数据挖掘与隐私保护

李涛

目录
CONTENTS

01 大数据隐私

大数据隐私

- 到了大数据时代，你就成了皇帝！ “穿新衣的皇帝”
  - 说过什么话，做过什么事，有什么爱好，生过什么病，家住哪里，亲朋好友都有谁……
  - 你自己知道的，它几乎都知道

大数据隐私

- 你自己都不知道的事情，大数据也可能知道
  - 集体照相时喜欢站在哪里，跨门槛时喜欢先迈左脚还是右脚，喜欢与什么样的人打交道，性格都有什么，哪位朋友与你观点不相同……
- 今后将要发生的事情，大数据也有可能知道
  - “饮食多、运动少”等信息→可能会“三高”
  - 许多人都在购买感冒药→流感即将爆发

大数据隐私

- 对象范围广泛
  - 个人
  - 家庭
  - 单位
  - 民族
  - 国家
- 大数据正在影响着世界的发展

大数据隐私

- 韩国的朴槿惠，被来自网络的导弹击中，身败名裂
  - 梨花女子大学走后门破格录取了一位富二代—郑维罗
  - 几位抗议者对郑维罗“人肉搜索”，发现郑同学的父亲曾是朴总统担任议员时的秘书长，秘书长的前期是朴总统的闺蜜，闺蜜的老爸曾被认为是总统的“导师”
  - 网友深挖硬刨，地毯搜索，发现处级干部都不是的闺蜜，竟然提前收到，并无偿修改过至少44份“正国级”总统演讲稿
  - 总统道歉，大检察厅设立特别检查组，核心幕僚辞职，总统府改组，闺蜜以“亲政干政”被逮捕，民众游行，总统下台

大数据隐私

- 希拉里邮件门
  - 支持率比对方高出12%，即将成为美国历史上首任女总统
  - 维基解密揭露了隐私：通过截获并分析她的私人邮件，发现在担任国务卿期间，竟然“假私济公”，用私人邮箱收发公家邮件
  - 赶紧删除了相关邮件，但被删除的邮件任然被恢复了
  - 美国国务院规定：国务卿的日常工作的相关业务，应该在经过授权的服务器上处理
  - 再接再厉穷追猛打，挖掘出更多隐私炸弹
  - 支持率塌方，总统梦破碎

什么是大数据?

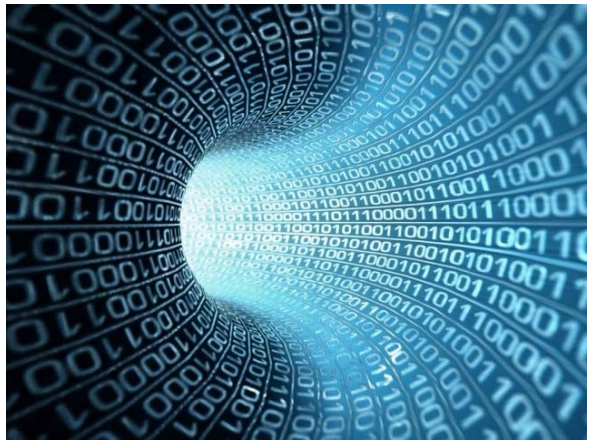
- 国际权威机构Gartner
  - 大数据，就是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产
- 麦肯锡全球研究院
  - 大数据是一种规模大到在获取、存储、管理、分析方面大大超出了传统的数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征
- 更多权威专家们总结了大数据的其它特性
  - 容量的超大性，种类的多样性，获取的快速性，管理的可变性，质量的真实性，渠道来源的复杂性，价值提取的重复性……

什么是大数据?

- 通俗的，形象的理解：许多千奇百怪的数据，被杂乱无章地堆积在一起
  - 主动的：网上说的话、发的微博微信、存放的照片、收发的电子邮件、留下的诸如上网记录等行动痕迹等
  - 被动采集的：马路摄像头捕获的视频、手机定位系统留下的路线图、录下的语音、驾车时的GPS信号、电子病历档案、公交刷卡记录等
  - 各种传感器设备采集到的万物信息：温度、湿度、速度等
- 每个人、每种通信和控制类设备，无论是软件还是硬件，都是大数据之源
- “不对外提供信息”本身，就是一条重要的信息，说明此人必有超级秘密

什么是大数据?

- 无论你是否喜欢，大数据就在那里；无论主动还是被动，你都在为大数据做贡献。大数据是人类的必然!



