

《网络数据挖掘》课程介绍

徐 君

课程基本信息

- 课程编号：091M5042H
- 课程属性：专业普及课
- 上课时间：周二下午5/6/7，40学时
- 周次：第2周—第17周
- 教室：教1-107
- 授课老师
 - 徐 君 (junxu@ict.ac.cn, <http://www.bigdatalab.ac.cn/~junxu>)
 - 罗 平 (luop@ict.ac.cn, <http://mldm.ict.ac.cn/MLDM/luoping>)
 - 沈华伟 (shenhuawei@ict.ac.cn)
- 助教
 - 敖 翔 (aoux@ics.ict.ac.cn)

课程结构

- 两部分
 - 授课：讲授课程的主要内容，包括数据挖掘基础、推荐系统、社交网络、互联网搜索、互联网广告和专题讲座
 - 作业：随堂作业+大作业
- 成绩评定
 - 平时成绩(40%): 随堂作业(~20%)、1次大作业(~20%)、课堂纪律、考勤、讨论等
 - 闭卷考试(60%): 数据挖掘基础、推荐系统、社交网络、互联网搜索、互联网广告

预备知识

- 数学
 - 概率统计
 - 线性代数
- 计算机
 - 数据结构、算法

教学内容

- 第一部分：数据挖掘基础（徐君，9学时）
 - 互联网数据挖掘的概念与发展简介
 - 关联规则
 - 监督学习
 - 无监督学习
- 第二部分：网络数据挖掘（27学时）
 - 推荐系统（罗平，9学时）
 - 图数据挖掘（沈华伟，9学时）
 - 互联网搜索/广告（徐君，9学时）
- 第三部分：专题讲座（3学时）
 - 时间：2016年10月25日
 - 主讲人：奇好云宝CEO胡云华博士

参考书

- 参考书
 - 数据挖掘基础：《Web数据挖掘》
 - 推荐系统：《推荐系统：技术、评估及高效算法》
 - 社交网络：《链接：商业、科学与生活的新思维》
 - 互联网搜索：Chris Manning et al., Introduction to Information Retrieval, Cambridge University, 2008.
- 其他推荐书籍
 - 《Mining massive data》

反馈与建议

- 欢迎大家对课程内容、教学方式等任何与课程相关的事情进行反馈和建议，帮助授课老师及时调整
 - 邮件 (junxu@ict.ac.cn)
 - 面对面交流
 - 课程网站互动
 - 其他任何形式
- 与你的同学共享问题和心得，共同进步

纪律

- 作弊：对作弊零容忍，包括作业和考试

Not everything that can be counted counts, and not everything that counts can be counted.

Albert Einstein, (attributed)

US (German-born) physicist (1879 - 1955)

- 作业提交
 - 可以与同学讨论，但是必须独立完成
 - 请按时提交，延迟提交将酌情扣分
- 考试不是目的，希望同学们通过课程，
 - 学习到在今后的实际项目开发中需要用到的基本数据分析方法和工具
 - 锻炼自己的写作和逻辑思维能力，为今后的科研、论文写作和展示报告打下基础

对课件的说明

- 课件将会共享在课程主页上
- 在准备课件的过程中，参考了国内外相关课程的课件以及互联网上的相关内容，不在课件中一一指出
 - 模型算法的描述
 - 图片、举例等
- 可能会含有错误或者不清楚的内容，如果发现请同学们指出；已根据去年的教学情况对课件进行了调整
- 根据实际情况，授课内容和课程大纲稍有出入

