数据挖掘只言片语

**1** **什么是数据挖掘**

业界多将“数据挖掘”定义得很玄乎，个人觉得多有炒作概念之嫌。的确，数据挖掘给使用者带来很多“意想不到”的结果（惊喜 or 惊吓，嘿嘿），但并不是只有利用了高深、复杂理论的技术才能称之为“数据挖掘”。有些人种花，可能只要挖个半米就能将花种下，有些人打井，需要挖地百尺，但我们不能否认种花的人那不叫“挖”。所以，只要你“挖”了，并且达到了你想要的目的，这都属于数据挖掘范畴，哪怕你只是在数据库中很简单的写了一条select的SQL查询语句。

因此，数据挖掘我觉得可以定义为：从纷乱芜杂的数据中提取到自己感兴趣的东西。

**2** **我们常听到OLAP、数据仓库、数据集市、知识发现，它们到底和数据挖掘是什么关系**

OLAP：online analysis process, 联机分析处理，一般具有汇总、合并和聚集功能，方便使用者能从不同角度观察信息。

可以认为OLAP是数据挖掘技术组成的一部分，或者说是整个数据挖掘流程中的一部分。因为OLAP还是有一些局限性的，例如虽然它支持多维分析和决策，但对于更深层次的分析，比如分类，聚类等，还需要其他分析工具支持。

数据仓库：是一种数据存储库，它支持多个异构数据源在单个站点以统一的模式组织存储库，数据仓库可以提供数据清理、数据集成和OLAP等服务供用户使用。

同样，数据仓库准确说应该属于数据库系统技术演变中的一员，它为数据挖掘提供了数据存储、清理、集成以及OLAP等服务，但数据挖掘可选择的数据库系统技术并不仅仅只有数据仓库，例如：hadoop、基于Web的数据库、甚至更高级的对象-关系数据库和时间序列数据库等，都可以为数据挖掘提供类似数据仓库的服务。

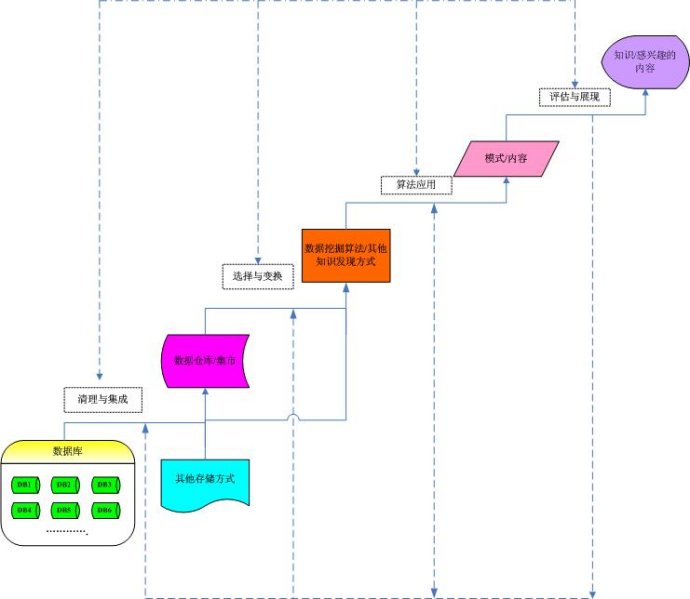
目前提供数据仓库的公司有IBM、ORACLE、TERADATA等。

数据集市：这是我在网上看到的一句话觉得形容的很贴切。

“从字义上看， ‘仓库’可以想像成一所大房子，高高的货架，合理的出入路线，是一种集中存储货物的地方，一般顾客是不来参观访问的； 而说到‘集市’，就容易联想到空旷的场地，川流不息，大小商户摆出摊子，卖衣物的、卖烧饼及卖艺的，是让顾客来消费的地方。 具体来说，数据仓库多用于数据存储，面向企业范围，提供一种面向通用数据管理的服务，不面向最终分析用户；而数据集市是面向分析应用的，会聚焦在选定的主题上，面向最终用户。”

因此，可以认为数据集市是一种小型的数据仓库，但是它面向主题和用户更明确。

**3** **数据挖掘过程**

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4882f26d0100uo5k&url=http://s11.sinaimg.cn/orignal/4882f26dgb1a9801619aa)

如上图所示，数据挖掘是一个循环迭代的过程，主要由以下几个步骤组成：

1）  数据清理：消除干扰数据和有误数据

2）  数据集成：多重异构数据源的组合

3）  数据选择：从数据源中提取与分析任务相关的数据

4）  数据转换：将源数据统一变换成适合挖掘的数据形式

5）  数据挖掘：使用算法提取数据模式

6）  模式评估：根据客观和主观方式，识别能成为知识的真正有趣的模式

7）  知识表达：使用可视化或其他表达技术，想用户提供他们真正感兴趣的东西

**4数据挖掘的功能分类**

1）  概念/类描述：特征化和区分：

特征化是指对数据的一般特性或特征进行汇总，比如，研究上半年交易增加10%的女装特征。目前常用的特征化化手段有：基于统计度量和图的简单数据汇总、基于OLAP的统计（上卷操作）、面向属性的归纳技术等。 

