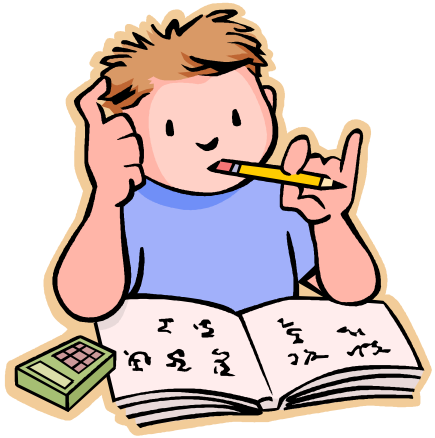
第一章 绪论

内容提要

- 为什么要做数据挖掘？
- 什么是数据挖掘？
- 数据挖掘的主要技术
- 数据挖掘的主要研究内容
- 数据挖掘的主要问题



- 几个需要思考的问题

在座的哪位同学不用手机上网？

你会用手机上网做哪些事情？

- 从几个热点概念说起
  - 为什么移动互联网时代对我们的生活有如此之大的影响？
  - 什么是“物联网”？
  - “大数据”与“云”计算的关系？
  - 什么是“互联网+”？

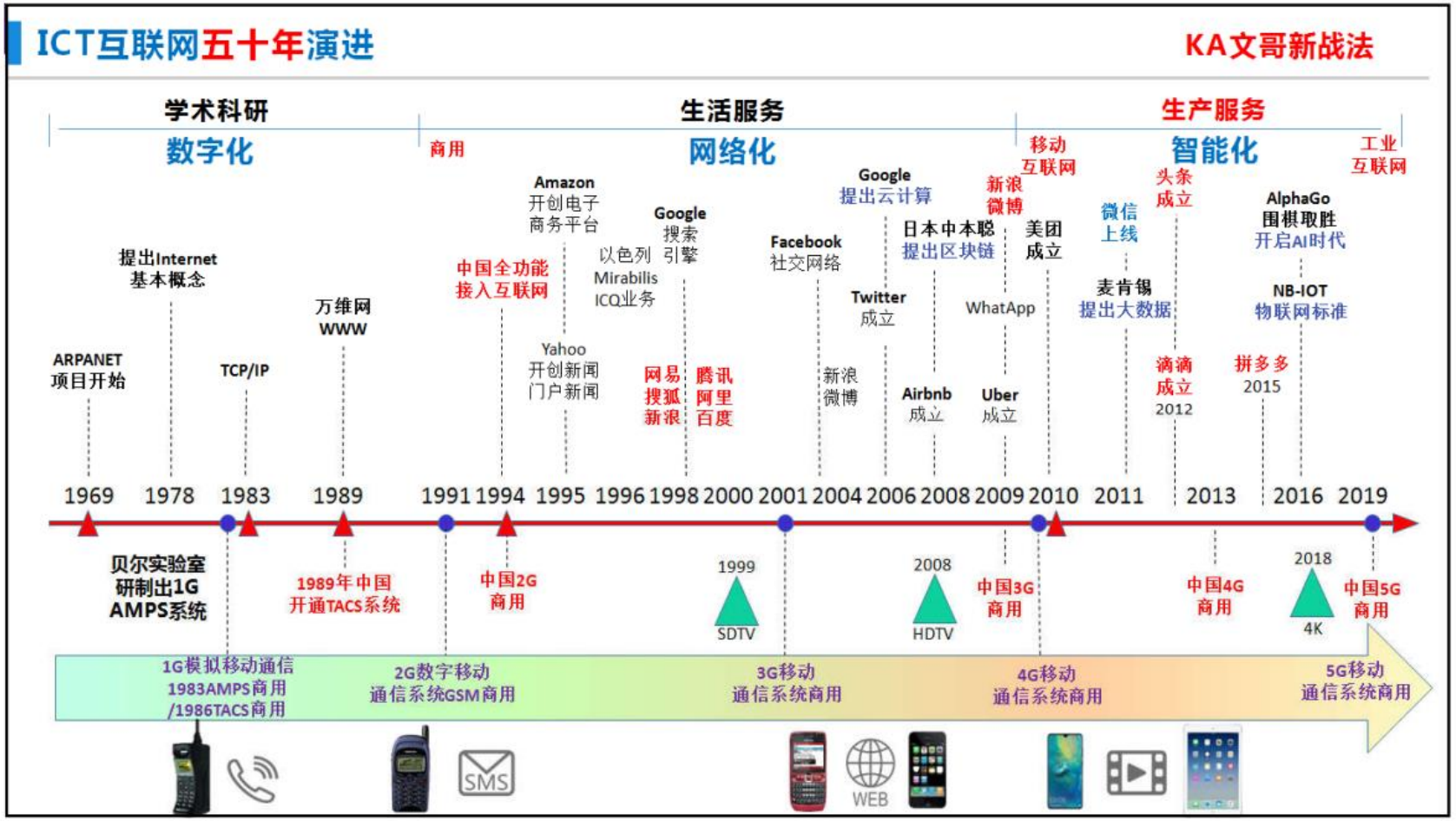
关于移动互联网

背景回顾

- 2010年：3G牌照发放，“中国移动互联网元年”
- 2014年：4G牌照发放，**手机上网人数超过计算机上网人数**
- 2015年：华为在日本实验5G网络通讯技术
- 2016年：美国5G试用
- 2017年：美国11城市5G试商用，9月1日 欧盟开始6G研发计划
- 2018年：中国5G试商用（2018.12已经开放5G频段）
- 2019年：中国5G手机发布，下半年**5G**开始商用（**视频1，视频2**）

“十年一G”（《大败局III》(吴晓波 著)）

- 1G：语音时代
- 2G：文本时代
- 3G：图片时代
- 4G：视频时代，开始“改变生活”
- 5G：万物互联，开始“改变社会”
- 6G：全域覆盖，场景智联（2030年）


