数据仓库ppt复习提纲

- 第一章：数据仓库的概念和体系结构
  - 概述 1
    - 数据仓库
    - 数据挖掘
  - 数据库的兴起
    - 人工管理
    - 文件系统
    - 数据库系统
    - 数据仓库 3
      - 联机事务处理（针对事务A、B） -> 联机分析处理（针对场景A+B）
      - 面向分析决策型应用的数据仓库
      - 数据仓库和数据库的区别和联系 3
  - 数据仓库的基本概念
    - 元数据
      - def：对数据进行描述的数据
      - 按照应用场合分类
        - **数据元数据（数据源信息）**
        - **过程元数据（软件接口功能）**
      - 按照用途分类 4
        - 技术元数据
        - 业务元数据
    - 数据粒度
      - 粒度越大，细节程度越低，综合性越高
      - 四种粒度级别 1.2.2
    - 数据模型
      - 概念数据模型：现实-信息
        - 星型模型
          - **事实表**
          - **维表**
        - 雪花模型
        - 星系模型
      - 逻辑数据模型：信息-数据
      - 物理数据模型：数据-计算机
    - ETL 5-1.2.4
      - **ETL的定义：抽取、转换、加载**
      - 数据抽取：**数据提取、数据清洁、数据转换、生成衍生数据**
      - 数据转换：**字段级转哈努、清洁和净化、数据派生、数据聚合和汇总**
    - 数据集市
      - def：关于少数几个主题的小型数据仓库
      - **面向分析决策型应用**
      - 两种构建方式

        - 自下而上
        - 自上而下
    - **5层模型 5-1.2.5** 书p16
        - **数据源（最基本）**
        - **数据预处理**
        - **数据存储于管理（最关键）**
        - **OLAP服务器：从不同维度来分析数据**
        - **数据处理**
- 数据仓库的特点和组成 6
    - **四个基本特征：面向主题、数据集成、数据非易失、数据随时间变化**
        - 基本原则：面向主题
        - 最重要特征：数据是集成的
    - **组成：加载管理器、仓库管理器、查询管理器**
- 数据仓库体系结构 9
    - 大数据5V特征 p18
    - 各种挑战
    - 大数据时代的数据仓库
- 第二章：数据 11
    - 数据的概念
        - 数据是信息的表现形式和载体
        - 数据经过加工后变为信息
    - 数据分类
    - 数据属性
        - 标称
        - 序数
        - 区间
        - 比率
    - **数据集的三个重要特性：维度、稀疏性、分辨率 12**
    - 数据预处理
        - 意义 13
        - 数据清洗
            - 空缺值处理
            - 属性选择
            - 噪声处理14
                - **分箱**；聚类；回归
                - **均值；最近边界；中值**
        - 数据集成 16
            - 出现的问题：模式匹配、数据值冲突、数据冗余
        - 数据变换 17
            - 平滑、聚集、概化、**规范化**、属性构造
        - 数据归约
            - 缩小数据范围
            - **数据压缩：哈弗曼编码 23**

        - 基于关系数据库
        - 星型模型
        - 面对多层次的复杂维度，使用**雪花模型**：一个复杂的维度通过多张表来描述
        - 优势&劣势 38-4.3.1
            - 没有大小限制
            - 但是响应速度差
            - 。。
    - MOLAP和ROLAP的对比 40-4.3.3
    - 数据立方体 40
        - 数据仓库针对数据立方体进行查询
        - 维度：观测角度
        - 测度：观察到的值
- 第五章：数据挖掘基础 44
    - 数据挖掘的定义
    - 数据库中的知识发现KDD
    - 数据仓库与数据挖掘的关系
        - 数据仓库是一种解决方案，是对原始操作数据进行各种处理并转换为有用信息的处理过程
        - 数据仓库是数据挖掘的数据源；数据挖掘是数据仓库的应用
    - 数据挖掘任务 45
        - 关联规则
        - 聚类分析
            - 划分聚类；层次聚类；基于密度的聚类
            - **通常作为数据挖掘的第一步**
        - 分类分析
        - 回归分析
        - 相关分析
        - 异常检测
    - 数据挖掘流程 49
        - 数据挖掘对象
            - 结构化数据：关系数据库；数据仓库
                - **数据库中的数据需要先进行预处理**
                - **数据仓库是数据挖掘的最佳环境**
            - 非结构化数据
        - 数据挖掘分类
        - 知识发现 50
            - 数据挖掘只是知识发现的一个步骤
            - **相关数据的收集和提取是知识发现的关键性工作**