数据挖掘背景简介



数据中蕴含知识与价值

数据挖掘

- 从数据中发现知识的步骤

- 采集数据

- 存储数据

- 管理数据

- **分析**数据
 - 结果**应用**

} 课程重点

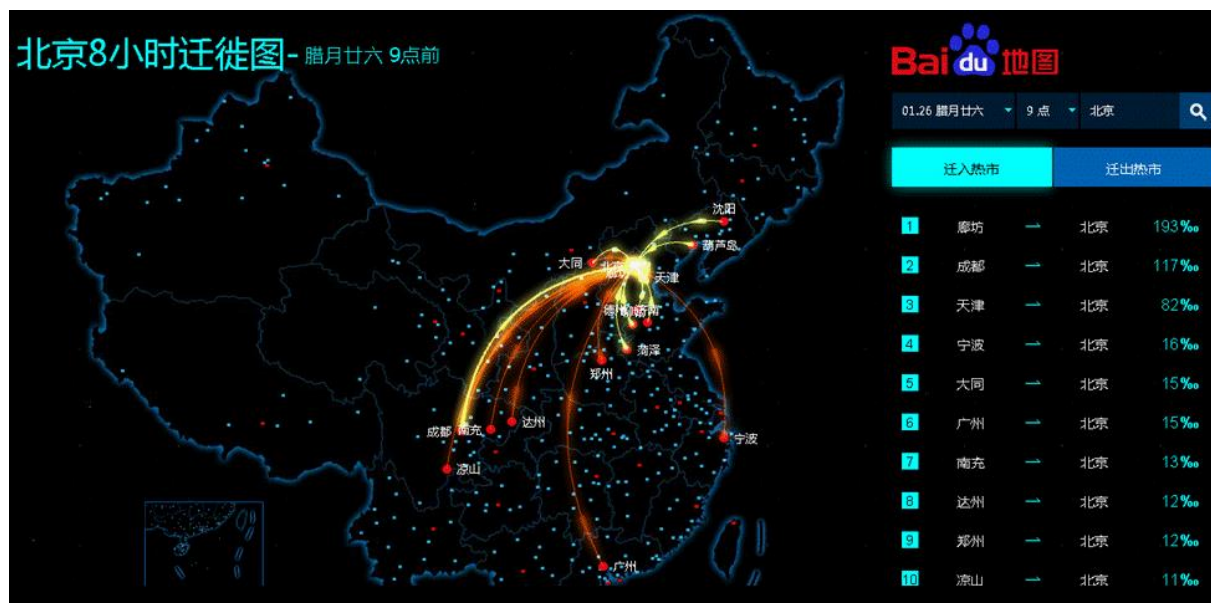
数据挖掘、(大)数据分析、统计分析

什么是数据挖掘？

- 给定大量的数据（数据库、文本、网络、图像等）
- 从数据中发现有用的模式
 - 有效性(valid): 在新的数据上任然有效
 - 新颖性(novel): 模式不是显然的，具有信息量
 - 可用(potentially useful): 能够应用于某些场合
 - 可理解(understandable): 能够被合理的解释

