

An Approach to Subject Database Planning Based on K-means

Qinglan Fan^{1†}, Lin Wang¹, Yaqing Liu², Dun Bai¹, Mingyu Lu²

1. Research Institute of Highway Ministry of Transport, Beijing 100088, China

2. School of Information Science & Technology, Dalian Maritime University, Dalian 116026, China

[†]Email: titi1900@163.com

Abstract

Subject database planning is always the emphasis of information resource planning. Algorithm for entities aggregation has heavy impact on the quality of subject database planning. However, the existing approaches to entities aggregation computation are inclined to fall into cluster offset, which does great harm to the quality of subject database planning. Against the problem, we firstly calculate the degree of aggregation between entities. Secondly, we view the relations of aggregation as the relations of links between web pages. We apply PageRank algorithm to sort all entity pairs by importance. At last, we exploit K-means algorithm to aggregate entities iteratively. The results of experiments show that our approach avoids cluster offset and improves the quality of subject database planning.

Keywords: Subject Database; Information Resource Planning; Cluster Offset

一种基于 K-means 算法的主题数据库规划方法^{*}

范青蓝¹, 汪林¹, 刘亚清², 白惇¹, 鲁明羽²

1. 交通运输部公路科学研究院, 北京 100088

2. 大连海事大学信息科学技术学院, 辽宁 大连 116026

摘要: 主题数据库规划一直是信息资源规划领域研究的重点, 而实体聚合算法是影响主题数据库规划质量的关键。但是现有的计算实体聚合度方法很容易陷入聚簇偏置, 影响了规划质量。针对这一问题, 作者首先计算实体对的亲和度, 然后将实体对的亲和关系看作网页之间的链接关系, 使用 PageRank 算法对实体对重要性排序, 进而使用 K-means 算法迭代来聚合实体。实验结果表明本文提出的方法能够避免聚簇偏置, 进而改善了主题数据库规划质量。

关键词: 主题数据库; 信息资源规划; 聚簇偏置

引言

数据库技术的发展与成熟极大地方便了交通运输企事业单位对数据的管理与维护。但是数据库的数量往往与这些企事业单位的规模成正比, 单位规模越大, 职能域越多、业务过程越复杂、涉及的业务活动越多, 用于存储相关数据的数据库也就越多, 从而造成业务活动间、业务过程间甚至职能域间数据库的耦合程度和冗余程度越来越高。不规范的数据环境导致“信息孤岛”问题的出现, “信息孤岛”已成为交通运输企事业单位信息化的瓶颈问题。

企业信息资源规划 IRP(Information Resource Planning)是指对企业生产经营活动所需要的信息, 从产生、获取, 到处理、存储、传输及利用进行全面规划^[1]。信息资源规划的一个核心内容就是主题数据库规划, 目的是打破“信息孤岛”, 实现数据标准的统一, 从而实现应用系统的集成和数据资源的共享^[3]。

^{*}基金资助: 受“面向 ITS 体系框架的交通运输数据资源规划研究”支持资助。

主题数据库规划早期完全是由人工来完成。但随着企业规模的增加，完全依靠人工来规划主题数据库变得越来越困难，相关学者开始探寻半自动或者自动的主题数据库规划方法^[4]。1997 年王玉书等人用形式化的语言定义了主题数据库规划中的若干核心概念和定理，这为后期主题数据库规划方法的发展奠定了基础^[5]。2005 年周炎涛等人针对主题数据库规划中存在的“双向性”问题，给出了基于亲合性和聚合性分析的单向性模型和计算公式，并通过实例对企业活动之间的相关性进行了具体分析^[6-7]。2008 年刘文远等人分析了主题数据库间存在的实体依赖和操作依赖关系，提出了依赖程度的量化公式。然后根据主题数据库之间的依赖度与被依赖度，提出了主题数据库的分解算法^[8]。实体聚合算法是影响主题数据库规划质量的关键，但是现有方法在计算实体聚合度时很容易陷入聚簇偏置。针对这一问题，本文以实体对的亲和度作为基础，将实体对的亲和关系看做网页之间的链接关系，进而使用 K-means 算法迭代来聚合实体，避免聚簇偏置，改善了主题数据库规划质量。