ICT互联网五十年演进
KA文哥新战法
学术科研
数字化
生活服务
网络化
生产服务
智能化
商用
ARPANET
项目开始
提出Internet
基本概念
TCP/IP
万维网
WWW
中国全功能
接入互联网
Amazon
开创电子
商务平台
Yahoo
开创新闻
门户新闻
以色列
Mirabilis
ICQ业务
Google
搜索
引擎
网易
搜狐
新浪
腾讯
阿里
百度
Facebook
社交网络
新浪
微博
Google
提出云计算
Twitter
成立
日本中本聪
提出区块链
Airbnb
成立
新浪
微博
WhatApp
Uber
成立
移动
互联网
美团
成立
微信
上线
麦肯锡
提出大数据
头条
成立
滴滴
成立
2012
AlphaGo
围棋取胜
开启AI时代
NB-IOT
物联网标准
拼多多
2015
工业
互联网
1969
1978
1983
1989
1991
1994
1995
1996
1998
2000
2001
2004
2006
2008
2009
2010
2011
2013
2016
2019
贝尔实验室
研制出1G
AMPS系统
1989年中国
开通TACS系统
中国2G
商用
1999
SDTV
2008
HDTV
中国3G
商用
中国4G
商用
2018
4K
中国5G
商用
1G模拟移动通信
1983AMPS商用
/1986TACS商用
2G数字移动
通信系统GSM商用
3G移动
通信系统商用
4G移动
通信系统商用
5G移动
通信系统商用
SMS
WEB

移动互联网影响我们的生活

- **人与人之间互联**：社交（微信，微博）等
- **人与物之间互联**：电商购物（淘宝，亚马逊），共享经济（滴滴）等
- **物与物之间互联**：物联网（智能家居，工业4.0数据）等

移动互联网带给我们什么？

- 人与人之间的关联数据（社交数据）
- 人与物之间的关联数据（行为数据）
- 物与物之间的关联数据（环境数据）

图1-3 数据挖掘：在你的数据中搜索知识（有趣的模式）

让数据产生价值

- **“大数据”**：1）**规模大（Volume）**，从TB级别跃升到PB级别，甚至ZB级别；2）**数据类型繁多（Variety）**，如文本、视频、音频、图片等及其变化组合，**多模态数据挖掘**；3）**速度快（Velocity）**，数据高速持续生成要求实时处理；4）**不确定性（Veracity）**，数据不确定，来源不可信；5）**有价值（Value）**，大量的数据中存在极有价值的信息。
- **从“大数据”中寻找金子——数据挖掘：**从大量的数据中提取出有价值的（ 非平凡的，隐含的，事先未知的，潜在的）模式或者知识。
- **数据挖掘：**
  - 模式发现：沃尔玛的购买模式分析（啤酒+小孩纸尿布），亚马逊图书推荐
  - 趋势预测：天气预报、交易预测

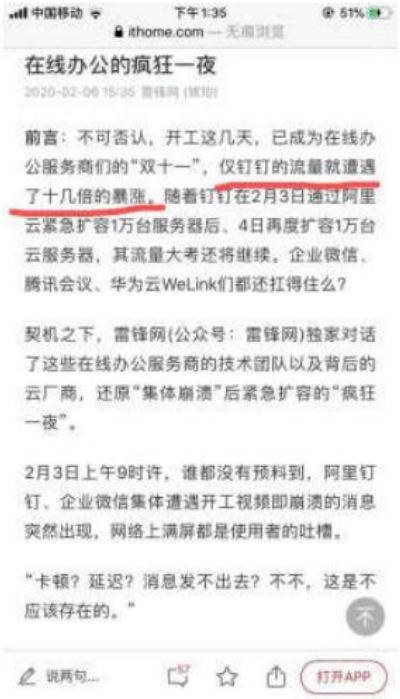
关于“云计算”

- **“云”存储与“云”计算**
  - “云”存储解决大规模数据存储与管理问题：可靠（容灾备份）、稳定（高并发）、低功耗（ARM低功耗技术）等等
  - “云”计算解决大规模数据的挖掘与分析问题：计算量巨大、计算资源的支持

关于“云计算”

- 2019猪年春晚 **百度** 红包（22:50左右百度云红包崩溃，208亿次互动/1.9亿次攻击)

- 钉钉、腾讯会议、Zoom等



什么是互联网+?

- “互联网+”就是“互联网+各个传统行业”，但这并不是简单的两者相加，而是利用信息技术以及互联网平台，让互联网与传统行业进行深度融合，创造新的发展生态。
- **“互联网+”实现：**
  - **消费者**实现消费升级：物美价廉，方便
  - **企业**实现产业升级与转型：为消费者提供更加方便和优质的服务，但是不能背离商业的本质。

互联网+生活（衣食住行）——O2O模式生活方式

- O2O模式的外卖平台
  - 美团外卖 / 饿了么+百度外卖
  - 基于地理位置的推荐：大众点评
- O2O模式的出行工具
  - 出行工具：**滴滴**出行（快的、Uber）
  - MoBike, **OfO**
- O2O模式的短租公寓
  - AirBnB 与 Booking、Agoda
  - **小猪短租** 与 携程、艺龙
- O2O模式的服装定制
  - O2O电商平台
  - O2O服装定制
  - O2O服装众筹

■ 互联网+金融——P2P互联网金融、众筹

- **P2P模式的互联网理财**
  - 移动支付：非银行机构的交易数已经是**970**亿笔（2016年底），超过了传统商业银行的**257**亿笔。
  - 理财产品：余额宝(**余额宝是今天世界上最大的货币基金**)
  - 机器人投资顾问：为金融投资服务降低了门槛，使美国过去在50万～100万美元为起点的理财降低到今天的5万美元，费用从5%降到0.3%～0.5%。
  - 问题：风控（与传统金融行业相比，2018年 国内P2P暴雷）、基础设施标准
- **P2P模式的互联网众筹**
  - 众筹购买新产品、众筹拍摄电影、股权众筹
  - 区块链与数字货币

概念梳理

- “**移动互联网**”是实现人-物两两之间互联互通的基础设施
- “**物联网**”是在移动互联的基础上解决物和物之间互联互通的实际应用
- “**大数据**”和“**云**”技术是在移动互联的基础上让数据产生价值的技术，以“数据挖掘”方法为技术手段来实现
- “**互联网+**”是在上述工作的基础上实现传统产业的转型升级

