第九课（第25-27课时）

事务型数据和关联分析

- 事务型数据和关联规则
- 频繁项集
- Apriori
- FP-Growth树
- 关联规则的产生
- 关联分析的评价

前情回顾

- **分析单个变量：各种方法**
- **分析多个变量：各种方法**
- **回归分析和广义线性模型：确认变量之间的关系**
  - 解释和预测
- **分类分析：预测类别型因变量，有监督学习**
- **聚类分析：无监督学习，发现数据点之间的关系**
- **基于重抽样：**
  - 统计量的显著性检验和区间估计（permutation test, Bootstrap）
  - 增强训练效果和评价的稳定性（CV，Bagging， Boost..）
- **模型选择：**
  - 拟合度，查准率，查全率，ROC

事务型数据和关联分析：任务描述

- 理解、掌握从事务型数据中确认频繁项集
- 理解Apriori和FP-Growth Tree算法
- 掌握提取频繁项集
- 了解对关联分析的评价方法

关联规则的挖掘

- 根据事务记录中某些项的出现来预测其他项的出现
  - 用于发现隐藏在大型数据集中的有意义的联系

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

关联规则示例

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

暗示了某些项的同时出现并非随机

定义：频繁项集

- **项集 Itemset**
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - 包含k 个项的集合
- **支持度计数 Support count (σ)**
  - 项集在数据记录中出现的频数
    - E.g.   σ({Milk, Bread,Diaper}) = 2
- **支持度 Support**
  - 项集在数据中出现的比例
    - E.g.   s({Milk, Bread, Diaper}) = 2/5
- **频繁项集 Frequent Itemset**
  - 支持度超过***minsup*** 的项集
    - **最小支持度阈值**

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

定义：关联规则 Association Rule

小象学院 ChinaHadoop.cn

- **关联规则Association Rule**
  - X → Y, X 和 Y 均为项集
  - 关联规则的表示形式：X，Y无交集
  - Example:
    {Milk, Diaper} → {Beer}

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

- **关联规则的度量**
  - Support (s) 支持度
    - 同时包含 X 和Y的项集的比例
  - Confidence (c) 置信度
    - 包含X的项集中同时出现Y的比例

例:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

关联规则挖掘的任务

- **找到所有的规则，满足支持度和置信度的要求**
  - support ≥ *minsup* threshold
  - confidence ≥ *minconf* threshold

- **暴力方法：逐个规则**
  - 找出所有的关联规则
  - 逐一计算每个规则的支持度和置信度
  - 去除不满足要求的规则（支持度和置信度分别小于*minsup*和*minconf*阈值）

  ⇒ 然而是高消费计算资源的!

挖掘关联规则

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

Example of Rules:

{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)

Observations:

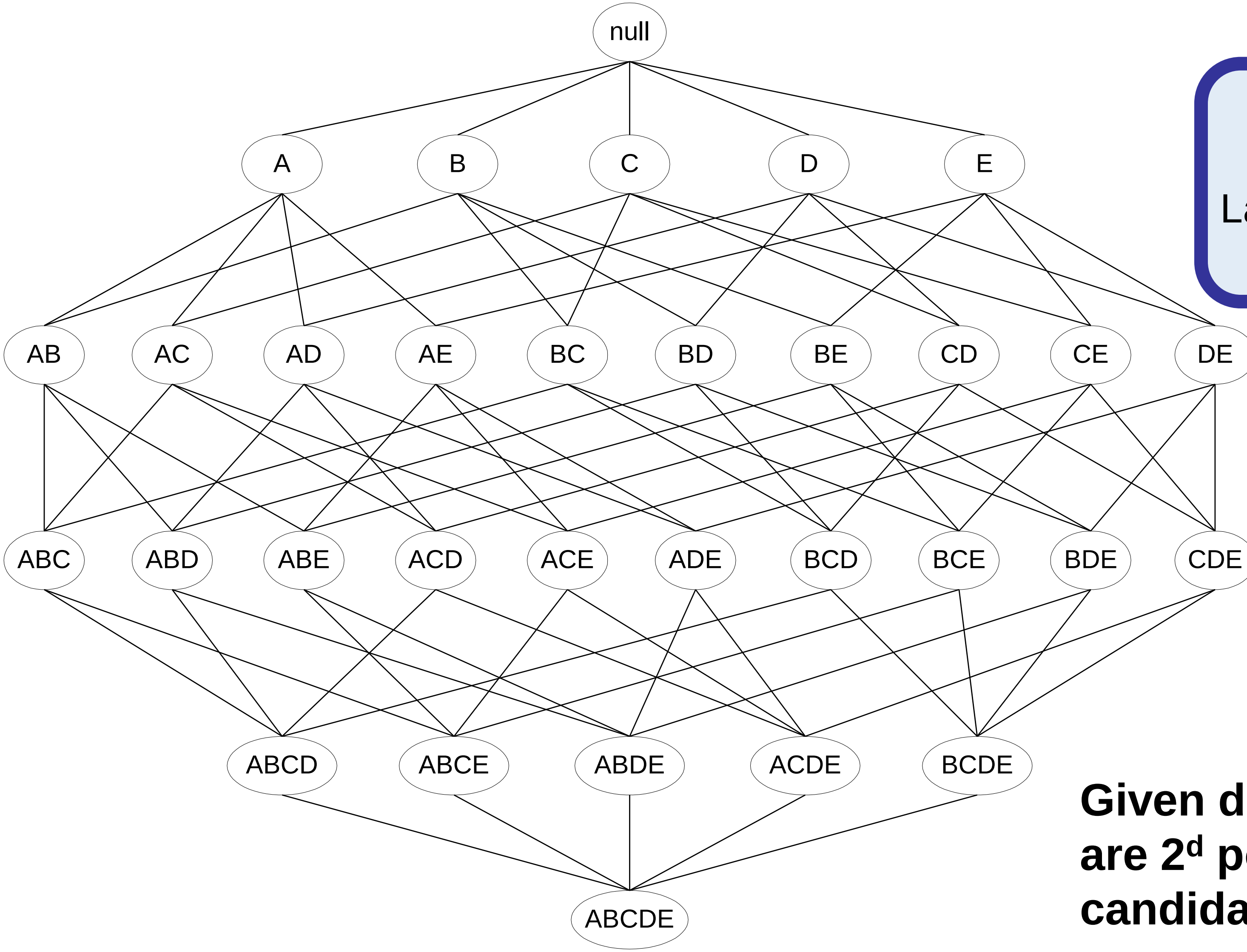
- 以上规则均源于同一个项集 {Milk, Diaper, Beer}
- 以上规则支持度相同，置信度不同
- 所以，要分开处理支持度和置信度问题

挖掘关联规则

- **Two-step approach:**
  1. 频繁项集产生 Frequent Itemset Generation
     – Generate all itemsets whose support ≥ minsup
  2. 规则的产生 Rule Generation
     – Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- **Frequent itemset generation is still computationally expensive**

频繁项集的产生 Frequent Itemset Generation



格结构
Lattice Structure

Given d items, there are 2^d possible candidate itemsets

频繁项集的产生

- **暴力方法 Brute-force approach:**
  - 数据集中的每个项集都可以是频繁项集
  - 对每个项集，计算其在数据记录中的出现频次

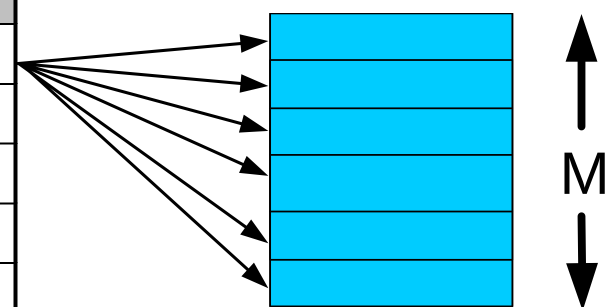
Transactions

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

N

w

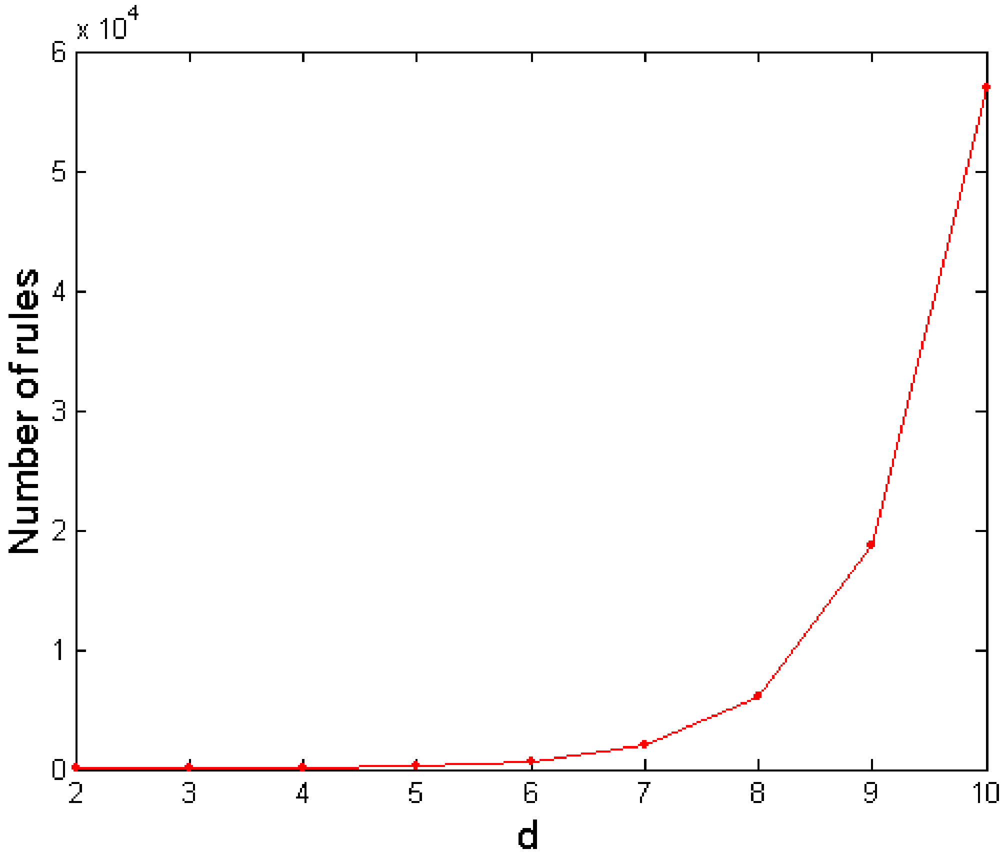
List of Candidates



  - 对每个可能的项集在每条记录中进行比对
  - 复杂度 ~ O(NMw) => 高，$M = 2^d$!!! (d是“商品”的种类数)

计算复杂度 Computational Complexity

- **给定d个不同的项:**
  - 项集的个数= 2^d
  - 则关联规则的个数 为:



$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$

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

If d=6, R = 602 rules

频繁项集产生策略
Frequent Itemset Generation Strategies

- **减少备选项集的数量 (M)**
  - 暴力法: $M=2^d$
  - 如何减少 M
- **减少 交易数目 (N)**
  - ？
- **减少 比对次数 (NM)**
  - 使用更有效的数据结构来存储项集和交易
  - 不需要一 一进行比对

