第一章 引论

笔记本： 数据挖掘：概念与技术
创建时间： 2017/12/20 8:53 **更新时间：** 2017/12/26 11:30
作者： Passero

1.1 什么是数据挖掘

许多人把数据挖掘被视为数据中的知识发现（KDD）的同义词，另一些人把数据挖掘视为知识发现过程的一个基本步骤。
知识发现过程由以下步骤的迭代序列组成：

1. 数据清理（消除噪声和删除不一致数据）
2. 数据集成（多种数据源可以组合在一起）
3. 数据选择（从数据库中提取与分析任务相关的数据）
4. 数据变换（通过汇总或聚集操作，把数据变换和统一成适合挖掘的形式 ）
5. 数据挖掘（基本步骤，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（根据可视化和知识表示技术，向用户提供挖掘的知识）

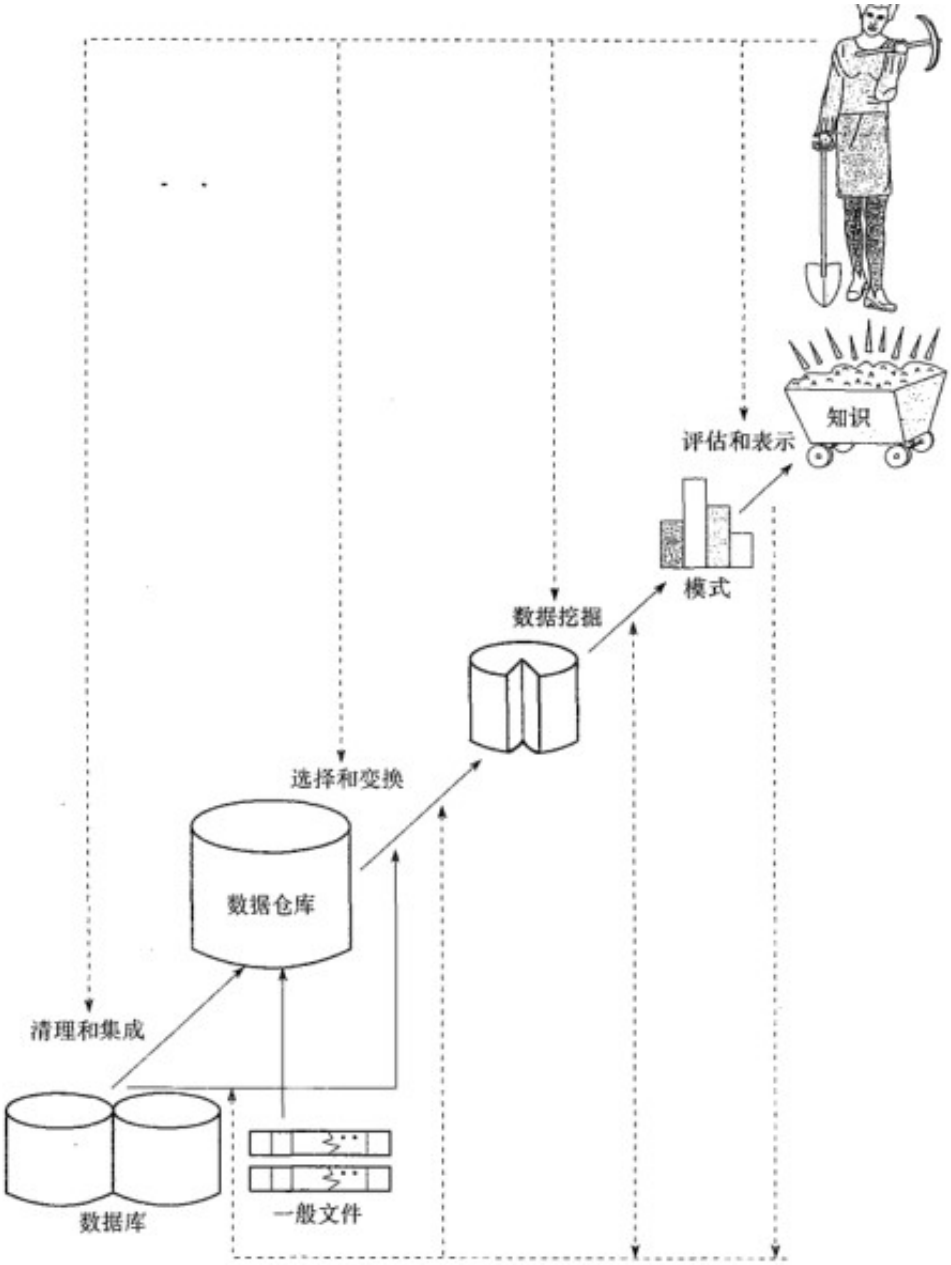

图 1.4　数据挖掘视为知识发现过程的一个步骤

数据挖掘是从大量数据中挖掘有趣模式和知识的过程。数据源包括数据库，数据仓库，web，其他信息存储库或动态流入系统的数据。

1.2可以挖掘什么类型的数据

- 数据库数据
- 数据仓库（数据仓库是一个从多个数据源收集的信息存储库，存放在一致的模式下，并且通常驻留在单个站点上。数据仓库通过数据清理，数据变换，数据集成，数据装入和定期数据刷新来构造。）通常数据仓库用称作数据立方体的多维数据结构建模，其中每个维对应于模式中的一个或一组属性，而每个单元存放某种聚集度量值，数据立方体提供数据的