应用背景

- 各行各业积累了一定规模或超大规模的数据信息
- 数据的爆炸性增长：从 terabytes (TB) 到 petabytes (PB)
- 数据采集与数据的实用性
- 例. 中国网页规模的变化(2012-01)

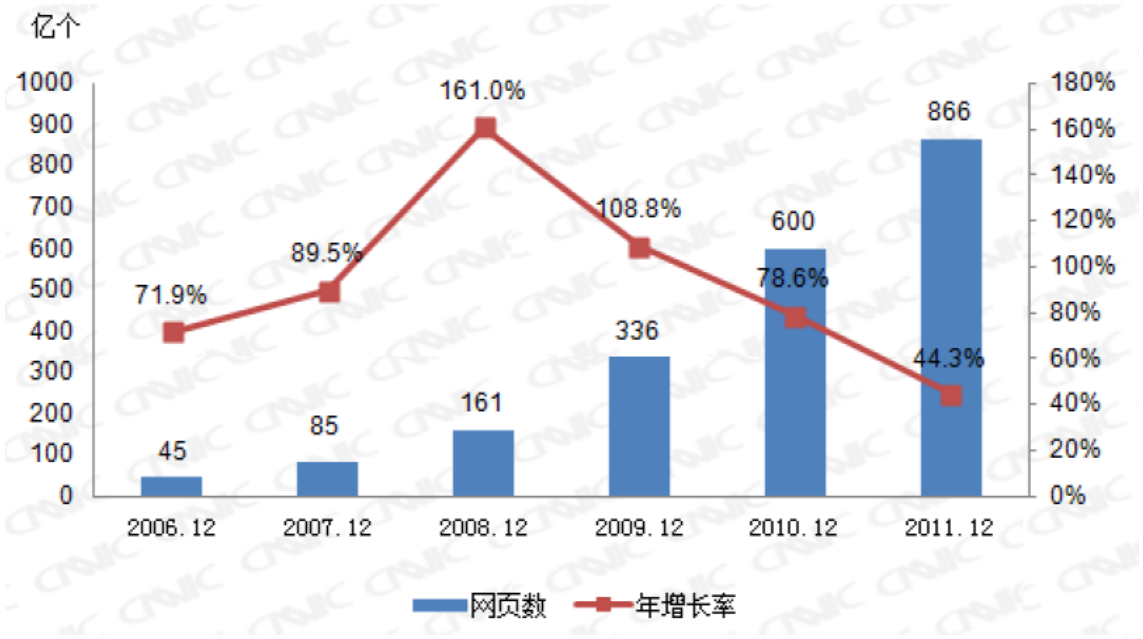


图 18 中国网页规模变化情况

CNNIC 统计的中国网页数量 (2012-01)

应用背景

- 一门跨越数据库技术、信息检索技术、算法、统计学和机器学习等领域的新兴研究领域—“数据挖掘”应运而生。

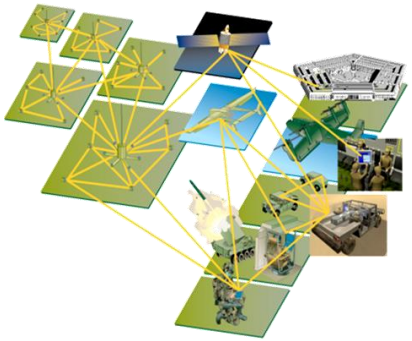
- 数据挖掘方法与技术发展的几个关键因素：
  - 商业上的驱动
  - 科学研究上的驱动
  - 数据挖掘与数据库技术

商业动机的观点

- **商业观点—数据挖掘重要的应用领域之一**
  - 数据来源：网页数据，电子商务，在商场/杂货店的购物统计，银行/信用卡，交易记录
  - 电脑变得越来越便宜，性能也越来越高
  - 竞争压力大
    - 提供更好、更个性化的服务以取得优势（例如：在客户关系管理方面）

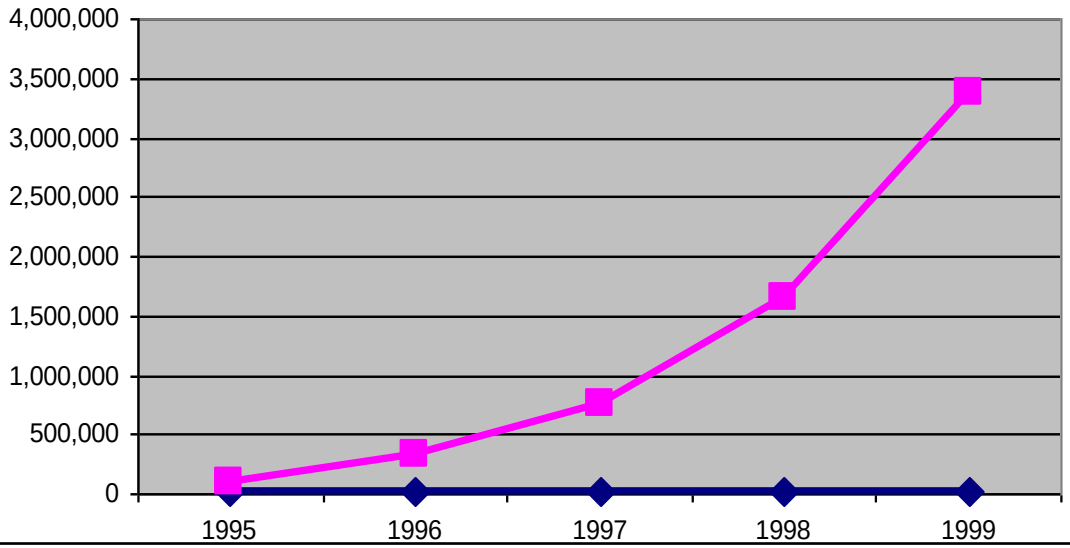
科学动机的观点

- 科学观点
  - 数据在以非常高的速度进行采集和储存(GB/小时)
    - 卫星上的远程传感器
    - 扫描天空的望远镜
    - 产生遗传表达数据的微振列芯片
    - 产生terabytes数据量的科学模拟
  - 传统技术处理原始数据不可行
  - 数据挖掘或许可以帮助科学家
    - 在数据分类和数据细分方面
    - 在假说的形成方面