通过减少候选项集数目
Reducing Number of Candidates

- **Apriori 原则:**
  - 频繁项集的子集也频繁

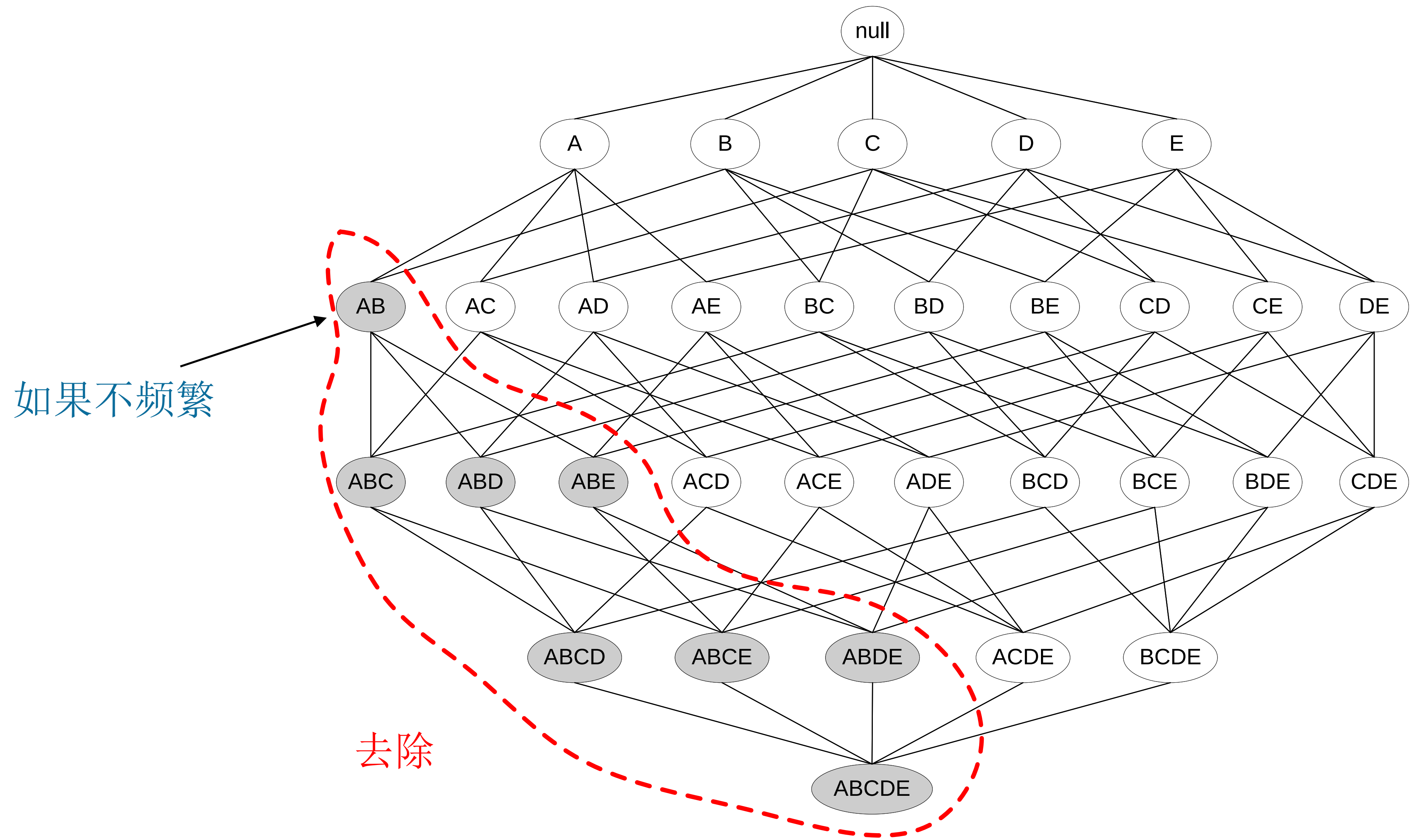
- **Apriori 基于项集支持度support的属性:**

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

  - 项集的支持度总不会大于其子集的支持度
  - support的反单调性

图示：Illustrating Apriori Principle

小象学院 ChinaHadoop.cn



Apriori 原理

小象学院 ChinaHadoop.cn

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

Items (1-itemsets)

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

Pairs (2-itemsets)

(舍弃包含Coke或Eggs的项集)

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	3

MinSup= 3

If every subset is considered,

$^6C_1 + {}^6C_2 + {}^6C_3 = 41$

With support-based pruning,

$6 + 6 + 1 = 13$

算法：Apriori Algorithm

小象学院 ChinaHadoop.cn

- **Method:**
  - Let k=1
  - 生成1-频繁项集
  - 重复直至无频繁项集被发现
    - 从k频繁项集产生 k+1频繁项集
    - 去除包含k-不频繁项集的项集
    - 计算支持度
    - 去除非频繁项集

算法实现

小象学院
ChinaHadoop.cn

- **示例代码和数据**
  - 链接: http://pan.baidu.com/s/1o8fDDZO 密码: uzki
  - 代码来自 《机器学习实战》11/12章

小象学院 ChinaHadoop.cn

减少比较次数
Reducing Number of Comparisons

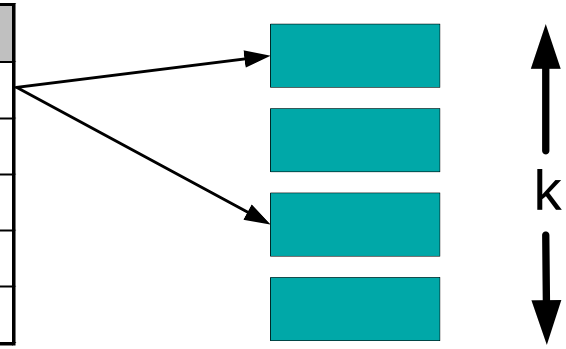
- **Candidate counting:**
  - 扫描数据记录，统计每个项集的支持度
  - 为了减少比对次数，将数据存储与哈希表中

Transactions

N

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

Hash Structure



Buckets

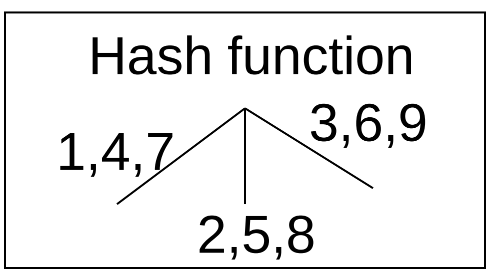
Hash(哈希）树的产生

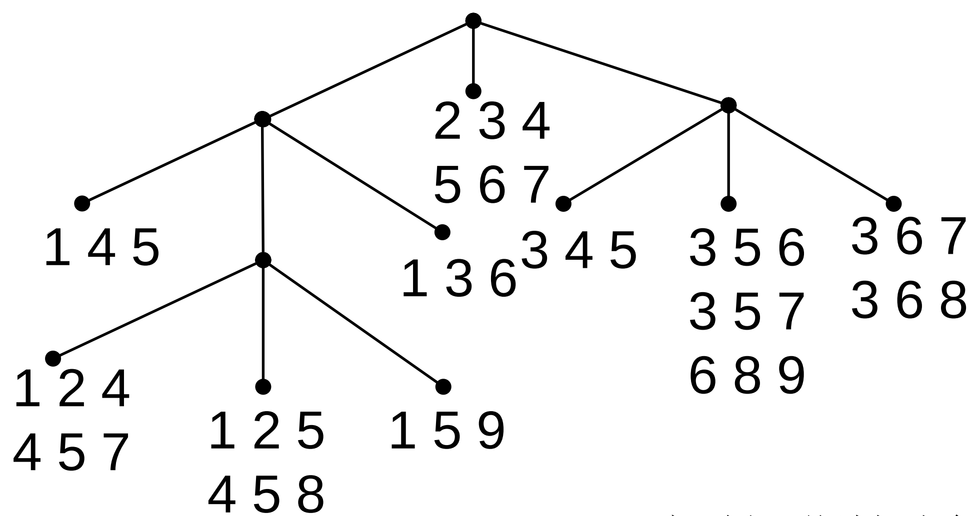
15 个长度为3的备选项集:

{1 4 5}, {1 2 4}, {4 5 7}, {1 2 5}, {4 5 8}, {1 5 9}, {1 3 6}, {2 3 4}, {5 6 7}, {3 4 5}, {3 5 6}, {3 5 7}, {6 8 9}, {3 6 7}, {3 6 8}

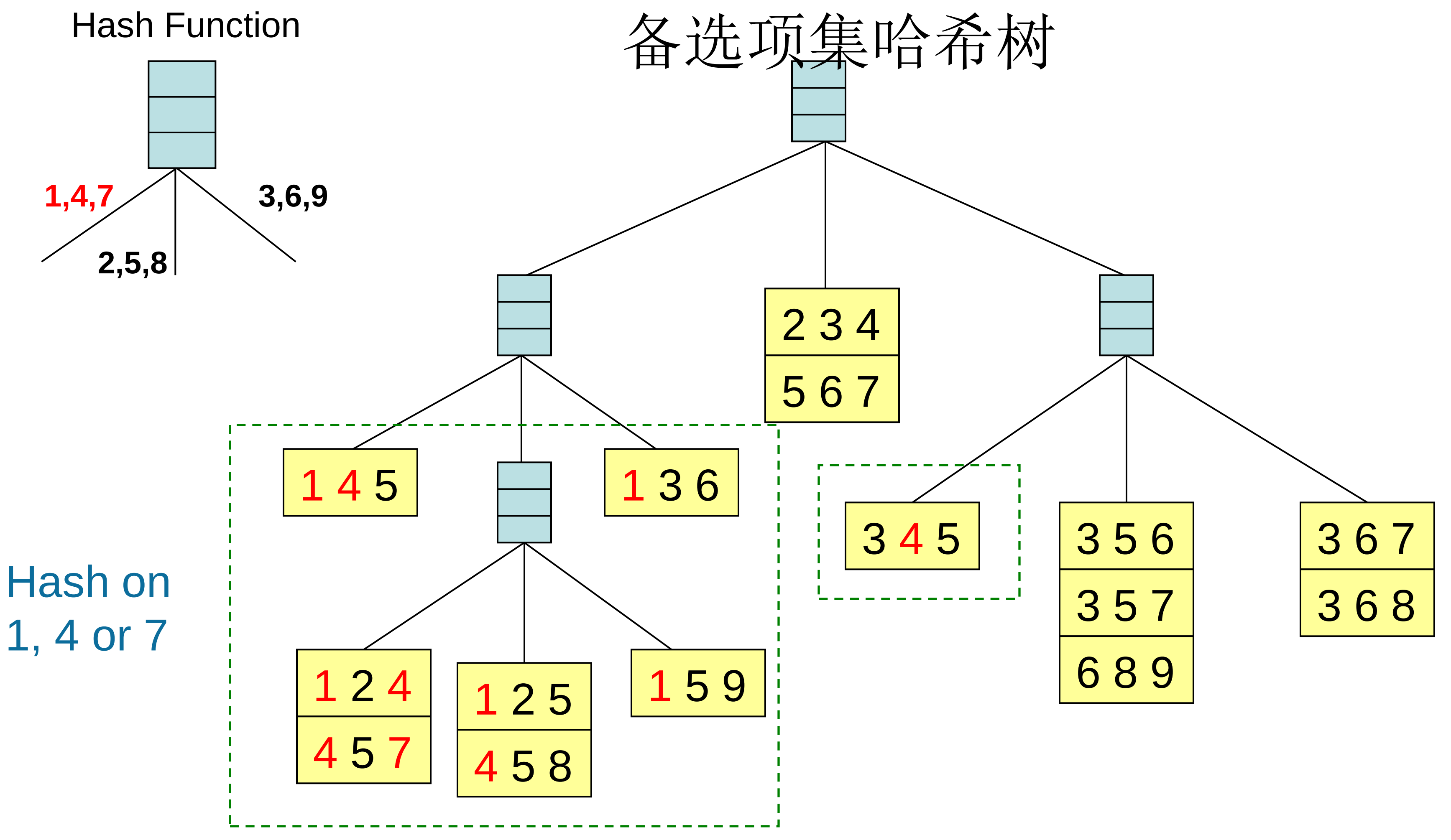
需要:

