

数据集成环境下基于相似度的数据库聚类算法

郑 凯^{a,b}, 梁卓明^b, 郑文栋^b

(华南师范大学 a. 教育信息技术学院; b. 网络中心, 广州 510631)

摘 要: 数据集成环境中的全局数据规划方法复杂度很高, 且需要经历较长的周期。针对该问题, 提出一种基于相似度集合运算的数据库聚类算法, 利用自定义的相似数据库、数据库聚类和聚类距离描述数据库的聚类过程, 并给出聚类效果的评价方法。实例分析结果证明, 该算法简单且具有通用性。

关键词: 数据集成; 数据库相似度; 语义缺失; 数据库聚类; 聚类距离

Database Clustering Algorithm Based on Similarity in Data Integration Environment

ZHENG Kai^{a,b}, LIANG Zhuo-ming^b, ZHENG Wen-dong^b

(a. School of Education Information Technology; b. Network Center, South China Normal University, Guangzhou 510631, China)

【Abstract】 The current methods in the plan of global-data in data integration should abstract a realistic model first, which is very complicated and needs a long period. In order to solve the problem, this paper presents a database clustering algorithm based on similarity. It defines similar database, database clustering and clustering distance, describes the database clustering process and gives evaluation method for clustering effect. Analysis on the case proves that the algorithm is concise and general.

【Key words】 data integration; database similarity; lack of semantic; database clustering; clustering distance

DOI: 10.3969/j.issn.1000-3428.2011.19.022

1 概述

数据集成已经成为企业信息集成过程中一个关键的基础性工作。由于企业数据集成对查询效率、资源挖掘和科学决策等方面存在潜在的要求, 因此一般采取在信息集成环境中建立一个全局性数据副本的物化集成方法。基于数据仓库的数据集成是物化集成技术中的主流发展方向, 数据仓库从分布异构的数据库中抽取具有全局意义的数据集中存储, 并以此为基础在分布的数据库中进行数据同步和决策分析, 同时保持数据源的自治性和独立性, 集成的全局数据与各个应用系统操作的数据分离。

如何科学地规划和管理数据仓库中的全局数据是企业数据集成中需要解决的一个关键问题。鉴于数据仓库“面向主题”的特点, James Martin 博士提出了主题数据库的概念。主题数据库是建立在各信息系统数据库的基础上, 面向一类对象或业务逻辑的结构且独立于具体应用的全局性数据资源。把主题数据库作为基本单元来划分数据仓库中的数据成为数据仓库中数据规划的主要研究方向, 并已经形成初步的理论基础和划分方法。文献[1]提出了一个基于实体-活动理论的实体集的概念和相关模型。在此基础上, 文献[2]提出了基于实体间聚合度的主题数据库构造方法。燕山大学刘文远教授带领的研究小组通过引入广义聚合度的概念^[3]对这种方法进行了改进。2008 年至 2009 年间, 该研究团队又陆续提出了基于实体依赖关系^[4]、基于实体亲密度^[5]和基于 K-means 聚类分析^[6]等一系列主题数据库规划方法, 这些方法的共同特点是以现实的实体-活动联系作为理论依据。但在目前的企业集成环境下, 抽象出准确的实体-活动联系模型非常困难, 需要投入大量的时间和人力, 因此, 这些方法仍缺乏足够的实用性。在关联信息缺失的数据集成环境下, 构造通用的数

据库聚类算法, 以形成的聚类数据库作为数据仓库中数据结构的基本单元, 是规划数据仓库中全局数据的新方向。

2 数据库聚类的形式化描述

建立数据集成环境下异构数据库之间量化的相似度计算是进行数据库聚类的前提。文献[7]定义了数据库间的相似度, 本文在此基础上给出一些新的定义。

定义 1(数据集成环境) 包含 K 个数据库的集成环境表示为 $E = \{D_1, D_2, \dots, D_K\}$, 其中, D_i 可描述为具有 L 个属性的集合, $D_i = \{A_1, A_2, \dots, A_L\}$ 。数据库进行聚类后, 包含 T 个聚类数据库的数据仓库表示为 $E' = \{Class_1, Class_2, \dots, Class_T\}$ 。

定义 2(数据库相似度) 数据库 D_i 和 D_j 之间相似度的度量值记为 $Sim(D_i, D_j)$ 。如果用 $|D_i \cap D_j|$ 表示 D_i 和 D_j 的属性交集个数, $|D_i \cup D_j|$ 表示 D_i 和 D_j 的属性并集个数, 则定义 $Sim(D_i, D_j) = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}$, $Sim(D_i, D_j)$ 的取值范围为 $[0, 1]$ 。

定义 3(相似数据库) 给出一个处于 $[0, 1]$ 区间的相似度阈值 α , 若 $Sim(D_i, D_j) \geq \alpha$, 则称数据库 D_i 和 D_j 相互为 α 相似数据库, 记为 $D_i \stackrel{\alpha}{\cong} D_j$ 。

定义 4(数据库聚类) T 个 α 相似数据库的聚类记为 $Class^\alpha$ 。根据定义, $Class_1^\alpha \cup Class_2^\alpha \cup \dots \cup Class_T^\alpha = E'$; 对于 $Class^\alpha$ 中的任意 2 个数据库 D_i 和 D_j , 有 $Sim(D_i, D_j) \geq \alpha$ 。

基金项目: 国家科技支撑计划基金资助项目(2008BAH37B05084)

作者简介: 郑 凯(1978—), 男, 高级实验师、硕士, 主研方向: 数据库技术, 教育信息技术; 梁卓明、郑文栋, 工程师

收稿日期: 2011-03-04 **E-mail:** david@scnu.edu.cn

