2013 教育部-IBM 产学合作专业综合改革项目
厦门大学《数据挖掘原理及实践》课程习题

第 4 章 数据仓库与数据的概念描述

1. 数据仓库的定义是什么？数据仓库有哪些显著特征?

2. 请简述数据概化的过程和基本方法。

3. 假定数据仓库包含三维: time, doctor, patient, 和两个度量: count 和 charge, 其中 charge 是医生对病人一次诊治的收费。

(1) 列举三种流行的数据仓库建模模式。

(2) 使用(1) 列举的模式之一, 画出上面的数据仓库的模式图。

(3) 由基本方体[day, doctor, patient]开始, 为列出 2004 年每位医生的收费总数, 应当执行哪些 OLAP 操作。

4. 假定 BigUniversity 的数据仓库包含如下 4 维: student, course, semester 和 instructor；2 个度量: count 和 avg_grade。在最低的概念层 (例如: 对于给定的学生. 课程. 学期和教师的组合), 度量 avg_grade 存放学生的实际课程成绩。在较高的概念层, avg_grade 存放给定组合的平均成绩。

(1) 该数据仓库画出雪花型模型图。

(2) 由基本方体[student, course, semester, instructor]开始, 为列出 BigUniversity 每个学生的 CS 课程的平均成绩, 应当使用哪些特殊的 OLAP 操作。

(3) 如果每维有 5 层 (包含 all), 如“student <major <status <university <all”, 该立方体包含多少方体 (包括基本方体和顶点方体)?

5. 数据仓库可以用星形模式或者雪花模式建模, 简略讨论这两种模式的相似点

和不同点，然后分析它们的相对优缺点。哪种模式更实用？给出你的观点并陈述你的理由。

6. 数据仓库实现的一个流行方法是构造一个称为数据立方体的多维数据库。不幸的是，这常常产生巨大的．稀疏的多维矩阵。给出一个例子，解释这种大型稀疏数据立方体。

7. 三种主要的数据仓库应用：信息处理，分析处理和数据挖掘的区别是什么？讨论 OLAP 挖掘 (OLAM)的动机。

8. 考虑下表显示的购物篮事务：

事务 ID	购物项
1	{牛奶，啤酒，尿布}
2	{面包，黄油，牛奶}
3	{牛奶，尿布，饼干}
4	{面包，黄油，饼干}
5	{啤酒，饼干，尿布}
6	{牛奶，尿布，面包，黄油}
7	{面包，黄油，尿布}
8	{啤酒，尿布}
9	{牛奶，尿布，面包，黄油}
10	{啤酒，饼干}

(1) 从这些数据中，能够提取出的关联规则的最大数量是多少 (包括零支持度的规则)?

(2) 能够提取的频繁项集的最大长度是多少?

(3) 写出从该数据集中能够提取的 3-项集的最大数量的表达式。

(4) 找出一个具有最大支持度的项集 (长度为 2 或者更大)。

(5) 找出一对项 a 和 b, 使得规则{a}-{b}和{b}-{a}具有相同的置信度。

9. 请比较 OLAP 和 OLTP 系统的区别。

10. 请简述数据挖掘中关联规则 Apriori 算法的思想。

11. 请举出至少 4 个数据挖掘的统计图形描述方式。