商务智能
Business Intelligence

数据挖掘

朱卫平 博士
计算机学院
武汉大学

目录

- 为什么进行数据挖掘
- 什么是数据挖掘
- 什么样的数据能被挖掘
- 可以挖掘什么类型的模式
- 数据挖掘使用哪些技术
- 数据挖掘应用
- 数据挖掘的主要挑战

为什么进行数据挖掘

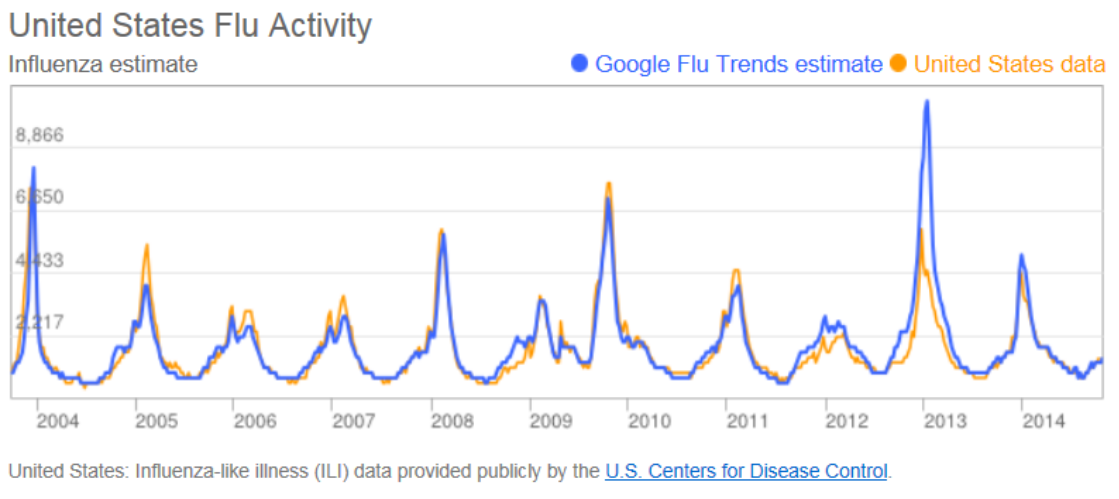
- 强大的、万能的能够自动从大量数据中挖掘有价值的信息的工具被急切的需要。这种需求催生了数据挖掘。
- 这个领域是年轻、动态变化并且前景乐观的。
- 数据挖掘正在并且将会持续的将我们大踏步的从数据时代跃入即将到来的信息时代。

为什么进行数据挖掘

- 举例 数据挖掘将一个大数据集转化成知识
  - 搜索引擎（例如google, baidu, YAHOO!)每天接收到数以亿计的查询请求。每一个请求都看成是用户描述他（她）需要的信息的一个事务。搜索引擎能从如此巨大的查询请求的数据集中学习到什么样新的有用的知识呢?
  - 令人感兴趣的是，我们能从这些数据中发现一些揭示有价值信息的模式。而靠单个查看每个数据记录是无法做到这点的
  - 举个例子，谷歌的Flu Trends使用一些特定的词语作为流感的指示器。它能够发现搜索流感信息的人群的数量与真正有流感症状的人群的数量之间的紧密关系。当所有的关于流感的信息聚集在一起时，就能呈现某种模式。使用聚集的谷歌搜索数据，Flu Trends能比传统系统提早两周估计到流感的发生。

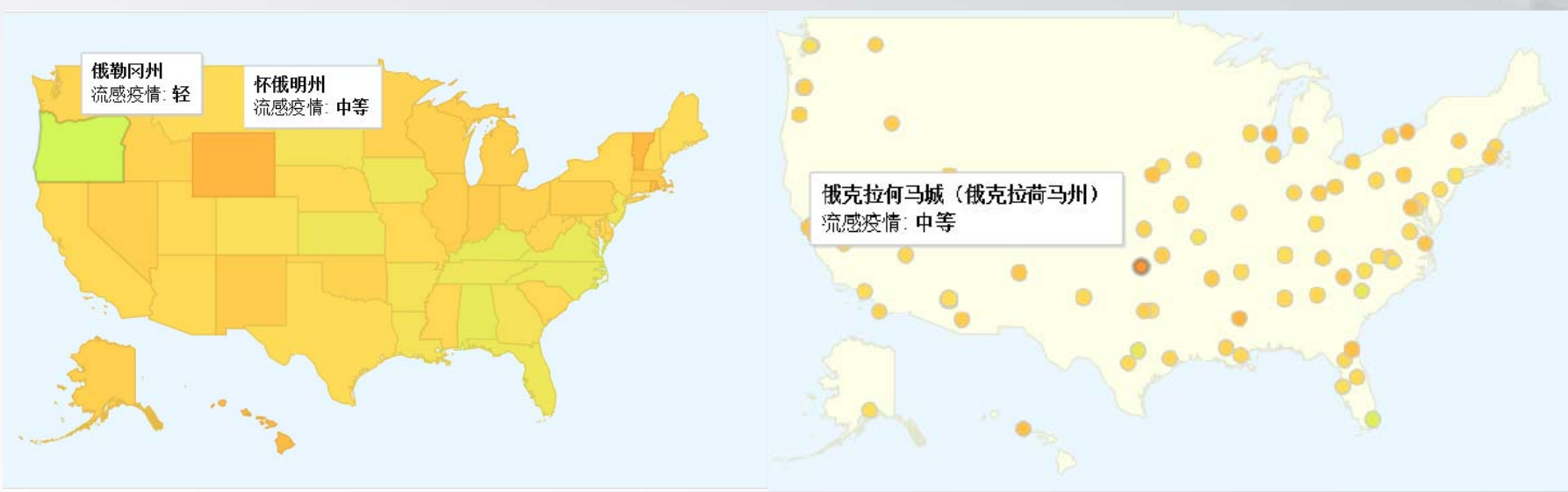
基于大数据的疾病预测

- 谷歌发现，疾病的爆发与人们在网页上的相关搜索高度相关，如流感高发期关于流感的搜索显著增多，过敏高发期关于过敏的搜索显著增多
- 传统的疾病监测数据需要1周左右的时间发布，谷歌设计的Flu Trends 可实时的对疾病爆发进行预测，提升反应速度
- 已在美,法,德，日等20个国家使用



