商务智能

2013年被称为大数据元年，2014年则将成为大数

据BI爆发年，BI工具将大数据的内在意义挖掘和展

现出来，并指导决策分析，将给企业带来大数据的

真正价值。

数据仓库（或数据集市）、查询报表、数据分析、数据

挖掘、数据备份和恢复等部分组成的、以辅助企业决策

为目的一类技术及其应用

商务智能（Business intelligence）

§ 工业界

- 商务智能可以被看作是一类技术或工具，利用它们可以

对大量的数据进行收集、管理、分析和挖掘，以改善业

务决策水平，增强企业的竞争力

§ 学术界

- 商务智能是一套理论、方法和应用，通过它们可以快速

地发现海量数据中隐含的各种知识，有效地解决企业面

临的管理和决策问题，支持企业的战略实施。

结构化数据：表格

非结构化数据：文本

信息（information）

§ 通过一定的技术和方法，对数据进行集成、分析，

挖掘其潜在的规律和内涵，得到的结果是信息。

§ 信息是具有商务意义的数据

知识（knowledge）

§ 当信息用于商务决策，并基于决策开展相应的商务

活动时，信息就上升为知识

§ 信息转化为知识的过程不仅需要信息，而且需要结

合决策者的经验和能力，用以解决实际的问题

数据仓库

§ 各种数据源的数据经过抽取、转换之后需要放到一

个供分析使用的环境，以便对数据进行管理，这就

是数据仓库

§ 数据集市（data mart）：通常针对单个部门的数

据仓库，区别于企业范围内的数据仓库。

§ 数据仓库可以将分析数据与实现业务处理的操作型

数据隔离，一方面不影响业务处理系统的性能，另

一方面为数据的分析提供了一个综合的、集成的、

统一的数据管理平台

在线分析处理：业务性能度量可以通过多个维度、

多个层次进行多种聚集汇总，通过交互的方式发现

业务运行的关键性能指标的异常之处。

数据挖掘是从大量数据中自动发现隐含的信息和知

识的过程，属于主动分析方法，不需要分析者的先

验假设，可以发现未知的知识

常用的分析方法包括分类、聚类、关联分析、数值

预测、序列分析、社会网络分析等

数据挖掘：分类

§ 分类（classification）是通过对具有类别的对象的

数据集进行学习，概括其主要特征，构建分类模型，

根据该模型预测对象的类别的一种数据挖掘和机器

学习技术。

§ 例如，电信公司的客户可以分为两类，一类是忠诚

的，一类是流失的。根据这两类客户的个人特征方

面的数据以及在公司的消费方面的数据，利用分类

技术可以构建分类模型

数据挖掘：聚类

§ 聚类（clustering）是依据物以类聚的原理，将没

有类别的对象根据对象的特征自动聚集成不同簇的

过程，使得属于同一个簇的对象之间非常相似，属

于不同簇的对象之间不相似。

§ 其典型应用是客户分群

数据挖掘：关联分析

§ 关联分析最早用于分析超市中顾客一次购买的物品

之间的关联性

§ 例如，发现关联规则（association rule）“尿不湿

®啤酒（0.5%，60%）”，其含义为，0.5%的交

易中会同时购买尿不湿和啤酒，且买尿不湿的交易

中有60%会同时买啤酒

数据挖掘：数值预测

§ 数值预测用于预测连续变量的取值。

§ 常用的预测方法是回归分析

§ 例如，可以根据客户个人特征，如年龄、工作类型、

受教育程度、婚姻状况等，来预测其每月的消费额

度。

数据挖掘：序列分析

§ 序列分析是对序列数据库进行分析，从中挖掘出有

意义模式的技术。

§ 序列模式（sequential pattern）的发现属于序列

分析，它是从序列数据库中发现的一种有序模式

§ 例如，《赤壁，鸿门宴，见龙卸甲》，

意味着“看了赤壁之后会接着看鸿门宴，过段时间会

看见龙卸甲”

