数据挖掘第四次作业

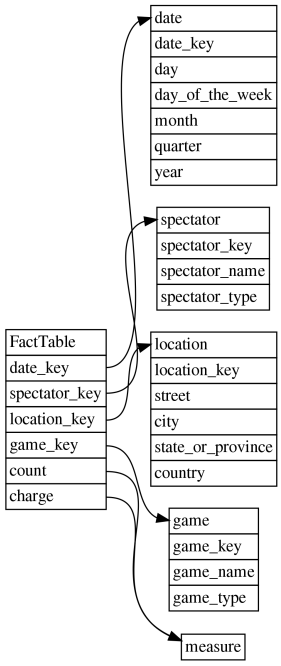
李晨昊 2017011466

2020 年 4 月 5 日

1 数据立方体练习

假定某一数据仓库包含 4 个维：date(日期), spectator(观众), location(地点) 和 game(节目)；2 个度量 count 和 charge，count 是观众的人数，charge 是观众在某日期某地点观看某节目的费用。观众分三类：学生、成年人和老人，每类观众有不同的收费标准。

1. 画出该数据仓库的星型模式图 StarSchema（自己定义维表属性）(2 分)。



2. 从基本方体 [date, spectator, location, game] 开始，为列出 2018 年学生观众在清华大学大礼堂的总付费，应当执行哪些 OLAP 操作，并说明原因（2 分）。

   (a) drill-down on spectator to spectator_type

   (b) drill-down on date to year

(c) drill-down on location to street

(d) dice for spectator_type= 学生 and date=2018 and location= 清华大学大礼堂

原因：需要先执行 drill-down 讲需要检测的维度暴露出来，然后用 dice 讲需要的数据方块选择出来。

2 频繁项集挖掘

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

1. 写出表 1 中事务数据集的极大频繁项集（最小支持度为 2）(2 分)。

   $\{CE, DE, ABC, ACD\}$

2. 写出表 1 中事务数据集的闭频繁项集（最小支持度为 2）(2 分)。

   $\{C, D, E, AC, BC, CE, DE, ABC, ACD\}$

3. 简述频繁项集、极大频繁项集和闭频繁项集之间的关系（2 分)。

   极大频繁项集是闭频繁项集的子集，闭频繁项集是频繁项集的子集。