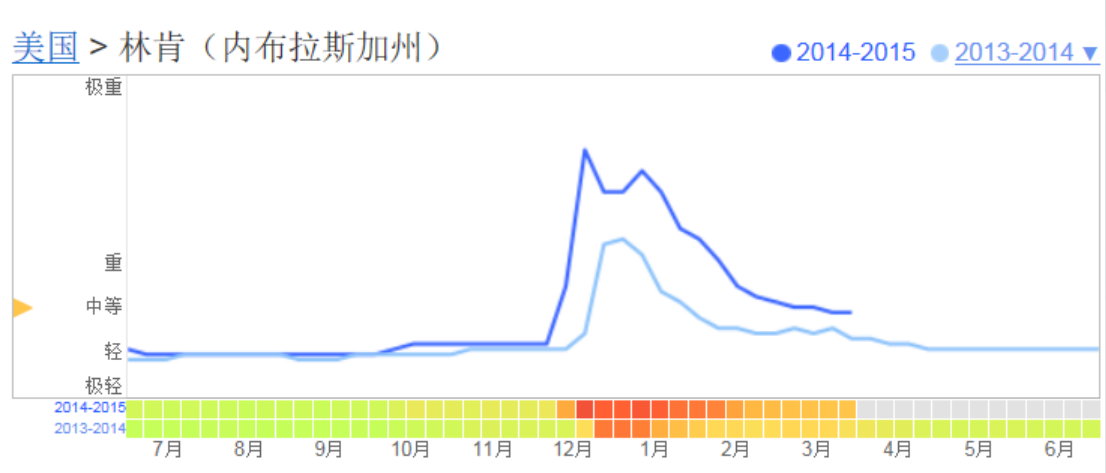
流感预测数据与官方提供数据高度一致

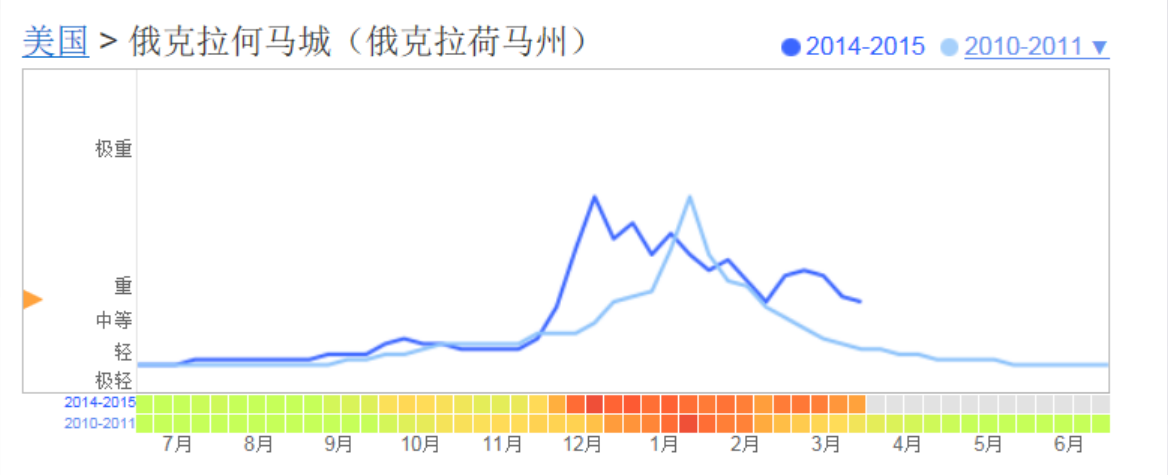
基于大数据的疾病预测



美国各州及城市流感疫情分析, 数据截止2015年4月1日

基于大数据的疾病预测





林肯与俄克拉何马城当前流感趋势与历史趋势数据

(截止2015年4月1日)

为什么进行数据挖掘

- 数据挖掘技术可以被看做是信息技术自然进化的产物
  - 数据库和数据管理技术发展的几个阶段：数据收集和数据库创建、数据管理（数据存储，检索和数据库事务处理）、高级数据分析（数据仓库和数据挖掘）。
  - 从1960年开始，数据库和信息科技开始从最初的文件处理系统进化到更复杂和功能更强大的数据库系统。
  - 从1970年开始，对数据库系统的研究从早期的层式结构和网状结构发展到关系数据库系统。
  - 数据库管理系统建立之后，数据库技术发展到高级数据库、数据仓库和数据挖掘阶段

为什么进行数据挖掘

- 丰富的数据、对多种数据分析工具的需求，被称为是“数据丰富但是信息量少”的环境，这种巨大的鸿沟催生了数据挖掘工具的系统化发展，把数据坟墓转化为知识金矿。

目录

- 为什么进行数据挖掘
- 什么是数据挖掘
- 什么样的数据能被挖掘
- 可以挖掘什么类型的模式
- 数据挖掘使用哪些技术
- 数据挖掘应用
- 数据挖掘的主要挑战

什么是数据挖掘?

- **数据挖掘**是从大量数据中发掘有趣的模式和知识的过程。
  - 很多词语有和数据挖掘类似的含义：数据知识挖掘、知识抽取、数据/模式分析、数据考古……
  - 很多人把数据挖掘看做和一个流行的词汇knowledge discovery from data(KDD，知识发现)一样的含义。