数据挖掘的任务

- 描述现象(规律发现)
 - 寻找给定数据中可解析的模式
 - 聚类、关联规则

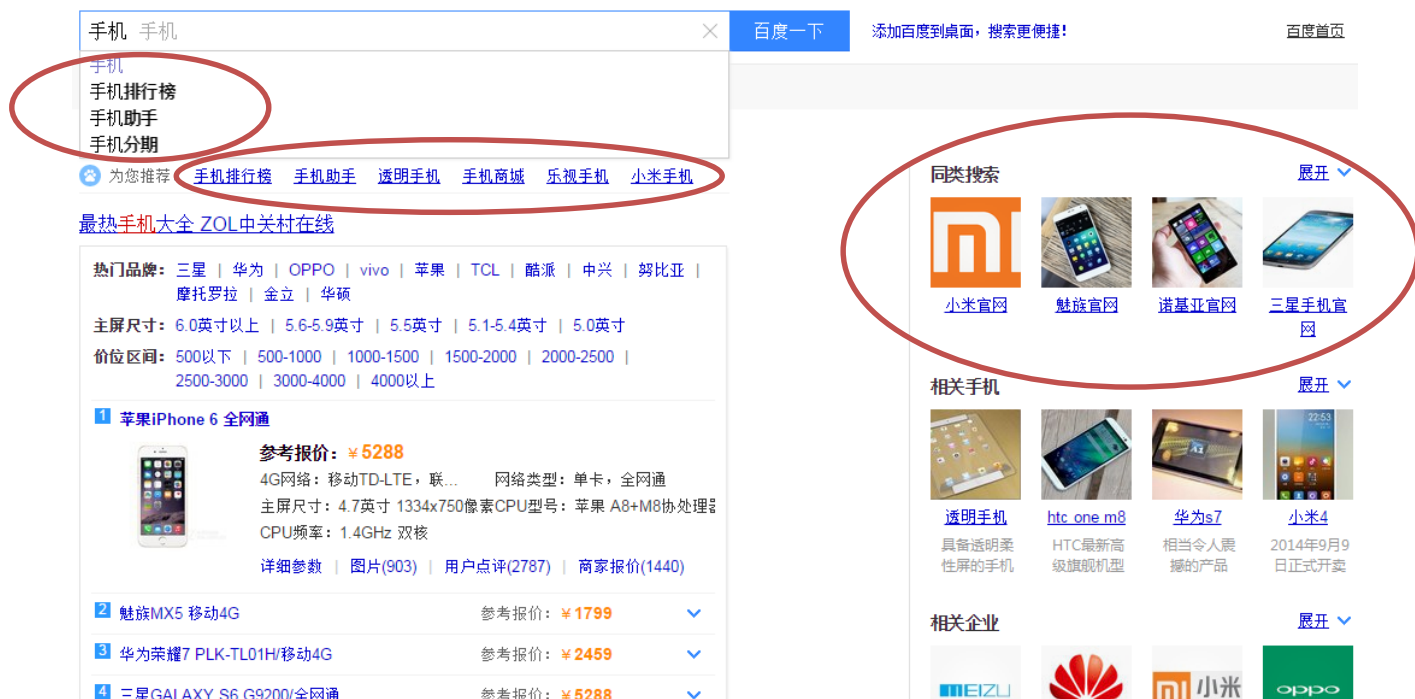


央视《据说春运》截图

数据挖掘的任务（续）

- 预测

- 从给定数据中发现规律并在未来的观测数据上进行预测
- 更加具有商业价值



手机 手机

手机

手机排行榜

手机助手

手机分期

为您推荐 手机排行榜 手机助手 透明手机 手机商城 乐视手机 小米手机

最热手机大全 ZOL中关村在线

热门品牌: 三星 | 华为 | OPPO | vivo | 苹果 | TCL | 酷派 | 中兴 | 努比亚 | 摩托罗拉 | 金立 | 华硕

主屏尺寸: 6.0英寸以上 | 5.6-5.9英寸 | 5.5英寸 | 5.1-5.4英寸 | 5.0英寸

价位区间: 500以下 | 500-1000 | 1000-1500 | 1500-2000 | 2000-2500 | 2500-3000 | 3000-4000 | 4000以上

1 苹果iPhone 6 全网通

参考报价: ¥5288

4G网络: 移动TD-LTE, 联... 网络类型: 单卡, 全网通

主屏尺寸: 4.7英寸 1334x750像素CPU型号: 苹果 A8+M8协处理器

CPU频率: 1.4GHz 双核

详细参数 | 图片(903) | 用户点评(2787) | 商家报价(1440)

2 魅族MX5 移动4G 参考报价: ¥1799

3 华为荣耀7 PLK-TL01H/移动4G 参考报价: ¥2459

4 三星GALAXY S6 G9200/全网通 参考报价: ¥5288

同类搜索

小米官网 魅族官网 诺基亚官网 三星手机官网

相关手机

透明手机 htc one m8 华为s7 小米4

相关企业

MIUI HUAWEI 小米 oppo

传统数据挖掘任务

- 分类
- 聚类
- 关联规则挖掘
- 序列挖掘
- 离群点发现
- 数据可视化

互联网时代(新)的数据挖掘任务

- 文本/网页分析
 - 知识库（实体、属性、关系抽取）
 - 关键词发现
- 社交网络分析
- 商品推荐
- 互联网搜索排序
- 互联网广告

数据挖掘的重要性

- 学科价值
 - 是数据分析工具的实战战场（理论联系实际）
 - 数据挖掘工具/算法广泛应用于其他学科
- 商业价值
 - 提升传统业务的效率（例如：提升电信行业的服务质量）
 - 孵化催生新的业务和商业模式（例如：互联网广告）

科学发现与数据分析（挖掘）

- 天文学发现

- 托勒密观点：认为天体的运动是完美的圆周运动；
第谷：一位杰出的观测家，通过二十多年的精确的观察天体的运动，发现似乎可以证明是圆周运动，但只是有点小小的误差；
- **开普勒**：算出观察的结果与理论值有8分的误差，导致了开普勒三大定律的发现(椭圆定律、面积定律、调和定律)，认为天体的运动是椭圆而不是圆；
- 牛顿：在开普勒三大定律的基础上结合牛顿三定律，推导出了万有引力定律。

