一、数据仓库

1.1 数据挖掘相关概念

支持数据挖掘技术的基础：

1. 海量数据搜集
2. 强大的多处理器计算机
3. 数据挖掘算法

从商业数据到商业信息的进化

进化阶段	商业问题	支持技术	产品厂家	产品特点
数据搜集（60年代）	“过去五年中我的总收入是多少？”	计算机、磁带和磁盘	IBM,CDC	提供历史性的、静态的数据信息
数据访问（80年代）	“在新英格兰的分部去年三月的销售额是多少？”	关系数据库（RDBMS），结构化查询语言（SQL），ODBC Oracle、Sybase、Informix、IBM、Microsoft	Oracle、Sybase、Informix、IBM、Microsoft	在记录级提供历史性的、动态数据信息
数据仓库；决策支持（90年代）	“在新英格兰的分部去年三月的销售额是多少？波士顿据此可得出什么结论？”	联机分析处理（OLAP）、多维数据库、数据仓库	Pilot、Comshare、Arbor、Cognos、Microstrategy	在各种层次上提供回溯的、动态的数据信息
数据挖掘（正在流行）	“下个月波士顿的销售会怎么样？为什么？”	高级算法、多处理器计算机、海量数据库	Pilot、Lockheed、IBM、SGI、其他初创公司	提供预测性的信息

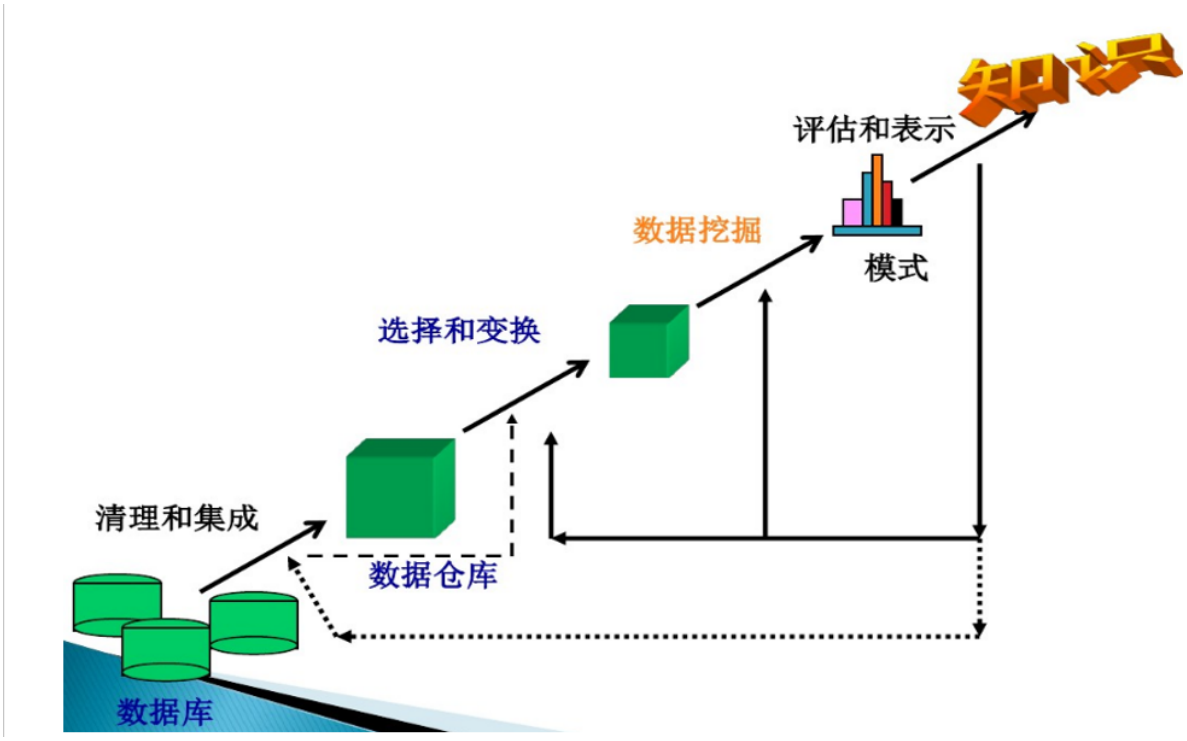
数据挖掘概念：

1. 从大型数据集中提取有趣的（非平凡的、蕴涵的、先前未知的、潜在有用的）信息或模式。
2. 数据挖掘是一项探测大量数据以发现有意义的模式和规则的业务流程。
3. 从大量数据中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。
4. 从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但是又是潜在有用的信息和知识的过程。
5. 商业定义：数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。