数据挖掘的步骤

1 **数据清理**（去除噪声和不一致的数据）

2 **数据集成**（多种数据源的融合）

3 **数据选择**（从数据库中提取和分析任务相关的数据）

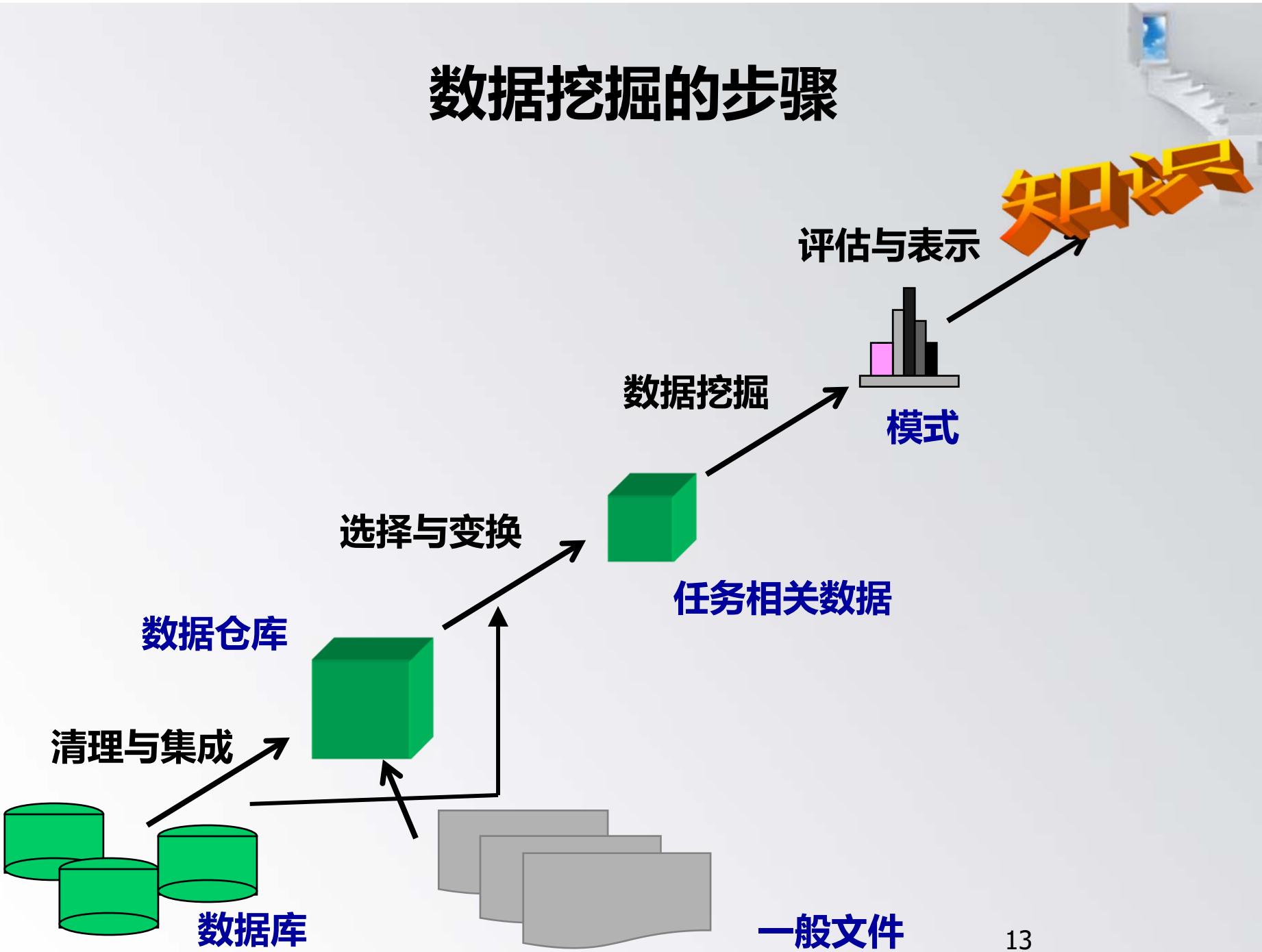
4 **数据转换**（把数据变换和统一成成适用于挖掘的形式）

5 **数据挖掘**（使用智能方法提取数据模式）

6 **模式评估**（根据某种兴趣度度量，识别代表知识的真正有趣的模式）

7 **知识表示**（使用可视化和知识表示技术，向用户提供挖掘的知识）

数据挖掘的步骤



目录

- 为什么进行数据挖掘
- 什么是数据挖掘
- 什么样的数据能被挖掘
- 可以挖掘什么类型的模式
- 数据挖掘使用哪些技术
- 数据挖掘应用
- 数据挖掘的主要挑战

什么样的数据能被挖掘?

- 数据挖掘能被应用于任何对目标应用有意义的数据类型。
  - 数据库数据
  - 数据仓库数据
  - 交易事务数据
  - 其他的类型，例如数据流、序列数据、图数据、空间数据、文本数据、多媒体数据等

什么样的数据能被挖掘-数据库

- DBMS（数据库管理系统）包含一系列相互关联的数据。
- 关系数据库是一系列的表，表都有表名，一系列的属性，和一系列的记录。关系数据库可以通过数据库查询语句来检索记录。
- 对关系数据库挖掘时，是想要发现趋势或者数据模式。比如，分析客户数据预测新用户的信用风险，基于他们的收入、年龄和以前的信用信息。还可以用来发现差异，比如，发现包装商品或者显著提升价格的变化。