- 填空题
    - 第一章
        - **数据仓库的特点分别是：**

        - 元数据是描述数据仓库内数据的结构和建立方法的数据。根据元数据用途的不同可将元数据分为 **技术 元数据和 业务 元数据**两类
        - OLAP技术多维分析过程中，多维分析操作包括 **切片 、 切块 、 钻取 、 旋转** 等
        - 数据库中的知识挖掘(KDD)包括以下七个步骤：**数据清理，数据集成，数据选择，数据变换，数据挖掘，模式评估，知识表示**
        - 数据挖掘的性能问题主要包括：算法的效率、可扩展性和并行处理
        - 当前的数据挖掘研究中，最主要的三个研究方向是：统计学、数据库技术和机器学习
        - 在万维网(WWW)上应用的数据挖掘技术常被称为：WEB挖掘
    - 第二章
        - 进行数据预处理时所使用的主要方法包括：数据清理、数据集成、数据变换、数据规约
        - 处理噪声数据的方法主要包括：分箱、聚类、计算机和人工检查结合、回归
        - 模式集成的主要问题包括：整合不同数据源中的元数据，实体识别问题
        - 数据概化是指：沿概念分层向上概化
        - 数据压缩可分为：有损压缩，无损压缩
        - 进行数值归约时，三种常用的有参方法是：线性回归方法，多元回归，对数线性模型
    - 第三章
        - 概念分层有四种类型，分别是：模式分层，集合分组分层，操作导出的分层，基于规则的分层
        - 同时满足最小置信度临界值、最小支持度临界值的关联规则称为强关联规则
    - 第四章
        - 关联规则挖掘中，两个主要的兴趣度度量是：支持度和置信度
        - Aprior算法包括和两个基本步骤：连接和剪枝
        - 大型数据库中的关联规则挖掘包含两个过程：找出所有频繁项集、由频繁项集产生强关联规则
    - 第五章
        - 通过对数据进行预处理，可以提高分类和预测过程的准确性、有效性和可伸缩性
    - 第六章
        - 在数据挖掘中，常用的聚类算法包括：划分方法、层次的方法、基于密度的方法、基于网格的方法和基于模型的方法
        - 许多基于内存的聚类算法所常用的两种数据结构是和数据矩阵、相异度矩阵

何谓数据挖掘它有哪些方面的功能

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、但又是潜在有用的信息和知识的过程称为数据挖掘。相关的名称有知识发现、数据分析、数据融合、决策支持等。

数据挖掘的功能包括：概念描述、关联分析、分类与预测、聚类分析、趋势分析、孤立点分析以及偏差分析等。

何谓数据仓库为什么要建立数据仓库

数据仓库是一种新的数据处理体系结构，是面向主题的、集成的、不可更新的(稳定性)、随时间不断变化(不同时间)的数据集合，为企业决策支持系统提供所需的集成信息。

建立数据仓库的目的有3个：

一是为了解决企业决策分析中的系统响应问题，数据仓库能提供比传统事务数据库更快的大规模决策分析的响应速度。

二是解决决策分析对数据的特殊需求问题。决策分析需要全面的、正确的集成数据，这是传统事务数据库不能直接提供的。

三是解决决策分析对数据的特殊操作要求。决策分析是面向专业用户而非一般业务员，需要使用专业的分析工具，对分析结果还要以商业智能的方式进行表现，这是事务数据库不能提供的。

何谓聚类它与分类有什么异同

聚类是将物理或抽象对象的集合分组成为多个类或簇(cluster)的过程，使得在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。

聚类与分类不同，聚类要划分的类是未知的，分类则可按已知规则进行；聚类是一种无指导学习，它不依赖预先定义的类和带类标号的训练实例，属于观察式学习，分类则属于有指导的学习，是示例式学习。

结构化数据和非结构化数据之间的主要区别是什么?