大数据挖掘

- “大数据挖掘”技术：采用了神经网络、遗传算法、决策树方法、粗糙集方法、覆盖正例排斥反例方法、统计分析方法、模糊集方法等
- “大数据挖掘”过程：数据收集、数据集成、数据规约、数据清理、数据变换、挖掘分析、模式评估、知识表达

大数据挖掘

- “大数据产业”和“垃圾处理废品回收产业”存在相似性
- 工作原理：
  - 数据收集—废品收购和垃圾收集
  - 数据集成—将废品和垃圾送往集中处理工厂
  - 数据规约—将废品和垃圾初步分类
  - 数据清理—将废品和垃圾适当清洁和整理
  - 数据变换—将破沙发拆成木、铁、皮等原料
  - 数据挖掘—认真分析如何将这些原料卖个好价钱
  - 模式评估—不断总结经验，选择并固定上下游卖家和买家
  - 知识表示—把这些技巧整理成口诀

大数据挖掘

- 原料结构
  - 大数据具有异构特性—生活垃圾、工作垃圾、建筑垃圾、可回收垃圾和不可回收垃圾等，在外形、质地，还是内涵等方面看，都是完全不相同的
  - 数量非常多，产生的速度也很快，处理起来也很困难
  - 本质区别：垃圾由原子组成，处理一次后，就没得处理了；大数据由电子组成的，可以反复处理，反复利用

大数据挖掘

- 利润率
  - 只要垃圾分类专家们愿意认真分拣，利润率可以超过任何相关行业。易拉罐转手卖掉，胜过铝矿利润率，旧家具拆成木材和皮料，利润也高于木材商和皮货商
  - 大数据专家将数据中挖掘出的旅客出行规律卖给航空公司，将某群体的消费习惯卖给百货商店，将网络舆情卖给相关的需求方
  - 大数据专家可以“一菜多吃”，反复卖钱，不断“冶金”，而且一次比一次赚钱，时间越久，价值越大

大数据隐私

- 大数据挖掘，从正面来说是创造价值，从负面来说就是泄露隐私
- 分解经典的“人肉搜索”
  - 一大群网友处于某种目的，充分利用自己的一切资源渠道，尽可能多的收集当事人或物的所有i洗脑洗，包括网络搜索到的信息、道听途说的信息、线下知道的信息、各种猜测信息等
  - 按照自己的目的精炼成新信息，反馈到网上与其同志们分享
  - 完成了第一次“人肉迭代”
  - 在此基础上交叉重复进行信息的收集、加工、整理等工作，诞生了第二批“人肉迭代”
  - 循环往复，经过N次迭代后（发酵），当事人的丑闻（善良）画像就跃然纸上
  - 构成“满意画像”的素材已经坐实，“人肉搜索”成功

大数据隐私

- 大数据挖掘，就是由机器自动完成特殊的“人肉搜素”，人肉的目的不再限于抹黑或颂扬某人，而是有更加广泛的目的
  - 为商品销售寻找买家
  - 为某类数据寻找规律
  - 为某些事物之间寻找关联等
- 只要目的明确，大数据就有用武之地

大数据隐私

- “人肉”与大数据挖掘存在类似
  - 网友被电脑替代
  - 网友们搜集的信息，被数据库中的海量异构数据替代
  - 网友寻找各种人物关联的技巧，被相应的智能算法替代
  - 网友们互相借鉴、彼此启发的做法，被各种同步运算替代
  - 机器迭代次数更多，速度更快，每次迭代就是一次机器的“学习”过程
  - 网友们最终“满意画像”被暂时的挖掘结果替代
  - 对于大数据挖掘来说，永远没有尽头，结果会越来越精确，智慧程度越来越高，用户根据自己的标准随时选择满意的结果

大数据隐私

- 就当前的情况来说，“大数据隐私挖掘”的杀伤力远远超过了“大数据隐私保护”所需要的能力
- “隐私保护”和“隐私挖掘”之间对抗演化，对隐私的挖掘获得好处的同时又产生了更多需要保护的隐私

大数据隐私

- 如何进行隐私保护？需要多管齐下
  - 法律上禁止“人肉搜索”为目的的挖掘行为
  - 增加“网民的被遗忘权”等法律条款
  - 对一些恶意的大数据行为进行发现、监督和管控
  - 重塑“隐私”的概念，

如何保护隐私?

- 澡堂着火的故事：匿名
- 只要做好匿名工作，“对大家都一样”的东西就无从谈论“隐私泄露”，就是无本之木、无水之源了

如何保护隐私?

- 匿名的重点:
  - 身份匿名
  - 属性匿名
  - 关系匿名
  - 位置匿名
- 在大数据之前，隐私保护要求：把“私”藏起来，身份可公开
- 大数据隐私保护要求：把“私”公开，身份藏起来，即匿名

如何保护隐私?

- 主要的匿名技术包括:
  - 基于数据失真的匿名技术
  - 基于数据加密的匿名技术
  - 基于限制发布的匿名技术
- 在隐私保护方面，任何手段都有其局限性，隐私保护的最高境界是：没有隐私（但这是不可能的）
- 不愿意告诉别人就想办法让别人无从知晓，但很难，“若想人不知，除非己莫为”

隐私的历史

- 整个人类史，就是一部“大数据隐私史”，分为两个主旋律：保护隐私和发现隐私
- 隐私=隐+私
  - 有“私”之后，才有需要“隐”的对象，才产生了隐私
  - 有了“隐”之后，才有去发现被隐之“私”的动力，才诞生了“挖掘”去发现的隐私
- 人体自身的隐私
  - 大家都光着屁股，没有隐私，但有大数据挖掘
  - 树叶遮挡身体，产生了人类身体上第一片“隐私”
  - 隐私越来越多.....