什么样的数据能被挖掘-数据库

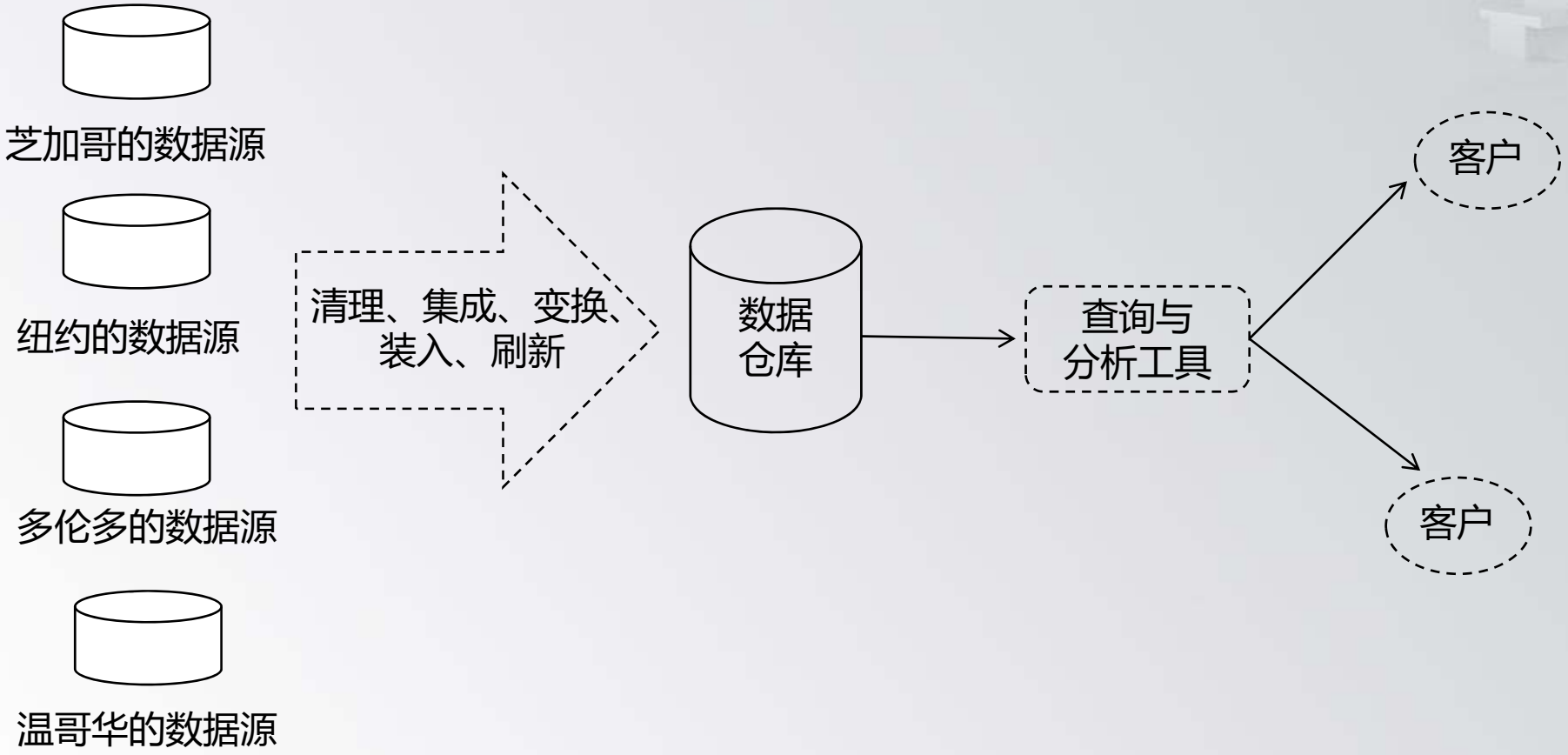
customer	(cust_ID, name, address, age, occupation, annual_income, credit_information, category, ...)
item	(item_ID, brand, categoty, type, price, place_made, supplier, cost, ...)
employee	(empl_ID, name, catagory, group, salary, commisson, ...)
branch	(branch_ID, name, address, ...)
purchases	(trans_ID, cust_ID, empl_ID, data, time, method_paid, amount)
items_sold	(trans_ID, item_ID, qty)
works_at	(empl_ID, branch_IDsss)

AllElectronics关系数据库的关系模式

什么样的数据能被挖掘-数据仓库

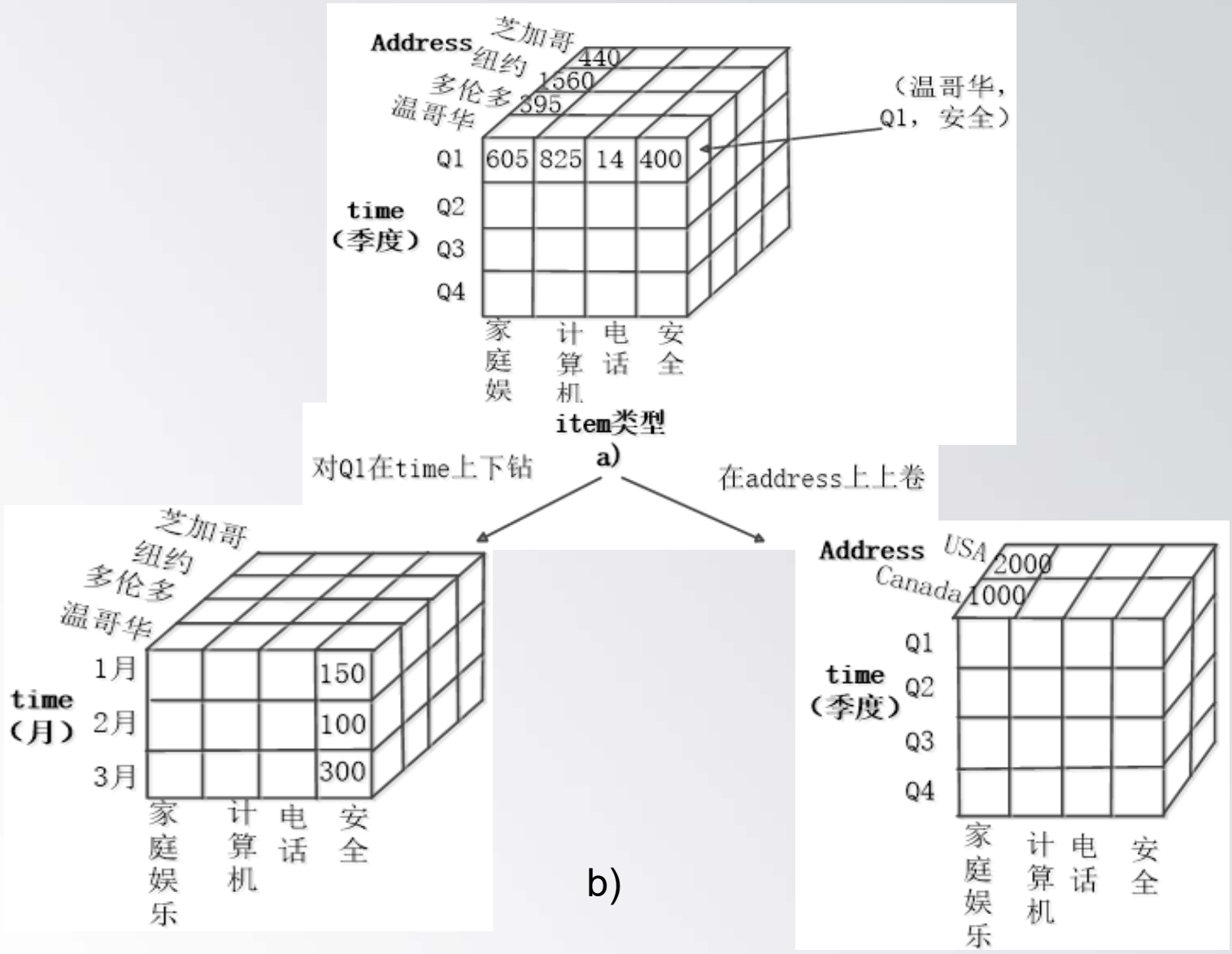
- 数据仓库是多种数据来源的信息仓库，以统一的模式存放，通常是在一个站点。数据仓库通过一系列的数据清洗、聚合、转换、加载和周期性的更新构建。
- 数据仓库以重要的主题组织，从历史的视角提供信息，常常是概要型的。数据仓库模型是高维数据结构，每一维对应于相应的一个或者一组属性。称为数据立方。
- 通过提供高维数据视角和概要数据，数据仓库为OLAP联机处理提供支持。高维数据挖掘以OLAP的方式在高维空间挖掘。