动机：为什么需要数据挖掘？

- 数据里经常有一些并不是很明显的“隐藏”的信息
- 人们可能会花费数周的时间才能发现有用的信息
- 许多数据根本就没有被分析。“我们淹没在数据里，却没获取到足够的知识
- “需要是发明之母”——数据挖掘——大量数据集的自动分析



From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”

数据库技术的演化

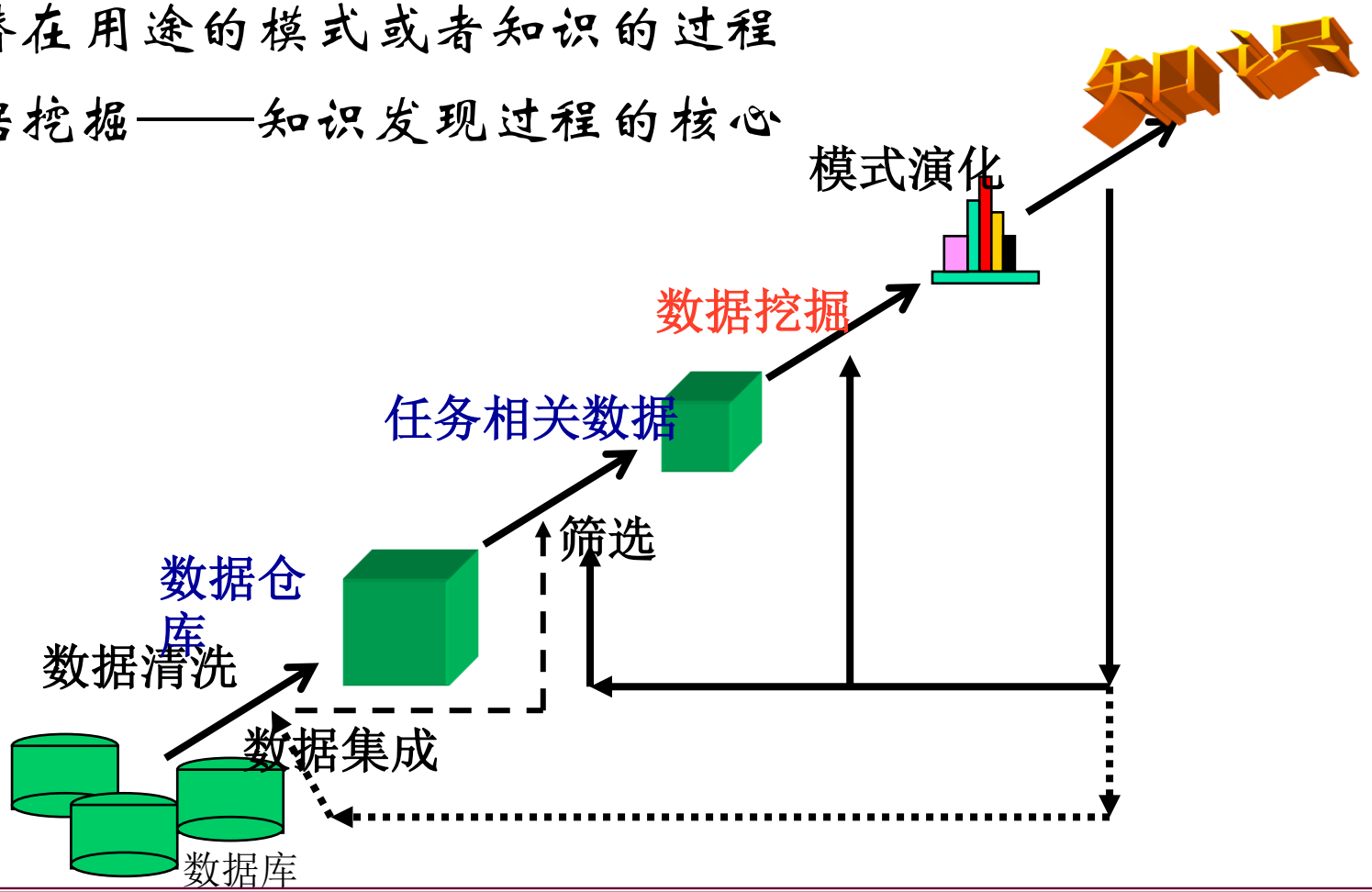
- 1960s:
  - 数据采集，数据库创立，IMS和网络数据库管理系统
- 1970s:
  - 关系型数据模型，关系型数据库(DBMS)实现
- 1980s:
  - 高级数据模型RDBMS （扩展关系，面向对象，演绎等）
  - 应用为导向的DBMS（空间的，科学的，工程的，等）
- 1990s:
  - 数据挖掘，数据仓库，多媒体数据库，网页数据库
- 2000s
  - 流数据管理和挖掘
  - 数据挖掘与应用
  - Web技术（XML，数据整合）和全球信息系统

什么是数据挖掘？

- **数据挖掘（从数据中发现知识）**
  - 从大量的数据中提取出有趣的（ 非平凡的，隐含的，事先未知的，潜在的）模式或者知识
- **别称**
  - 从数据库发现知识（KDD）
  - *知识抽取*
  - *数据/模式分析*
  - *数据考古*
  - *数据捕捞*
  - *信息收获*
  - *商业智能*
- **注意：所有东西都是“数据挖掘”么？**
  - 简单搜索和查询处理
  - （演绎）专家系统