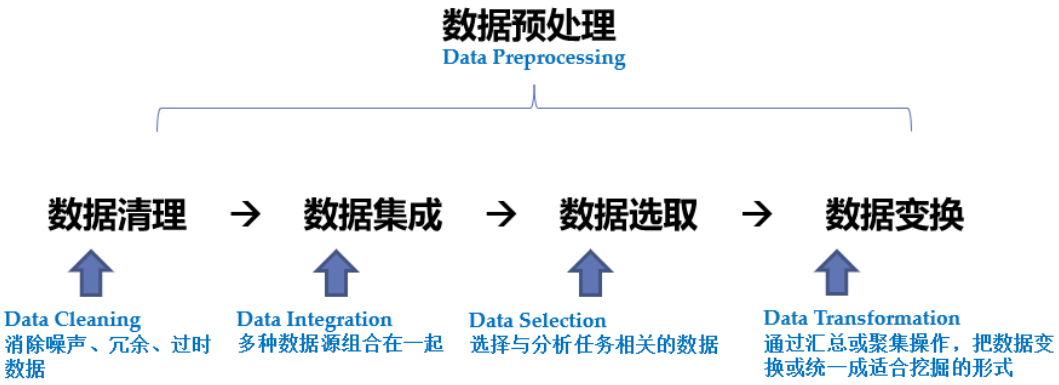
数据挖掘的目的：

从所获取的数据中发现新的、规律性信息和知识，辅助科学决策。

知识发现过程：



1. 数据准备



2. 数据挖掘阶段，确定挖掘的任务，决定使用的算法，提取数据模式
3. 模式评估和表示，

   模式评估（Data Evaluation):根据某种兴趣度度量，识别表示知识的真正有意义的模式。

   知识表示（Knowledge presentation):使用可视化和知识表示技术，向用户提供挖掘的知识。

通过对获取的知识来看数据挖掘和数据查询的区别：

1. 浅知识本质是真实的。可以很容易地在数据库中存储和操作浅知识，数据库查询是提取数据中浅知识的优秀工具。
2. 多维知识也是真实的。然而这种数据以多维格式存储，联机分析处理（OLAP）工具用于处理多维数据。
3. 隐含知识表示数据中的模式或规则，这些模式或规则不容易用数据库查询语言查询出来。然而，数据挖掘算法却可以轻易地找到它们。
4. 深知识是存储在数据库中，仅仅在给出要查找内容的方向时，才能找到的知识。目前数据挖掘工具还不能定位深知识。

1.2 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。

特点：面向主题、集成的数据、不可更新、随时间变化。

1.3 数据挖掘功能

1. 通过对某类数据对象进行汇总、分析和比较，获得对此类对象内涵的描述，并概括这类对象的有关特征。
2. 关联分析，从大量的数据中发现项集之间有趣的联系、相关关系或因果结构，以及项集的频繁模式。
3. 用于预测分析的分类与回归。
4. 聚类分析，将数据划分或分割成相交或者不相交的群组的过程，通过确定数据之间在预先指定的属性上的相似性就可以完成聚类任务。
5. 离群点分析。

1.4 数据挖掘要解决的问题

•**可伸缩**：着眼于数据量剧烈增长的问题

•**高维性**：对象拥有数量不少的属性

•**异种数据和复杂数据**：数据来源广泛，且结构复杂（XML格式，文本格式，流格式等）

•**数据的所有权与分布**：分布式数据处理

•**非传统的分析**：数据挖掘要求自动产生和评估假设，并且数据挖掘数据集多是时机性样本，而非随机性样本

1.5 数据仓库进一步了解

数据仓库是一个过程而不是一个项目；数据仓库是一个环境，而不是一件产品。

1.5.1 数据仓库的数据组织

1. 粒度与分割，

   粒度：分为两种形式，第一种粒度是对数据仓库中数据的综合程度高低的一个度量，它既影响数据仓库中的数据量的多少，也影响数据仓库所能回答询问的种类。第二种粒度形式，即样本数据库，它根据给定的采样率从细节数据库中抽取出一个子集。