什么样的数据能被挖掘-数据仓库



ALLElectronics数据仓库的典型框架

什么样的数据能被挖掘-数据仓库



一个ALLElectronics数据仓库的多维数据立方体：a)显示ALLElectronics的汇总数据；b)显示图a)中数据立方体上的下钻和上卷的结果。

什么样的数据能被挖掘-事务数据

- 事务数据库存放交易记录，例如顾客的一次购买，机票的预订，或者用户点击了一个web页面。交易数据被存放在表中，每条记录表示一次交易记录。
- 假如我们想知道哪些商品放在一起出售更好，如果我们知道打印机通常会和电脑一起被购买，则可以对买电脑的顾客提供打印机购买折扣，或者完全免费，以期销售更多电脑。
- 传统的数据库系统不能做这种商业分析。但是基于交易数据的数据挖掘能够发现这种频繁模式，即发现那些商品会被一起经常购买。

什么样的数据能被挖掘-事务数据

Trans_ID	商品ID的列表
T100	I1，I3，I8，I16
T200	I2, I8
….	….

AllElectronics 销售事务数据库片段

什么样的数据能被挖掘-其它类型数据

- 其他数据如和时间相关的数据，序列数据，流数据，空间数据，工程设计数据，超链接和多媒体数据，图数据和网络数据，web数据等。
  - 通过挖掘股票交易数据中未被发现的趋势帮助你计划投资策略；
  - 通过挖掘计算机网络数据流进行入侵检测；
  - 对于空间数据，基于主要高铁线路和城市的距离描述城市贫困率的变化；
  - 通过挖掘文本数据来识别某领域的热点演化。
  - 通过挖掘用户对于产品的评论，获得客户情绪和了解产品在市场上的接受度。
  - 通过挖掘多媒体数据，来对图像进行目标识别和对其进行语义标签和分类。
  - 挖掘WWW上的信息，可以发现网页的变化以及不同网页之间的关联关系，或者用户，社区以及活动之间的关联关系。

什么样的数据能被挖掘

- 需要注意的是，在许多应用领域，数据是多种类型共存的
  - 比如，web挖掘中，包括文本数据和多媒体数据在网页上。对于多种类型数据融合的应用的数据挖掘，数据清洗、聚合是很困难的事情，因为多种数据源的复杂交互。

目录

- 为什么进行数据挖掘
- 什么是数据挖掘
- 什么样的数据能被挖掘
- 可以挖掘什么类型的模式
- 数据挖掘使用哪些技术
- 数据挖掘应用
- 数据挖掘的主要挑战

可以挖掘什么类型的模式?

- 数据挖掘任务可以被归类为两种类别：描述性的和预测性的。
  - 描述性的挖掘任务是描述目标数据集的数据属性。
  - 预测性的挖掘任务是归纳现有数据以用来做预测。
- 具体的模式分析
  - 类/概念描述：特征化和区分
  - 挖掘频繁模式、关联规则和相关性
  - 用于预测分析的分类和回归
  - 聚类分析
  - 离群点分析

类/概念描述：特征化和区分

- 对这些单个的类别和概念进行描述非常有用。这种描述称为类别/概念描述。
- 描述可以通过：
  （1）通过总结目标类别特征的数据特征化；
  （2）把目标类和一个或多个可对比类做比较的数据区分；
  （3）同时使用上面2种方法。

类/概念描述：特征化和区分

- 数据特征化是总结目标类别数据的一般特征。
  - 数据一般通过查询来收集。例如，想研究上一年花掉5000美元以上的客户，可以通过SQL查询语句来进行
  - 有多种数据描述的方法。可以使用基于统计测量和散点图的简单数据总结。基于数据立方的OLAP操作可以使用在特定维度空间的数据摘要。面向属性的归纳技术也可以用来描述数据。
  - 描述的结果可以通过多种图表展现，包括饼图、柱状图、曲线、高维数据立方体和多维表、交叉表等。也可以使用规则形式的广义关系来表示。

类/概念描述：特征化和区分

举例：总结去年购买商品花掉5000美元以上的客户特征。

描述结果可能是这些客户的一般信息，如他们是40-50岁之间，有工作，有很高信用度。

类/概念描述：特征化和区分

- 数据区分是比较目标类别数据对象和一个或者多个对比类对象的一般特征。
  - 举例如，客户关系经理想比较那些经常购买计算机产品和很少购买这类产品的客户特征。描述结果给出这些客户的一般对比信息，比如经常购买电脑产品的80%的客户是20到40岁之间的有大学文凭的，很少买这类产品的人中60%是老年人或者青少年，没有大学学历。