虽然结构化（定量）数据提供了客户的"鸟瞰图"，但非结构化（定性）数据提供了对客户行为和意图的更深入了解。让我们探讨一些关键的差异领域及其影响：

- **来源：**结构化数据来源于 GPS 传感器、在线表格、网络日志、Web 服务器日志、OLTP 系统等，而非结构化数据源包括电子邮件、文字处理文档、PDF 文件等。
- **形式：**结构化数据由数字和数值组成，而非结构化数据由传感器、文本文件、音频和视频文件等组成。
- **模型：**结构化数据具有预定义的数据模型，并在放入数据存储之前被格式化为一组数据结构（例如，写入时模式），而非结构化数据以其本机格式存储并且在使用之前不会被处理（例如，读取模式）。
- **存储：**结构化数据以需要较少存储空间的表格格式（例如，Excel 表或 SQL 数据库）存储。它可以存储在数据仓库中，这使其具有高度可扩展性。另一方面，非结构化数据存储为需要更多空间的媒体文件或 NoSQL 数据库。它可以存储在数据湖中，这使得它难以扩展。
- **用途：**结构化数据用于机器学习 (ML) 并驱动其算法，而非结构化数据用于自然语言处理(NLP) 和文本挖掘。

相似点如下：

1. 目标：结构化数据挖掘和非结构化数据挖掘都是通过分析数据来发现隐藏在数据中的有用信息和模式，以帮助人们做出更好的决策。
2. 技术：结构化数据挖掘和非结构化数据挖掘都使用类似的数据挖掘技术，例如聚类、分类、关联规则挖掘等。
3. 数据预处理：结构化数据挖掘和非结构化数据挖掘都需要进行数据清洗、特征选择、数据转换等预处理步骤，以提高数据的质量和可挖掘性。
4. 数据可视化：结构化数据挖掘和非结构化数据挖掘都需要使用数据可视化技术，以帮助人们更好地理解和解释数据挖掘的结果。
5. 应用场景：结构化数据挖掘和非结构化数据挖掘都可以应用于各种领域，例如商业、金融、医疗、教育、社交媒体、搜索引擎等。

文本分类和文本聚类的异同

文本分类和文本聚类都是文本挖掘的主要技术，它们之间有一些重要的异同点。

异同点如下：

1. 目标：文本分类和文本聚类的目标不同。文本分类的目标是将文本分为不同的预定义类别，例如新闻分类、垃圾邮件过滤等。而文本聚类的目标是将文本聚集成不同的组，每个组内的文本具有相似的主题或内容。

2. 数据处理：文本分类和文本聚类的数据处理方式也不同。在文本分类中，需要进行特征提取和选择，以提取最相关的特征并减少噪声和冗余。在文本聚类中，需要进行相似度计算和聚类算法，以确定文本之间的相似度和聚类的结果。
3. 算法：文本分类和文本聚类使用的算法也不同。文本分类常使用的算法包括朴素贝叶斯、支持向量机（SVM）和决策树等。文本聚类常使用的算法包括K均值、层次聚类和密度聚类等。
4. 性能评估：文本分类和文本聚类的性能评估方式也不同。文本分类通常使用准确率、召回率、F1得分等指标来评估分类器的性能。文本聚类通常使用类内相似度、类间距离等指标来评估聚类的性能。
5. 应用场景：文本分类和文本聚类都有广泛的应用场景。文本分类常用于新闻分类、情感分析、垃圾邮件过滤等领域。文本聚类常用于文本聚合、信息检索、社交媒体分析等领域。

Web挖掘是指从Web中提取有用的信息和知识的过程，其中包括内容挖掘、结构挖掘和使用挖掘等三种方法。

这三种方法的特点如下：

1. 内容挖掘：内容挖掘是指从Web页面中提取文本、图像、视频等非结构化数据的过程。内容挖掘的特点是需要使用自然语言处理和机器学习等技术来处理非结构化数据，例如文本分类、情感分析、实体识别等。内容挖掘的应用场景包括搜索引擎、新闻聚合、社交媒体分析等。
2. 结构挖掘：结构挖掘是指从Web页面的结构中提取信息的过程，例如HTML标记、链接、网页层次结构等。结构挖掘的特点是需要使用数据挖掘和机器学习等技术来识别和提取网页的结构信息，例如网页聚类、网页分类、网页去重等。结构挖掘的应用场景包括信息检索、网页去重、网页分类等。
3. 使用挖掘：使用挖掘是指从Web使用记录中提取有用的信息和知识的过程，例如用户行为、点击率、购买记录等。使用挖掘的特点是需要使用大数据技术和机器学习等技术来分析和挖掘用户行为数据，例如推荐系统、广告定向投放、用户画像等。使用挖掘的应用场景包括电商平台、社交媒体、在线广告等。