- **哈希函数**
- **最大叶子数目: 当某叶节点备选项集数目超过此数，分裂该叶节点**

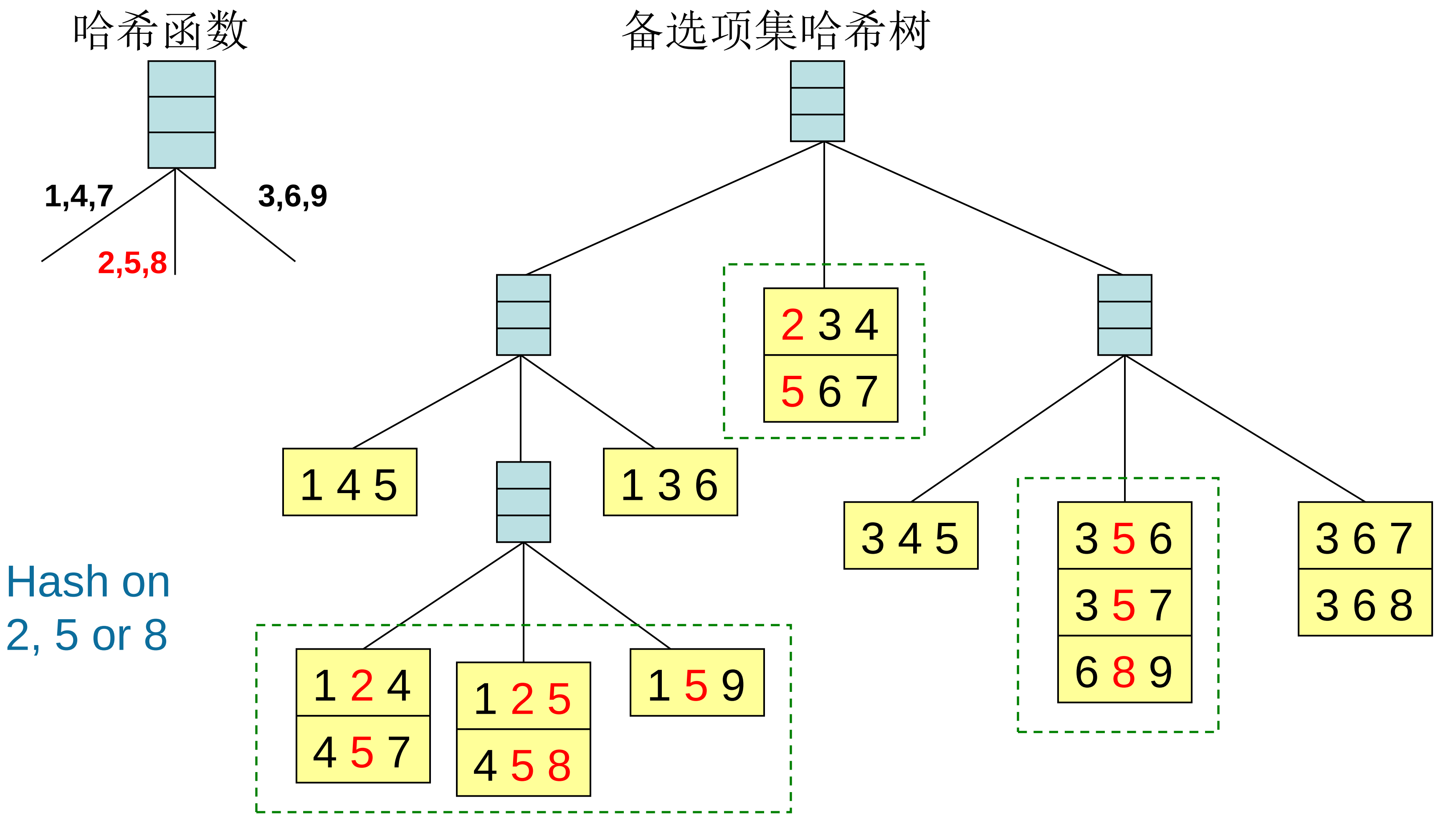




哈希树

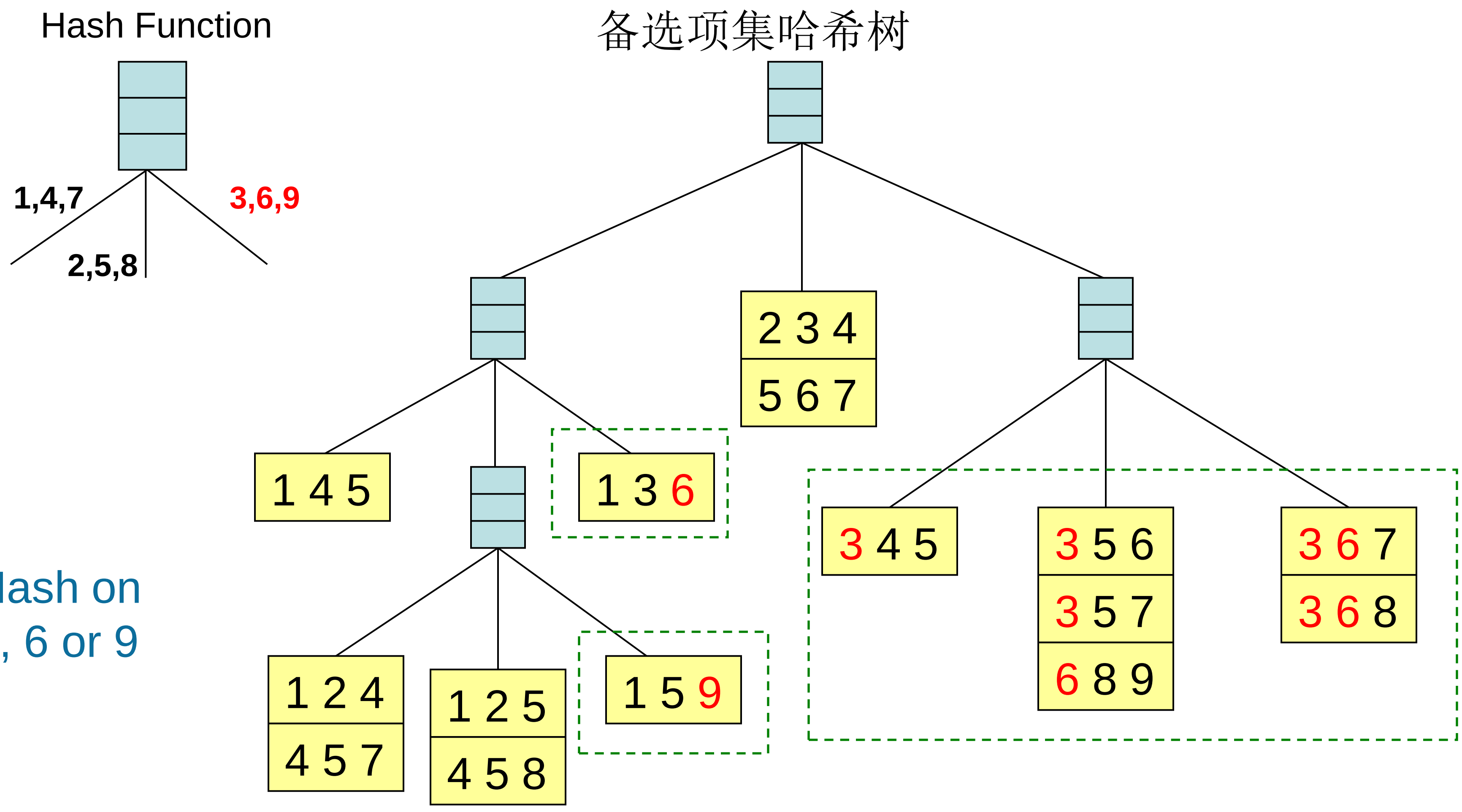

Hash Function
1,4,7
2,5,8
3,6,9
备选项集哈希树
2 3 4
5 6 7
1 4 5
1 3 6
3 4 5
3 5 6
3 5 7
6 8 9
3 6 7
3 6 8
Hash on
1, 4 or 7
1 2 4
4 5 7
1 2 5
4 5 8
1 5 9

哈希树



哈希函数
1,4,7
2,5,8
3,6,9
备选项集哈希树
2 3 4
5 6 7
1 4 5
1 3 6
3 4 5
3 5 6
3 5 7
6 8 9
3 6 7
3 6 8
1 2 4
4 5 7
1 2 5
4 5 8
1 5 9
Hash on
2, 5 or 8