隐私的历史

- 身外隐私
  - 人类第一个身外隐私：多余的食物，悄悄挖个坑埋起来，藏就是隐私，藏的越多，隐私越多
  - 家庭诞生，“隐私”的保护和发现不再局限于个体，家庭也有隐私
  - 部落出现，语言出现，“隐私”从物质形态，扩展到声音形态，即信息形态
  - 国家诞生了，文字发明了，隐私信息能够传代，“隐”的方法也更丰富，包括各种密码、咒语、符号等

隐私的历史

- 历史上有多种隐私保护方法
  - 把隐私信息一分为多，合并起来才是“私”，例如兵符，这是大数据挖掘的逆过程
  - “隐私”编制成各种独特的符号，没有额外知识就读不懂
- 法律的制定、道德的建设、文化的形成、武器的发明、技术的进步等，从某种意义上说，都是为了保护各种各样的“隐私”

隐私的历史

- 人类对“隐私挖掘”的兴趣，一点不亚于“隐私保护”
  - 哲学家，挖掘世界的隐私，个体的隐私，寻找理性的本质，物在得关系
  - 科学家，挖掘大自然的隐私
  - 文学家，用艺术去挖掘人类情感的隐私
  - 胡同里的大爷、大妈，挖掘隐私的高手
  - 领导，获得信息的渠道多，通过对数据的挖掘，所得出的决定更加全面
  - 人类各种各样的活动，都在挖掘隐私：星云的隐私、地球的隐私、海洋的隐私、大陆的隐私、动物的隐私、植物的隐私、遗传的隐私、量子的隐私，等等

02 数据挖掘

为什么进行数据挖掘

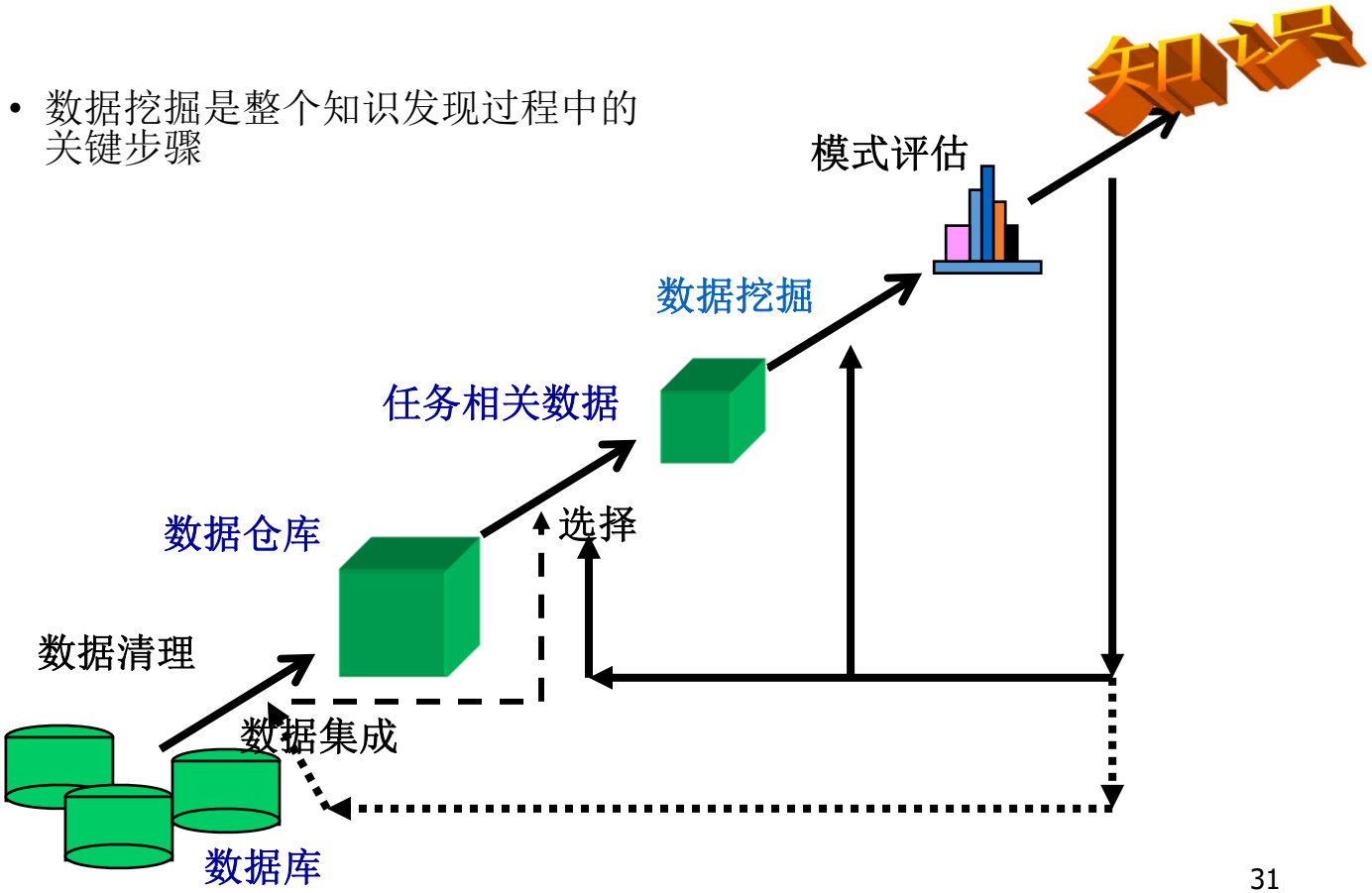
- 数据大爆炸：数兆兆字节（Tera-Byte，TB）或数千兆兆字节（Peta-Byte，PB）
  - 数字搜集和获取工具快速发展的结果
    - 自动化数据搜集工具，数据库系统，Web，社交系统
  - 充足的数据来源
    - 商业: Web, 电子商务, 交易信息, 股票, …
    - 科学: 远程传感器, 生物信息, 科学情报分析, …
    - 社会和每个个体, 数字摄像头，社交网站
- 数据的爆炸式增长、广泛可用和巨大数量使得我们的时代正成为真正的数据时代，急需功能强大和通用的工具，从这些海量数据中发现价值的信息，把这些数据转化成有组织的知识