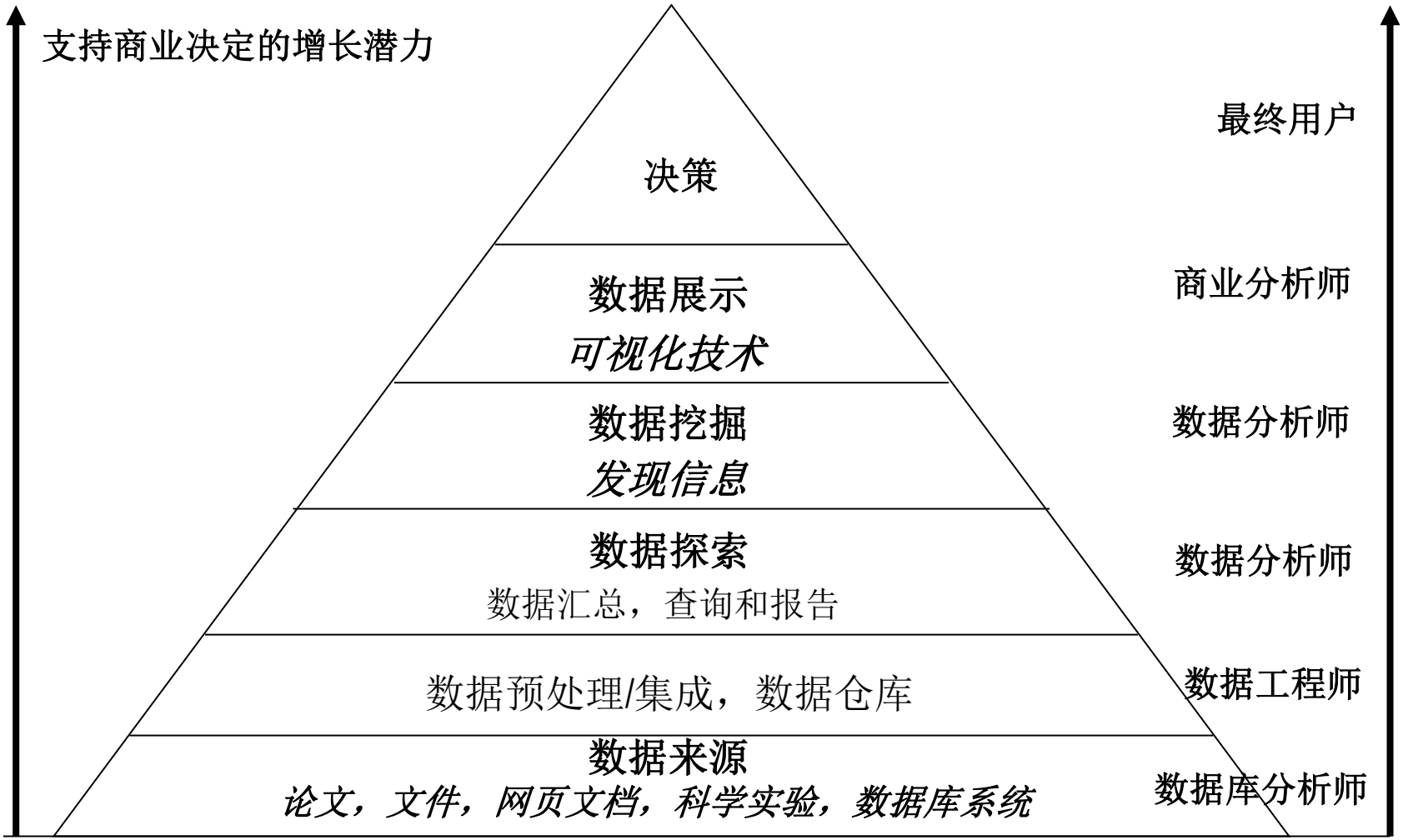
知识发现过程

- 知识发现—从一组大规模或海量数据中发现和挖掘新的具有潜在用途的模式或者知识的过程
- 数据挖掘——知识发现过程的核心



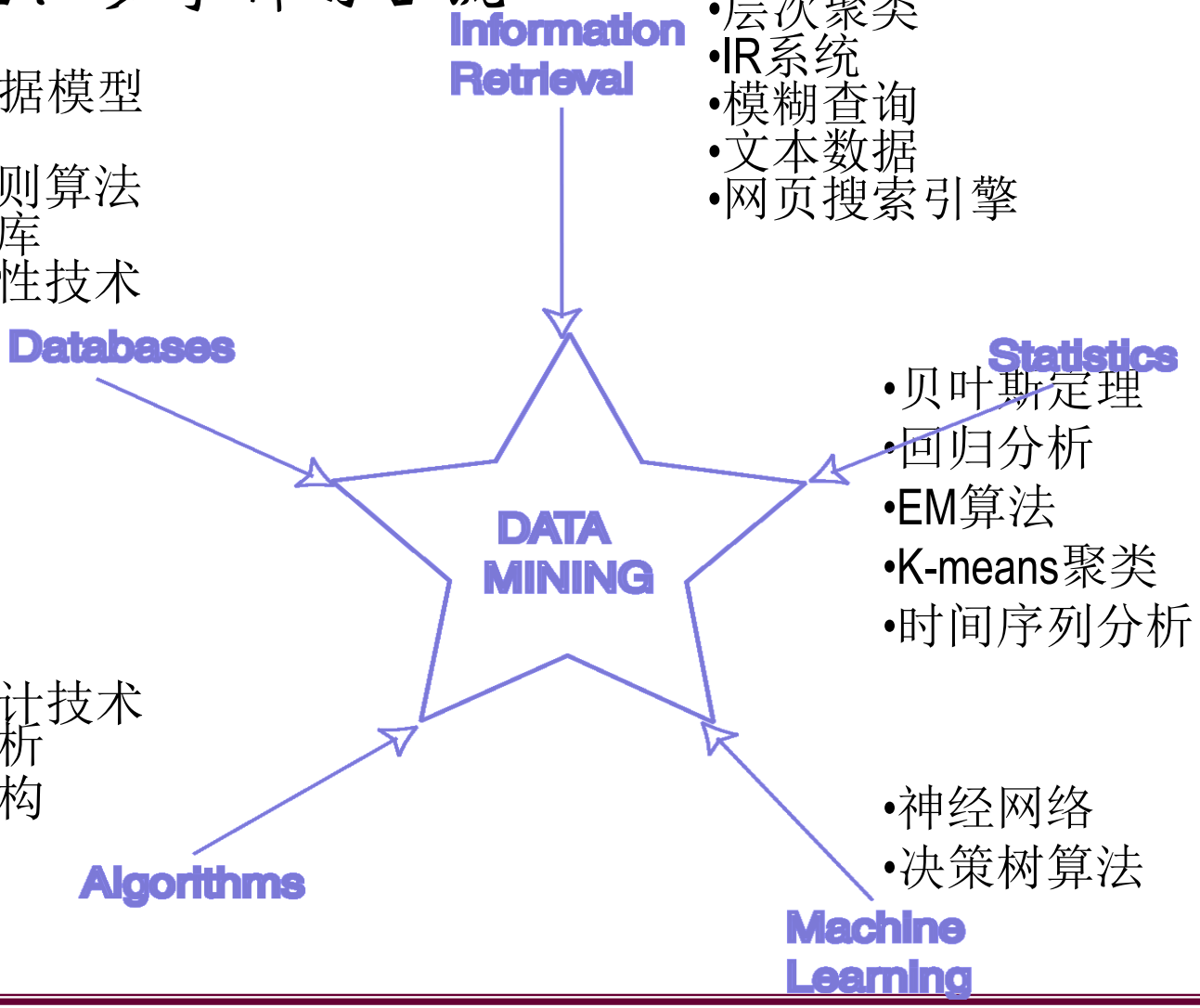
数据挖掘 vs KDD

- **数据库中的知识发现(KDD):**
  - 在数据中发现有用信息和模式的过程.
- **数据挖掘:**
  - 用算法抽取从KDD过程中衍生出的信息和模式.


支持商业决定的增长潜力
决策
最终用户
数据展示
可视化技术
商业分析师
数据挖掘
发现信息
数据分析师
数据探索
数据汇总，查询和报告
数据分析师
数据预处理/集成，数据仓库
数据工程师
数据来源
论文，文件，网页文档，科学实验，数据库系统
数据库分析师

数据挖掘的主要技术

数据挖掘：多学科的合流



数据挖掘的主要技术

- 为何不能用传统的数据分析方法？
- 数据量太大
  - 算法必须具有高扩展性以处理tera-bytes的数据
- 数据的高维度性
  - 微阵列可能有成千上万个维度
- 数据的高复杂性
  - 流数据和传感数据
  - 时间序列数据，时空数据，序列数据