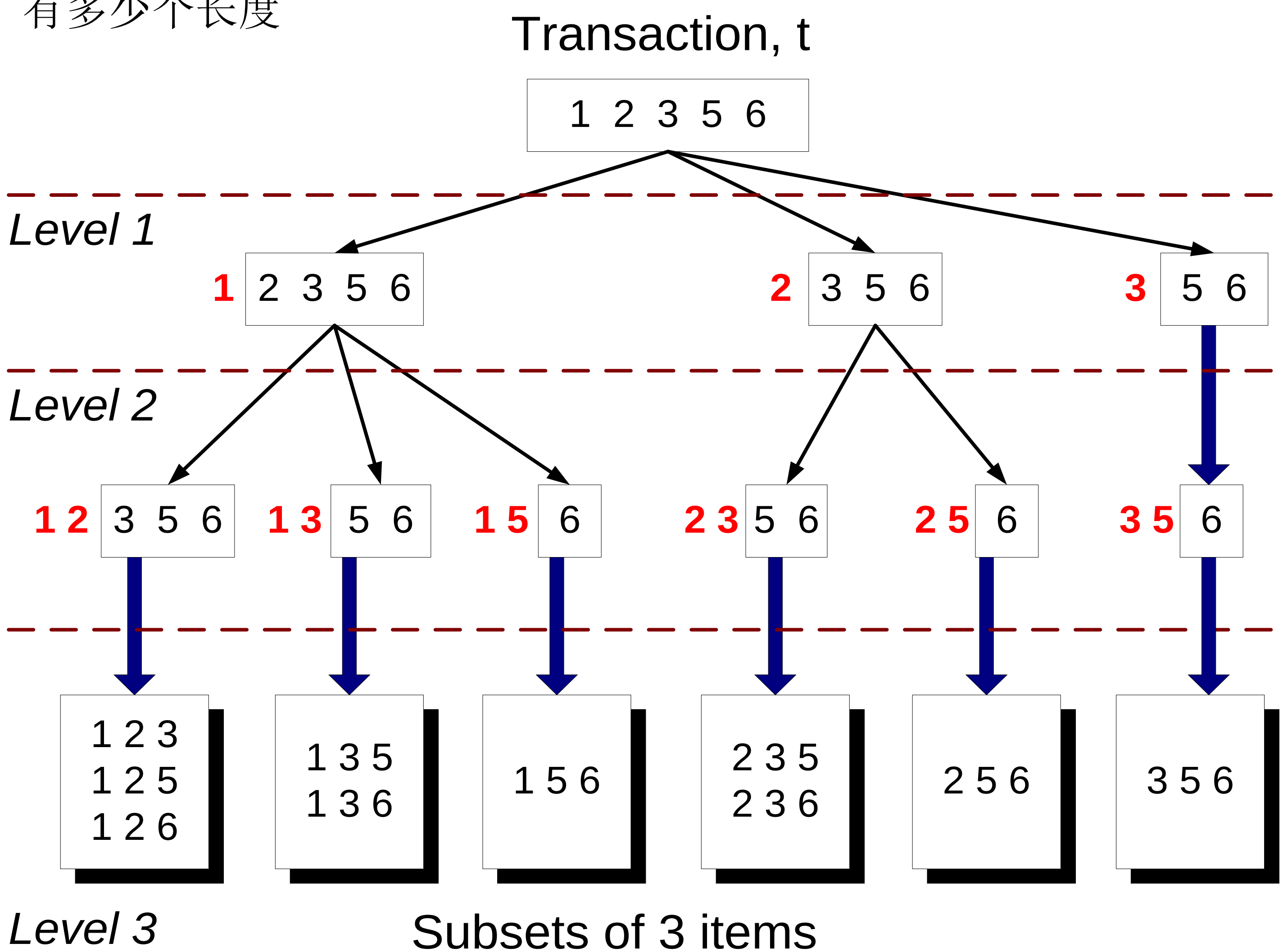
哈希树



Hash on
3, 6 or 9

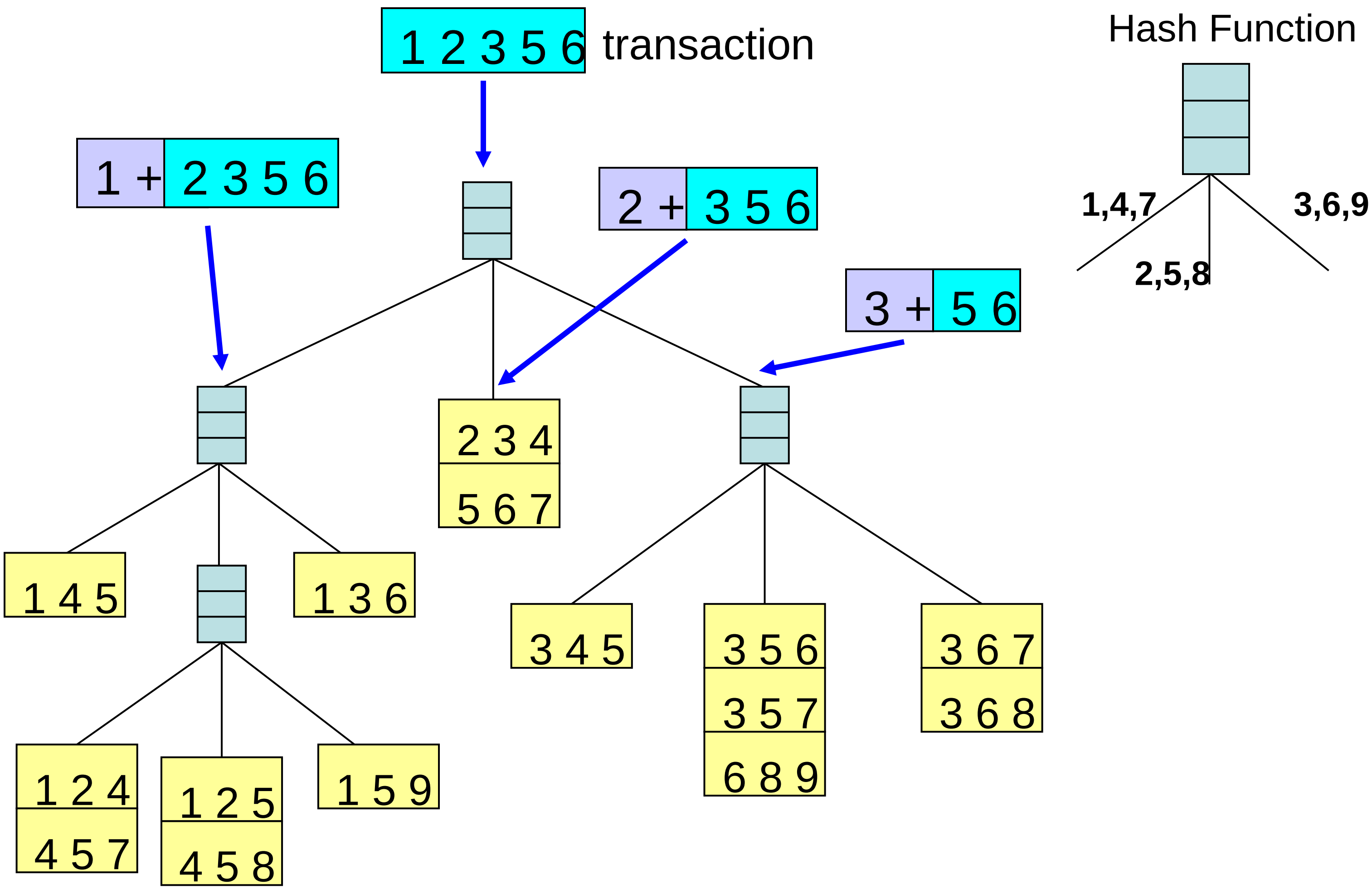
子集操作

对特定的交易，有多少个长度为3的项集？



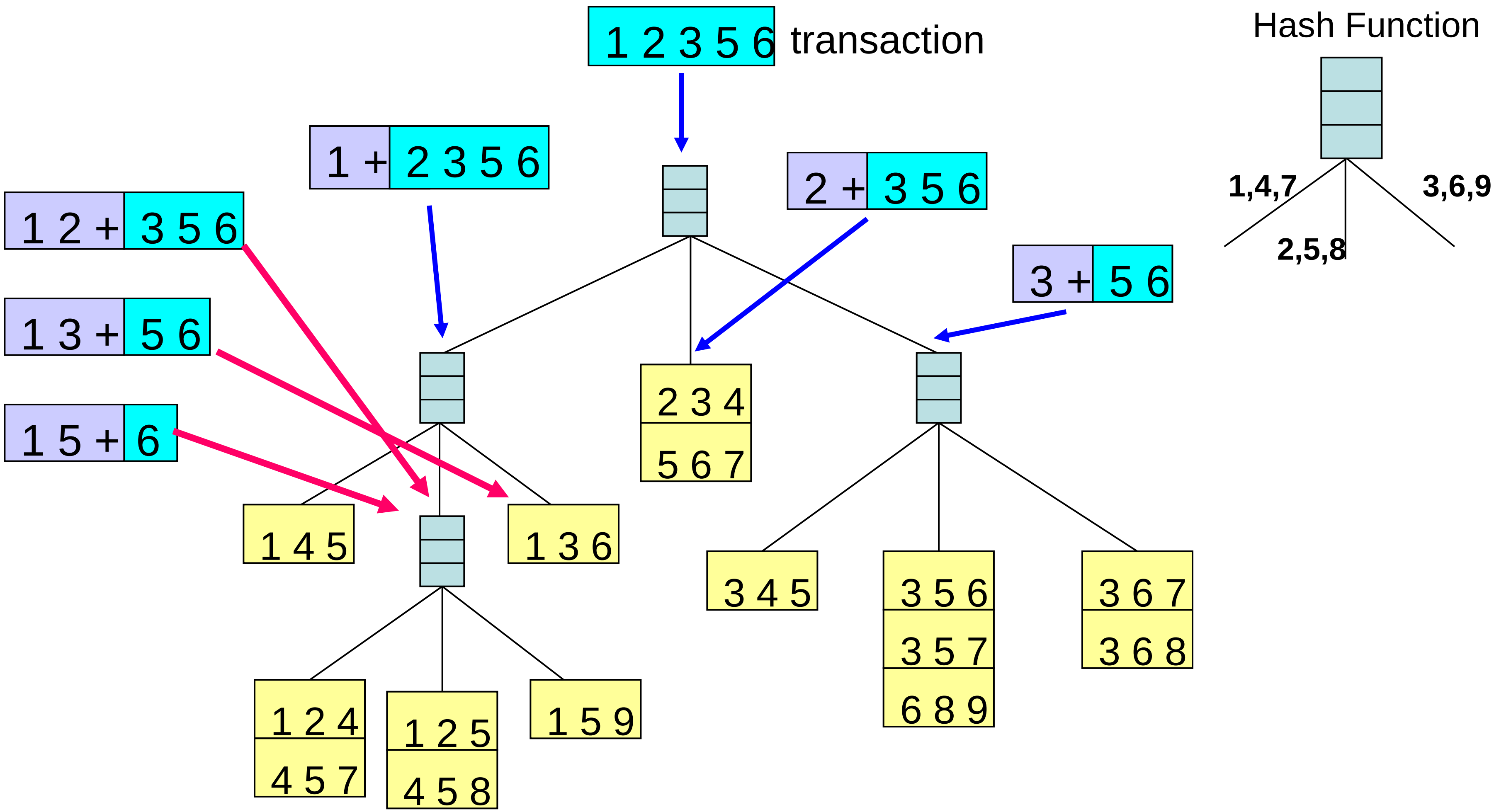
使用哈希树对子集进行操作

小象学院 ChinaHadoop.cn

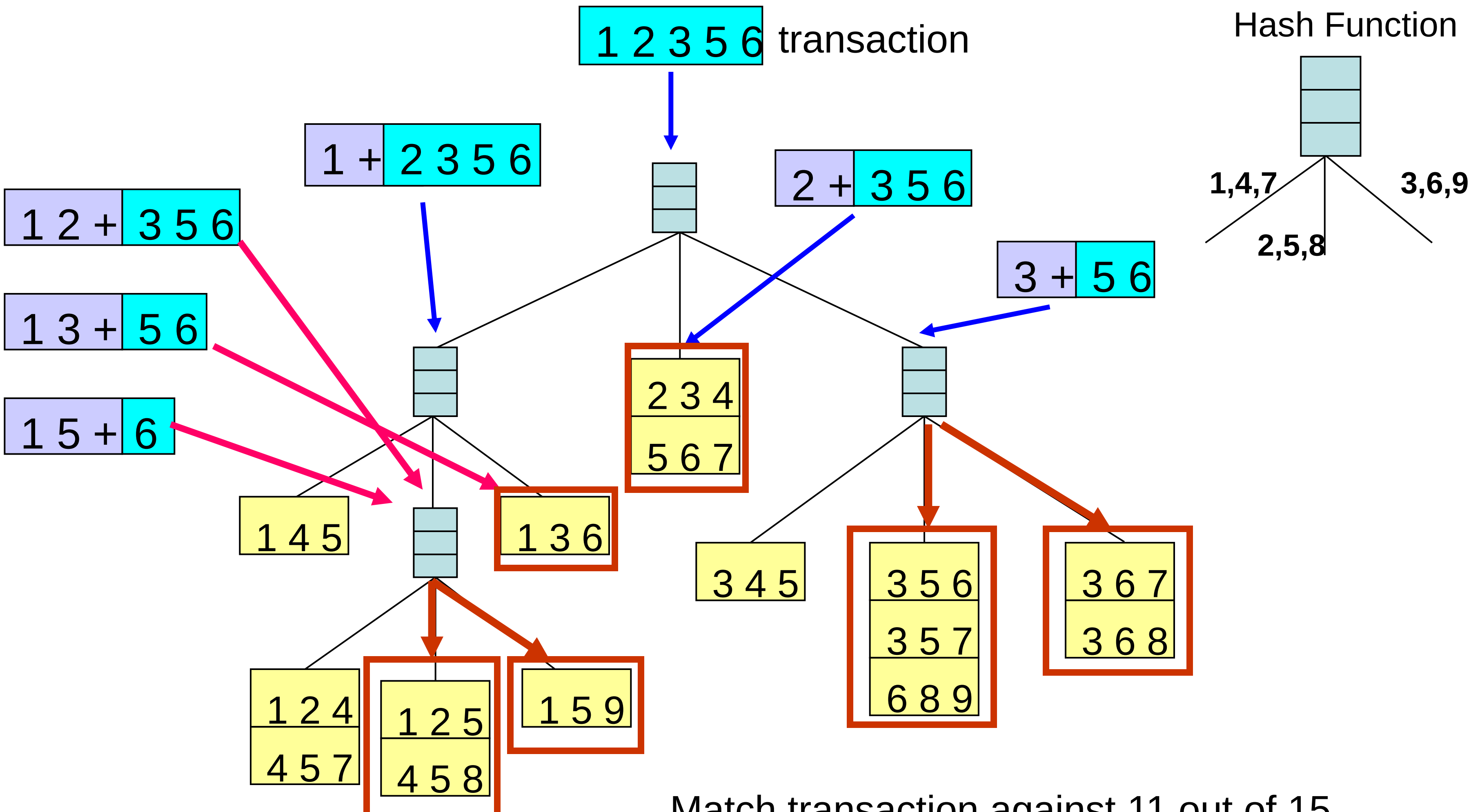


使用哈希树对子集进行操作

小象学院 ChinaHadoop.cn



使用哈希树对子集进行操作



Match transaction against 11 out of 15 candidates

有关计算复杂度

- **最小支持度的选择**
  - 降低minsup：产生更多的频繁项集，增加频繁项集的最大长度
- **维度 (项的数目)**
  - 影响存储、计算和I/O
- **数据库的大小（记录数目）**
  - 运行时间
- **平均记录的宽度**
  - 密集的数据集具有较大的W
  - 影响频繁项集的最大长度，和哈希树遍历的时间