什么是数据挖掘

- 数据挖掘 (从数据中发现知识)
  - 从大量数据中挖掘有趣模式和知识的过程
  - 将数据坟墓转换成知识金块
- 其他的名字
  - 数据中的知识发现 (KDD), 知识提取, 数据/模式分析, 数据考古, 数据捕获等等

知识发现过程

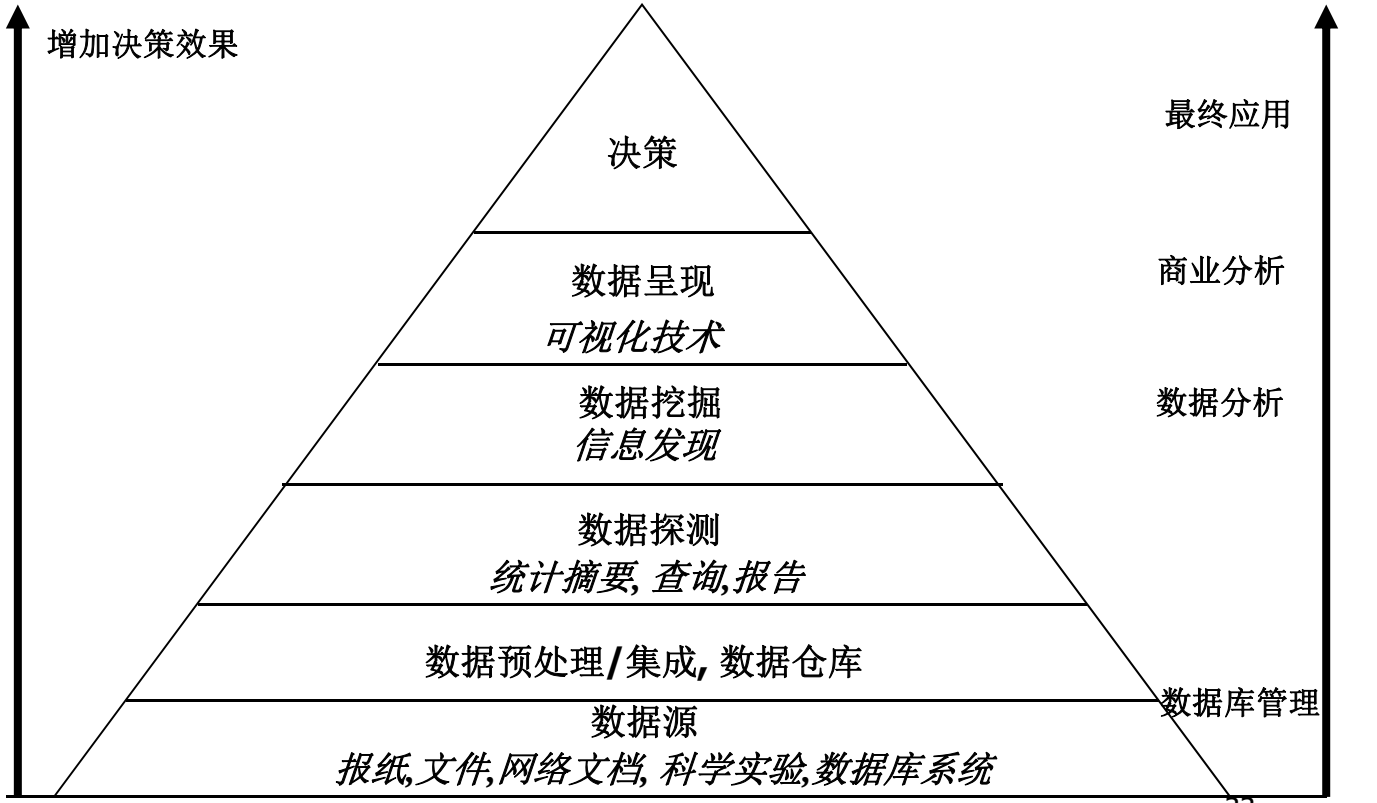
- 数据挖掘是整个知识发现过程中的关键步骤



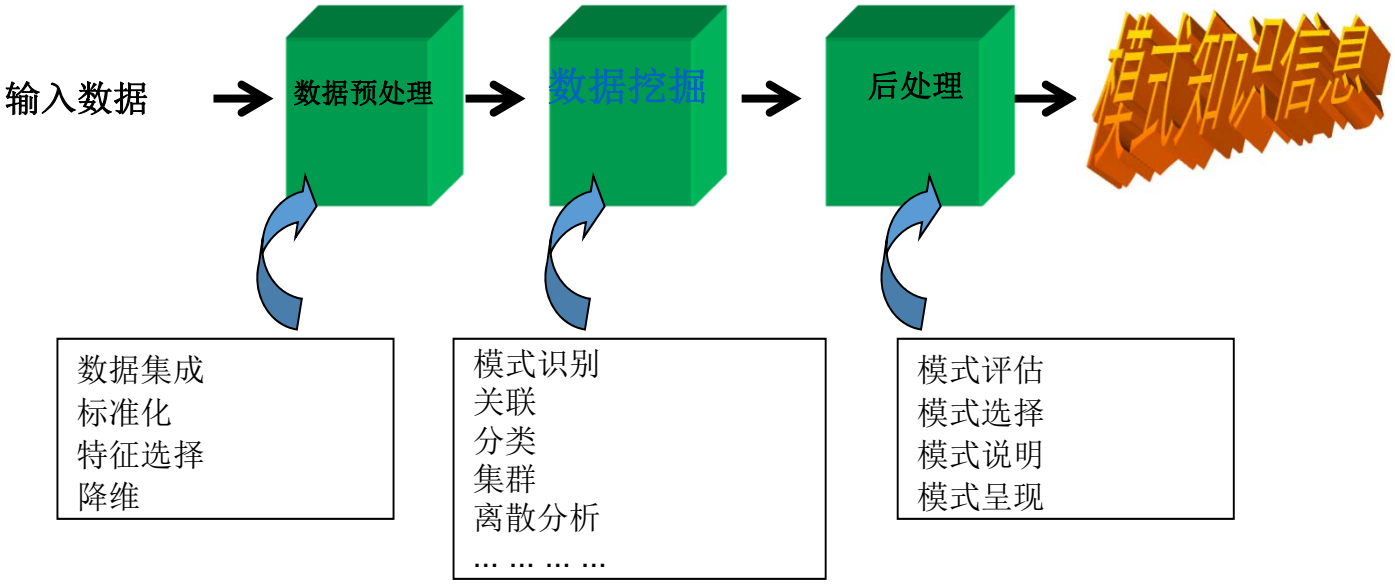
案例：Web挖掘框架

- Web挖掘通常包括
  - 数据清理
  - 对多种数据来源进行数据集成
  - 建立数据仓库
  - 建立数据立方体
  - 数据选择
  - 数据挖掘
  - 呈现挖掘结果
  - 模式和知识存进知识库

数据挖掘在智能商务中的应用



知识发现过程：机器学习和统计分析



- 典型的机器学习和统计分析过程

案例：医疗数据挖掘

- 健康服务和医疗数据挖掘 – 通常采用机器学习和统计分析方法
- 数据预处理(包括特征提取和降维)
- 分类或者聚合
- 后处理，呈现

数据挖掘的多维度

- 待挖掘的数据
  - 数据库, 数据仓库, 事务数据, 数据流（视频监控和传感器数据），时间相关或序列数据（历史纪录、股票交易记录），空间数据，超文本和多媒体数据，万维网.....
- 可挖掘的模式（数据挖掘的功能）
  - 特性，区别，关联，分类，聚合，趋势，等等
  - 当前分析和预测分析
  - 多维度功能，多种平面的挖掘
- 使用的技术
  - 数据密集，数据仓库，机器学习，统计分析，模式识别，可视化，高性能计算，等等
- 应用范围
  - 零售，通信，银行，诈骗分析，股票分析，网络挖掘，等等

对什么样的数据进行挖掘?

