本文主要内容来自 SpriCoder的博客，更换了更清晰的图片并对原文的疏漏做了补充和修正。

本文提供了 pdf 版，以供打印： 商务智能-06-数据挖掘 | EagleBear2002 的博客。

1. 数据挖掘

在数据库及数据仓库中存贮有大量的数据，它们具有规范的结构形式与可靠的来源，且数量大、保存期间长，是一种极为宝贵的数据资源。充分开发、利用这些数据资源是目前计算机界的一项重要工作

1.1 数据资源的利用有三种方式

1. 数据资源的查询服务
2. 数据资源的演绎
    1. 知识的利用与搜索（AI）
    2. 演绎数据库
    3. 统计分析软件（SAS，SPSS）
    4. OLAP
3. 数据资源的归纳
    1. 数据挖掘：数据资源的归纳

1.2 数据的三种利用方式之间的区别

1. 可以从文具盒（数据库）中找到橡皮和铅笔，不可能得到橡皮要和铅笔配合使用等这样的信息，使用数据挖掘技术可以发现一些用户未知的信息。
2. 可以从一张家族谱中找到“甲”是“乙”的后代（知识库），但无法据此推断出“丙”的祖先是谁，使用数据挖掘则技术可以寻找到哪些具有普遍意义的信息（知识），并可以将其应用到其它同类应用中，以帮助用户进行决策。

2. 什么是数据挖掘

数据挖掘（DM：Data Mining）又称为数据库中的知识发现（KDD：Knowledge Discovery in Database）

1. 起源于 80 年代初
2. 机器学习和数据分析的理论及实践是数据挖掘研究的基础，极大的商业应用前景又是数据挖掘研究工作的巨大推动力

传统的数据库查询和统计只能提供想要的信息，而数据挖掘技术则可以发现没有意识到的未知信息

2.1 数据挖掘定义

什么是数据挖掘?

1. 定义一：数据挖掘就是对数据库（数据仓库）中蕴涵的、未知的、非平凡的、有潜在应用价值的模式（规则）的提取
2. 定义二：数据挖掘就是从大型数据库（数据仓库）的数据中提取人们感兴趣的知识。这些知识是隐含的、事先未知的潜在有用信息

因此，数据挖掘必须包括三个因素：

1. 数据挖掘的本源：大量、完整的数据
2. 数据挖掘的结果：知识、规则
3. 结果的隐含性：因而需要一个挖掘过程

2.2 数据挖掘描述

1. 应该是在一个大量的、完整数据集中进行数据的挖掘工作，例如：从一个没有同名的人群中可以抽取出关键字“姓名”没有同名现象，但我们并不能据此推断出“所有人都不会取相同的名字”。
2. 归纳结果应该是具有普遍性意义的规则，从一万条数据中找出的规律也应该能够适用于十万、一百万……的情况。
3. 数据挖掘的目的：用归纳出的规律来指导客观世界。

2.3 数据挖掘的几个基本概念（了解）

2.3.1 模型

用高级语言表示的表达一定逻辑含义的信息，这里通常指数据库中数据与数据之间的逻辑关系

例如：在某超市的商品销售数据库中，我们可以找到以下信息：

1. 男性顾客在购买婴儿尿布时也往往同时购买啤酒
2. 在购买面包和黄油的顾客中，大部分的人同时也买了牛奶

2.3.2 知识

满足用户对客观评价标准（例如：兴趣度/置信度）和主观评价标准要求的模式

2.3.3 置信度

在某一数据集上，模式成立的程度。

例如：模式 R1：在购买面包和黄油的顾客中，大部分的人同时也买了牛奶。该模式的置信度为：同时购买“面包、黄油、牛奶”的顾客人数占同时购买“面包、黄油”的顾客人数的百分比，即：

$$\frac{\text{同时购买面包、黄油和牛奶的顾客人数}}{\text{同时购买面包和黄油的顾客人数}}$$

通过数据挖掘所发现的模式的置信度大小涉及到许多因素：如数据的完整性、样本数据的大小、领域知识的支持程度等。

如果没有足够的置信度，模式便不能成为知识。因此，在数据挖掘过程中，通常要规定模式的最小置信度。

2.3.4 兴趣度

在某一数据集上，模式被用户关注的程度（也被称为支持度）；

例如：模式 R1 的支持度为“同时购买‘面包，黄油和牛奶’的顾客人数占总顾客人数的百分比”，即：

$$\frac{\text{同时购买面包，黄油和牛奶的顾客人数}}{\text{总的顾客人数}}$$

只有当一个模式的“兴趣度”达到一定的程度时，那么该模式才是一个有意义的模式，才能引起用户的注意，有助于用户的决策制订过程。因此，在数据挖掘过程中也要规定模式的“最小兴趣度”，以淘汰哪些在极少情况下才会出现的模式。