频繁项集的压缩

- 某些频繁项集的支持度如果与其超集相同，则它是冗余的

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

- **Number of frequent itemsets** $= 3 \times \sum_{k=1}^{10} \binom{10}{k}$

- **Need a compact representation**

FP-growth Algorithm

小象学院 ChinaHadoop.cn

- **Apriori：**
  - 不断的构造候选集、筛选候选集挖掘出频繁项集
  - 多次扫描原始数据，当数据较大时，磁盘I/O次数太多，效率比较低下。
- **FP-growth：**
  - 仅扫描原始数据两遍，通过FP-tree数据结构对原始数据进行压缩
- **1. FP-tree构建（类似于前缀树）**
  - 通过两次数据扫描，将原始数据中的事务压缩到一个FP-tree树
- **2. 递归挖掘FP-tree**
  - 通过FP-tree找出每个item的条件模式基、条件FP-tree，递归的挖掘条件FP-tree得到所有的频繁项集

FP-tree构建

➢ **第一遍扫描数据，找出频繁1项集L，按降序排序**

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

1-频繁项集

Item	频次
B	8
A	7
C	7
D	5
E	3

FP-Tree的构建

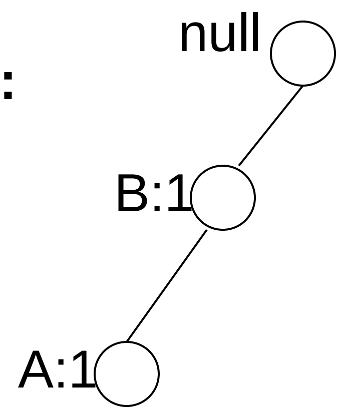
- **第二遍扫描数据：**
  - 对每条记录，过滤其中不频繁集合（minsup），剩下的频繁项集按L顺序排序
    - 阈值假设为3，过滤掉E
  - 把条记录的频繁1项集插入到FP-tree中，相同前缀的路径可以共用
  - 增加一个header table，把FP-tree中相同item连接起来，也是降序排序

FP-tree 的构建

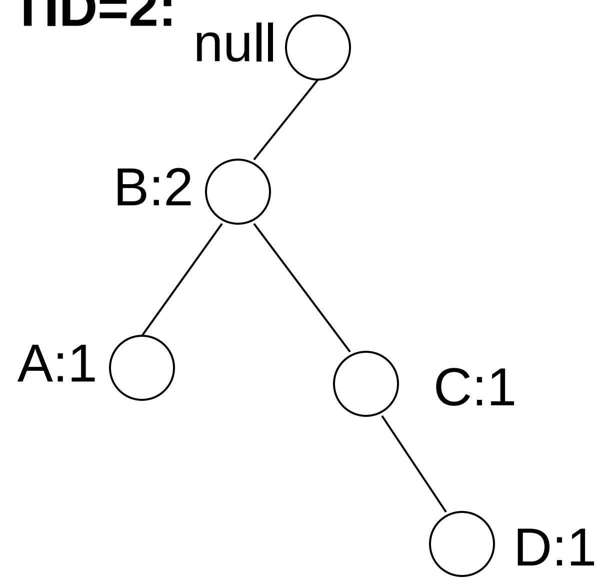
小象学院 ChinaHadoop.cn

TID	Items
1	{B,A}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{B,A,C}
6	{B,A,C,D}
7	{B,C}
8	{B,A,C}
9	{B,A,D}
10	{B,C,E}

After reading TID=1:



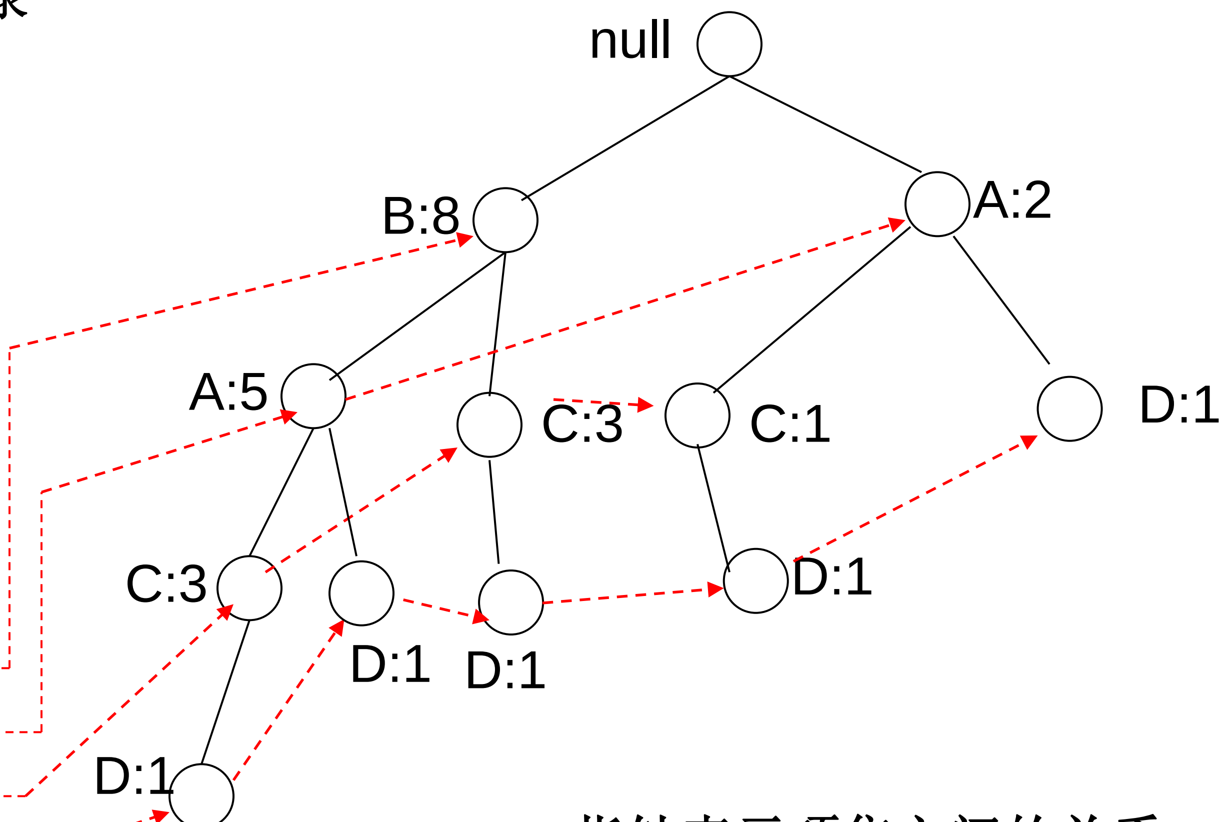
After reading TID=2:



FP-Tree 构建

TID	Items
1	{B,A}
2	{B,C,D}
3	{A,C,D}
4	{A,D}
5	{B,A,C}
6	{B,A,C,D}
7	{B,C}
8	{B,A,C}
9	{B,A,D}
10	{B,C}

交易数据记录



Header table

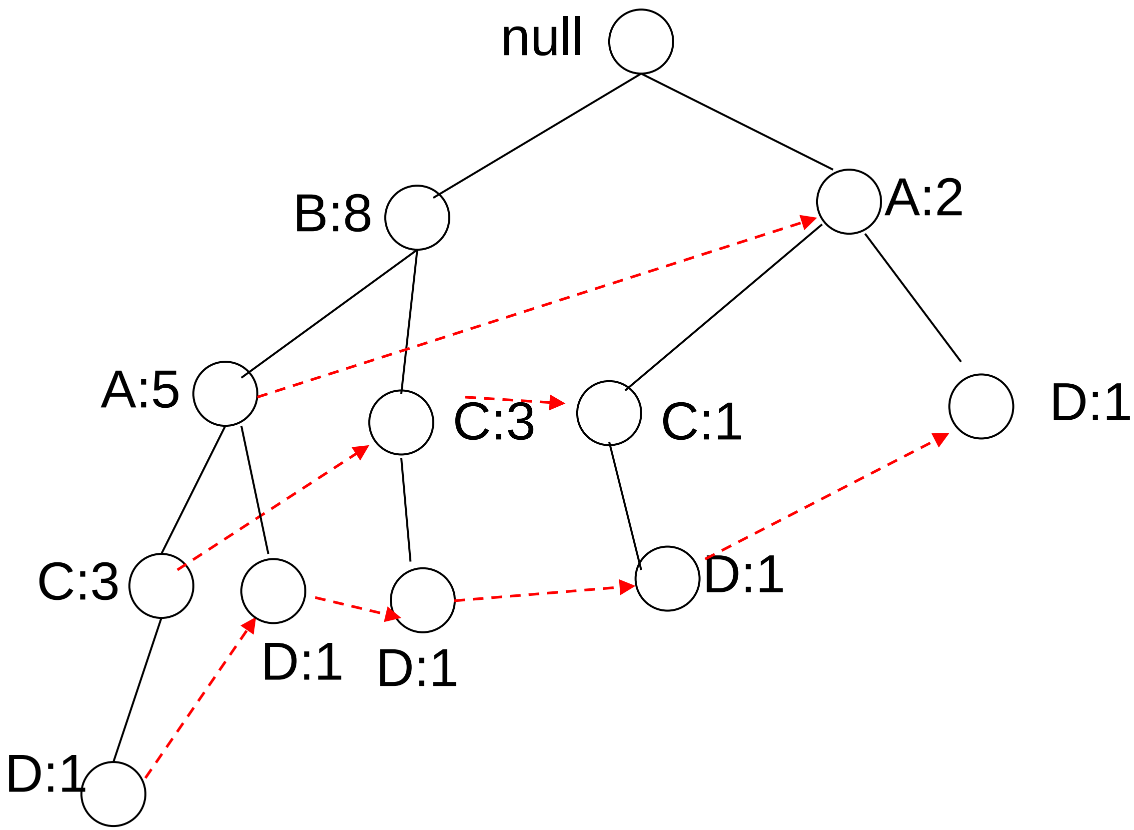
Item	Pointer
B	8
A	7
C	7
D	5
~~E~~	~~3~~

指针表示项集之间的关系

FP-growth:频繁项挖掘

小象学院 ChinaHadoop.cn

从header table的最下面的item开始，
构造每个item的条件模式基（conditional pattern base）



条件模式基 Conditional Pattern	
D	BAC:1,BA:1,BC:1,AC:1,A:1
C	BA:3,B:3,A:2
A	B:5
B	{}

FP-growth:频繁项挖掘

- **构造条件FP-tree（conditional FP-tree）**
  - 累加每个CPB上的item的频数，过滤低于阈值的item，构建条件FP-tree
  - 阈值假设为3

条件模式基 Conditional Pattern	
D	BAC:1,BA:1,BC:1, AC:1,A:1
C	BA:3,B:3,A:2
A	B:5
B	{}

{}

B:3

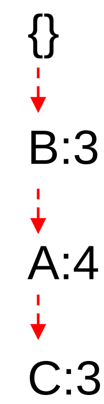
A:4

C:3

FP-growth:频繁项挖掘

- **递归的挖掘每个条件FP-tree**
  - 累加后缀频繁项集，直到找到FP-tree为空或者FP-tree只有一条路径（只有一条路径情况下，所有路径上item的组合都是频繁项集）

条件模式基 Conditional Pattern	
D	BAC:1,BA:1,BC:1, AC:1,A:1
C	BA:3,B:3,A:2
A	B:5
B	{}



规则的产生 Rule Generation

- **给定频繁项集L, 需要找到f ⊂ L ，是的f → L – f 满足最小置信度的要求**
  - 例： 频繁项集{A,B,C,D} 对应的规则

    ABC →D, ABD →C, ACD →B, BCD →A,
    A →BCD, B →ACD, C →ABD, D →ABC
    AB →CD, AC → BD, AD → BC, BC →AD,
    BD →AC, CD →AB,