- 数据库相关的类型和应用
  - 关系型数据库，数据仓库，事物数据
- 其它类型的数据和应用
  - 数据流和传感器数据
  - 时间相关数据，序列数据
  - 地图，社会和信息网络数据
  - 异构数据和遗留数据
  - 空间数据和时间数据
  - 多媒体数据
  - 文本数据
  - 万维网

数据挖掘功能：泛化

- 数据集成和数据仓库构建
  - 数据清洗，转变，集成，多维度数据建模
- 数据立方体技术
  - 计算的简要方法，多维度聚合
  - OLAP (在线分析过程，online analytical processing)
- 多维度概念描述：特征和异常
  - 泛化，摘要，构建数据的特征，例如辨别干燥和潮湿的地域

数据挖掘功能：关联和相关性分析

- 频率项挖掘
  - 沃尔玛超市哪些商品经常一起被购买？
- 关联、相关和因果的关系
  - 典型的关联规则
    - 尿布 → 啤酒[0.5%, 75%] (支持度, 置信度)
- 怎么从大数据中去有效挖掘这些模式和规则？
- 怎么使用这些模式去分类、聚合，还有其它的应用么？

数据挖掘功能：分类

- 分类和预测
  - 基于一些训练样例构建模型
  - 为未来的预测描述和区分类别
    - 例如：根据气候将国家归类，根据油耗将汽车归类
  - 预测一些未知的类别
- 典型的方法
  - 决策树，贝叶斯分类，支持向量机，神经网络，基于规则的分类，基于模式的分类，等等
- 典型应用
  - 信用卡欺骗侦测，定向营销，疾病分类，网页分类，等等

数据挖掘功能：聚类分析

- 无监督学习（例如分类标签未知）
- 一组数据组成新的类别
- 原则：类内的相似性最强，类间的相似性最小
- 应用范围广泛，例如识别顾客的同类子群，城市内顾客位置

数据挖掘功能：离群点分析

- 离群分析
  - 离群：数据对象与数据的一般行为或模型不一致
  - 噪声或者异常? — 一个人的垃圾可能是另一个人的财富
  - 方法: 基于聚类分析方法，基于分类的分析方法, ...
  - 在欺骗探测，罕见事件分析方面很有用

时间和规则：顺序模式，趋势和估计分析

- 顺序，趋势和估计分析
  - 趋势，时间序列和偏差分析: 例如，回归和价值预测
  - 顺序模式挖掘
    - 例如，首先购买相机，再购买大容量SD存储卡
  - 周期分析
  - 相似性分析
- 挖掘数据流
  - 有序的，随机变化，潜在无限的数据流

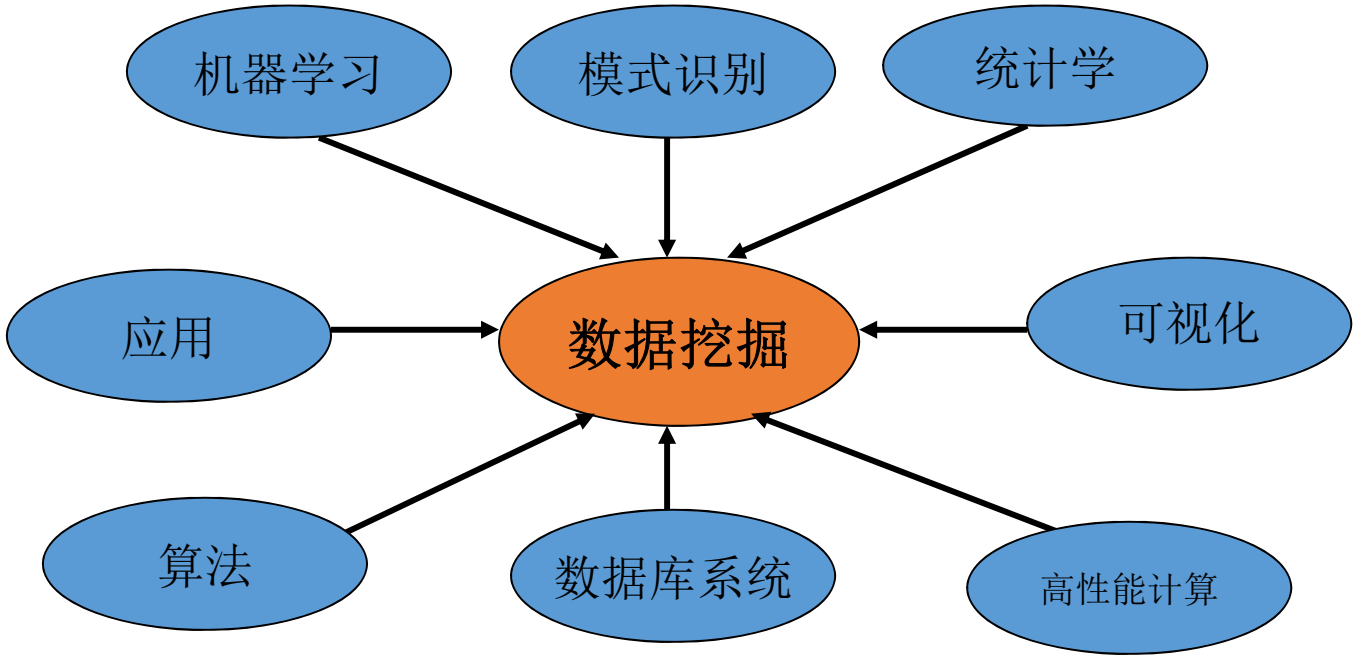
结构和网络分析

- 图片分析
  - 找到频率高的子图（例如，化合物），树状（XML），子结构（Web 片段）
- 信息网络分析
  - 社交网络: 行动者 (对象，节点) 和关系 (边缘)
    - 例如，微博中的人物关系
  - 多种类网络
    - 一个人可以是多种类信息网络：朋友，家人，同学.....
  - 连接承载着语义信息：连接挖掘
- Web 挖掘
  - Web是一个大信息网络：从网页排名到百度
  - 分析Web信息网络
    - Web社区发现，观点挖掘，用途挖掘 ...

