules like “ A up B up in a day then E up in the following day, what should we do?
- 请思考并设计一种能够有效发现此类时序关联规则的方法
- （提示：最好能够利用已有的Apriori算法）

Day	stock
day1	A, B, C, D
day2	C, D, E
day3	A, B
day4	A, B, E, F
day5	C, D, E

思考题 & 选做作业

A shop is selling four types of electronic products: TV, DVD, WM (Washing Machine), and LC (Laptop Computer). The sales records for a same customer were maintained as shown in Figure 3：

Transaction	Product
DAY#1	WM，TV，DVD
DAY #2	WM
DAY #3	WM，TV，LC，DVD
DAY #4	TV
DAY #5	WM，DVD
DAY #6	LC，DVD
DAY #7	WM，TV，DVD
DAY #8	DVD
DAY #9	WM，TV，DVD
DAY #10	LC

If the manager is interested in the association rules with time lag, such as $X \Rightarrow_t Y$ (expressing that "buying X is associated with buying Y on the next t^{th} day"), what is the value for Dconf($DVD \Rightarrow_2 DVD$)? Please briefly describe a possible way to find such rules using the Apriori method (note: please only describe the ideas, and it is not necessary to find all rules)?

本次课程小结

- 关联规则挖掘相关的基本概念
- 关联规则挖掘的有效方法
- 不同种类的关联规则
- 关联规则的应用

期末课程论文说明

- **主题要求**
  - 必须与“大数据管理”相关
  - 建议围绕所学专业背景下的“大数据管理问题”展开
- **内容要求**
  - 不少于4000字，版式：word中正文小四字体，1.5倍行距
  - 独立完成，不得大段拷贝或直接引用网上、书上及他人已发布内容，需要适当引用时请在引用位置注明参考文献来源（查重）
  - 论文内容框架（建议）：
    - 1. 学习本课程的心得体会、感受，对本课程教学的建议和意见（必有）
    - 2. 论文背景介绍
    - 3. 论文涉及的大数据问题及管理需求、策略和意义（可举实例说明）
    - 4. 本人对该大数据问题的看法、观点及讨论
    - 5. 总结
    - 6. 参考文献和资料

期末课程论文说明（续）

- **论文提交要求**
  - 需要以电子版提交，建议提交word版本
  - 作业提交邮箱：bigdata_homework@163.com
  - 作业提交截止时间：**第19周周日（2015.01.11）24时**
- **其他说明**
  - **电子版论文文件请务必按照“学号_班级_姓名.docx”命名，例如“2014211234_2014212103_张三.docx”，也请在邮件中留下姓名、学号及联系方式，以备论文有问题时能够联系到；**
  - 请在截止时间之前提交论文（不要在截止时间附近，以避免系统原因过期），过期将不再接收论文提交，成绩为0，请务必注意；
  - 每次提交论文后，作业邮箱都会有“已收到邮件”的自动回复，如未收到自动回复，表示发送不成功，请在截止时间内重新提交；
  - 论文评分的关注重点
    - 有效的课程建议和意见
    - 关注问题的新颖度
    - 个人分析和讨论的深度
    - 论文的整体工作量