多维视图，并允许预计算和快速访问汇总数据。

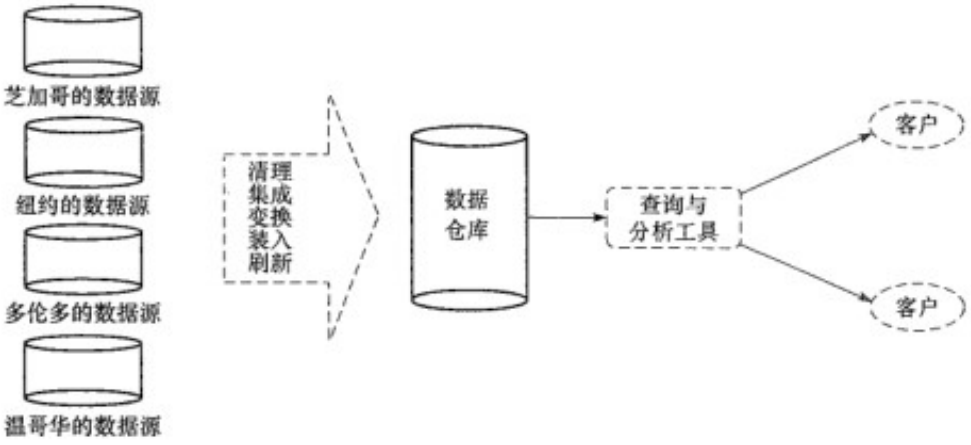

图 1.6　AllElectronics 数据仓库的典型框架

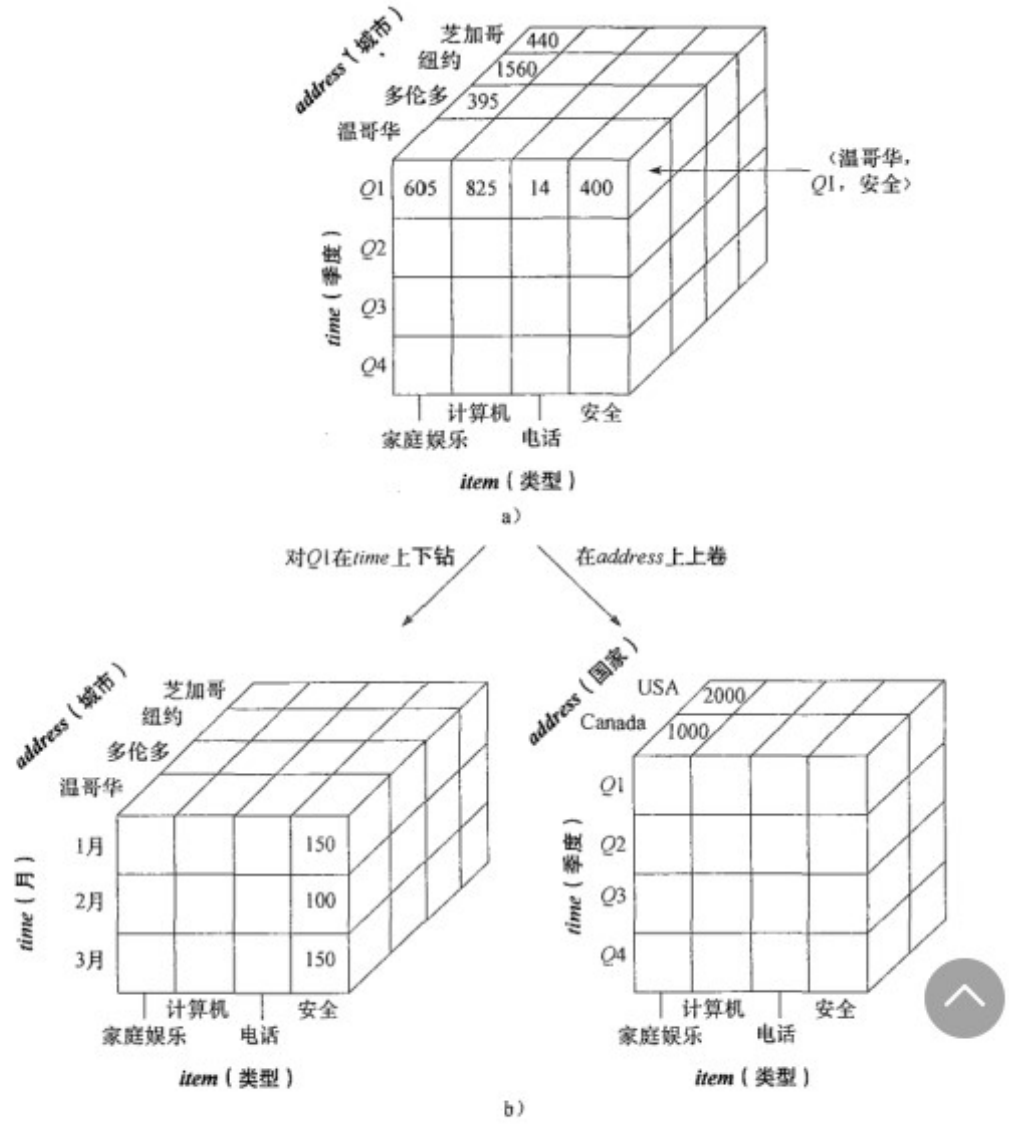

图 1.7　一个通常用于数据仓库的多维数据立方体：a）显示 AllElectronics 的汇总数据；b）显示图 a）中数据立方体上的下钻和上卷的结果。为便于观察，只给出部分立方体单元值

- 事务数据（一般来说，事务数据库的每个记录代表一个事务，如孤苦的一次购物，一个航

班机票，或一个用户的网页点击。通常，一个事务包含一个唯一的事务标识号，以及一个组成事务的项[如交易中购买的商品]的列表。事务数据库可能有一些与之相关联的附加表，包含事务的其他信息，如商品描述，关于销售人员或部门等的信息。）

trans_ID	商品*ID*的列表
T100	I1, I3, I8, I16
T200	I2, I8
…	…

图 1.8　AllElectronics 销售事务数据库的片段

- 其他类型的数据（比如时间相关或序列数据[如历史记录，股票交易数据，时间序列和生物学序列数据等]，数据流，超文本和多媒体数据等。可以从这些类型的数据中挖掘各种知识，例如，就时间数据而言，可以挖掘银行数据的变化趋势，这可以帮助银行根据顾客流量安排出纳员。）

1.3可以挖掘什么类型的模式

数据挖掘功能用于指定数据挖掘任务发现的模式。一般而言，这些任务可以分为两类：描述性和预测性。描述性任务刻画目标数据中数据的一般性质，预测性挖掘任务在当前数据上进行归纳，以便做出预测。

1. 类/概念描述：特征化与区分（数据特征化-一般地汇总所研究类[通常称为目标类]的数据 数据区分-将目标类与一个的或多个可比较类[通常称为对比类]进行比较 数据特征化和区分）
   数据特征化是目标类数据的一般特性或特征的汇总。数据区分是将目标类数据对象的一般特性与一个或多个对比类对象的一般特性进行比较。
2. 挖掘频繁模式，关联和相关性
   频繁模式是在数据中频繁出现的模式。支持度，置信度，单维关联规则，多维关联规则。
3. 用于预测分析的分类与回归
   分类：找出描述和区分数据类或概念的模型（或函数），以便能够使用模型预测类标号未知的对象的类标号；回归建立连续值函数模型，也就是说，回归用来预测缺失的或难以获得的数值数据值，而不是（离散的）类标号。相关分析可能需要在分类和回归之前进行，它试图识别与分类和回归过程显著相关的属性。