2.3.5 非平凡性

平凡知识：

1. 能够以确定的计算过程提取的模式称为平凡知识，例如：根据数据库中的薪水字段求得职员的平均薪水
2. 平凡的知识不是数据挖掘的目标

在数据挖掘中，知识的发现过程都应具有某种不确定性和一定的自由度，也就是要发现不平凡的知识

2.3.6 有效性

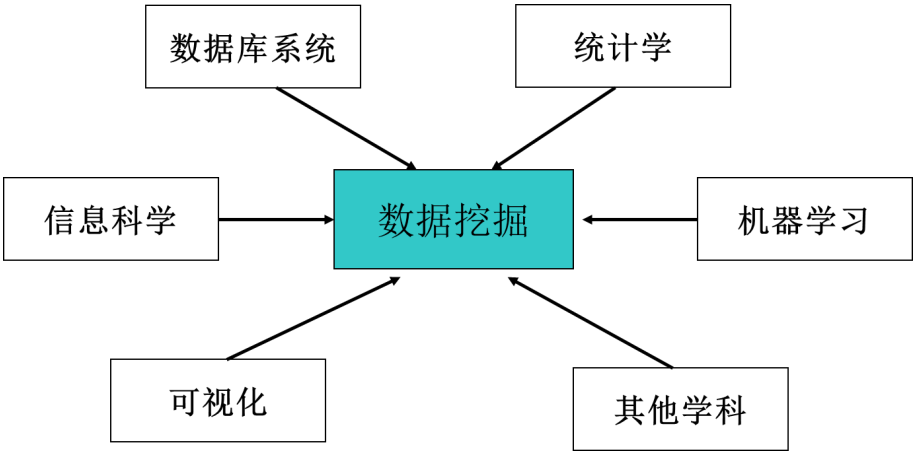
1. 知识的发现过程必须能够有效地在计算机上实现
2. 时间有效性
3. 空间有效性

2.4 数据挖掘的特点

区分什么是数据挖掘，什么不是数据挖掘：

1. 数据挖掘要处理大量的数据，处理的数据规模十分庞大，达到 GB、TB，甚至更大
2. 由于用户不能形成精确的查询要求，因此要依靠数据挖掘技术为用户找寻他可能感兴趣的东西
3. 在数据挖掘过程中，规则的发现基于统计规律：所发现的规则不必适用于所有数据，而是当达到一定的“门槛”时，即认为具有此规则。因此，利用数据挖掘技术可能会发现大量的规则
4. 数据挖掘所发现的规则是动态的，只反映了当前状态的数据集合具有的规则：随着不断地向数据库（数据仓库）中加入新数据，需要不断地重新进行数据挖掘以更新所发现的规则

2.5 数据挖掘的相关领域

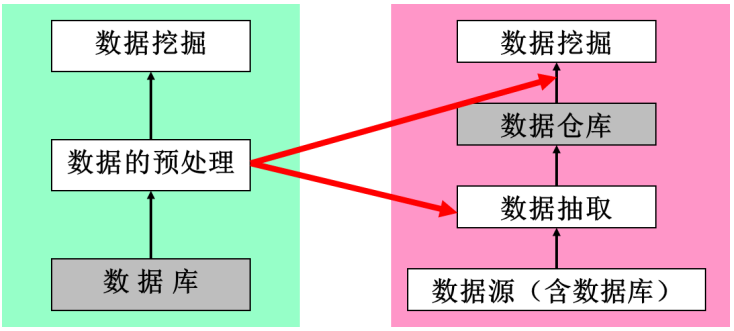


3. 数据仓库与数据挖掘

3.1 概述

在传统的决策支持系统中，数据挖掘技术是建立在数据库的基础上的，数据挖掘只是其中的一个部分，在这之前需要大量的数据查询和预处理。

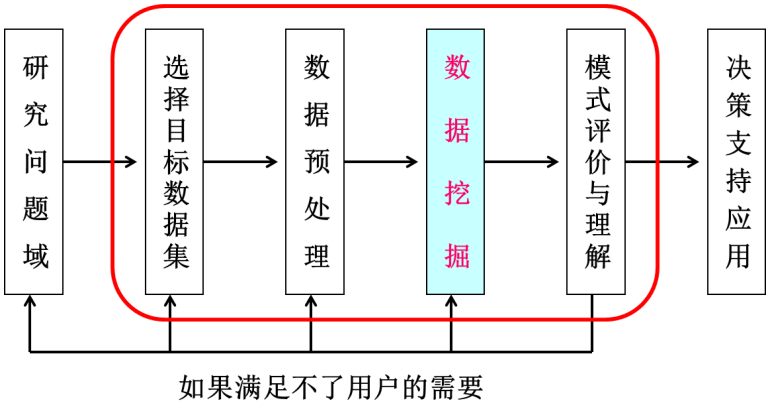
有了数据仓库技术之后，由于数据仓库中的数据都是经过抽取、整理和预处理后的综合数据，因而数据挖掘工作可以在数据仓库上直接运行。



3.2 利用数据库系统进行数据挖掘的缺点

1. 动态数据：大多数数据库的基本特点是内容将经常变化。在一个在线系统中，必须采用预警机制来保证数据库中的这些变化不会导致错误的数据挖掘结果。
2. 噪声和不确定性：
    1. 噪声数据：数据库中的错误数据和异常现象
    2. 不确定性：发现的模式可能只在一部分数据上有效
3. 冗余信息：
    1. 同一数据在操作型数据环境中的多处出现
    2. 这种冗余信息有时会误导知识的发现过程
    3. 有可能会“夸大”某个模式的置信度，从而导致发现大量的无意义的模式：$\frac{A+B}{A} < \frac{(A+B)+n}{A+n}$
    4. 也有可能“低估”某个模式的兴趣度，从而导致丢失一些有意义的模式：$\frac{A+B}{A} > \frac{A+B}{A+n}$
4. 不完整数据：
    1. 由于不完整的数据域和数据域上值的缺少造成的不完整数据当然会影响发现的结果
    2. 数据库的最初设计并没有考虑知识发现的应用，模式的发现、评价、解释很可能需要在当前数据库中并不存在的信息
5. 稀疏数据：数据库中的信息在实例空间中可能是稀疏的，这会严重影响发现的效率