  - 结构数据，图片，社会网络和多关联联数据
  - 异构数据库和遗留数据库
  - 空间的，时空的，多媒体的，文本和网页数据
  - 软件程序，科学模拟
- 新的和复杂的应用

数据挖掘的功能（一）

- 多维概念描述：特征抽取和识别
  - 归纳，总结，对比数据的特点，例如，干与湿地区
- 频繁模式，关联，相关性 vs 因果关系
  - 尿布→啤酒[0.5%, 75%] （相关性 还是 因果关系？）
- 分类和预测
  - 构建描述和区分类别或者概念的模型以预测未来数据
    - 例如，基于气候对国家进行分类，或者基于每英里汽油损耗量对汽车进行分类
  - 预测一些未知的或者丢失的数值

数据挖掘功能（二）

- **聚类分析**
  - 分类标签未知：组合数据形成新的类别，例如，通过聚类房屋寻找分布模式
  - 最大化类内部的相似性，最小化类间的相似性
- **离群点分析**
  - 离群点：与其它数据的一般行为不一致的数据
  - 噪音或者是例外？在欺诈检验和小概率事件分析中很有用
- **趋势和演化分析**
  - 趋势和偏差：例如，回归分析
  - 序列模式挖掘：例如，数码相机—>大SD存储卡
  - 周期分析
  - 基于相似性分析
- **其他模式指导或统计分析**