- **如果|L| = k，则2^k – 2 条候选规则！！**

规则的产生 Rule Generation

- **是否有更高效的方法？**
  - 通常置信度没有反单调性

    $c(ABC \rightarrow D)$， $c(AB \rightarrow D)$关系不定

  - 但是同一个频繁项集产生的关联规则具有反单调性！
  - e.g., L = {A,B,C,D}:

$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

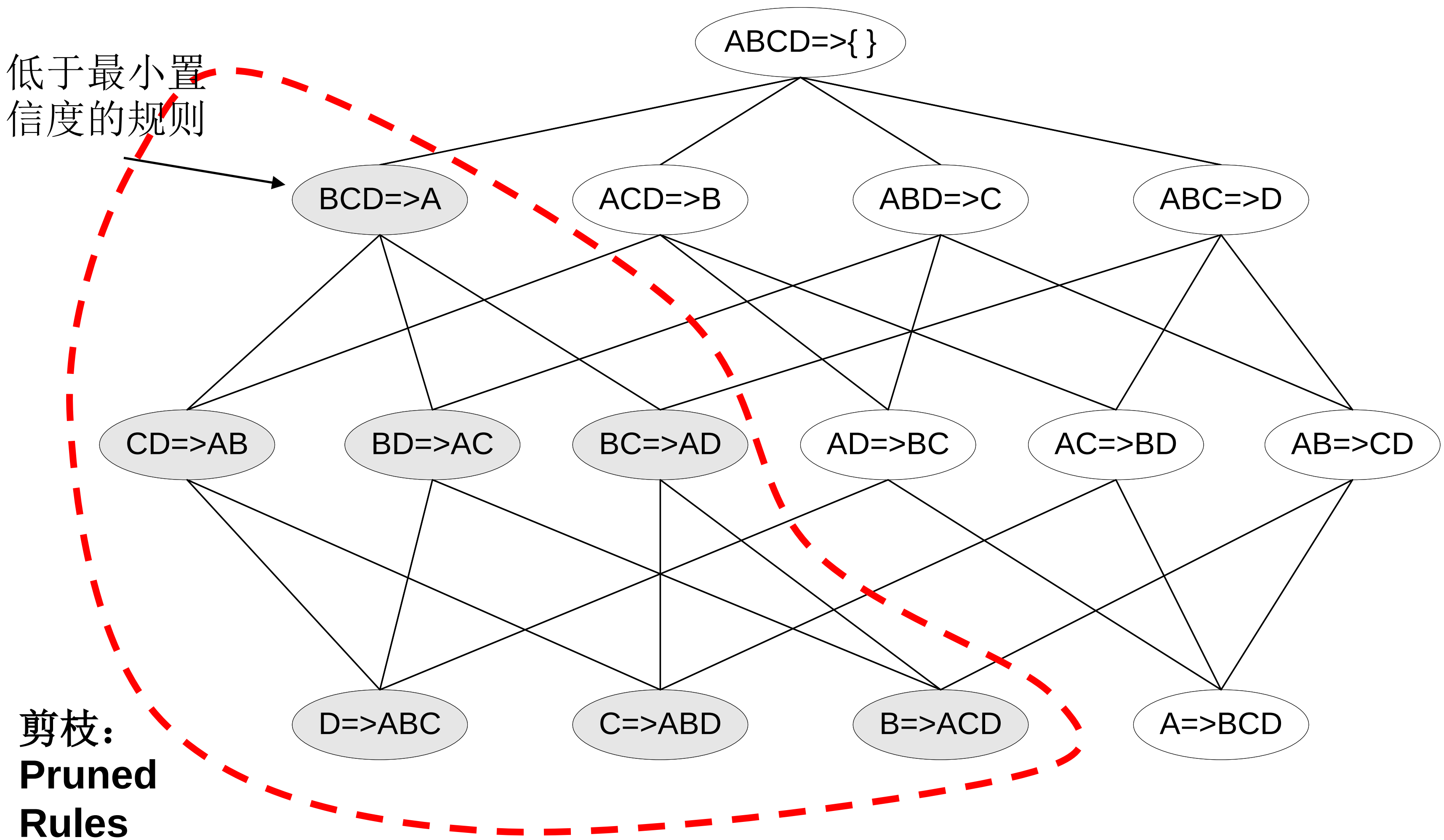
•

$$c = \frac{\sigma(A, \mathrm{B}, \mathrm{C}, D)}{\sigma(\mathrm{A}, \mathrm{B}, \mathrm{C})} \qquad c = \frac{\sigma(A, \mathrm{B}, \mathrm{C}, D)}{\sigma(\mathrm{A}, \mathrm{B})}$$

规则产生的Apriori算法
Rule Generation for Apriori Algorithm

小象学院
ChinaHadoop.cn

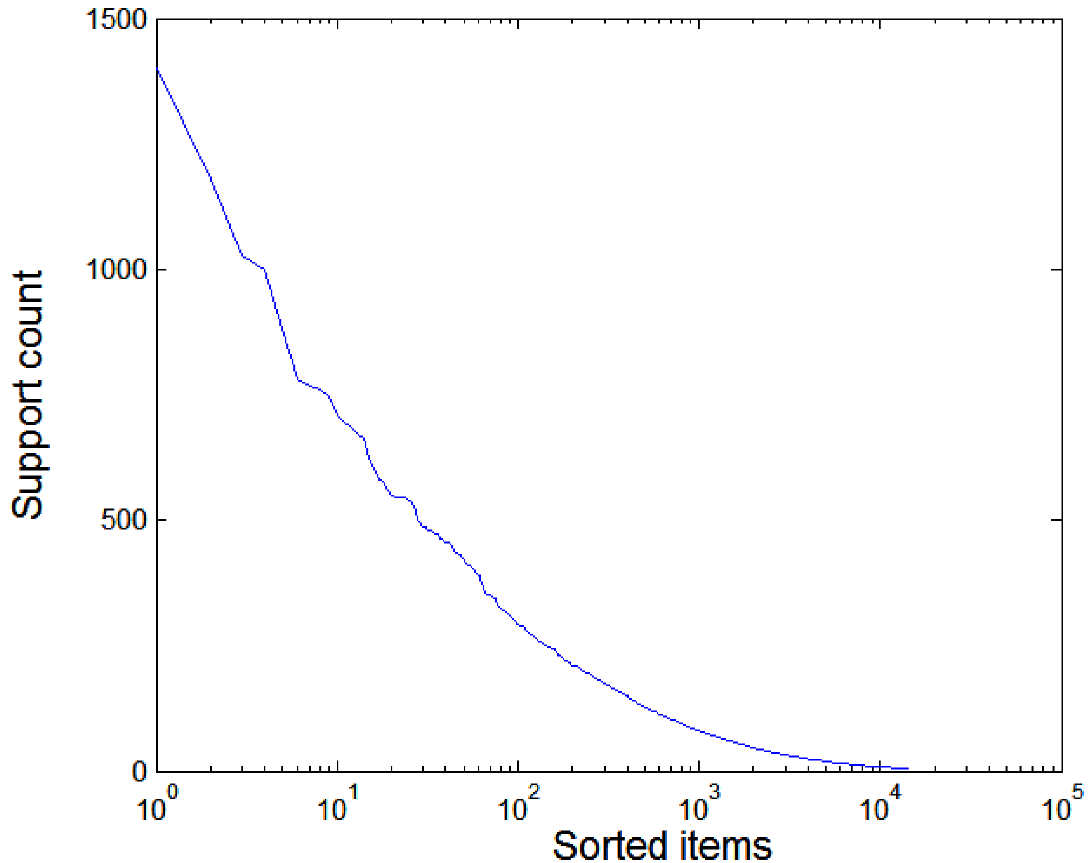
Lattice of rules



支持度的分布

- Many real data sets have skewed support distribution

Support distribution of 零售数据



Effect of Support Distribution

- 设置合理的最小支持度阈值
  - 太高：遗漏
  - 太低：计算复杂
- 或许单一支持度阈值并非高效？

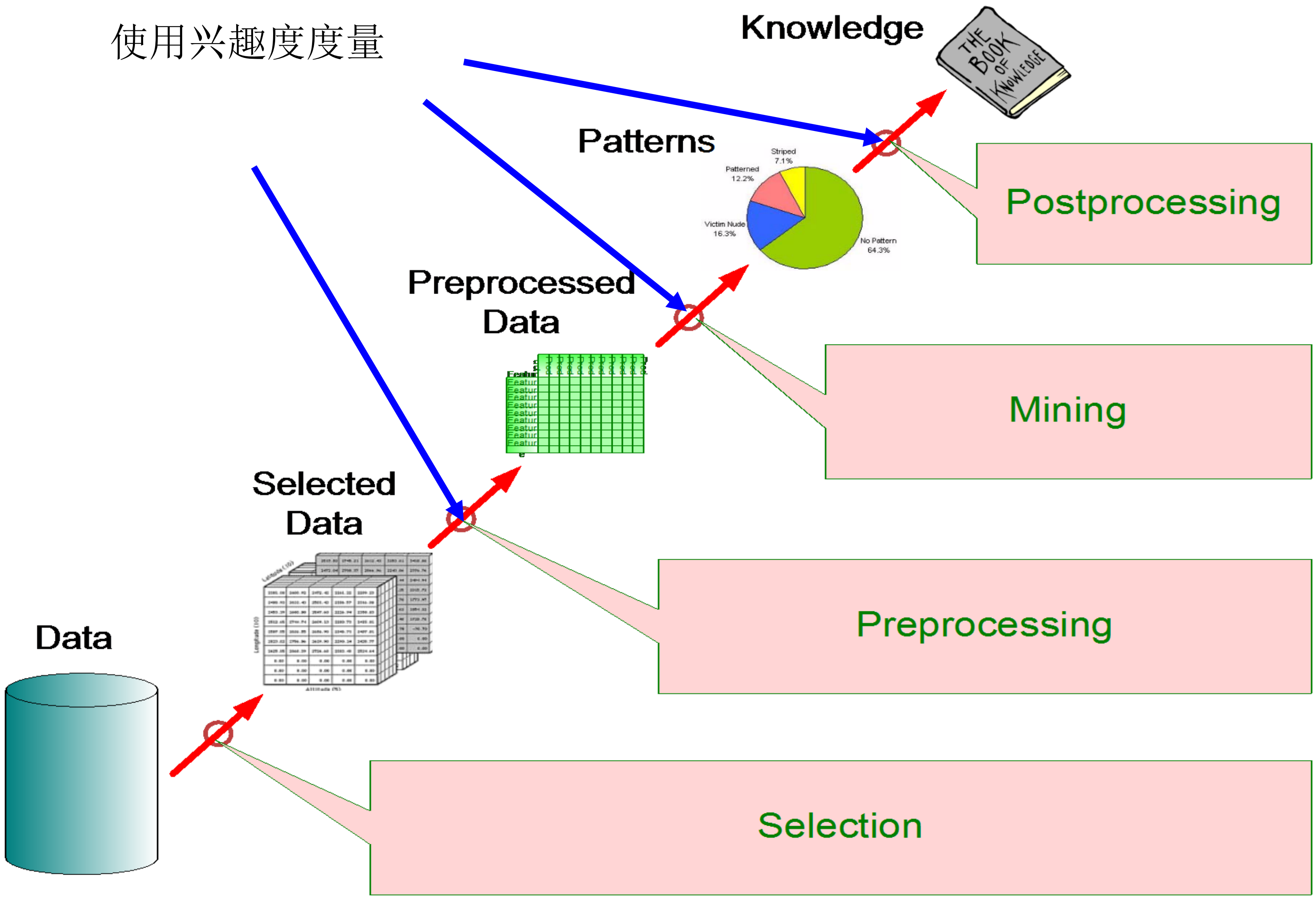
多个Minimum Support

- **多个最小支持度?**
  - MS(i): 不同的物品设置不同的最小支持度
  - e.g.: MS(Milk)=5%, MS(Coke) = 3%, MS(Broccoli)=0.1%, MS(Salmon)=0.5%
  - MS({Milk, Broccoli}) = min (MS(Milk), MS(Broccoli)) = 0.1%
  - 障碍: 支持度不再具有反单调性
    - Suppose: Support(Milk, Coke) = 1.5% and Support(Milk, Coke, Broccoli) = 0.5%
    - {Milk,Coke} is infrequent but {Milk,Coke,Broccoli} is frequent