数据区分，将目标类与一个或多个可比较类进行比较，从而得到所需的模式。比如，分析定期购买数码产品的顾客和偶尔购买这种产品的顾客，结果发现，频繁购买者的年龄集中在20-30岁之间，而偶尔购买者年龄偏大或者偏小。因此通过年龄这个属性，变可发现者两类人的区别。 

2）  频繁模式、关联和相关

频繁挖掘是指在大量数据中发现频繁出现或者有趣关联的模式。

最常见的就是购物篮分析，例如某商场市场部经理，想在大量的交易中发现哪些商品会经常被一起购买。常见的关联规则表现形式如下：

Buys(X ,”computer”)=>Buys(X,”software”) [support=1%,confidence=50%]

X是变量，表示顾客。

Buys表示行为“买”。

support=1% 表示支持度，意思是在所有交易中有1%的交易是既买了computer,又买了software。

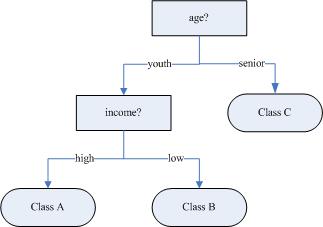
confidence=50% 表示置信度，意思是如果有一位顾客买了计算机，则购买软件的可能性是50%。

目前常见的频繁关联模式挖掘算法有Apriori算法、FP增长算法、基于约束的关联挖掘等。 

3）  分类和预测

分类是这样一个过程，找出描述和区分数据类型的模型，以便能够使用该模型对未知的数据进行描述和分类。目前分类算法比较多，例如：分类（IF-THEN）规则、决策树、数学公式、神经网络、贝叶斯分类、支持向量机和K最近邻分类。

以决策树为例，它是一种类似于流程图的树结构，每个树节点代表在一个维度上的判断，每个分支代表判断后的一个输出，而叶子节点代表数据的最后类别。如下图，便是一棵简单的决策树，表示对顾客分类，首先对年龄进行判断，如果年老，则顾客属于C 类，如果年轻，则再判断收入，如果收入偏高，那么顾客属于A类，如果收入偏低，则顾客属于B类。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4882f26d0100uo5k&url=http://s12.sinaimg.cn/orignal/4882f26dgb1a982d40e8b)

预测和分类的主要区别是，分类是预测分类的类别，这个类别是离散和无序的。而预测是建立连续值函数模型。也就是说，它用来预测空缺的或者不知道数值数据值，而不是一个类别。目前最常见的预测算法有回归分析。 

4）  聚类分析

不像分类和预测分析数据的类别，聚类分析的数据对象它事先是不考虑已知的类别。一般情况下，聚类分析的训练数据中不提供数据类别，因为开始并不知道类别。对象根据最大化类内部的相似性、最小化类之间的相似性原则进行聚类。也就是说，同一个聚类中的对象是具有最高相似性的，而不同聚类之间的对象是不相似的。

目前常见的聚类分析算法有划分式聚类（K均值、K中心点算法）、层次方法（BIRCH算法、ROCK算法）、基于密度的方法（DBSCAN、OPTICS等）、基于约束的聚类算法等。 

5）  离群点分析

数据源中可能会包含一些数据对象，它们与大多数数据的一般行为不一致。这些数据对象变称为离群点。大部分数据挖掘方法都将离群点作为噪声丢弃，然而，在一些应用中，比如欺诈交易、反洗钱过程里，罕见的事件可能比正常的事件更令人感兴趣。

目前常见的离群点分析算法有：基于统计分布的离群点分析、基于距离的离群点分析、基于密度的局部离群点分析、基于偏差的离群点分析。 

6）  演变分析

演变分析描述行为随时间变化对象的规律或趋势，并对其建模。

比如，假如你有股票交易市场过去几年的数据，并希望投资高科技产业公司的股票。那么通过演变分析可以识别整个股票市场和特定公司的股票演变规律。这种规律可以帮助预测股票市场价格的未来走向。

演变分析和分类预测的区别主要在于，前者是基于时间序列的，而后者没有这个特定的要求。

**5** **挖掘得到的结果都是用户感兴趣的吗？**

答案肯定是“NO”。实际上，结果当中只有一小部分是用户感兴趣的。

这个问题通常还会引出下面3个问题：“什么样的结果才是有趣的？数据挖掘能产生所有的有趣结果吗？还有它能仅产生有趣模式吗？”

什么样的结果才是有趣的？a) 易于理解；b) 对于数据的校验是正确有效的；c) 潜在有用，即有相关的业务需求； d) 新颖，即不在目前业务范围之内，但确实有用。

衡量这四个标准，有主观和客观两类度量方式。客观度量比如关联分析中的支持度和置信度；而主观是依据用户的聚焦点和长期以往的经验积累。

数据挖掘能产生所有的有趣结果吗？目前看来这也是不可能的。

还有它能仅产生有趣模式吗？这是数据挖掘优化的一个方向，但准确率和召回率之间往往是互相制约的，准确率高了召回率或许就降低，反之依然。也就是说，假设一个业务场景下有10条有趣模式，为了仅产生有趣结果，避免无趣结果干扰，算法最后可能只得到5条有趣模式；为了能将更多的有趣模式纳入进来，修改算法后，可能将有趣模式增加到8条，但可能又纳入了2条无趣结果。所以召回率和准确率之间的平衡是数据挖掘优化的一个方向。