商业模式与数据分析（挖掘）

- 一家创业公司的模式
 - 商机：手机apps开发者对自己开发出的app的推广使用情况（下载、安装、卸载、日活、月活）缺乏了解
 - 无力建立7*24的数据中心
 - 方案：提供免费手机apps开发工具包，将用户使用情况发回数据中心，并提供apps使用情况分析报表
 - 纵向比较：app随时间变化使用情况
 - 横向对比：与别的（类似）apps的数据对比
 - 数据挖掘挑战：报表生成、剔除作弊数据

相关领域

- 机器学习
- 概率统计
- 数据库
- 信息检索
- 推荐系统
- 自然语言处理

学习资源

- 学术会议
 - 数据挖掘：SIGKDD、ICDM
 - 信息检索：SIGIR、WWW、CIKM、WSDM
 - 自然语言处理：ACL、EMNLP、NAACL
 - 机器学习：ICML、NIPS
- 学术期刊
 - 数据挖掘：TKDD、TKDE、TOIS、TIST、IPM
 - 机器学习：JMLR
- 软件系统
 - Azure Machine Learning、ICT-BDA
 - Weka
 - Apache Lucene、Mahout、SparkML

数据挖掘方法举例：关联规则

提纲

- 关联规则简介
- Apriori算法
- 总结

关联规则挖掘由来



- 由Agrawal等在1993 SIGMOD中发表的论文提出，谷歌学术统计引用量16000+

Mining Association Rules between Sets of Items in Large Databases

Rakesh Agrawal

Tomasz Imielinski*

Arun Swami

IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

- 是数据挖掘/数据库领域最重要的论文之一

关联规则挖掘

- 假设所有的数据类型为类别型
- 不适合直接应用于数值型数据
 - 数值型可以转换为类别型
- 算法目的：基于购物篮数据，分析用户的购买习惯
 - 购物篮数据：{牛奶、面包、啤酒}{手机、耳机}
 - 购买习惯：面包 \rightarrow 牛奶 [sup=5%, conf=85%]

数据描述

- 项目集合: $I = \{i_1, i_2, \dots, i_m\}$
- 一次交易 t :
 - 交易的项目集合: $t \subseteq I$
- 交易数据集合 T
 - 一系列交易: $T = \{t_1, t_2, \dots, t_n\}$

数据举例1：超市购物篮数据

- 超市购物篮数据

t1: {面包, 牛奶, 芝士}

t2: {苹果, 鸡蛋, 盐, 乳酪}

... ..

tn: {饼干, 鸡蛋, 牛奶}

- 对应的概念

- I: 超市中所有（被购买过）的商品

- t: 一次交易中被购买的商品

- T: 搜集到的所有交易

数据举例2： 文本文档

- 一个文本文档可以被表达为一个“词袋”，文档集合可以表达为

d1: {学生, 教学, 学校}

d2: {教学, 学校, 城市, 篮球}

... ...

dn: {篮球, 教练, 球队}

- 对应的概念

– I : 在文档曾经中出现过的词

– t : 一个文本文档

– T : 文档集合

关联规则

- 关联规则的形式

$X \rightarrow Y$, 其中 $X, Y \subset I$ 并且 $X \cap Y = \phi$

- 称 X 和 Y 为 **项目集(itemset)**

– 例如: $X = \{\text{牛奶}, \text{面包}\}$, $Y = \{\text{啤酒}\}$

- k -itemset: 含有 k 个物品的 itemset

– 例如: 2-itemset $\{\text{牛奶}, \text{面包}\}$

- 关联规则刻画了用户潜在的**购买行为习惯**

支持度和置信度：对规则定量描述

- 支持度sup: T中的交易同时包含X和Y的百分比（概率）

$$\text{sup} = \Pr(X \cup Y) = \frac{|\{t \in T | X \cup Y \subseteq t\}|}{|T|}$$

- 置信度conf: T中包含X的事物同时也包含Y的百分比

$$\text{conf} = \Pr(Y|X) = \frac{|\{t \in T | X \cup Y \subseteq t\}|}{|\{t \in T | X \subseteq t\}|}$$