关联模式的评估 Pattern Evaluation

- **大多数规则无趣、冗余**
  - 若 {A,B,C} → {D} and {A,B} → {D} 具有相同的支持度和置信度，则冗余
- **可用兴趣度来剪枝和排序**

Application of Interestingness Measure



兴趣度的客观度量

小象学院 ChinaHadoop.cn

➢ 使用规则的列联表计算兴趣度

规则$X \rightarrow Y$的列联表

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	$\lvert T \rvert$

f_{11}: support of X and Y
f_{10}: support of X and $\overline{Y}$
f_{01}: support of $\overline{X}$ and Y
f_{00}: support of $\overline{X}$ and $\overline{Y}$

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

置信度的缺陷（支持度的缺陷？）

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

$\Rightarrow$ 置信度高，但是规则并不正确

$\Rightarrow$ $P(\text{Coffee}|\overline{\text{Tea}}) = 0.9375$

统计独立性（相关性）

- **1000 名学生**
  - 600 人会游泳(S)
  - 700 人会骑车(B)
  - 420 人会游泳和骑车(S,B)

  - $P(S \wedge B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$

  - $P(S \wedge B) = P(S) \times P(B)$ => 独立
  - $P(S \wedge B) > P(S) \times P(B)$ => 正相关
  - $P(S \wedge B) < P(S) \times P(B)$ => 负相关

基于统计框架的度量

小象学院 ChinaHadoop.cn

➤ **考虑到相关性**

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

提升度

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

兴趣

$$PS = P(X,Y) - P(X)P(Y)$$

相关分析

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

Example: Lift/Interest

	$Coffee$	$\overline{Coffee}$	
Tea	15	5	20
$\overline{Tea}$	75	5	80
	90	10	100

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

$\Rightarrow$ Lift = 0.75/0.9= 0.8333 (< 1, therefore is negatively associated)

Lift & Interest的缺陷

小象学院 ChinaHadoop.cn

	Y	$\overline{Y}$	
X	10	0	10
$\overline{X}$	0	90	90
	10	90	100

	Y	$\overline{Y}$	
X	90	0	90
$\overline{X}$	0	10	10
	90	10	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Statistical independence:

If P(X,Y)=P(X)P(Y) => Lift = 1

There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

What criteria should we use to determine whether a measure is good or bad?

What about Apriori-style support based pruning? How does it affect these measures?

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{A}\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A}\overline{B})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{A}\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A}\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i),-\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\Big(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\Big)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{A}\overline{B})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{A}\overline{B})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

好的度量的标准

- **Piatetsky-Shapiro**:
  好的度量方式M的三条标准:
  - M(A,B) = 0 if A and B are statistically independent
  - M(A,B) increase monotonically with P(A,B) when P(A) and P(B) remain unchanged
  - M(A,B) decreases monotonically with P(A) [or P(B)] when P(A,B) and P(B) [or P(A)] remain unchanged

Property under Variable Permutation

小象学院 ChinaHadoop.cn

	B	$\overline{\mathbf{B}}$
A	p	q
$\overline{\mathbf{A}}$	r	s

	A	$\overline{\mathbf{A}}$
B	p	r
$\overline{\mathbf{B}}$	q	s

Does M(A,B) = M(B,A)?

Symmetric measures:

- support, lift, collective strength, cosine, Jaccard, etc

Asymmetric measures:

- confidence, conviction, Laplace, J-measure, etc

Property under Row/Column Scaling

小象学院 ChinaHadoop.cn

Grade-Gender Example (Mosteller, 1968):

	Male	Female	
High	2	3	5
Low	1	4	5
	3	7	10

	Male	Female	
High	4	30	34
Low	2	40	42
	6	70	76

2x 10x

Mosteller:

Underlying association should be independent of the relative number of male and female students in the samples

Property under Inversion Operation

小象学院 ChinaHadoop.cn

	A	B	C	D	E	F
Transaction 1	1	0	0	1	0	0
.	0	0	1	1	1	0
.	0	0	1	1	1	0
.	0	0	1	1	1	0
.	0	1	1	0	1	1
.	0	0	1	1	1	0
	0	0	1	1	1	0
	0	0	1	1	1	0
	0	0	1	1	1	0
Transaction N	1	0	0	1	0	0

(a) A, B (b) C, D (c) E, F

Example: ϕ-Coefficient

小象学院 ChinaHadoop.cn

- ϕ-coefficient is analogous to correlation coefficient for continuous variables

	Y	$\overline{\text{Y}}$	
X	60	10	70
$\overline{\text{X}}$	10	20	30
	70	30	100

	Y	$\overline{\text{Y}}$	
X	20	10	30
$\overline{\text{X}}$	10	60	70
	30	70	100

$$\phi = \frac{0.6 - 0.7\times 0.7}{\sqrt{0.7\times 0.3\times 0.7\times 0.3}} = 0.5238$$

$$\phi = \frac{0.2 - 0.3\times 0.3}{\sqrt{0.7\times 0.3\times 0.7\times 0.3}} = 0.5238$$

ϕ Coefficient is the same for both tables

Property under Null Addition

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	s

⇒

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	s + k

Invariant measures:

- support, cosine, Jaccard, etc

Non-invariant measures:

- correlation, Gini, mutual information, odds ratio, etc

Different Measures have Different Properties

小象学院
ChinaHadoop.cn

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 … 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 … 1 … ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	0 … 1	Yes	Yes	Yes	Yes	No	No*	Yes	No
J	J-Measure	0 … 1	Yes	No	No	No	No	No	No	No
G	Gini Index	0 … 1	Yes	No	No	No	No	No*	Yes	No
s	Support	0 … 1	No	Yes	No	Yes	No	No	No	No
c	Confidence	0 … 1	No	Yes	No	Yes	No	No	No	Yes
L	Laplace	0 … 1	No	Yes	No	Yes	No	No	No	No
V	Conviction	0.5 … 1 … ∞	No	Yes	No	Yes**	No	No	Yes	No
I	Interest	0 … 1 … ∞	Yes*	Yes	Yes	Yes	No	No	No	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 … 0 … 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 … 0 … 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 … 1 … 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 … 1 … ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}-1}\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)\ldots 0 \ldots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

关联分析练习：

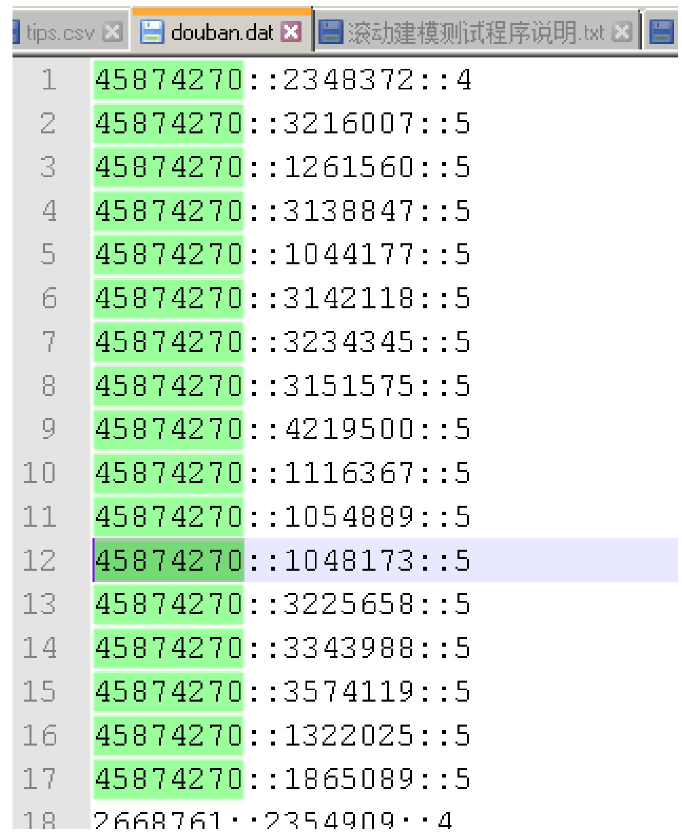
- **购物链数据**
  - 如需使用MLiA代码，需要对数据进行转换

association.xls	2016/7/9 12:18	Microsoft Excel 97-…	413 KB
jieba-master.zip	2016/7/9 0:03	Compressed (zippe…	12,109 KB
历史日线数据_样本(2013 2014年数据)2.zip	2016/7/9 10:06		

	A	B	C	D	E
1	CUSTOMER	TIME	PRODUCT		
2	0	0	hering		
3	0	1	corned_b		
4	0	2	olives		
5	0	3	ham		
6	0	4	turkey		
7	0	5	bourbon		
8	0	6	ice_crea		
9	1	0	baguette		
10	1	1	soda		
11	1	2	hering		
12	1	3	cracker		
13	1	4	heineken		
14	1	5	olives		
15	1	6	corned_b		
16	2	0	avocado		
17	2	1	cracker		
18	2	2	artichok		
19	2	3	heineken		
20	2	4	ham		
21	2	5	turkey		
22	2	6	sardines		
23	3	0	olives		

大作业1：豆瓣书评数据（7月16日讲解）

- 数据：
  - 链接: http://pan.baidu.com/s/1o8fDDZO 密码: uzki
- 书的信息样例：https://book.douban.com/subject/1048173/
- 目标：
- 1.找出跟某一用户相似的若干读者
- 2.找出跟某一本书有类似读者群的书
- 3. 为读者划分类型
- 4. 为书划分类型
- 5. 为某一读者推荐他没有读过的书
- 6. 使用关联分析进行推荐并且比较



大作业2：股票日数据挖掘（7月23日讲解）

➢ **综合股指数据（1年期）**

- sh000001.csv
- sh000016.csv
- sh000300.csv
- sz399001.csv
- sz399005.csv
- sz399006.csv
- sz399905.csv