知识评估

- 所有挖掘的知识都有有趣的么？
  - 一个人可以挖掘大量的模式和知识
  - 一些可能适合特定维度空间（时间，地点……）
  - 一些可能是非典型的，可能是暂时的 …
- 评估挖掘的知识 → 直接挖掘有用的知识?
  - 描述的 vs. 预测的
  - 覆盖范围
  - 典型的 vs. 新颖的
  - 精度
  - 及时
  - …

数据挖掘：多种技术的集合



为何需要多种技术?

- 巨大的数据
  - 算法必须具有高适应性、可扩展性来处理海量数据
- 高维度数据
  - 基因芯片就具有成千上万的维度
- 高复杂数据
  - 数据流和传感器数据
  - 时间线数据，临时数据，线性数据
  - 结构数据，地图，社交数据，多连接数据
  - 多种类数据库和传统数据库
  - 空间数据，时空数据，多媒体，文本和Web数据
  - 软件代码，科学仿真
- 新颖的复杂的应用不断出现

面向什么应用?

- 网页分析：网页分类，聚合，网页排名
- 协作分析和推荐系统
- 特定市场的购物篮数据分析
- 生物和医学数据分析: 分类和聚合分析（基因芯片分析），生物序列分析，生物网络分析
- 数据挖掘和软件工程
- 从主要的特定数据挖掘系统(e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) 到无形的数据挖掘

数据挖掘的主要问题

- 挖掘方法
  - 挖掘各种新的知识类型
  - 挖掘多维空间中的知识
  - 数据挖掘：跨学科的努力
  - 提升网络环境下的发现能力
  - 处理不确定性、噪声或不完全数据
  - 模式评估和模式或约束指导的挖掘
- 用户界面
  - 交互挖掘
  - 结合背景知识
  - 数据挖掘结果的表示和可视化

数据挖掘的主要问题

- 有效性和可伸缩性
  - 数据挖掘算法的有效性和可伸缩性
  - 并行、分布式和增量挖掘算法
- 数据库类型的多样性
  - 处理复杂的数据类型
  - 挖掘动态的、网络的、全球的数据库
- 数据挖掘与社会
  - 数据挖掘的社会影响
  - 保护隐私的数据挖掘
  - 无形的数据挖掘

03 隐私保护

隐私保护的需求

- 数据挖掘
  - 发现知识的同时，也给数据的隐私带来了威胁，例如疾病预防的数据挖掘将会暴露“病人所患疾病”的敏感数据
- 数据发布
  - 将数据库中数据呈现给用户过程中，不采取适当保护措施将可能造成敏感数据泄露，例如公司的财务年报

隐私保护的需求

- 实施隐私保护技术主要考虑
  - 如何保证数据应用过程中不泄露隐私
  - 如何更有利于数据应用
- 当前，隐私保护领域的研究工作主要集中于如何设计隐私保护原则和算法更好地达到这两方面的平衡

隐私的度量

- 数据隐私的保护效果是通过攻击者披露隐私的多寡来侧面反映的
- 统一用“披露风险”（disclosure risk）来描述
- 披露风险表示为攻击者根据所发布的数据和其它背景知识可能披露隐私的概率。关于隐私数据的背景知识越多，披露风险越大
  - $r(s,K) = P_r(S_k)$
  - s表示敏感数据，事件S_k表示“攻击者在背景知识K的帮助下揭露敏感数据s”，$r(s,K)$表示披露风险

隐私的度量

- 对数据集而言，若数据所有者最终发布数据集*D*的所有敏感数据的披露风险都小于阈值*a*，*a*在0和1之间，则称该数据集的披露风险为a
- 不做任何处理所发布数据集的披露风险为1
- 当所发布数据集的披露风险为0时，这样发布的数据被称为实现了完美隐私

隐私保护技术

- 基于数据失真
  - 使敏感数据失真但同时保持某些数据或数据属性不变
- 基于数据加密
  - 采用加密技术在数据挖掘过程中隐藏敏感数据
- 基于限制发布
  - 根据具体情况有条件地发布数据

基于数据失真的隐私保护技术

- 通过扰动原始数据来实现隐私保护
- 随机扰动，采用随机化过程修改敏感数据
  - 对外界而言，只可见扰动后的数据，实现了对真实数据的隐藏
  - 扰动后的数据仍然保留着原始数据分布X的信息
  - 对扰动后的数据进行重构，可以恢复原始数据分布X的信息，但不能重构原始数据的精确值