支持度和置信度

TID	网球拍	网球	运动鞋	羽毛球
1	1	1	1	0
2	1	1	0	0
3	1	0	0	0
4	1	0	1	0
5	0	1	1	1
6	1	1	0	0

- 规则 $X \rightarrow Y$, 其中 $X = \{\text{网球拍}\}$, $Y = \{\text{网球}\}$

- 支持度 sup

$$\text{sup} = \frac{|\{1, 2, 6\}|}{|T|} = 0.5$$

- 置信度 conf

$$\text{conf} = \frac{|\{1, 2, 6\}|}{|\{1, 2, 3, 4, 6\}|} = 0.6$$

$X = \{\text{网球拍}, \text{网球}\}$, $Y = \{\text{运动鞋}\}$,
计算 $X \rightarrow Y$ 的支持度和置信度。

关联规则挖掘的目标

- 规定交易集合，找到所有满足最小支持度(minsup)和最小置信度(minconf)的规则
- 对算法的要求
 - 找到所有满足条件的规则(completeness)
 - 不能将数据全部读入内存进行计算

举例

TID	商品
1	Beef, Chicken, Milk
2	Beef, Cheese
3	Cheese, Boots
4	Beef, Chicken, Cheese
5	Beef, Chicken, Clothes , Cheese, Milk
6	Chicken, Clothes, Milk
7	Chicken, Milk, Clothes

- 假设 $\text{minsup}=0.3$, $\text{minconf}=0.8$
- 其中一个频繁集(frequent itemset)为
 $\{\text{Chicken, Milk, Clothes}\}$
[sup=3/7]
- 基于此频繁集的关联规则
 $\{\text{Clothes}\} \rightarrow \{\text{Milk, Chicken}\}$
[sup = 3/7, conf = 3/3]
... ..
 $\{\text{Clothes, Chicken}\} \rightarrow \{\text{Milk}\}$
[sup = 3/7, conf = 3/3]

关联规则挖掘算法

- 基于不同的数据结构和策略，研究者们提出了不同的关联规则挖掘算法
 - Apriori算法
 - F-P算法
 - Eclat算法
 -
- 注意：在给定交易集合 T 、最小支持度 minsup 和最小置信度 minconf 的情况下， T 中存在**唯一**的关联规则集合

提纲

- 关联规则简介
- Apriori算法
- 总结

Apriori算法

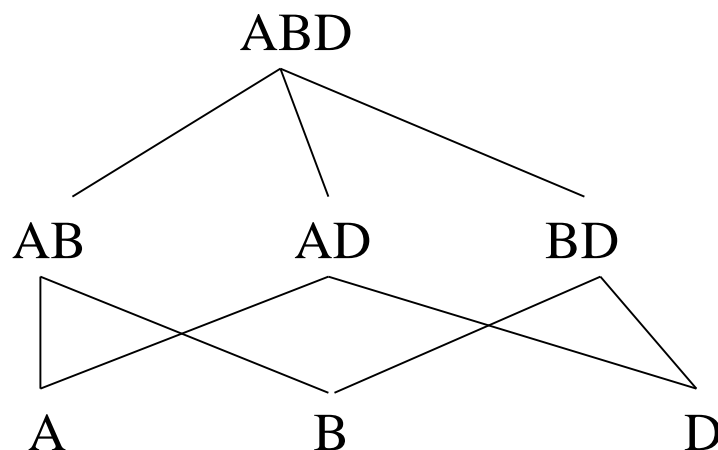
Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. VLDB 1994.

- 关联规则挖掘中最著名的算法
- 算法分为两个步骤
 - 寻找所有满足最小支持度minsup的频繁集(frequent itemset)
 - 基于频繁集产生规则
- 例如:
 - 1. 寻找频繁集
 {Chicken, Clothes, Milk} [sup=3/7]
 - 2. 产生一条规则
 {Clothes} → {Milk, Chicken} [sup = 3/7, conf = 3/3]

步骤1：频繁集挖掘

- 频繁集(frequent itemset): 满足最小支持度 ($\text{sup} \geq \text{minsup}$) 的项目集合
- 向下闭包性质: 频繁集的子集都是频繁集

$$\text{sup} = \text{Pr}(X \cup Y) = \frac{|\{t | t \in T \wedge X \cup Y \subseteq t\}|}{|T|}$$