	A	B	C	D	E	F	G	H	I
1	index_cod	date	open	close	low	high	volume	money	change
2	sh000001	2014/12/31	3172.6	3234.68	3157.26	3239.36	4.06E+10	4.32E+11	0.021752
3	sh000001	2014/12/30	3160.8	3165.82	3130.35	3190.3	3.98E+10	4.37E+11	-0.00069
4	sh000001	2014/12/29	3212.56	3168.02	3126.94	3223.86	5.1E+10	5.56E+11	0.003298
5	sh000001	2014/12/26	3078.01	3157.6	3064.18	3164.16	4.61E+10	4.89E+11	0.027686
6	sh000001	2014/12/25	2992.46	3072.54	2969.87	3073.35	3.77E+10	3.79E+11	0.033643
7	sh000001	2014/12/24	3039.21	2972.53	2934.91	3050.51	3.77E+10	3.79E+11	-0.01981
8	sh000001	2014/12/23	3085.08	3032.61	3025.67	3136.84	4.38E+10	4.19E+11	-0.03032
9	sh000001	2014/12/22	3129.27	3127.45	3090.51	3189.87	6.79E+10	6.24E+11	0.006064
10	sh000001	2014/12/19	3053.08	3108.6	3018.42	3117.53	5.21E+10	5.16E+11	0.016705
11	sh000001	2014/12/18	3062.8	3057.52	3030.32	3089.79	4.36E+10	4.67E+11	-0.00114
12	sh000001	2014/12/17	3031.95	3061.02	2993.33	3076.6	5.43E+10	5.80E+11	0.013074
13	sh000001	2014/12/16	2953.81	3021.52	2943.91	3021.9	4.54E+10	4.93E+11	0.023057
14	sh000001	2014/12/15	2921.45	2953.42	2890.9	2960.23	4E+10	4.11E+11	0.00519
15	sh000001	2014/12/12	2929.36	2938.17	2914.96	2962.51	4.09E+10	4.20E+11	0.004248
16	sh000001	2014/12/11	2912.35	2925.74	2892.61	2965.68	4.83E+10	4.80E+11	-0.00485
17	sh000001	2014/12/10	2855.94	2940.01	2807.68	2946.71	5.13E+10	5.35E+11	0.029317
18	sh000001	2014/12/9	2992.49	2856.27	2834.59	3091.32	7.72E+10	7.93E+11	-0.0543
19	sh000001	2014/12/8	2907.82	3020.26	2879.85	3041.66	5.88E+10	5.93E+11	0.028121
20	sh000001	2014/12/5	2926.57	2937.65	2813.05	2978.03	6.41E+10	6.39E+11	0.013172
21	sh000001	2014/12/4	2783.47	2899.46	2772.43	2900.51	5.33E+10	5.09E+11	0.043148
22	sh000001	2014/12/3	2768.68	2779.53	2733.87	2824.18	5.62E+10	5.30E+11	0.005782
23	sh000001	2014/12/2	2667.82	2763.55	2665.69	2777.37	4.38E+10	3.97E+11	0.031114
24	sh000001	2014/12/1	2691.73	2680.16	2668.84	2720.74	4.47E+10	4.01E+11	-0.001
25	sh000001	2014/11/28	2629.63	2682.84	2622.06	2683.18	4.66E+10	4.02E+11	0.019901
26	sh000001	2014/11/27	2615.37	2630.49	2599.11	2631.4	3.64E+10	3.39E+11	0.010037
27	sh000001	2014/11/26	2572.65	2604.35	2570.4	2605.07	3.37E+10	3.17E+11	0.014312
28	sh000001	2014/11/25	2532	2567.6	2527.08	2568.38	3.14E+10	2.82E+11	0.013707
29	sh000001	2014/11/24	2505.53	2532.88	2495.52	2546.75	3.63E+10	3.30E+11	0.018533
30	sh000001	2014/11/21	2452.64	2486.79	2446.65	2488.2	2.12E+10	1.98E+11	0.013916

sh000001

大作业2：股票日数据挖掘（7月23日讲解）

小象学院 ChinaHadoop.cn

➢ 个股日线（1年期）

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	code	date	open	high	low	close	change	volume	money	traded_ma	market_val	turnover	adjust_pric	report_typ	report_dat	PE_TTM	PS_TTM	PC_TTM	PB
2	sh600000	2014/12/31	15.45	15.79	15.11	15.69	0.021484	4.69E+08	7.27E+09	2.34E+11	2.93E+11	0.031417	130.2546	#######	#######	6.375742	2.515326	3.401388	1.265
3	sh600000	2014/12/30	14.95	15.5	14.83	15.36	0.027425	4.44E+08	6.77E+09	2.29E+11	2.87E+11	0.02973	127.5151	#######	#######	6.241646	2.462423	3.32985	1.238
4	sh600000	2014/12/29	15.4	15.88	14.71	14.95	0.012187	6.25E+08	9.56E+09	2.23E+11	2.79E+11	0.041897	124.1114	#######	#######	6.075039	2.396694	3.240966	1.205
5	sh600000	2014/12/26	14.3	14.84	14.19	14.77	0.036491	4.67E+08	6.79E+09	2.2E+11	2.76E+11	0.031293	122.617	#######	#######	6.001893	2.367837	3.201944	1.19
6	sh600000	2014/12/25	13.75	14.27	13.53	14.25	0.058692	4.56E+08	6.36E+09	2.13E+11	2.66E+11	0.030539	118.3001	#######	#######	5.790589	2.284474	3.089216	1.149
7	sh600000	2014/12/24	14.11	14.2	13.31	13.46	-0.04607	4.18E+08	5.72E+09	2.01E+11	2.51E+11	0.027983	111.7418	#######	#######	5.469569	2.157827	2.917955	1.085
8	sh600000	2014/12/23	14.4	14.98	14.08	14.11	-0.03883	4.42E+08	6.4E+09	2.11E+11	2.63E+11	0.029591	117.138	#######	#######	5.733704	2.262032	3.058868	1.137
9	sh600000	2014/12/22	14.18	15.2	14.14	14.68	0.041874	6.83E+08	1E+10	2.19E+11	2.74E+11	0.045799	121.8699	#######	#######	5.965325	2.35341	3.182436	1.183
10	sh600000	2014/12/19	13.96	14.2	13.63	14.09	0.014399	4.36E+08	6.1E+09	2.1E+11	2.63E+11	0.029233	116.9719	#######	#######	5.725573	2.258824	3.05453	1.136
11	sh600000	2014/12/18	14.22	14.34	13.71	13.89	-0.01559	5E+08	7.01E+09	2.07E+11	2.59E+11	0.033481	115.3115	#######	#######	5.6443	2.226761	3.011172	1.119
12	sh600000	2014/12/17	13.49	14.39	13.33	14.11	0.061701	8.7E+08	1.2E+10	2.11E+11	2.63E+11	0.05828	117.1379	#######	#######	5.7337	2.262031	3.058866	1.137
13	sh600000	2014/12/16	12.7	13.3	12.65	13.29	0.041536	4.39E+08	5.71E+09	1.98E+11	2.48E+11	0.029397	110.3304	#######	#######	5.400485	2.130572	2.881099	1.071
14	sh600000	2014/12/15	12.8	12.82	12.46	12.76	-0.01695	3.35E+08	4.23E+09	1.9E+11	2.38E+11	0.022451	105.9305	#######	#######	5.185116	2.045606	2.766202	1.02
15	sh600000	2014/12/12	13.05	13.38	12.78	12.98	-0.00536	3.11E+08	4.07E+09	1.94E+11	2.42E+11	0.020811	107.7569	#######	#######	5.274514	2.080875	2.813895	1.046
16	sh600000	2014/12/11	12.96	13.55	12.85	13.05	-0.00836	4.43E+08	5.85E+09	1.95E+11	2.43E+11	0.029657	108.338	#######	#######	5.302959	2.092097	2.82907	1.052
17	sh600000	2014/12/10	12.7	13.25	12.21	13.16	0.040316	6.2E+08	7.9E+09	1.96E+11	2.45E+11	0.041565	109.2512	#######	#######	5.34766	2.109732	2.852918	1.061
18	sh600000	2014/12/9	13.56	14.16	12.46	12.65	-0.084	8.69E+08	1.18E+10	1.89E+11	2.36E+11	0.058246	105.0173	#######	#######	5.140419	2.027972	2.742357	1.019
19	sh600000	2014/12/8	13.39	14.04	13.2	13.81	0.020695	7.04E+08	9.62E+09	2.06E+11	2.58E+11	0.047171	114.6474	#######	#######	5.611792	2.213936	2.99383	1.113
20	sh600000	2014/12/5	13.42	14	12.9	13.53	0.019593	8.6E+08	1.16E+10	2.02E+11	2.52E+11	0.057606	112.3228	#######	#######	5.498011	2.169048	2.933129	1.090
21	sh600000	2014/12/4	12.58	13.3	12.37	13.27	0.054849	7.2E+08	9.31E+09	1.98E+11	2.48E+11	0.048251	110.1644	#######	#######	5.392359	2.127366	2.876764	1.069
22	sh600000	2014/12/3	12.84	13.29	12.32	12.58	-0.02253	7.31E+08	9.38E+09	1.88E+11	2.35E+11	0.048956	104.4362	#######	#######	5.111972	2.01675	2.727181	1.014
23	sh600000	2014/12/2	12.03	13.08	12.03	12.87	0.058388	5.89E+08	7.4E+09	1.92E+11	2.4E+11	0.039495	106.8437	#######	#######	5.229815	2.063241	2.790049	1.03
24	sh600000	2014/12/1	12.45	12.98	12.12	12.16	-0.01936	6.07E+08	7.59E+09	1.81E+11	2.27E+11	0.040694	100.9494	#######	#######	4.941303	1.949418	2.636131	0.980
25	sh600000	2014/11/28	11.53	12.48	11.48	12.4	0.080139	8.32E+08	9.95E+09	1.85E+11	2.31E+11	0.055766	102.9419	#######	#######	5.038829	1.987894	2.68816	0.999
26	sh600000	2014/11/27	11.48	11.72	11.28	11.48	0.014134	4.4E+08	5.06E+09	1.71E+11	2.14E+11	0.029468	95.30429	#######	#######	4.664982	1.840405	2.488717	0.925
27	sh600000	2014/11/26	11.25	11.43	11.1	11.32	0.021661	4.52E+08	5.09E+09	1.69E+11	2.11E+11	0.030307	93.97604	#######	#######	4.599966	1.814756	2.454032	0.912
28	sh600000	2014/11/25	10.81	11.09	10.75	11.08	0.020258	2.89E+08	3.16E+09	1.65E+11	2.07E+11	0.019399	91.98358	#######	#######	4.502439	1.77628	2.402002	0.893
29	sh600000	2014/11/24	10.57	10.99	10.45	10.86	0.006487	4.74E+08	5.09E+09	1.62E+11	2.03E+11	0.031767	90.15718	#######	#######	4.41304	1.74101	2.354308	0.87
30	sh600000	2014/11/21	10.58	10.82	10.46	10.79	0.020814	2.1E+08	2.23E+09	1.61E+11	2.01E+11	0.014072	89.5761	#######	#######	4.384597	1.729789	2.339134	0.870
31	sh600000	2014/11/20	10.45	10.66	10.37	10.57	0.009551	1.62E+08	1.71E+09	1.58E+11	1.97E+11	0.010883	87.74968	#######	#######	4.295197	1.694519	2.29144	0.852
32	sh600000	2014/11/19	10.43	10.54	10.39	10.47	0.001914	1.45E+08	1.52E+09	1.56E+11	1.95E+11	0.009739	86.91951	#######	#######	4.254561	1.678488	2.269762	0.844

sh600000

大作业2：股票日数据挖掘（7月23日讲解）

小象学院 ChinaHadoop.cn

- **目标：**
  - 1. 根据股指进行预测
  - 2. 找出权重股，与真实权重股进行对比
  - 3. 根据个股数据对个股进行聚类，形成“板块”
  - 4. 尝试挖掘板块之间的关系

中文分词

结巴中文分词

- https://github.com/fxsjy/jieba
- 支持三种分词模式：
  - 精确模式，试图将句子最精确地切开，适合文本分析；
  - 全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；
  - 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词

安装说明

代码对 Python 2/3 均兼容

- 全自动安装：`easy_install jieba` 或者 `pip install jieba` / `pip3 install jieba`
- 半自动安装：先下载 http://pypi.python.org/pypi/jieba/ ，解压后运行 `python setup.py install`
- 手动安装：将 jieba 目录放置于当前目录或者 site-packages 目录
- 通过 `import jieba` 来引用

中文分词

代码示例

```
# encoding=utf-8
import jieba

seg_list = jieba.cut("我来到北京清华大学", cut_all=True)
print("Full Mode: " + "/ ".join(seg_list))  # 全模式

seg_list = jieba.cut("我来到北京清华大学", cut_all=False)
print("Default Mode: " + "/ ".join(seg_list))  # 精确模式

seg_list = jieba.cut("他来到了网易杭研大厦")  # 默认是精确模式
print(", ".join(seg_list))

seg_list = jieba.cut_for_search("小明硕士毕业于中国科学院计算所，后在日本京都大学深造")  # 搜索引擎模式
print(", ".join(seg_list))
```

输出:

```
【全模式】: 我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学

【精确模式】: 我/ 来到/ 北京/ 清华大学

【新词识别】：他, 来到, 了, 网易, 杭研, 大厦    (此处，“杭研”并没有在词典中，但是也被Viterbi算法识别出来了)

【搜索引擎模式】: 小明, 硕士, 毕业, 于, 中国, 科学, 学院, 科学院, 中国科学院, 计算, 计算所, 后, 在, 日本, 京都, 大学, E
```

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop
- 网站：http://chinahadoop.cn
- 问答社区：http://wenda.ChinaHadoop.cn