输入	1. 原始数据为 $x_1, x_2, \cdots, x_n$，服从于未知分布 X； 2. 扰动数据为 $y_1, y_2, \cdots, y_n$，服从特定分布 Y
输出	随机扰动后的数据：$x_1+y_1, x_2+y_2, \cdots, x_n+y_n$

（a）随机扰动过程

输入	1. 随机扰动后数据：$x_1+y_1, x_2+y_2, \cdots, x_n+y_n$ 2. 扰动数据的分布 Y
输出	原始数据分布 X

（b）重构过程

基于数据加密的隐私保护技术

- 分布式环境下实现隐私保护要解决的首要问题是通信的安全性
- 基于数据加密的隐私保护技术多用于分布式应用中，例如分布式数据挖掘、分布式安全查询等
- 一种能确保分布式环境下隐私安全的模型是k-TTP(k-Trusted Third Party)，k-TTP利用信任第三方,每个站点加密私有数据传递给第三方，当且仅当至少有k个站点的信息改变时，所有站点的相关统计信息才能被披露。k-TTP模型的约束，使我们不能揭露少于k 个站点的统计信息。

基于限制发布的隐私保护技术

- 有选择的发布原始数据，不发布或者发布精度较低的敏感数据
- 数据匿名化一般采用两种基本操作
  - 抑制：抑制某数据项，即不发布该数据项
  - 泛化：对数据进行更概括、抽象的描述。例如，对整数5的一种泛化形式是[3,6]，因为5在区间[3,6]内

数据匿名化

- 数据匿名化所处理的原始数据，如医疗数据、统计数据等，一般为数据表形式：表中每一条记录对应一个个人，包含多个属性值，这些属性值可分为3类：
  - 显示标识符：能唯一标识单一个体的属性，如身份证号、姓名等
  - 准标识符：联合起来能唯一标识一个人的多个属性，如邮编、生日、性别等联合起来可能是准标识符
  - 敏感属性：包含隐私数据的属性。如疾病、薪资等。

数据匿名化

- 表格中为一原始医疗数据，每一条记录对应一个唯一的病人，其中“姓名”为显示标识符属性，{“年龄”，“性别”，“邮编”}为准标识符属性，“疾病”为敏感属性

姓名	年龄	性别	邮编	疾病
Andy	4	M	12000	胃溃疡
Bill	5	M	14000	消化不良
Ken	6	M	18000	肺炎
Nash	9	M	19000	支气管炎
Alice	12	F	22000	流感
Betty	19	F	24000	肺炎

k-匿名

- *k*-匿名原则：所发布的数据表中的每一条记录不能区分与其它*k*-1条记录
- *k*值越大，对隐私的保护效果越好，但丢失的信息越多

年龄	性别	邮编	疾病
[1, 5]	M	[10*k*, 15*k*]	胃溃疡
[1, 5]	M	[10*k*, 15*k*]	消化不良
[6, 10]	M	[15*k*, 20*k*]	肺炎
[6, 10]	M	[15*k*, 20*k*]	支气管炎
[11, 20]	F	[20*k*, 25*k*]	流感
[11, 20]	F	[20*k*, 25*k*]	肺炎

l-diversity

- l-diversity保证每一个等价类的敏感属性至少有*l*个不同的值
- 攻击者最多以1/*l*的概率确认某个体的敏感信息
- 表格中发布的数据也是满足2-diversity的：每一个等价类中至少有2个不同的敏感属性值

年龄	性别	邮编	疾病
[1, 5]	M	[10*k*, 15*k*]	胃溃疡
[1, 5]	M	[10*k*, 15*k*]	消化不良
[6, 10]	M	[15*k*, 20*k*]	肺炎
[6, 10]	M	[15*k*, 20*k*]	支气管炎
[11, 20]	F	[20*k*, 25*k*]	流感
[11, 20]	F	[20*k*, 25*k*]	肺炎

基于聚类的匿名化算法

- 将原始记录映射到特定的度量空间中，再对空间中的点进行聚类来实现数据匿名
- r-gather：以所有聚类中的最大半径为度量，需要达到的目标是：对所有数据进行聚类，在保证每个聚类至少包含k个数据点的同时，也使所有聚类中的最大半径越小越好

基于聚类的匿名化算法

- 发布的结果只包含聚类中心、半径及相关的敏感属性值，同一个等价类中的记录不可区分，因此对个人的敏感信息实现了隐藏

年龄	地址	疾病
a	b	胃炎
$a+2$	b	消化不良
c	$d+3$	感冒
c	d	肺炎
c	$d-3$	感冒

年龄	地址	记录数	疾病
$a+1$	b	2	胃炎 消化不良
c	d	3	感冒 肺炎 感冒

小结

- 数据匿名化由于能够处理多种类型的数据，并发布真实的数据，能满足众多实际应用的需求，因此受到广泛关注
- 数据匿名化是一个复杂的过程，需要同时权衡原始数据、匿名数据、背景知识、匿名化技术、攻击等众多因素