1 基本术语与定义

在主题数据库规划领域，有四个基本的术语：职能域、业务过程、业务活动和实体^[5]。职能域(Function Area)或职能范围、业务范围，是指一个企业或组织中的一些主要业务活动领域，例如人事、财务、科研、规划等。职能域可由多个业务过程(Business Process)组成。例如对于产品规划职能，它由产品设计、产品定价、产品说明书管理等处理组成。在每个业务流程中都包含一定数目的业务活动(Business Activity)。例如，产品说明书管理中登记说明书作为活动业务活动是一个部门或组织功能分解后最基本的功能单元。一个企业的构成除了上述活动外，还有就是企业中的数据。数据分为属性类，如员工的姓名、年龄等。若干个联系密切的属性类构成的一个集合，称作实体(entity)。

定义 1 实体-业务活动是一个三元组 $AE=(A, E, R)$ ，其中 $A=\{a_1, a_2, a_3, \dots, a_n\}$ 是业务活动集， $E=\{e_1, e_2, e_3, \dots, e_m\}$ 是实体集， $R=\{r_{ij}|1 \leq i \leq n, 1 \leq j \leq m\}$ ，如果实体 e_j 不参与业务活动 a_i ， r_{ij} 的取值为 0，否则 r_{ij} 的取值为 1。

定义 2 对于任意的 $e_j \in E$ ，称 $t(e_j)=\{a_i|r_{ij}=1, 1 \leq i \leq n\}$ 为 e_j 在 AE 上的投影。对于任意的 $e_{j1}, e_{j2} \in E$ ，称 $t(e_{j1}, e_{j2})=\{a_i|r_{ij1}=1 \text{ 且 } r_{ij2}=1, 1 \leq i \leq n\}$ 为 e_{j1}, e_{j2} 在 AE 上的投影。

定义 3 对于任意的 $e_{j1}, e_{j2} \in E$ ， e_{j1} 对 e_{j2} 的实体相关系数被定义为 $p(e_{j1}, e_{j2})=|t(e_{j1}, e_{j2})|/|t(e_j)|$ 。不难理解， $p(e_{j1}, e_{j2})$ 的取值范围在 0-1 之间。如果两个实体 e_{j1} 和 e_{j2} 从来没有被同一个业务活动使用过，那么 $p(e_{j1}, e_{j2})$ 的取值为 0；如果 e_{j1} 和 e_{j2} 总是被同一个业务活动使用，那么 $p(e_{j1}, e_{j2})$ 的取值为 1。

例如，对于一个实体-业务活动 $AE=(A, E, R)$ ，记 $A=\{a_1, a_2, a_3\}$ ， $E=\{e_1, e_2, e_3, e_4\}$ ， $R=\{r_{11}=0, r_{12}=1, r_{13}=0, r_{14}=0, r_{21}=1, r_{22}=1, r_{23}=0, r_{24}=0, r_{31}=0, r_{32}=0, r_{33}=0, r_{34}=0\}$ 。则 $t(e_1)=\{a_2\}$ ， $t(e_2)=\{a_1, a_2\}$ 。则 $t(e_1, e_2)=\{a_2\}$ 。 $p(e_1, e_2)=1$ ， $p(e_2, e_1)=0.5$ 。

2 基于 PageRank 的实体对重要性排序算法

为了使实体对聚簇时更好地确定主题数据库的数量，更重要的是为了能更好地避免出现聚簇偏置问题，我们首先使用 PageRank^[2]算法（见公式(1)）对实体对的重要性进行排序，根据排序结果来决定主题数据库的数量。

$$I(e_j) = \beta + (1 - \beta) \sum_{i \in B_j} \frac{I(e_i)}{|F_i|} \quad (1)$$

对于公式(1)， $I(e_j)$ 是实体 e_j 的重要性度量， β 是一个调节因子，通常取 0.15， $B_j=\{p(e_k, e_j)|1 \leq k \leq m\}$ ， $F_i=\{p(e_i, e_k)|1 \leq k \leq m\}$ 。根据公式(1)， $I(e_j)$ 越大说明 e_j 越重要，反之 e_j 越不重要。