挖掘频繁模式、关联规则和相关性

- 频繁模式，含义是数据中经常发生的模式。包括频繁项集，频繁序列，频繁子结构。
  - 频繁项集指的是在交易数据集中经常同时发生的商品。
  - 频繁序列，比如顾客先买了笔记本电脑，再买了数码相机，接着买了内存卡，这是一个序列模式。
  - 频繁子结构指的是结合项集或者子序列的不同的结构形式（图、树、或者格）。
  - 挖掘频繁模式，会发现有趣的数据之间的关联和相关度

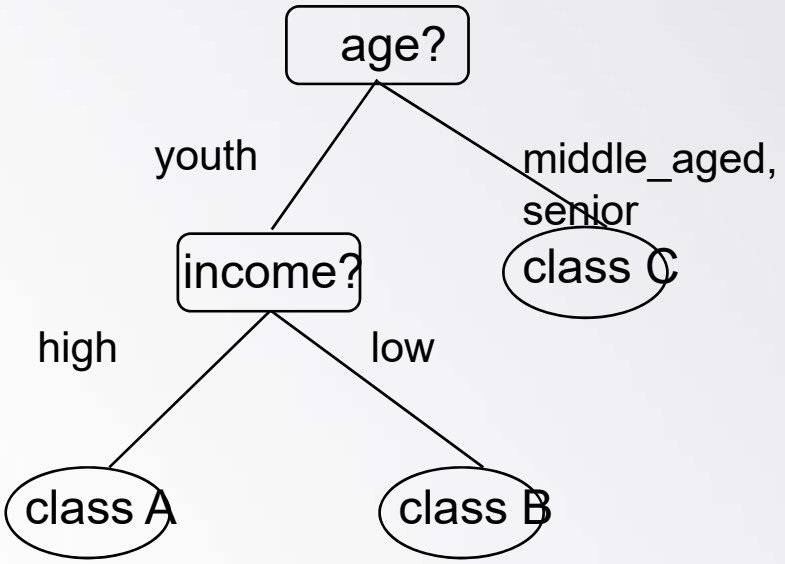
挖掘频繁模式、关联规则和相关性

- 分类是找到模型可以描述和区分数据类别或者概念的方法。模型从一系列的训练数据中分析获得，用于预测未知类别的数据标签。
  - 主要技术如：分类规则、决策树、神经网络、朴素贝叶斯、支持向量机、k-means等

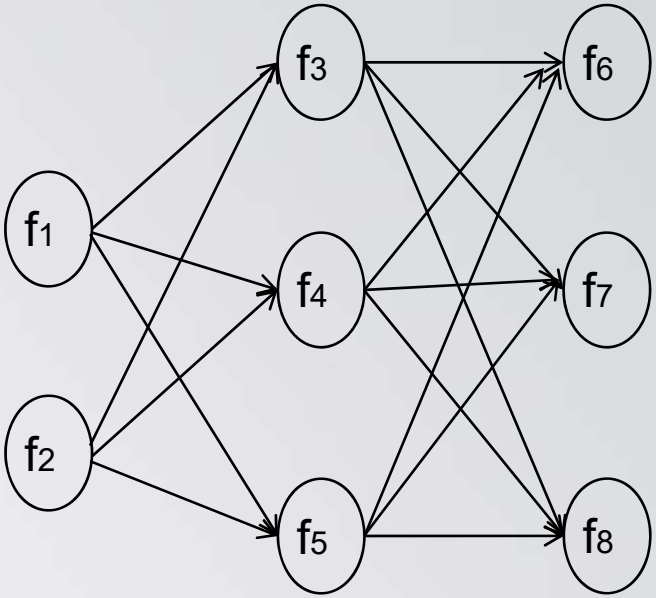
用于预测分析的分类和回归

age(X, “youth”) AND income(X, “high”) → class(X, “A”)
age(X, “youth”) AND income(X, “low”) → class(X, “B”)
age(X, “middle_aged”) → class(X, “C”)
age(X, “senior”) → class(X, “D”)

a)



b)



c)

分类模式可以用不同的形式表示：a) IF-THEN规则；b)决策树；c)神经网络

用于预测分析的分类和回归

- 回归是建立连续值函数模型，预测缺失或难以获得的数值型数据。
- 相关性分析是在分类和回归之前的步骤，我们需要选择那些属性跟分类和回归的过程显著相关。不相关的属性不被包含在考虑之列。

聚类分析

- 聚类分析针对没有标签的数据进行。基于最大化类别内部的相似度，最小化类别之间的相似度的原则来分组。
- 举例如，从电商数据中识别同类型的顾客人群。

离群点分析

- 数据集可能包含不遵守一般行为和模型的数据。这些目标称为离群点。
  - 检测离群点可以使用统计检验方法、距离测量、或者基于密度的方法。
  - 举例，通过与常规的消费相比较发现大笔金额的异常消费，可以发现信用卡的盗刷问题。离群值可能跟消费的地点、支付类型或者频率有关。

所有的模式都很有趣吗?

- 一般来说，答案是否定的。只有一小部分模式在实际上对特定的用户是有用的。
- 一个模式是有趣的有如下几个条件:
  - **1）能很容易被人理解**
  - **2）对于新的或者测试数据以一定的确信度也是合理的**
  - **3）潜在有用的**
  - **4）新奇的**

  一个有趣的模式能表达**知识**。

如何衡量模式是否有趣?

- 一些有关模式是否有趣的**客观测量方法**如:
  - 关联规则挖掘的客观衡量是规则的支持度，表示给定的规则在交易数据库中所占的百分比。另一个是置信度，表示关联规则的确定程度。
  - 另一种客观的有趣度的衡量包括准确度和覆盖率。准确率告诉我们被一个规则正确分类的数据所占的百分比。覆盖率告诉我们规则可以作用的数据所占的百分比。
  - 一般来说，每一个有趣程度的测量方法都有一个用户能控制的阀值。

如何衡量模式是否有趣?