3.3 数据挖掘的步骤



数据挖掘的步骤：

1. 数据集成
2. 数据规约
3. 挖掘
4. 评价
5. 表示

3.4 数据集成

数据挖掘的基础是数据，因此在挖掘前必须进行数据集成，这包括：

1. 首先，从各类数据系统中提取挖掘所需的统一数据模型，建立一致的数据视图
2. 其次，完成数据加载，从而形成挖掘的数据基础

鉴于前述原因，目前一般都用数据仓库以实现数据集成

在数据仓库数据的加载过程中，一般需要需要对数据作以下的预处理：

1. 数据清理
    1. 填补丢失的数据

    2. 清除噪声数据
    3. 修正数据的不一致性
2. 数据集成
3. 数据转换：收集到的数据并不一定适合数据挖掘的需要。如已有的挖掘方法可能无法处理这些数据，存在一些不规则的数据，或者数据本身不够充分等，因此需要对收集到的数据进行转换

3.5 数据规约

用于数据挖掘的数据量是非常巨大的，通过数据归约技术可以减低数据量，提高数据挖掘操作的性能：如果在归约后的数据集上进行数据挖掘可以获得与原来一样或几乎一样的挖掘结果，就可以考虑采用一定的数据归约技术来减少数据量，提高数据挖掘的效率

常见的数据归约技术有：

1. 数据立方体计算
2. 挖掘范围的选择
3. 数据压缩
4. 离散化处理

挖掘范围的选择：在不影响挖掘结果的前提下，尽可能地选取哪些与挖掘操作有关的属性集

数据压缩：

1. 减低数据的规模，节省存储空间开销和数据通讯开销
2. 如果采用的数据挖掘算法不需要解压就可以直接利用那些压缩数据进行数据挖掘，那么数据压缩技术将是非常有用的

离散化处理：

1. 如果一个属性的值域是一个连续区域，可以将它划分为若干个区域，然后用每个区域的标识值来代替原来的值。用以减低该属性上属性值的个数
2. 也可以利用这种数据归约技术来自动地建立该属性的概念层次树

3.6 挖掘

根据挖掘要求选择相应的方法与相应的挖掘参数（如最小置信度、最小兴趣度参数等），在挖掘结束后即可得到相应的规则。

3.7 评价

经过挖掘后所得结果可能有多种，此时可以对挖掘的结果按一定标准作出评价，并选取评价较高者作为最终结果。

3.8 表示

数据挖掘结果的规则可在计算机中用一定形式表示出来，它可以包括文字、图形、表格、图表等可视化形式，也可同时用内部结构形式存储于知识库中供日后进一步分析之用。

4. 常用的数据挖掘方法

目前一般常用的数据挖掘方法很多，它们大多属于数学统计方法或人工智能中的机器学习算法，以及人工神经网络/遗传算法等

在数据库中常用的几种数据挖掘方法包括：

1. 特征规则挖掘
2. 关联规则挖掘
3. 序列模式分析

4. 分类分析
5. 聚类分析

4.1 特征规则挖掘

特征规则：

1. 是一种常见的知识形式，它用于描述一类数据对象的普遍特征，是普化知识的一种
2. 特征规则的数据挖掘方法有两类：
    1. 面向属性归约方法
    2. 数据立方方法

4.1.1 面向属性规约方法

这是一种常用的特征规则的挖掘方法：通过对属性值间概念的层次结构进行归约，以获得相关数据的概括性知识，通常又称为普化知识。

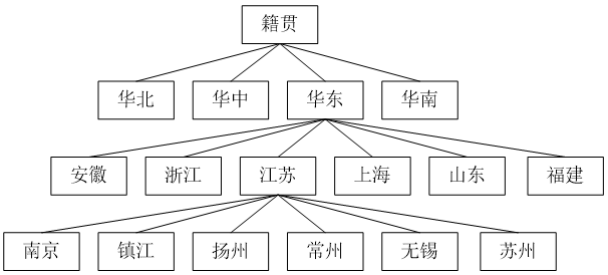
在实际情况中，许多属性都可以进行数据归类，形成概念汇聚点。

1. 这些概念依抽象程度的不同可构成描述它们层次结构的**概念层次树**
2. 根据概念层次树可以对供挖掘用的数据进行预处理，以生成一个适合于进行数据挖掘工作的数据集。因此“面向属性”的数据规约过程实际上可以作为数据挖掘工作而进行的数据预处理

4.1.1.1 概念层次树

指某属性值所具有的从具体的概念值到概念类的层次关系树

1. 一般由用户提供，或者从领域知识中得到相关属性的概念层次树
2. 也可以通过多属性体系结构自动构建
3. 例：属性“籍贯”的概念层次树



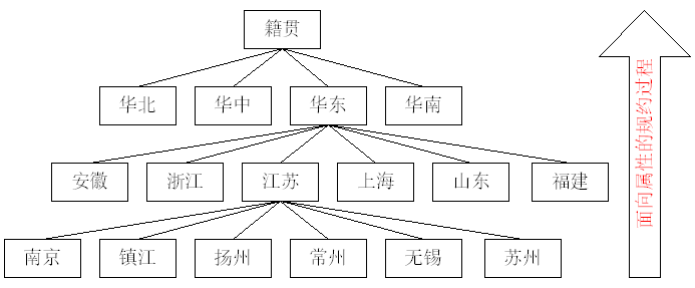
4.1.1.2 规约

用概念层次树上高层的属性值去替代低层的属性值，又称为概念提升。如：

1. 用“江苏”去代替“南京”
2. 用“华东”去代替“江苏”（或代替“南京”）

目的：

1. 规范化一个属性的取值
2. 提高模式的置信度和兴趣度（从而达到知识的阈值）



4.1.1.3 基本关系表

待挖掘的原始细节数据，以关系（二维表格）的形式出现，通常来自于准备好的数据库或数据仓库中

学　号	姓　名	系　别	书　名	借阅日期
9932007	颜立	经济	大趋势	2000. 3. 16
9833090	王家卫	金融	大趋势	2000. 3. 16
9813105	王向东	医学院	大趋势	2000. 5. 8
9928073	朱小明	企管	大趋势	2000. 5. 20
9822041	刘伟	历史	大趋势	2000. 6. 30
9932056	陈立业	经济	大趋势	2000. 9. 19
9923143	刘英	新闻	大趋势	2000. 12. 3