根据公式(1)，实体对的重要性可由公式(2)给出。

$$I(e_i, e_j) = I(e_i) + I(e_j) \quad (2)$$

显然，根据公式(2)可知， $I(e_i, e_j)$ 越大说明实体对 (e_i, e_j) 越重要，反之 (e_i, e_j) 越不重要。

3 基于 K-means 的实体聚簇算法

主题数据库规划就是要确定企业需要建立哪些主题数据库，即对企业进行自顶向下的全局规划，把企业的全部数据资源划分成一些易于管理的单元，每一个单元就是一个主题域。根据实体之间的关系紧密程度对实体进行聚类分析，从而得到主题数据库的初步规划结果^[9-10]。

本文提出的实体对聚簇算法如算法 1 所示。首先确定簇的个数，即主题数据库的个数。然后从实体相关系数集中选择 N 个元素作为簇中心；然后计算余下的元素与每个簇中心的差异度，差异度公式见公式(3) (x 是待聚簇的元素， c 是簇中心)，将元素归入与其差异度最小的簇中，当所有的元素入簇后计算每个簇中元素的平均值作为新的簇中心，如果聚簇前后簇集不发生变化则实体聚簇过程结束，否则使用当前的聚簇结果重新聚簇。

$$d(x, c) = \sqrt{(x - c)^2} \quad (3)$$

算法 1: cluster(P, N)
输入: P , 实体相关系数集合,
 N , 最终簇的个数。
输出: CS , 簇集。
1. $CS_1 \leftarrow \text{extractByImportance}(P, N)$ //从 P 中根据实体对重要性的排序抽取前 N 个元素作为簇中心。
2. **while**(true)
3. **for each** $p \in P - CS_1$ **do**
4. $\text{degreeOfDiversity}(p, CS_1)$ //计算 p 与各个簇中心的差异度
5. $\text{minimumDegreeOfDiversity}(p)$ //将 p 归入与其具有最小差异度的簇。
6. **end for**
7. $CS_2 \leftarrow \text{calculateCenterOfCluster}(CS_1)$ //重新计算簇中心
8. **if** $CS_1 == CS_2$ **then** //如果聚簇前后簇集 CS 无变化
9. **return** CS_1
10. **else**
11. $CS_1 \leftarrow CS_2$
12. **end if**

4 主题数据库规划方法在高速公路中的应用

我们从高速公路的数据库中选择了 6 个 MIS 系统，车辆档案管理系统，驾驶员档案管理系统，交通事故管理系统，车辆黑名单管理系统，停车场管理系统，服务区管理系统。通过对实际的业务活动的调研，确定了行驶里程数($e1$)、车辆型号($e2$)、车主($e3$)、驾驶员($e4$)、事故地点($e5$)、事故时间($e6$)、事故类型($e7$)、事故车辆($e8$)、伤亡人员($e9$)、停车场面积($e10$)、停车位数量($e11$)、服务区数量($e12$)、服务区位置($e13$)等 13 个实体。13 个实体间的相关系数如表 1 所示。

表 1 实体相关系数表

	$e1$	$e2$	$e3$	$e4$	$e5$	$e6$	$e7$	$e8$	$e9$	$e10$	$e11$	$e12$	$e13$
$e1$	1	0.9	0	0	0.6	0.6	0.6	0.6	0	0	0	0	0
$e2$	0	1	0	0	0.6	0.6	0.6	0.6	0	0	0	0	0
$e3$	0.5	0.5	1	0.9	0.5	0.5	0.5	0.5	0.6	0.2	0	0	0.1
$e4$	0	0	0.5	1	0.6	0.6	0.6	0.6	0	0	0	0	0
$e5$	0	0	0	0	1	0.9	0.9	0.9	0	0	0	0	0
$e6$	0	0	0	0	0	1	0.9	0.9	0	0	0	0	0
$e7$	0	0	0	0	0	0	1	0.9	0	0	0	0	0
$e8$	0	0	0	0	0	0	0.5	1	0	0	0	0	0
$e9$	0.6	0	0	0	0.9	0.9	0.9	0.9	1	0	0	0	0.9
$e10$	0	0	0	0	0	0	0	0	0.6	1	0.7	0	0
$e11$	0	0	0	0	0	0	0	0	0.6	0.9	1	0	0
$e12$	0	0	0	0	0.6	0.6	0.6	0.6	0	0.5	0.7	1	0
$e13$	0.5	0.5	0.5	0	0.6	0.6	0.6	0.6	0.6	0.1	0.5	0.9	1

