数据仓库与数据挖掘1-8讲提纲_snowball

Lecture 1 引言

- 问题的提出

技术发展的角度	• 信息体系（DIKW）：信息金字塔
	• 数据管理技术发展的脉络
	• 现有DBMS系统
	  ○ 关注的问题
	  ○ 面临的挑战

应用发展的角度	• 电子化、信息化、数字化
	  ○ 金融信息化
	  ○ 科学研究
	  ○ 企业生产人员：不同人员信息需求

- 数据环境理论

体系结构化环境的层次	• 操作层 • 原子/数据仓库层 ○ 报表区别：操作型 vs 数据仓库 • 部门层 • 个体层
数据仓库要解决的基本问题	
数据仓库需要建立，而不是购买	

- 几项技术之间的区别和联系
  - 数据库技术
  - 与数据仓库技术
  - 操作型系统（Operational Systems）
  - 联机事务（OLTP）系统
    - RDBMS通常应用于OLTP

Lecture 2 数据仓库基本概念及其特征

数据仓库的定义	• 定义 • 理解 • 回答的问题

数据仓库的特点	• 面向主题
	• 集成的
	  ○ 数据集成的方法：MQS
	  ○ 数据仓库vs. 联邦数据库
	• 相对稳定的
	• 反映时间变化

OLTP系统和数据仓库	全面比较

Lecture 3 数据仓库设计

- 数据仓库建设的目标

使组织机构的信息变得容易存取	• 容易理解，见名知义 • 存取工具必须简单易用，存取速度快
一致地展示组织机构的信息	• 数据具有可信性 • 高质量的数据：一致的、完整的、定义唯一理解的

具有广泛的适应性和便于修改	• 变化：用户需求、业务情形、数据内容和技术状况 • 新数据的加入，现有数据和应用不应该发生改变或者崩溃
发挥安全堡垒作用以保护信息资产	能够有效地控制对机构机密信息和个人隐私信息的访问
数据仓库必须在推进有效决策方面承担重要角色	
数据仓库建设成功的前提是为业务群体所接受	

- 数据仓库设计的基本思想
- 数据仓库设计方法概述

DB和DW设计方法的比较	• 处理类型
	• 面向需求
	• 设计目标
	• 数据来源
	• 设计方法
	  ○ DB : SDLC(System Development Life Cycle)
	  — 应用需求驱动
	  DW: CLDS
	  — 数据驱动＋需求驱动
	  ○ SDLC与CLDS方法比较
	• 数据仓库设计的原则
	• 在实际工程中的设计方法
	• 数据驱动系统设计方法的基本思路

DW设计的三级数据模型	• DW与DB的三级数据模型的区别
	  ○ 过程模型与数据模型
	• DW设计的三级数据模型
	  ○ 概念模型
	  ○ 逻辑模型
	  ○ 物理模型

性能问题	• 粒度划分 • 数据分片 • 合并表 • 选择冗余 • 进一步分离数据 • 导出数据 • 建立广义索引
数据仓库中的元数据	• 定义 • 重要性 • 内容

- 数据仓库设计中的性能考虑

Lecture 4 联机分析处理

- OLAP的提出

提出	• 关系数据库满足了联机事务处理（OLTP）的要求 • 存在着大量的分析型应用—— RDB无法适应 • 在RDBMS上开发前端产品，支持上述应用逻辑 ○ E.F.Codd把这类技术称为“OLAP”（1993年）
OLAP应用举例	• 不同时间段的比较（同期比） • 排序和统计分类(top N/bottom N) • 客户特定的即席分析(市场分割、即席分组的情况)

- 多维数据结构

数据立方体	多维数组 数据单元

维	• 组织方式：维层次路径
	• 维层次
	  ○ 维成员（DIMENSION VALUES）,维成员树
	  ▪ 维层次关系
	  ○ 维成员属性（ATTRIBUTES）

	○ 维层次和类的区别
事实（度量）	

- 多维数据操作（含义，举例）
  - 主要操作
    - 切片（Slice）
    - 切块（Dice）
    - 旋转
    - 钻取
  - 其他操作
    - Drill through（穿透）
    - Ranking（排序）
- 多维数据模型的实现

实现技术 （OLAP分类）	• Relational OLAP (ROLAP) • Multidimensional OLAP (MOLAP) ○ MOLAP 和ROLAP 的比较 • Hybrid OLAP (HOLAP)
多维数据库存取	• 多维查询语言----MDSQL • 用关系结构表示多维数据 ○ 事实表 / 维表 ○ 星型模式 / 雪花模式 • 事实的提取

Lecture 5 数据仓库系统

- 数据仓库系统概念
  - 统包括：数据、硬件、软件和用户
  - 数据库系统与数据仓库系统比较
- 数据仓库系统体系结构（图）
- 数据仓库系统组成

数据源	▪ 内部数据 ▪ 外部数据
数据存储及管理 （数据仓库管理系统）	• 在现有的数据库管理系统的基础上增加若干功能 • 构建数据仓库管理系统
OLAP引擎	OLAP引擎的分类 OLAP引擎的要求
工具	• 后端工具 ○ 数据仓库建模工具 ○ ETL工具 ○ 数据仓库监测工具 ○ 数据仓库运行与维护工具 • 前端工具： ○ 查询和报表工具 ○ 联机分析处理工具 ○ 数据挖掘工具