4.1.1.4 概括关系表

概括关系表通过基本关系表规约而来，其属性包括：

1. 目标数据集中参与数据挖掘的一个或多个属性：每一个属性都通过相关的概念层次树进行了规约
2. 系统为每个概括关系表新增加的一个 COUNT 属性

基本关系表中的元组被称为“基本元组”，而概括关系表中的元组则被称为“宏元组”：一个宏元组概括了多个基本元组，其中的 COUNT 属性被用来记录该宏元组所概括的基本元组数。

在概括关系表上进行数据挖掘的优点：

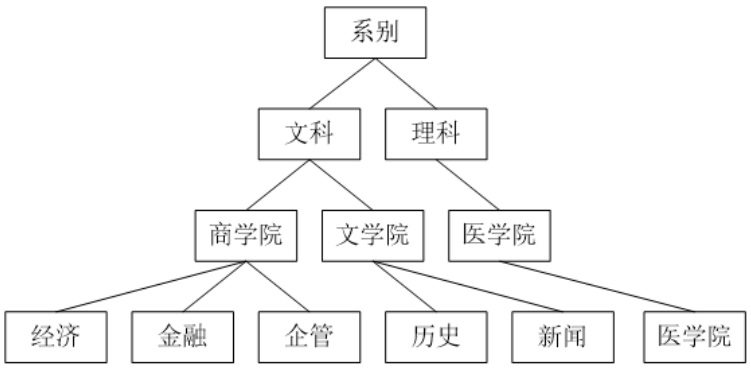
1. 可以通过面向属性的规约方法进行数据的预处理工作，以规范化属性的取值
    1. 填补缺少的属性值，剔除噪声数据
    2. 将不适宜数据挖掘工作的属性值进行转化
2. 根据概括关系表中的属性对基本关系表进行规约，可以限制每个属性可以取值的数量，从而将概括关系表中的宏元组的数量控制在一个合适的范围内，以提高数据挖掘的性能
3. 可以在不同的概念层次上进行数据挖掘，在不同概念层次上进行数据挖掘，完全可能获得不同的挖掘结果

4.1.2 面向属性规约方法例子

1. 数据挖掘的目的：寻找借阅《大趋势》一书的学生有什么特征？（就读专业的特色）

学　号	姓　名	系　别	书　名	借阅日期
9932007	颜立	经济	大趋势	2000. 3. 16
9833090	王家卫	金融	大趋势	2000. 3. 16
9813105	王向东	医学院	大趋势	2000. 5. 8
9928073	朱小明	企管	大趋势	2000. 5. 20
9822041	刘伟	历史	大趋势	2000. 6. 30
9932056	陈立业	经济	大趋势	2000. 9. 19
9923143	刘英	新闻	大趋势	2000. 12. 3

2. 构造“系统”属性的概念层次树



3. 依据上述的概念层次树对基本关系表进行规约

1. 在概括关系表中只保留三个属性：系别，书名，count
2. 在不同的概念层次上，经过规约可以得到不同的概括关系表

系 别	书 名	借阅次数（count）
商学院	大趋势	4
文学院	大趋势	2
医学院	大趋势	1

关系概括表（1）

系 别	书 名	借阅次数（count）
文 科	大趋势	6
理 科	大趋势	1

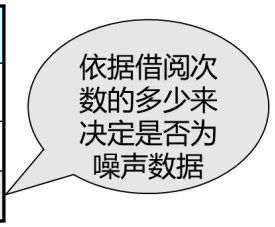
关系概括表（2）

4. 数据挖掘前的预处理：
    1. 在开始特征规则挖掘之前，需要从概括关系表中剔除那些出现频率过低的噪声数据（宏元组）。以减少数据挖掘所处理的宏元组的数量，提高挖掘的效率；同时也避免得到过多的兴趣度不满足要求的挖掘结果
    2. 通常我们会定义一个噪声数据的阈值 M
        1. M 通常指概括关系表中的 count 属性的值
        2. 如果某个宏元组在 count 属性上的取值小于或等于 M，则该宏元组将被看作为噪声数据，不参与后续的数据挖掘过程
            1. 虽然噪声数据不参与挖掘过程，但并不能就此从概括关系表中删除噪声数据
            2. 在计算规则的兴趣度（支持度）时需要包括这些噪声数据
5. 基于基本关系表的特征规则挖掘（M = 1）
    1. 以“灰色”为底色的宏元组为噪声数据

学 号	姓 名	系 别	书 名	借阅日期
9932007	颜立	经济	大趋势	2000. 3. 16
9833090	王家卫	金融	大趋势	2000. 3. 16
9813105	王向东	医学院	大趋势	2000. 5. 8
9928073	朱小明	企管	大趋势	2000. 5. 20
9822041	刘伟	历史	大趋势	2000. 6. 30
9932056	陈立业	经济	大趋势	2000. 9. 19
9923143	刘英	新闻	大趋势	2000. 12. 3

6. 所发现的特征规则是：借阅《大趋势》一书的是“经济系”的学生
7. 基于概括关系表（1）的特征规则挖掘（M = 1）
    1. 以“灰色”为底色的宏元组为噪声数据