步骤1：频繁集挖掘过程

- 迭代算法：生成1-频繁集；然后生成2-频繁集；依次迭代
 - 利用Apriori性质：在第k次迭代中（生成k-频繁集），只需要考虑以k-1频繁集为子集的集合
- 生成1-频繁集 F_1
 - $k=2, \dots$,
 - 利用Apriori性质，生成候选k-频繁集 C_k
 - 过滤 C_k 中不符合条件的候选频繁集($\text{sup} < \text{minsup}$), 得到 F_k

步骤1：频繁集挖掘过程举例

最小支持度：minsup=0.5,

最小出现次数： $|T| * \text{minsup} = 2$

数据格式：集合:出现次数

数据T

TID	商品集合
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

1. 扫描T $\rightarrow C_1 = \{\{1\}:2, \{2\}:3, \{3\}:3, \{4\}:1, \{5\}:3\}$
 $\rightarrow F_1 = \{\{1\}:2, \{2\}:3, \{3\}:3, \{5\}:3\}$
 $\rightarrow C_2 = \{\{1, 2\}, \{1, 3\}, \{1, 5\}, \{2, 3\}, \{2, 5\}, \{3, 5\}\}$
2. 扫描T $\rightarrow C_2 = \{\{1, 2\}:1, \{1, 3\}:2, \{1, 5\}:1, \{2, 3\}:2, \{2, 5\}:3, \{3, 5\}:2\}$
 $\rightarrow F_2 = \{\{1, 3\}:2, \{2, 3\}:2, \{2, 5\}:3, \{3, 5\}:2\}$
 $\rightarrow C_3 = \{\{1, 2, 3\}, \{1, 3, 5\}, \{2, 3, 5\}\}$
3. 扫描T $\rightarrow C_3 = \{\{1, 2, 3\}:1, \{1, 3, 5\}:1, \{2, 3, 5\}:2\}$
 $\rightarrow F_3 = \{\{2, 3, 5\}:2\}$

商品排序

- 为了计算方便，在表示频繁集时，集合中的商品被认为按照字典顺序排列显示
- 当集合表达为 $\{i_1, i_2, \dots, i_k\}$ 时，我们认为 $i_1 < i_2 < \dots, < i_k$
 - $\{3, 1, 5, 2\} \rightarrow \{1, 2, 3, 5\}$

步骤1：频繁集挖掘算法

Algorithm Apriori(T)

```
 $C_1 \leftarrow \text{init-pass}(T);$   
 $F_1 \leftarrow \{f \mid f \in C_1, f.\text{count}/n \geq \text{minsup}\};$  //  $n$ : no. of transactions in  $T$   
for ( $k = 2; F_{k-1} \neq \emptyset; k++$ ) do  
     $C_k \leftarrow \text{candidate-gen}(F_{k-1});$   
    for each transaction  $t \in T$  do  
        for each candidate  $c \in C_k$  do  
            if  $c$  is contained in  $t$  then  
                 $c.\text{count}++;$   
            end  
        end  
     $F_k \leftarrow \{c \in C_k \mid c.\text{count}/n \geq \text{minsup}\}$   
end  
return  $F \leftarrow \bigcup_k F_k;$ 
```

函数 $\text{candidate-gen}(F_{k-1})$

- 输入：k-1 频繁集
- 输出：候选k频繁集
- 分为两步
 - join：生成所有可能的候选k-频繁集
 - prune：剔除不可能为频繁集的候选项

函数 candidate-gen(F_{k-1})

Function candidate-gen(F_{k-1})

$C_k \leftarrow \emptyset$;

forall $f_1, f_2 \in F_{k-1}$

 with $f_1 = \{i_1, \dots, i_{k-2}, i_{k-1}\}$

 and $f_2 = \{i_1, \dots, i_{k-2}, i'_{k-1}\}$

 and $i_{k-1} < i'_{k-1}$ **do**

$c \leftarrow \{i_1, \dots, i_{k-1}, i'_{k-1}\}$;

// join f_1 and f_2

$C_k \leftarrow C_k \cup \{c\}$;

for each $(k-1)$ -subset s of c **do**

if ($s \notin F_{k-1}$) **then**

 delete c from C_k ;

// prune

end

end

return C_k ;