- 数据仓库系统工具

数据仓库建模工具	建立概念模型 建立逻辑模型 难点
ETL工具	• 数据抽取（Data Extraction） ○ 开放的数据源 ○ 数据的完整性和有效性检查 • 数据转换（Data Transformation） ○ 模式冲突 ○ 语义冲突 • 数据加载（Data Load）
数据仓库监测工具	• 监视数据内容 • 数据仓库性能的监测

	• 数据仓库“报警”时钟（主动）
数据仓库运行、维护工具	• 安全性管理 • 数据仓库的备份和恢复 • 如何保证数据仓库系统的高可用性 • 数据存放周期和过期数据的处理
数据仓库查询、报表工具	• 二维报表 • 交叉表 • 邮签报表 • 自由式报表
数据仓库前端分析展现工具	• 验证型（Verification）工具：OLAP工具 • 挖掘型（Discovery）工具：Data Mining工具

Lecture 6 数据挖掘技术概论

- 数据挖掘的提出
  - ”啤酒和尿布”的故事
  - 数据爆炸
  - 数据、信息和知识
- 数据挖掘基本概念
  - 定义
  - 感兴趣的模式
    - 有效
    - 新颖
    - 潜在有用
    - 最终可被理解
- 数据挖掘的应用与发展前景
  - 卫星遥感
  - 生物信息
  - 购物篮问题
  - Web日志分析
  - 市场营销
  - 风险管理
- 数据挖掘的技术分类

数据挖掘技术的要求	
不同角度的数据挖掘分类	• 数据源不同 • 不同分析方法 • 采用不同技术 • 不同应用领域
方法分类	描述型：概念/类描述、关联分析、聚类分析、异常点检测 预测型：分类分析、趋势分析
方法评估	兴趣度的度量 “查全率”和“查准率”

- 数据挖掘基本步骤

KDD处理的基本步骤	○ 数据准备 ▪ 数据选择 ▪ 数据清理与预处理 ○ 数据集成 ○ 数据挖掘 ○ 知识表达（挖掘结果的表述） ○ 模式评估 ○ 知识应用
数据挖掘系统的典型结构	
数据挖掘成功的关键	

- 数据挖掘的发展方向
  - 数据挖掘技术的研究历史

- 数据挖掘的问题
- 数据挖掘研究的主要方向

Lecture 7 数据预处理

- 数据质量的概念和内涵
  - 数据质量的基本概念
  - 数据质量主要问题
- 为什么需要数据预处理?
  - 脏数据是数据质量的主要问题
  - 数据预处理的主要任务

数据清洗	• 补充缺失数据 • 识别孤立点，平滑噪音数据 • 处理不一致的数据
数据集成	• 模式集成 • 冗余数据的处理 • 检测和解决数值冲突
数据转换	• 平滑处理：从数据中消除噪音数据 • 聚集操作：对数据进行综合，类似于Data Cube的构建 • 数据概化：构建概念层次 • 数据规范化：将数据集中到一个较小的范围之中 • 属性构造

数据归约	• 数据立方体聚集
	• 减少数据维度（维归约）
	  ◦ 基于判定树归纳的方法
	• 数据压缩：使用编码机制压缩数据集
	  ◦ 有损
	  ◦ 无损
	• 数值压缩：用替代的、较小的数据表示替换或估计数据
	  ◦ 有参：如线性回归
	  ◦ 无参：如直方图
	• 数据聚类

• 数据离散化与概念层次的构建	• 分箱（Binning） • 直方图分析 • 聚类分析的方法 • 基于熵的离散化 • 根据自然分类进行分割 • 概念层次树

- 在大型数据库中挖掘描述统计度量
  - 数据描述性分析
  - 聚集函数
  - 数字特征

Lecture 8 关联规则挖掘（不全）

- 什么是关联规则挖掘
  - 基本形式，种类
  - 相关概念
    - k-项集，频繁项集
    - 规则有效性和确定性的度量值：支持度、置信度
  - 形式化定义
  - 基本思路，基本过程
    - 频繁项集：基本特征
  - 优缺点，重要性

单维	• 查找频繁项集—Apriori算法
	• FP-growth 算法：不用生成候选集
	• 模式评价
	  ○ 兴趣度
	  ○ 置信度

多层	• 自上而下，深度优先的方法
	• 多层关联规则
	  ○ 支持度不变
	  ○ 支持度递减

多维	• 属性
	  ○ 分类属性
	  ○ 量化属性
	• 搜索频繁k阶-谓词集合
	  ○ 用量化属性的静态离散化挖掘多维关联规则
	  ○ 量化关联规则
	  ○ 基于距离的关联规则

基于约束	• 知识类型约束：指定要挖掘的知识类型
	• 数据约束：指定与任务相关的数据集
	• 维/层次约束：指定所用的维或概念结构中的层
	• 规则约束：指定要挖掘的规则形式(如规则模板)
	  ○ 单调性约束(monotone constraint)
	  ○ 反单调性约束(anti-monotone constraint)
	  ○ 简洁性约束(succinct constraint)
	  ○ 可转变的约束(convertibale constraint)
	  ○ 不可转变的约束(unconvertibale constraint)
	• 兴趣度约束：指定规则兴趣度阈值或统计度量