根据表 1 和本文提出的实体聚簇算法, 最终确立了如下 5 个主题数据库: 以实体 e_1 和 e_2 为核心的车辆主题数据库, 以实体 e_3 和 e_4 为核心的司机主题数据库, 以实体 e_5 , e_6 , e_7 , e_8 和 e_9 为核心的交通事故数据库, 以实体 e_{10} 和 e_{11} 为核心的停车场主题数据库, 以实体 e_{12} 和 e_{13} 为核心的服务区主题数据库。

5 结论

本文提出了一种基于 K-means 算法的主题数据库规划方法。首先根据实体之间共享的业务活动来计算实体对的相关系数, 根据相关系数, 借鉴 PageRank 公式来对实体对的重要性进行排序, 并由此来确定主题数据库的数量, 最后使用 K-means 算法来对主题数据库进行聚簇。本文提出的方法在高速公路 MIS 系统做了初步的应用, 应用结果显示本文提出的方法能够很好地避免聚簇偏置, 规划结果也更为合理。

REFERENCES

- [1] GAO Fu-xian. Transport Informatization and information resource plan. Computer and Communications, 2000, 18(06): 4-7
- [2] QIN Zheng, ZHANG Ling, LI Na. Application of an Improved PageRank in Web Crawler. Journal of Computer Research and Development, 2006, 43(06): 1044-149
- [3] CHEN Qian. LIU Wei. MENG Qingjiu. Study on topic database of information resource plan. Journal of Intelligence, 2006, (02): 136-137
- [4] CHEN Yong-yue, ZHOU Ning, XIA Huo-song. Research on the Construction and Quality Evaluation Method of Theme Database. QING BAO KE XUE, 2011, 29(02): 222-226
- [5] WANG Yu-shu, DONG Pi-ming. MATHEMATICAL FORMULA OF EVALUATING THE RATIONALITY OF SUBJECT DATABASES. JOURNAL OF SOFTWARE, 1997, 8(02): 93-98
- [6] ZHOU Yan-tao, XIE Dong, WU Zheng-guo. Research on Integration Analysis Methodology of Subject Database. Computer Engineering and Applications, 2005, 41(12): 168-170
- [7] ZHOU Yan-tao, XIE Dong, WU Zheng-guo. Aggregation Analysis of Subject Databases Based on Unilateral Model[J]. Journal of Hunan University (Natural Sciences), 2005, 32(03): 101-104
- [8] LIU Wen-yuan, YU Jia-xin, XU Li-na, CHEN Guo-ying. Mode of Large-scale Subject Database's Subdivision Based on Dependency Relations. Computer Science, 2008, 35(05): 223-225
- [9] ZHANG Wen-juan. RESEARCH AND APPLICATION OF SUBJECT DATABASE PLANNING IN IRP. Hebei: Yanshan University, 2010
- [10] LEI Xiao-feng, XIE Kun-qing, LIN Fan, XIA Zheng-yi. An Efficient Clustering Algorithm Based on Local Optimality of K-means. Journal of Software, 2008, 19 (07): 1683-1692

【作者简介】



¹ 范青蓝 (1983-), 女, 汉族, 硕士, 助理研究员, 研究方向: 智能交通。硕士毕业于北京工业大学控制工程专业。
Email: fql@itsc.cn

² 汪林 (1980-), 男, 汉族, 硕士, 副研究员, 研究方向: 智能交通, 交通气象, 交通安全。硕士毕业于交通运输部公路科学研究所。Email: wanglin@itsc.cn

³ 刘亚清 (1979-), 男, 汉族, 博士, 副教授, 研究方向: 大数据, 信息资源规划。博士毕业于大连海事大学计算机应用技术专业。Email: liuyaqing@dlmu.edu.cn