系 别	书 名	借阅次数（count）
商学院	大趋势	4
文学院	大趋势	2
医学院	大趋势	1



概括关系表（1）

系 别	书 名	借阅次数(count)
文 科	大趋势	6
理 科	大趋势	1

概括关系表（2）

8. “数据规约”与“挖掘结果”之间的关系：
    1. 在采用面向属性规约方法进行数据挖掘时，如果规约的概念层次过低或过高，可能会减少挖掘所发现的规则
        1. 过低：大量的宏元组会成为噪声数据，被剔除在规则的挖掘之外
        2. 过高：会减少概括关系表中宏元组的数量，从而减少挖掘结果中的规则数
    2. 因此，在开始挖掘之前需要选择一个合适的规约层次。同时挖掘所获得的结果规则的多少也与用户定义的噪声数据的阈值 M 有关

4.1.3 数据立方方法

可以发现，在面向属性规约方法中，经常要做各种统计查询。如果预先做好某些经常需要用到但花费较高的统计、求和等集成计算，并将统计结果存放在多维数据库中。那么在构造概括关系表时，就可以直接从多维数据库中获得所需要的统计结果，从而节省数据规约的时间，提高数据挖掘的效率

采用上述方法的特征规则挖掘方法被称为“数据立方方法”

在数据立方方法中，常用的分析方法有：

1. 数据概括（roll_up 上翻）：将属性值提升到较高的概念层次上。如：从“基本关系表”到“概括关系表一”，再到“概括关系表二”的分析过程
2. 数据细化（drill_down 下翻）：将属性值减低一些层次。如：从“概括关系表二”到“概括关系表一”，再到“基本关系表”的分析过程

4.1.3.1 特征规则挖掘 和 OLAP

特征规则挖掘是由参数主导的自动化过程，而 OLAP 是由分析人员主导的人工过程。

在特征规则挖掘过程中，算法可以在阈值的指导下：

1. 自动决定排除冗余以及和当前挖掘任务无关的属性
2. 自动决定各个属性规约的层次
3. 在对比集的指导下，在挖掘结果中去除与当前挖掘任务关联不大的属性

4.1.3.2 概念描述：特征与区分

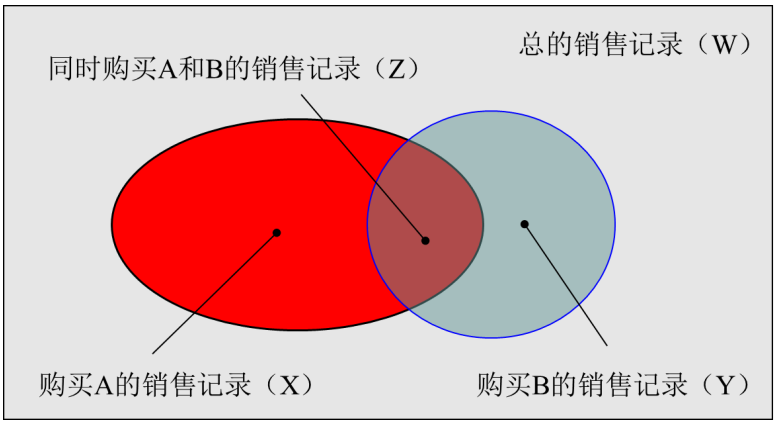
除了使用特征规则挖掘，发现目标集中蕴涵的数据特点外，还可以在引入对比集后进行区分规则挖掘。

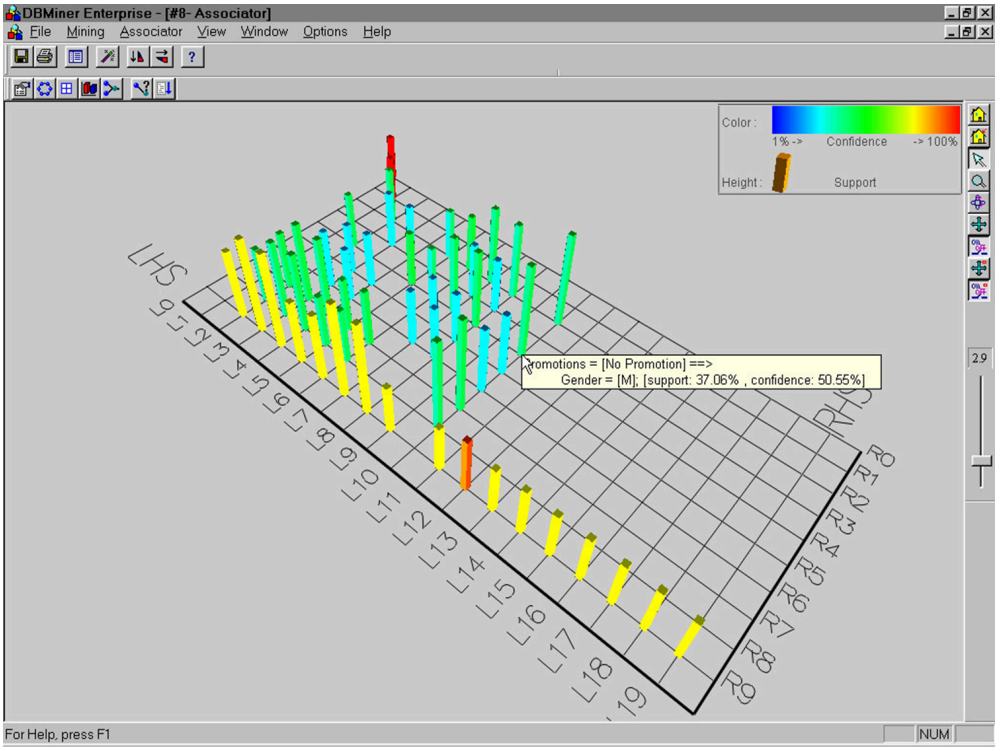
特征规则挖掘和区分规则挖掘，是描述型数据挖掘的“一体两面”，共同构成对一个目标集的概念描述。