- **主观的有趣度的衡量**基于用户对数据的看法
  - 如果模式是没有预料到的或者提供了可以指导用户行为的策略，则认为这些模式是有趣的。比如，“大量地震之后会常常有一系列小震”是很可行性的且基于这个信息能挽救生命。
  - “意料之内”的模式也可能是有趣的，如果它们验证了人们的假设，或与用户的预感类似。

数据挖掘能产生所有有趣的模式吗?

- 这是数据挖掘的完整性问题。
- 数据挖掘系统产生所有可能的模式是不现实和不高效的。
- 对一些数据挖掘任务来说，比如关联规则挖掘，能充分保证算法的完整性， 使用约束和有趣度测量能保证数据挖掘完整性。

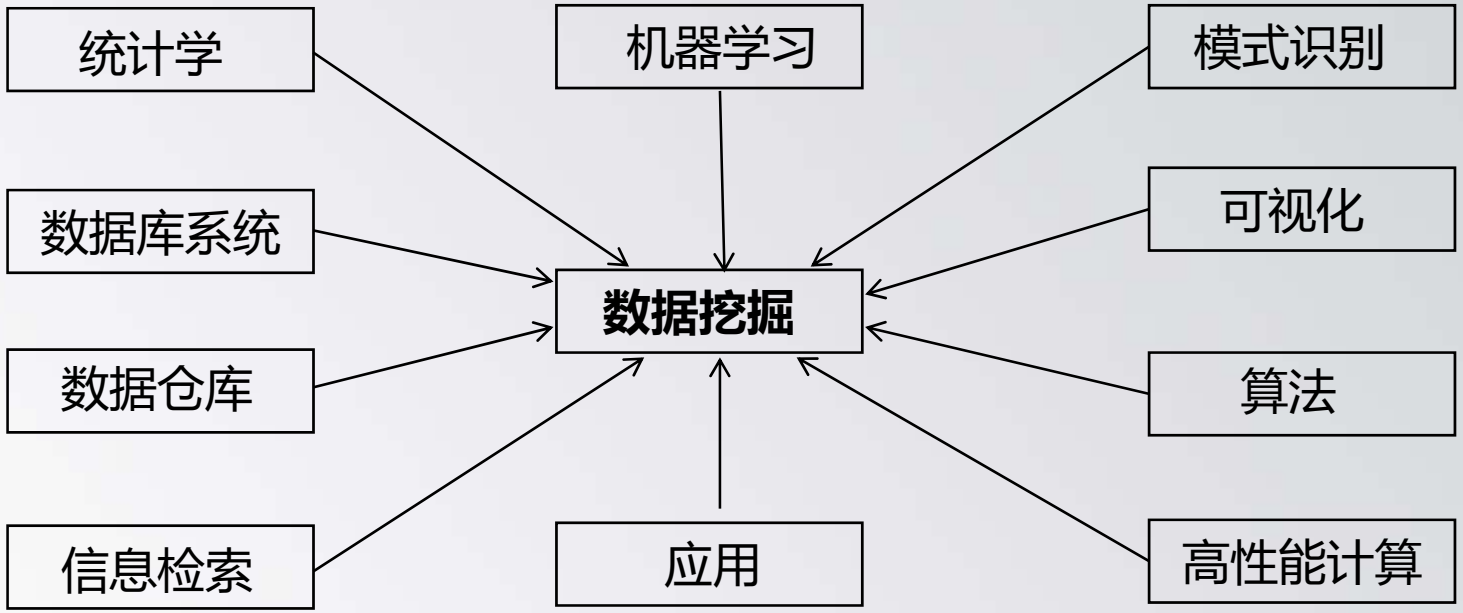
一个数据挖掘系统能只产生有趣的模式吗?

- 这是数据挖掘的优化问题。
- 只产生有趣的模式是会高度令人满意的。因为对于用户和挖掘系统来说，不需要从生成的模式中鉴别是否有趣，因此是很高效的。
- 但是，虽然这方面研究有进展，但优化问题仍然是一个挑战性的问题。

目录

- 为什么进行数据挖掘
- 什么是数据挖掘
- 什么样的数据能被挖掘
- 可以挖掘什么类型的模式
- 数据挖掘使用哪些技术
- 数据挖掘应用
- 数据挖掘的主要挑战

数据挖掘使用哪些技术?



数据挖掘从其他许多领域吸纳技术

目录

- 为什么进行数据挖掘
- 什么是数据挖掘
- 什么样的数据能被挖掘
- 可以挖掘什么类型的模式
- 数据挖掘使用哪些技术
- 数据挖掘应用
- 数据挖掘的主要挑战

数据挖掘应用

- 商务智能 （Business Inteligence）
  - 对商业机构来说，更好的了解组织的交易环境是非常重要的。比如他们的顾客、市场、供应、资源以及竞争者。
  - 商务智能提供商务运作的历史、现状和预测视图。
  - 如果没有数据挖掘，企业
    - 无法做出有效的市场分析
    - 无法比较客户对于相似产品的看法，
    - 无法发现竞争者的优点和弱点，
    - 无法留住有价值的顾客，
    - 无法做出敏捷的商业决策。

数据挖掘应用

- 商务智能 （Business Inteligence）
  - 数据挖掘是商业智能的核心。
  - 分类和预测技术是商业智能的预测分析的核心，因为有很多供需和销售的应用。
  - 聚类在客户关系管理上发挥中心作用。顾客依据相似性被聚类。使用描述化的数据挖掘技术，我们可以更好的理解不同顾客群的特征，发展不同的客户定制程序

数据挖掘应用

- Web搜索引擎
  - Web搜索引擎是在web上搜索信息的特殊的计算机服务器。搜索结果通常是一个列表，列表可能包含网页、图像或者其他类型的文件。
  - Web搜索引擎是很大的数据挖掘应用。大量的数据挖掘技术被应用到搜索引擎的多个方面，从**爬取**（决定哪些页面被爬取和爬取频率）、**索引**（选取建立索引的页面并决定索引被建立时扩充的范围）到**搜索**（页面如何被排序，哪些广告被加载，搜索结果如何被个性化和上下文感知）。