   分割：目的同样在于提高效率。它是将数据分散到各自的物理单元中去，以便能分别独立处理。分割的标准如日期、地域、业务领域等。

2. 数据仓库中的元数据，是对数据描述的基础，是数据的数据。

   元数据的作用：

   ①用来对数据仓库中的各种数据进行描述。

   ②用来组织和管理并挖掘信息资源。

   ③描述系统的具体功能要求、执行程序和系统的整体过程。

   ④为数据模型提供存储说明和存储格式，便于扩展。

   元数据记录的主要内容：

   ①对数据仓库中数据的描述。

   ②程序员和决策支持系统的分析员所熟知的数据结构。

   ③数据仓库的数据源。

④数据加入数据仓库时的转换。

⑤数据质量信息，空间数据组织信息，空间参考信息，实体属性信息。

⑥数据模型。

⑦数据模型和数据仓库的关系。

抽取、选择、查询数据的历史记录。

3. 数据仓库中常见的数据组织形式：

①简单堆积文件:它将每日由数据库中提取并加工的数据逐天积累并存储起来。

②轮转综合文件:数据存储单位被分为日、周、月、年等几个级别。

③简化直接文件: 它类似于简单堆积文件，但它是间隔一定时间的数据库快照。

④连续文件: 通过两个连续的简化直接文件，可以生成另一种连续文件。

4. 数据仓库的数据追加，

常用的数据追加技术和方法：

①时标方法

②DELTA文件

③前后映象文件的方法

④日志文件

1.5.2 关键技术

1. 数据的抽取，数据抽取在技术上主要涉及互连、复制、增量、转换、调度和监控等几个方面。
2. 数据的存储和管理

①对大量数据的存储和管理；

②并行处理。

③针对决策支持查询的优化

支持多维分析的查询模式

3. 数据的表现，数据表现是数据仓库的门面，主要集中在多维分析、数理统计和数据挖掘方面。
4. 数据仓库设计的技术咨询，在数据仓库的实施过程中，技术咨询服务至关重要。

1.5.3 与数据库的区别

对比内容	数据库	数据仓库
数据内容	当前值	历史的、归纳的、计算的数据
数据目标	面向业务操作程序，重复处理	面向主题域，面向分析应用
数据特性	动态更新，按字段变化	不能直接更新，只能定时添加、更新
数据结构	高度结构化，结构复杂，适合操作计算	简单、清晰，适合分析
使用频率	较高	相对较低
数据访问量	只访问少量记录	有可能需要访问大量记录
响应要求	以秒为单位	时间长

特点：

1. 数据仓库的数据是面向主题的
2. 数据仓库的数据是集成的
3. 数据仓库的数据是不可更新的
4. 数据仓库的数据是随时间不断变化的

二、分类

定义：

•分类是指将数据映射到预先定义好的群组或类。

•分类的是利用一个分类函数（分类模型、分类器），该模型能把数据库中的数据影射到给定类别中的一个。

•数据分类(Data Classification)：对于一个未知类别标签的数据对象Zu，给出它的类别名称或标签。

分类流程：

1. **•将样本转化为等维的数据特征（特征提取）。**
2. **•选择与类别相关的特征（特征选择）。**
3. **•建立分类模型或分类器（分类）。**
4. **使用模型，对将来的或未知的对象进行分类**

聚类算法：**解决的是事物分组的问题，目的是将类似的事物放在一起**

分类算法：**•是解决"这是什么?"的问题，分类所承担的角色就如同回答小孩子的问题"这是一只船"，"这是一棵树"等。**

•把每个数据点分配到合适的类别中，即所谓的"分类"

分类算法的好处：

1. **预测的准确率，正确地预测新的或先前未见过的数据的类标号的能力**
2. **速度构造模型的速度、利用模型进行分类的速度**
3. **强壮性给定噪声数据或具有空缺值的数据，模型正确预测的能力**
4. **可伸缩性当给定大量数据时，有效地构造模型的能力**
5. **可解释性涉及学习模型提供的理解和洞察的层次**

2.1 基于距离的分类方法

2.1.1 K最近邻居算法（KNN）

•计算每个训练实例到待分类实例之间的距离

•找出和待分类实例距离最近的k个训练实例

•找到的k个训练实例中哪个类别占的最多，待分类实例就属于哪个类别

1.要求的信息：

o**训练集**

o**距离计算值**

o**要获取的最邻近的邻居的数目**k

2.伪代码：

算法　K-近邻分类算法
输入：　训练数据T；近邻数目K；待分类的元组t。
输出：　输出类别c。
（1）$N=\Phi$;
（2）FOR each $d \in T$ DO BEGIN
（3）　IF $|N| \leq K$ THEN
（4）　$N=N \cup \{d\}$;
（5）　ELSE
（6）　IF $\exists u \in N$ such that $sim(t, u)<sim(t, d)$ THEN BEGIN
（7）　$N=N-\{u\}$;
（8）　$N=N \cup \{d\}$;
（9）　END
（10）END
（11）c=class to which the most u∈N.

3.**从k个最近邻居中决定分类结果**

1. 方案一，**选出k个最近的邻居中的数量最多的类标号**
2. 方案二，
   - o**k个最近邻居分别按距离计算权重，权重最大的类标号获胜。**
   - o**权重可以采用 1 / (d**2** + 1) 来计算，其中d为某个最近邻居到待分类实例的距离。**

2.1.2 决策树

是一种二叉树形式的用于预测分析的模型。

树中包含三类信息：

1. 根节点，内部节点
2. 叶子节点
3. 出边

决策树的优势：

- 学习速度较快（比其它的分类方法）
- 可转换为简单、易于理解的分类规则
- 可以使用SQL查询访问数据库
- 与其它方法相媲美的分类精度

使用步骤：

第1步：利用训练集建立并精化一棵决策树，建立决策树模型。这个过程实际上是一个从数据中获取知识，进行机器学习的过程。

第2步：利用生成完毕的决策树对输入数据进行分类。对输入的记录，从根结点依次测试记录的属性值，直到到达某个叶结点，从而找到该记录所在的类。

建立模型分为两个阶段：建树与剪枝