4.2 关联规则挖掘

关联规则挖掘是另外一种比较常用的数据挖掘方法：

1. 关联规则（Association Rule）
2. 关联规则用于表示事务数据库中诸多属性之间的关联程度。而关联规则挖掘则是利用数据库中的大量数据通过关联算法寻找属性间的相关性
    1. “属性”在这里也被称为“项”（Item），若干个属性所构成的一个属性集也被称为一个“项集”（Item Set）
3. 例：在购买商品 A 的客户中的大部分人会同时购买商品 B，则可用关联规则表示为：
    1. 规则 R1： $A \rightarrow B$





2. 如果不考虑关联规则的兴趣度和置信度，那么任意组合均构成关联规则
    1. 事实上，人们一般只对满足一定的兴趣度和置信度的关联规则感兴趣
3. 为了发现出有意义的关联规则，需要给定两个阈值：**最小兴趣度和最小置信度**
    1. 满足最小置信度和最小兴趣度的规则为强规则，否则为弱规则
    2. 关联规则挖掘的实质是在数据库（数据仓库）中寻找强规则

4.3 Aprior 算法

待补充。

4.4 发现频繁项集的例子

待补充。

4.5 生成关联规则的例子

待补充。

4.6 序列模式分析

序列模式分析与关联规则挖掘类似，也是为了找出数据对象之间的联系，但序列模式分析法的侧重点是为了找出数据对象之间的前因后果关系。被分析对象具有前后的时序关系。

例如：

1. 下雨 ---- 洪涝
2. 电筒 ---- 电池

4.7 分类分析

数据分类（data classification）是数据挖掘的主要内容之一，主要是通过分析训练数据样本，产生关于类别的精确描述。这种类别通常由分类规则组成，可以用来对未来的数据进行分类和预测。

首先为每一个数据（记录）打上一个标记，即按标记对数据（记录）进行分类，而分类分析则是对每类数据（具有相同标记的一组记录）找出其固有的特征与规律。

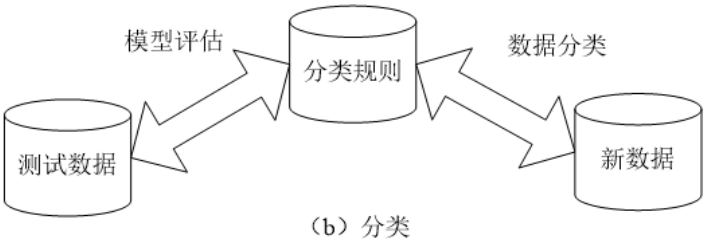
例如：

1. 信用卡公司对持卡人的信誉度标记按：优、良、一般，及差四档分类。这样，持卡人就被分成为四种类型，然后利用分类分析找出每一类持卡人的特征与规律
2. 电话计费系统可以根据在不同时间段内电话的使用频率来调整计费单价

数据分类是一个两个步骤的过程：

- 第 1 步：建立一个模型，描述给定的数据类集或概念集。通过分析由属性描述的数据库元组来构造模型
    1. 用于建立模型的元组集称为训练数据集，其中每个元组称为训练样本
    2. 每个训练样本属于一个预定义的类，由类标号属性确定
    3. 由于给出了类标号属性，因此该步骤又称为有指导的学习
    4. 如果训练样本的类标号是未知的，则称为无指导的学习（聚类）
    5. 学习模型可用分类规则、决策树和数学公式的形式给出
- 第 2 步：使用模型对数据进行分类。包括评估模型的分类准确性以及对类标号未知的元组按模型进行分类

（a）学习



（b）分类

训练数据集的分类标准可以是用户给定的，也可以从领域知识中获取。

分类分析法是一种特征归纳的方法，它将每类数据所共有的特性抽取以获得规律性的规则，目前有很多分析方法，它们大都基于：

1. 决策树方法
2. 贝叶斯方法
3. 人工神经网络方法
4. 约略集方法
5. 遗传算法

4.8 决策树方法

待补充。

4.9 聚类分析

1. 聚类分析又称集群分析，它是研究分类问题的一种多元统计方法
2. 聚类分析分为距离聚类和相似系数聚类
3. 聚类分析与分类分析相反：
    1. 首先输入的是一组没有被标记的记录，系统按照一定的规则合理地划分记录集合（相当于给记录打标记，只不过分类标准不是用户指定的）
    2. 然后可以采用分类分析法进行数据分析，并根据分析的结果重新对原来的记录集合（没有被标记的记录集合）进行划分，进而再一次进行分类分析，如此循环往复，直到获得满意的分析结果为止
4. 例如
    1. 信用卡的等级划分
    2. 学生的分类
5. 主要的聚类方法
    1. 划分方法
    2. 层次的方法
    3. 基于密度的方法
    4. 基于网格的方法
    5. 基于模型的方法
6. 聚类分析结果——聚类树

5. 数据挖掘的应用

5.1 金融业

1. 对帐户进行信用等级评估
2. 股票交易规律分析
3. 信用卡使用模式分析
4. 金融市场的分析和预测

5.2 保险业

1. 保险费率的确定：从大量客户投保数据中分析并取得不同条件、不同人员、不同险种、不同时间与年龄的保险费率，使保险业主能获得合理的利润
2. 险种关联分析：分析客户在购买了某种保险后是否同时还会购买另一种保险
3. 认购险种的预测：通过数据挖掘预测新险种的客户群以及新险种的前景