步骤1：频繁集挖掘过程举例

最小支持度：minsup=0.5,
 最小出现次数：|T| * minsup = 2
 数据格式：集合:出现次数

数据T

TID	商品集合
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

- 扫描T → $C_1 = \{\{1\}:2, \{2\}:3, \{3\}:3, \{4\}:1, \{5\}:3\}$
 → $F_1 = \{\{1\}:2, \{2\}:3, \{3\}:3, \{5\}:3\}$
 → $C_2 = \{\{1, 2\}, \{1, 3\}, \{1, 5\}, \{2, 3\}, \{2, 5\}, \{3, 5\}\}$
- 扫描T → $C_2 = \{\{1, 2\}:1, \{1, 3\}:2, \{1, 5\}:1, \{2, 3\}:2, \{2, 5\}:3, \{3, 5\}:2\}$
 → $F_2 = \{\{1, 3\}:2, \{2, 3\}:2, \{2, 5\}:3, \{3, 5\}:2\}$
 → $C_3 = \{\{1, 2, 3\}, \{1, 3, 5\}, \{2, 3, 5\}\}$
- 扫描T → $C_3 = \{\{1, 2, 3\}:1, \{1, 3, 5\}:1, \{2, 3, 5\}:2\}$
 → $F_3 = \{\{2, 3, 5\}:2\}$

步骤1：频繁集挖掘举例

- 假设 $F_3 = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{1, 3, 5\}, \{2, 3, 4\}\}$
- join结果
 - $C_4 = \{\{1, 2, 3, 4\}, \{1, 3, 4, 5\}\}$
- prune结果
 - $C_4 = \{\{1, 2, 3, 4\}\}$

由于 $\{1, 4, 5\}$ 不在 F_3 中，因此剔除 $\{1, 3, 4, 5\}$

步骤2：生成规则

- 频繁集 \neq 关联规则，规则生成过程如下：
- Foreach 频繁集 X
 - Foreach $A \subset X \wedge A \neq \phi$
 - $B = X - A$
 - If $\frac{\text{sup}(A \cup B)}{\text{sup}(A)} \geq \text{minconf}$
 - 生成规则 $A \rightarrow B$
 - $\text{sup}(A \rightarrow B) = \text{sup}(A \cup B) = \text{sup}(X)$
 - $\text{conf}(A \rightarrow B) = \frac{\text{sup}(A \cup B)}{\text{sup}(A)}$

步骤2：生成规则举例

- 假设频繁集 $\{2, 3, 4\}$ 的支持度 $\text{sup}=0.5$
- 非空子集及支持度为： $\{2, 3\}:0.5, \{2, 4\}:0.5, \{3, 4\}:0.75, \{2\}:0.75; \{3\}:0.75, \{4\}:0.75$
- 生成如下支持度 $\text{sup}=0.5$ 的关联规则
 - $\{2, 3\} \rightarrow \{4\}, \text{conf}=1$
 - $\{2, 4\} \rightarrow \{3\}, \text{conf}=1$
 - $\{3, 4\} \rightarrow \{2\}, \text{conf}=0.6667$
 - $\{2\} \rightarrow \{3, 4\}, \text{conf}=0.6667$
 - $\{3\} \rightarrow \{2, 4\}, \text{conf}=0.6667$
 - $\{4\} \rightarrow \{2, 3\}, \text{conf}=0.6667$

步骤2：小结

- 生成规则 $A \rightarrow B$ ，需要计算 $\text{sup}(A \cup B)$ 和 $\text{sup}(A)$
- 生成规则所需的所有信息保存在频繁集中，无需再次扫描数据 T
- 生成规则时间复杂度低，Apriori 算法主要时间耗费在频繁集生成过程中

关于Apriori算法

- $k=1$ 到 K ，逐层搜索 k -频繁集，最多搜索 K 轮
- 一般来说一条规则中涉及到的商品不会很多（例如 $K < 10$ ）
- 能够在大的数据集合上（并行）执行
- m 个商品的集合将生成 $O(2^m)$ 条规则
- 利用数据的稀疏特性，提高最小支持度和最小置信度将极大减少生成的频繁集和规则数目
 - 通常产生大量的关联规则

关联规则小结

- 关联规则
 - 关联规则简介
 - Apriori算法

Thanks!