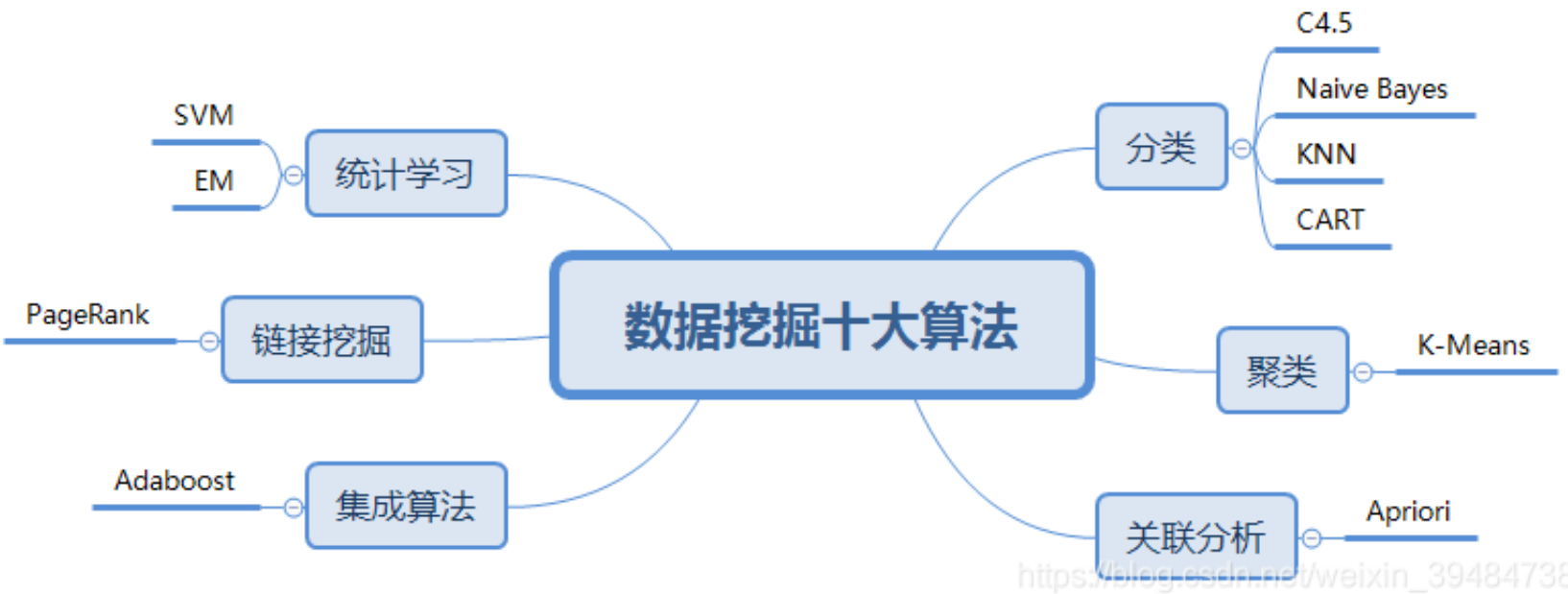
主要研究内容

- **数据挖掘的分类**
  - 描述性数据挖掘

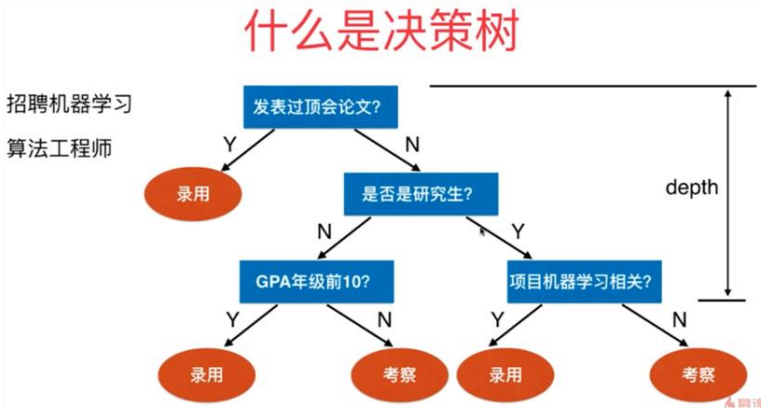
    分析其中隐含的规律性描述，例如频繁模式、关联规则
  - 预测性数据挖掘

    开展对于未知规律和知识的预测研究，比如分类、聚类
- **不同角度导致不同的分类**
  - 数据角度：被挖掘的数据种类，比如流数据、图数据、空间数据
  - 知识角度：被发现的知识的种类，比如广义知识 、关联知识
  - 方法角度：所用技术的种类，统计方法、聚类分析方法
  - 应用角度：采用的应用的种类

数据挖掘十大算法


数据挖掘十大算法
统计学习
SVM
EM
链接挖掘
PageRank
集成算法
Adaboost
分类
C4.5
Naive Bayes
KNN
CART
聚类
K-Means
关联分析
Apriori

- 分类算法



▪决策树分类器C4.5：

决策数包含：决策结点、分支、叶子

▪KNN—k近邻分类算法：

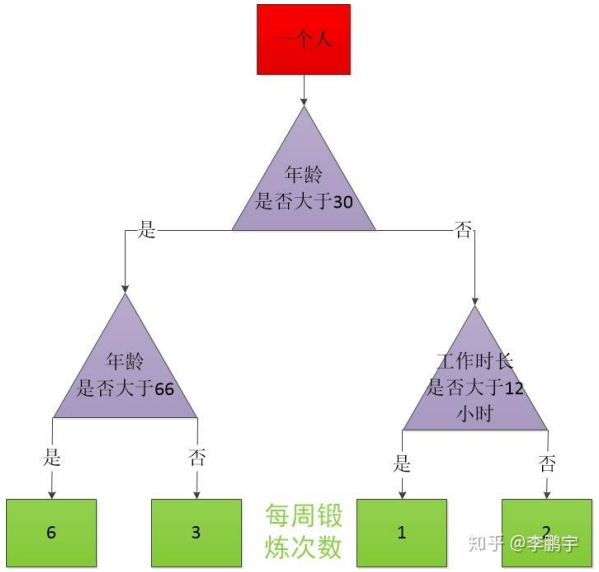
依据最邻近的一个或者几个样本类别来决定待分类样本

▪朴素贝叶斯算法：

分类问题转换为概率问题

▪CART-分类与回归树算法

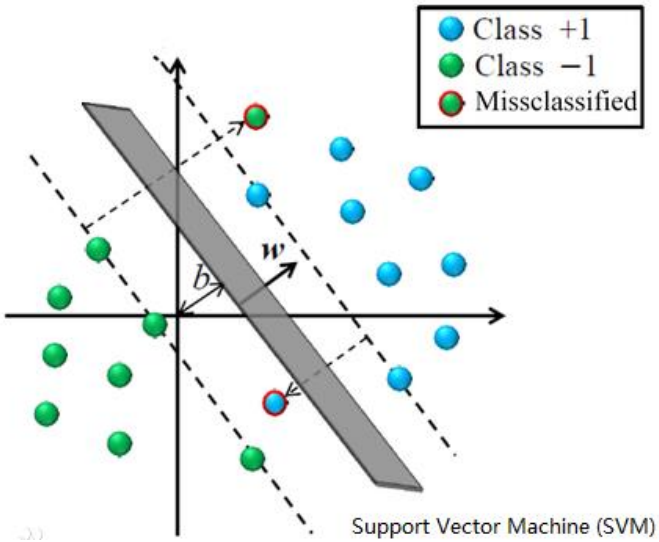
一种决策树分类方法，采用基于最小距离的基尼指数估计函数，用来决定由该子数据集生成的决策树的拓展形。