目录

- 为什么进行数据挖掘
- 什么是数据挖掘
- 什么样的数据能被挖掘
- 可以挖掘什么类型的模式
- 数据挖掘使用哪些技术
- 数据挖掘应用
- 数据挖掘的主要挑战

数据挖掘的主要挑战

- 挖掘方法
- 用户交互
- 效率和可扩展性
- 数据库类型的多样化

数据挖掘的主要挑战

挖掘方法

- 1、挖掘各种新的知识类型

  数据挖掘覆盖了数据分析和知识发现任务的广泛范围。这些任务基于同一种数据库使用不同的挖掘方法。因为应用类型非常多样化，新的挖掘任务不断出现，使数据挖掘成为一个动态和快速增长的领域。

- 2、挖掘多维空间中的知识

  在很多种情况下，数据能被看成是一个高维数据方块。挖掘数据方块能从本质上提升数据挖掘的功能和灵活性

数据挖掘的主要挑战

- **挖掘方法**
  - **3、多学科交叉的数据挖掘**

    数据挖掘能通过融合多种学科知识来得到本质提升。 例如，自然语言文本挖掘就是融合了数据挖掘技术到信息检索和自然语言处理技术。另外，在大型程序中挖掘软件错误，是结合了软件工程知识到数据挖掘过程中。
  - **4、提升网络环境下的挖掘能力**

    很多数据对象是互相链接和内在关联的。比如web， 数据库关系，文件或者文档。多种数据对象的语义关联可以被用来提升数据挖掘技术。在一种数据对象挖掘的知识能被用来提升到关联或者语义关联的数据对象的知识发现上

数据挖掘的主要挑战

- **挖掘方法**
  - **5、处理数据的不确定性、噪声和不完整性**

    数据清洗、预处理、离群点发现和删除、不确定性的质疑都是需要被融合到数据挖掘过程中的技术。
  - **6、模式评估和模式导向（或限制导向）的挖掘**

    需要使用使用一些主观测量技术去评估模式是否有趣。基于给定的用户分类和基本信仰和期望，来对模式给出一个评分，以此对挖掘过程给出导向，产生更有趣的模式和减少搜索空间。

数据挖掘的主要挑战

- **用户交互：**如何和挖掘系统交互，如何在挖掘中结合用户的背景知识，如何可视化和理解挖掘结果。
  - **1、交互挖掘**

    数据挖掘过程应该是高度交互性的。需要建立灵活的用户界面和探索性的挖掘环境，以便于用户的交互。

    用户可以抽样一些数据，然后描述数据的一般特征，评估可能的挖掘效果。交互式挖掘需要能够让用户能动态的改变搜索焦点，基于结果精化挖掘请求，挖掘，切块，旋转，在挖掘时动态的对数据立方进行探索

数据挖掘的主要挑战

- **用户交互**
  - **2、结合背景知识**

    背景知识、限制、规则以及其他的领域相关的信息需要被 融合到知识发现过程中。这些知识能被用于模式评估和为挖掘有趣模式作为向导
  - **3、特殊的数据挖掘和数据挖掘查询语言**

    高层次的数据挖掘查询语言或者其他的高层次的灵活的用户界面能给用户定义特殊无组织的数据挖掘任务的自由。对于挖掘请求的过程的优化是一个很有前景的研究方向。
  - **4、数据挖掘结果的展示和可视化**

    数据挖掘结果需要能生动灵活的展示，以便于发现的知识被更好的理解和直接应用。这需要系统能够采用更丰富的知识表达、更友好的用户界面和可视化技术

数据挖掘的主要挑战

效率和可扩展性

- 1、数据挖掘算法的效率和可扩展性

  数据挖掘算法的运行时间需要是可预测的、短的、可以被应用接受的。

- 2、并行的、分布式的和可增长的挖掘算法

  许多数据集的规模很大，分布式分布，很多数据挖掘算法的高复杂度催生了并行和分布式的数据集中式挖掘算法。

  云计算和集群计算，促进了并行数据挖掘的问题。

  数据挖掘过程的高代价和不断增长的输入促使了增量式数据挖掘，即能够合并新数据的更新而不需要从头开始从整个数据集挖掘

数据挖掘的主要挑战

- **数据库类型的多样化**
  - **对于复杂数据类型的处理**

    期望在多种数据类型和多种数据挖掘目标的情况下，使用一种数据挖掘系统能挖掘所有类型的数据是不现实的。可以**建立基于领域的或基于应用的精细数据挖掘系统，对特定数据类型做深度挖掘**。建立高效的和有效的针对各种应用的挖掘工具是一个有挑战性和活跃的研究领域。

数据挖掘的主要挑战

- **数据库类型的多样化**
  - **挖掘动态、网络化的和全局的数据仓库**

    网络把不同来源的数据连接在一起，形成了巨大的、分布式的、异质的全局信息系统。对多种数据来源的结构化、半结构化和非结构化并且内在连接的数据是对数据挖掘的巨大挑战。

    对这些数据的为挖掘将有助于发现比在小规模的孤立数据仓库中更多的异质网络中的模式和知识。Web挖掘、多数据源挖掘、信息网络挖掘将成为有挑战性和快速增长的数据挖掘领域。

数据挖掘和社会

- 我们如何利用数据挖掘造福社会？如何保护不被错误使用？
- **隐私保护的数据挖掘**

  隐私保护的数据发布和数据挖掘是正在进行的研究领域。原则是在成功的进行数据挖掘的同时察觉数据敏感性和保护个人隐私。对用户数据的不合适暴露或者潜在的侵犯用户隐私以及数据隐私权是需要被考虑的问题。

数据挖掘和社会

隐形数据挖掘

随着技术的发展，越来越多的系统开始收集用户的数据用于提高系统性能或更好的服务，但这些数据收集用户是不知道的。

智能搜索引擎和基于网络的商家都在使用这种隐形挖掘技术。比如，人们在线购物时，并不知道商家很可能在收集顾客的购买模式，这些将被用来在以后向其推荐其他商品。

数据挖掘与法律问题

课后思考题

- 请思考数据挖掘可能会遇到哪些法律问题，可能会和哪些法律有关，请举出具体例子并讨论。

Thank You!

Q&A