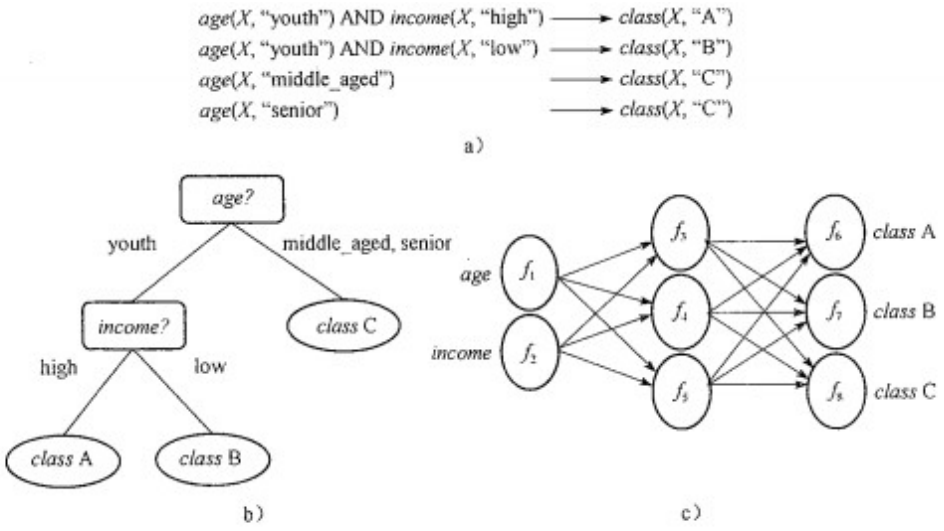

图 1.9　分类模型可以用不同形式表示：a）IF-THEN 规则；b）决策树；c）神经网络

4. 聚类分析

聚类分析数据对象，而不考虑类标号。对象根据最大化类内相似性，最小化类间相似性的原则进行聚类或分组，也就是说，对象的簇这样形成，使得相比之下在同一个簇中的对象具有很高的相似性，而与其他簇中的对象很不相似。

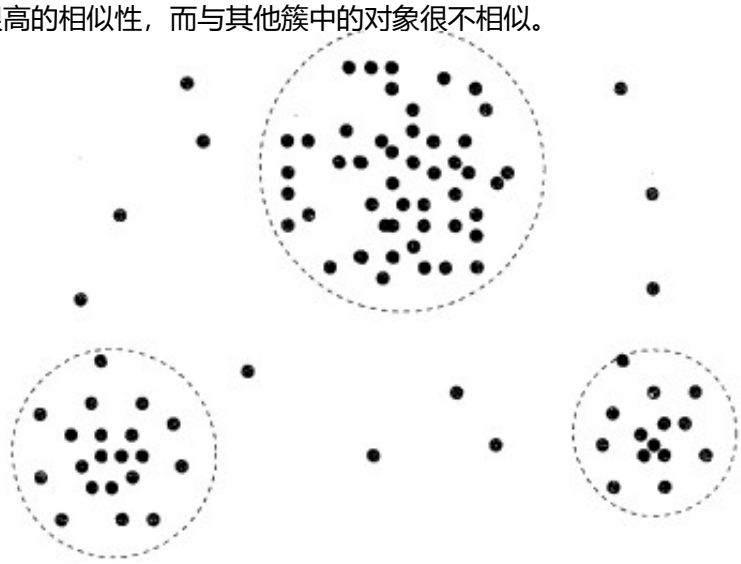

图 1.10　关于一个城市内顾客位置的二维图，显示了 3 个数据簇

5. 离群点分析

数据集中可能包含一些数据对象，它们与数据的一般行为或模型不一致。这些数据对象是离群点。大部分数据挖掘方法都将离群点视为噪声或异常而丢弃，然而，在一些应用中（例如，欺诈检测），罕见的时间可能比正常出现的事件更令人感兴趣。离群点数据分析称作离群点分析或异常挖掘。e.g.通过检测一个给定账号与正常的付费相比付款数额特别大，离群点分析可以发现信用卡欺骗性使用。离群点还可以通过购物地点和类型或购物频率来检测。

6. 所有模式都是有趣的吗

客观度量：支持度，置信度

$$support(X \Rightarrow Y) = P(X \cup Y)$$

$$confidence(X \Rightarrow Y) = P(Y \mid X)$$

1.4使用什么技术

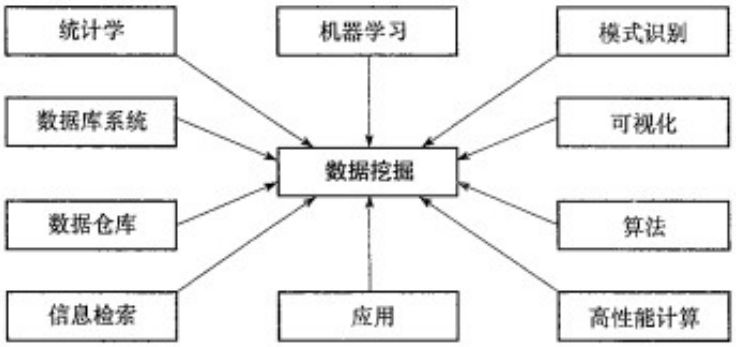

图 1.11　数据挖掘从其他许多领域吸纳技术