5.3 零售业

1. 可以分析顾客行为与习惯
2. 可以分析商场销售商品的构成
3. 数据挖掘还可用于商品销售预测、商品价格分析以及零售点设置布局等方面

5.4 科学研究

数据挖掘可以从大量的、漫无边际的实验数据与历史资料中提炼出对科学规则发现有用的信息，从而起到协助科学规律发现的作用。

5.5 其他行业

- 医疗
- 电信
- 司法
- 故障诊断

5.6 应用实例

我们将数据挖掘技术应用于某保险公司的业务数据库上，以挖掘该保险公司有关客户、业务员以及承保、理赔方面的规律。挖掘的部分结果如下：

关联规则发现：从 20912 条元组所构成的 524 个事务中，共发现了 4 条关联规则：

1. “递增型养老保险”和“少儿一生幸福”有关联
2. “递增型养老保险”和“为了明天终生幸福”有关联
3. “为了明天终生幸福”和“递增型养老保险”有关联
4. “为了明天终生幸福”和“少儿一生幸福”有关联

这四条知识说明，保户投保的险种之间可能有一定的关系。对于一个保户来说：

- 如果他投保了“递增型养老保险”，那么他很可能投保“少儿一生幸福”和“为了明天终生幸福”这两个险种
- 如果他投保了“为了明天终生幸福”，那么他很可能投保“递增型养老保险”和“少儿一生幸福”。

特征规则发现：从 3197 条元组中共发现投保客户的 2 条特征知识：

- 中等收入，中年人，特征明显
- 中等收入，青年人，特征明显

这两条知识说明：

保险公司投保的客户以中等收入的中、青年人为主。

关联规则发现：从 481 条元组中共发现了 3 条关联知识

1. 保费总额 高 → 学历：大专 可信度较高
2. 保费总额 中 → 学历：大专 可信度稍高
3. 保费总额 低 → 学历：高中 可信度较高

这三条知识说明，一个业务员的业绩（由保费总额代表）和他的学历有一定的关系，但和他/她的性别无关。

1. 如果一个业务员的业绩好，那么他的学历很有可能是大专
2. 如果一个业务员的业绩中等，那么他的学历有可能是大专
3. 如果一个业务员的业绩低，那么他的学历很有可能是高中。

分类规则发现：从 18075 个客户的个人情况记录中发现六条分类知识：

1. 少儿 → 受保人；可信度很高
2. 儿童，男 → 受保人；可信度较高
3. 收入高 → 投保人；可信度很高
4. 收入中等 → 投保人；可信度较高
5. 老年 → 投保人；可信度较高
6. 收入很高，中年 → 投保人；可信度很高

上述的六条知识说明，通过收入、年龄和性别可以区分出一部分保险客户是投保人还是受保人

这些知识在一定程度上也反映了投保人和受保人所具有的一些特征：

1. 如果客户是少年或儿童男性，那么他很可能是受保人
2. 如果客户的收入较高或收入中等，那么他很可能是投保人
3. 如果客户是老年人，那么他可能是投保人
4. 如果客户是收入很高的中年人，那么他很可能是投保人

5.7 复杂类型数据源的数据挖掘

数据挖掘可以作为数据仓库的一种分析手段，同时，数据仓库又可作为数据挖掘的数据准备工作，因此这两者间有着密切的联系

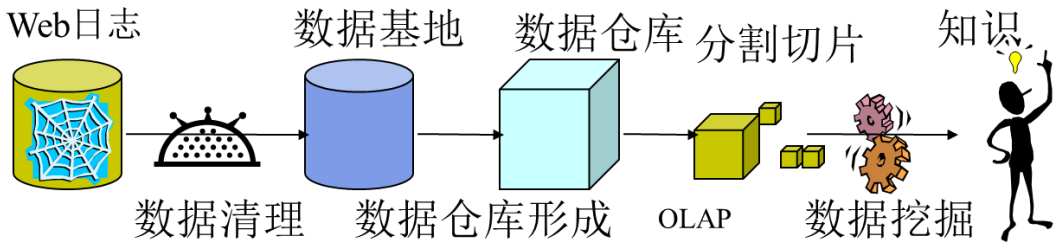
但是从研究角度看，这两者有着不同的研究目的与内容，因此属于两个不同的领域

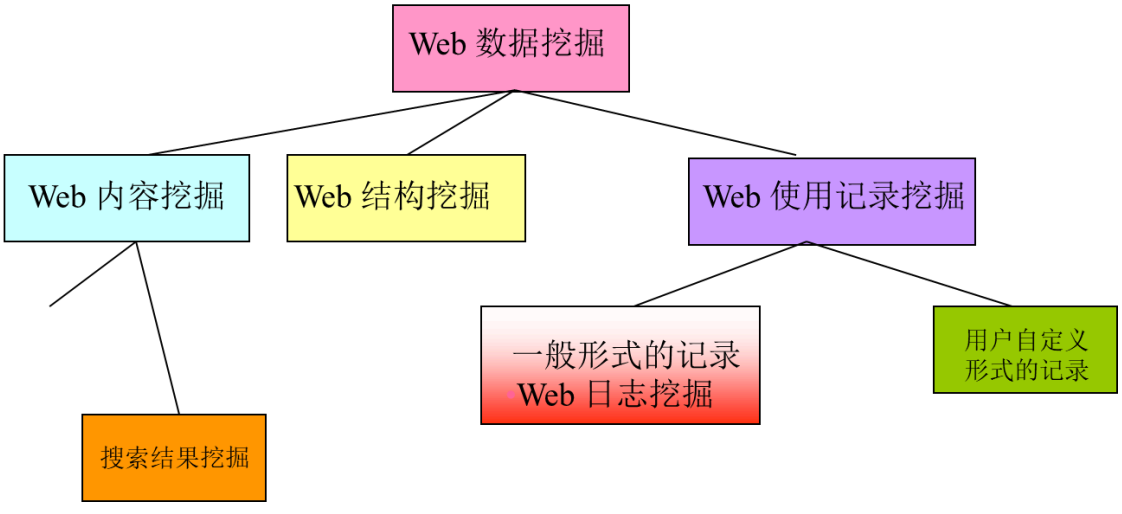
我们不仅可以在数据库（数据仓库）上进行数据挖掘，也可以将数据挖掘技术应用到其他领域