数据挖掘：社会网络分析

§ 社会网络（social network）是由个人或组织及其

之间的关系构成的网络

§ 社会网络分析（social network analysis）是对社

会网络的结构和属性进行分析，以发现其中的局部

或全局特点，发现其中有影响力的个人或组织，发

现网络的动态变化规律等。

2.2 数据仓库与数据库

§ 联系表现在两个方面

- 数据仓库中的大部分数据来自于业务系统的数据库中

- 当前绝大多数数据仓库都是利用数据库系统来管理的

§ 区别：构建目的、管理的数据、管理方法都不同

- 数据库主要用于实现企业的日常业务运营，提高业务运

营的效率；数据仓库的构建主要用于集成多个数据源的

数据，这些数据最终用于分析

- 数据库通常只包含当前数据，数据的存储尽量避免冗余，

数据的组织按照业务过程涉及的数据实现，是应用驱动

的。数据仓库中的数据是按照主题组织的，将某一主题

的所有数据集成在一起，数据存在冗余

- 数据库中的数据需要进行频繁的插入、删除、修改等更

新操作，需要复杂的并发控制机制保证事务运行的隔离

性。

- 数据仓库中的数据主要用于分析处理，除了初始的导入

和成批的数据清除操作之外，数据很少需要更新操作

- 数据库中数据的更新操作的时效性很强，事务的吞吐率

是个非常重要的指标。而数据仓库的数据量十分庞大，

分析时通常涉及大量数据，时效性不是最关键的。数据

仓库中的数据质量非常关键，不正确的数据将导致错误

的分析结果。

在线事务处理（online transaction processing），

简称OLTP，是数据库管理系统的主要功能，用于

完成企业内部各个部门的日常业务操作。

§ 在线分析处理（online analytical processing），

简称OLAP，是数据仓库系统的主要应用，提供数

据的多维分析以支持决策过程

DSS

§ 决策支持系统（decision support system）的概

念最早由Scott Morton在20世纪70年代初提出

- “一种交互式的基于计算机的系统，用于协助决策者使

用数据和模型解决非结构化的问题”

- 决策支持系统的用户主要是管理人员和业务分析人员，

主要目的是辅助决策者进行科学决策

关联规则的挖掘:

从交易数据库、关系数据库以及其他的数据集

中发现项或对象的频繁的模式(frequent patterns)

、关联(associations)的过程“啤酒尿布综合症”

项 集 ( Itemset): x={i j1 , i j2 , …, i jp }

l 每 个 项 集 包 含 的 项 的 个 数 ， 称 为 项 集 的 长 度 ， 一 个 长

度 为 k 的 项 集 又 称 为 k 项 集

支持度：

sup(X)=|X|/n

E.g. X={A} ,

Y={A, B}=AB

若 support(X) >=minsup ， 则 X 称 为 频 繁 项 集

(frequent itemset) ， 也 可 以 说 X 是 频 繁 的

置信度

发现所有的频繁集 (frequent itemsets): 支持

度>= minimum support的所有项集

统计每个k项候选集的支持度，找出频繁的k项

集： L k

- 利用频繁的k项集生成k+1项候选集( Candidate

itemset )： C k +1 ，k=k+1;

第四章，分类

分类任务就是通过学习得到一个目标函数（

分类模型） f， 把每个属性集x映射到一个预

先定义的类标号y。

Eager

- Two steps

ØModel construction

ØModel usage

- 决策树、 贝叶斯、 SVM、 logistic regression …..

§ Lazy

- K最近邻， 基于案例的推理

数值预测：

K nearest Neighbors

Regression

线性回归分析

（1）构建包含因变量和自变量的训练集；

（2）通过散点图，确认因变量和自变量之间的近似线

性关系；

（3）计算系数，构建模型；

（4）检验模型；

（5）利用模型进行预测

聚类：