分类算法

支持向量机

将低维空间的点映射到高维空间，使它们成为线性可分，再使用线性划分的原理来判断分类边界。



AdaBoost算法

针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)。

AdaBoost 算法广泛的应用于人脸检测、目标识别等领域。

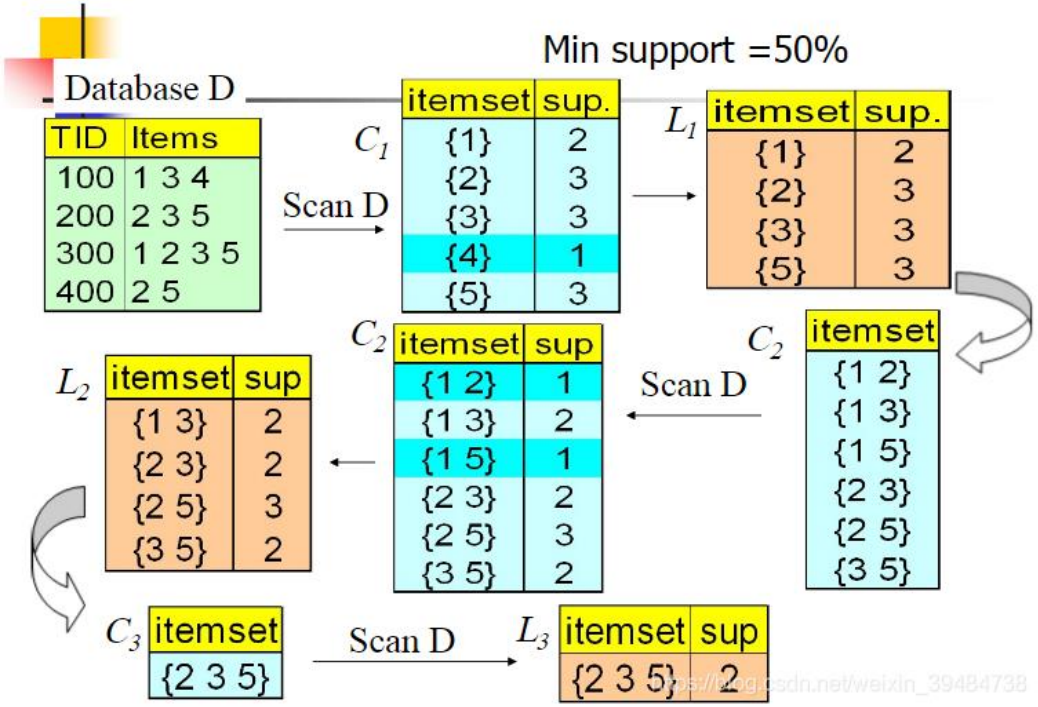
关联分析

Apriori算法

Apriori 算法分为两个阶段：

1）寻找频繁项集。

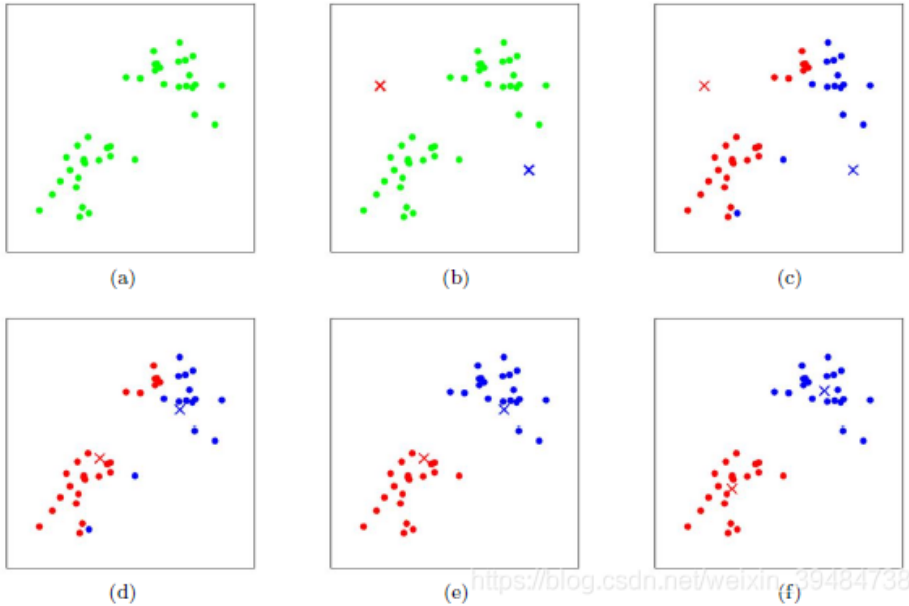
2）由频繁项集找关联规则



聚类算法

K-means算法—k均值算法:

按照样本之间的距离大小，将样本集划分为K个簇。



EM最大期望估计算法

基于模型的聚类方法，是在概率模型中寻找参数最大似然估计的算法。

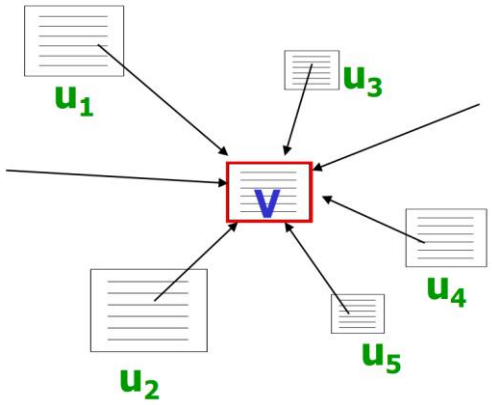
EM 经常用在机器学习和计算机视觉的数据集聚（Data Clustering）领域。

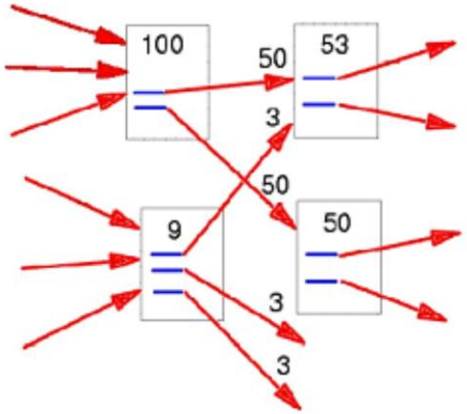
排序算法

- PageRank算法：

google 的页面排序算法，是基于从许多优质的网页链接过来的网页，必定还是优质网页的回归关系，来判定所有网页的重要性。

- 也就是说，一个人有着越多优质朋友的人，他是优质的概率就越大。
- 通过此算法调整网页搜索中的排序，提高搜索结果的相关性和质量。





数据挖掘中的主要问题

- **挖掘方法**
  - 从不同的数据类型挖掘不同类型的知识，例如，生物，流，网络
  - 性能：效率，效益和可扩展性
  - 模式评价：趣味性问题
  - 背景知识的合并
  - 噪音和不完整的数据处理
  - 并行，分布式和增量挖掘方法
  - 用现有的与发现的知识进行整合：知识融合

数据挖掘中的主要问题

- **用户互动**
  - 数据挖掘查询语言和即时挖掘
  - 数据挖掘结果的表达和可视化
  - 在多个抽象层次交互式挖掘知识
- **应用程序和社会影响**
  - 特定域的数据挖掘及无形数据挖掘
  - 数据安全性，完整性和隐私保护

Thank you !!!