如：网络信息、文本信息、图象识别……

5.8 Web 数据挖掘

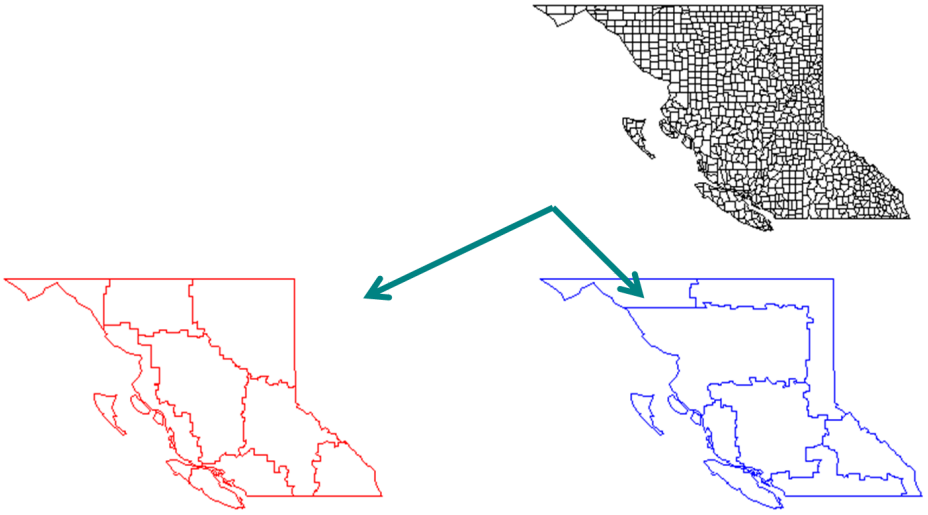
Web 上有海量的数据信息，怎样对这些数据进行复杂的应用成了现今数据库技术的研究热点。





5.9 空间数据库挖掘

空间数据库存储了大量与空间有关的数据，例如地图、遥感或医学图象数据。

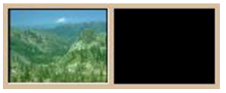


5.10 多媒体数据库挖掘

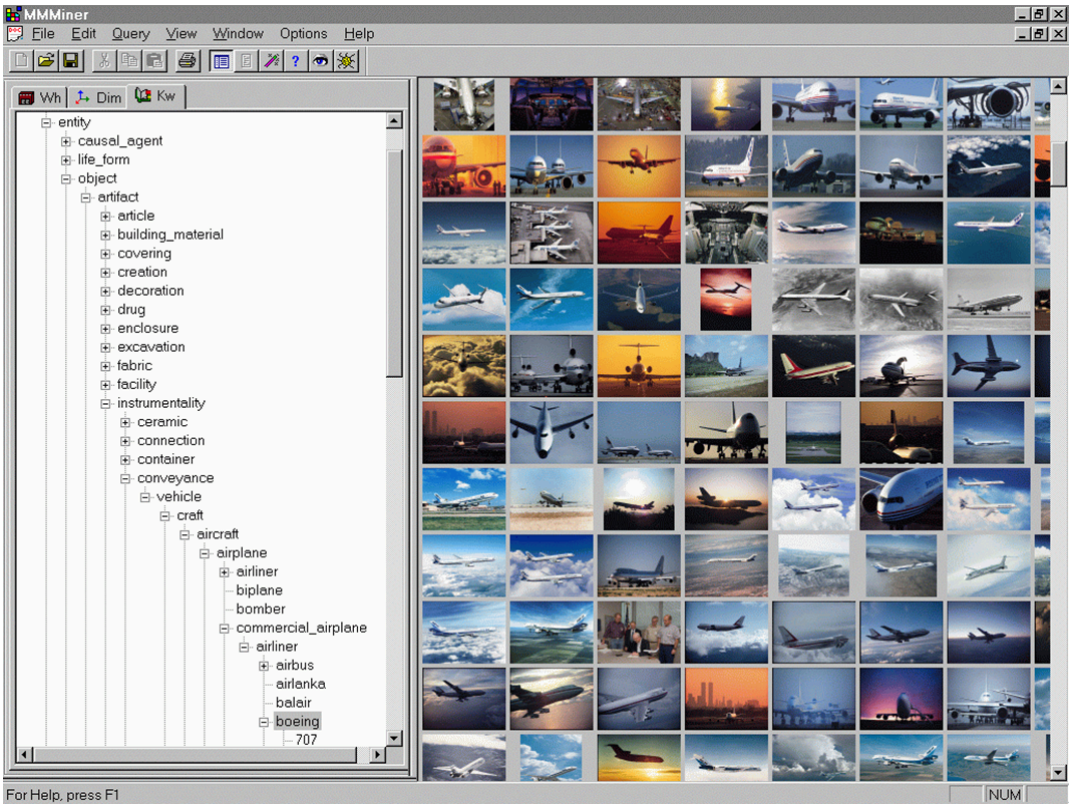
多媒体数据库是指存储和管理大量多媒体对象的数据库，如音频数据、图象数据、视频数据、人类基因数据、因特网数据等。

多媒体数据挖掘主要考虑的是图象数据的挖掘，包括多媒体数据中的相似搜索、多维分析、分类和预测，以及多媒体数据的关联挖掘。

查找蓝天中的飞机



查找“蓝天”与“绿地”



5.11 时间序列数据挖掘