数据自治开放应用平台设计与实践

陈德华，潘乔，王梅，乐嘉锦
东华大学计算机科学与技术学院，上海 201620

摘要
围绕数据自治开放的数据管理新模式，提出了一套面向数据自治开放应用的整体解决方案，讨论了该解决方案的整体框架、系统功能及接口。针对数据提供者和数据使用者，对数据自治的数据逻辑组织、数据安全描述接口、查询描述接口等应用系统构建的关键技术进行了介绍。最后以医疗领域三级诊疗应用系统为例，探索了数据自治开放应用实践。

关键词
大数据；数据自治开放；数据逻辑组织；需求接口

中图分类号：TP319 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018019

Platform design and practice for self-governing openness of data application

CHEN Dehua, PAN Qiao, WANG Mei, LE Jiajin
School of Computer Science, DongHua University, Shanghai 201620, China

Abstract
Asystemic solution for the application construction of self-governing openness of data based on the new data management pattern of self-governing openness of data was proposed. The whole framework, system functions and interfaces of the solution were provided. Then the key points about logical data organization, security requirements description and query requirements description technologies were designed and investigated for the data provider and data consumer, respectively. Finally, self-governing openness of data practices in the field of medical was discussed, and the framework of tiered diagnosis and treatment system application by using the proposed technologies was given.

Key words
big data, self-governing openness of data, logical data organization, requirement interface

1 引言

随着信息技术的飞速发展，全球数据资源正以前所未有的速度不断增长和积累，大数据以其具有的巨大潜在价值成为政府、学术界和企业界共同关注的焦点。数据资源的价值具有隐藏性，研究人员需进行深度分析和挖掘，方能发现其潜在的价值。然而目前这一过程受到数据的完整性以及数据的访问权限等条件的制约，使得数据的应用价值受到极大限制。很明显，只有开放的数据资源，才能释放数据的价值。数据开放是大数据发挥其作用的基础。

数据资源的战略性和商业价值正在日益显现，据麦肯锡公司预测，医疗数据的有效利用每年可为美国医疗健康体系带来3 000多亿美元的潜在价值，并贡献0.7%的年度生产力增长，可为加拿大医疗健康体系节省100亿美元的卫生费用[1]。因此，目前全球各国都将数据开放提升到国家战略层面。截至2014年4月，已经有63个国家制定了政府数据的开放计划，如美、英、法等8国签署的《开放数据宪章》①。国务院2015年印发《促进大数据发展行动纲要》，明确将政府数据的开放排在首位。然而数据开放带来的数据安全和隐私泄露的风险[2-4]使得数据资源的开放变得越来越困难。数据自治开放为数据开放提供了新思路和新的解决办法。数据管理方式的变化将带来数据应用技术的巨大变革。长久以来，针对面向联机事务处理（online transaction processing，OLTP）的关系数据库以及面向联机分析处理（online analytical processing，OLAP）的多维数据模型[5]产生了大量的研究成果，已建立成熟完备的应用解决方案。围绕数据开放自治新模式，尚缺乏完整的理论体系和应用解决方案。面向上层数据提供者和数据使用者，如何进行合理有效的数据组织，提供友好的上下层接口，建立一套面向数据自治开放应用的整体解决方案，提供数据自治开放需求描述方法、设计原则及实现部署，并将数据开放的新理论、新技术与新方法运用到数据自治开放实践中，探索面向特定领域的应用实践，是本文要解决的主要问题。

① https://www.gov.uk/government/publications/open-data-charter

2 数据自治开放应用平台框架

2.1 数据自治开放平台

数据自治开放平台实现了数据在自治环境下的完全开放、脱敏开放和原始开放。整个平台由平台基础层和平台应用层组成。其中，平台基础层从下到上又分为数据存储系统、数据资源管理系统、软件行为监控系统；平台应用层由数据自治开放应用系统和用户组成。平台的整体架构如**图1**所示。

平台基础层为整个数据自治开放平台提供数据存储与管理、数据使用的软件行为监控功能。其中，数据存储与管理功能分别由数据存储系统和数据资源管理系统实现，完成面向具体应用的数据安全存储和管理；数据使用行为的监控功能由软件行为监控系统实现，主要负责对数据使用者在访问及使用数据过程中的操作行为进行监控和验证。平台应用层在平台基础层上，为具体的数据自治开放应用系统提供数据设计模式和数据自治开放解决方案。

可见，在整个平台中，数据自治开放应用系统直接面向用户，为数据提供者和数据使用者提供数据开放服务，处于平台

的承上启下地位，直接影响平台的推广及应用。

2.2 数据自治开放应用平台功能

图2给出了数据自治开放应用平台的功能划分。按照用户的不同角色，数据自治开放应用平台的功能可分为面向数据提供者和面向数据使用者两大部分。

（1）面向数据提供者的功能

针对数据提供者对数据的安全需求，数据自治开放应用平台可参照数据库范式设计体系构建数据安全范式规则，建立数据逻辑模式，使之能够有效地存储和管理开放数据，具有防泄露、保护权益能力。

● 数据资源管理：面向上层应用，对不同业务的数据资源进行整合，设计不同类型数据的语义关联组织方法，建立易于不同领域数据使用者使用的开放数据资源。在此基础上，实现对各种数据资源的分类、查询、统计、汇总等，帮助数据提供者实现数据资源的统一管理。

● 数据逻辑模式：对开放数据的类型、内容、性质以及数据间的联系进行描述，定义数据上的操作类型和操作方式。从数据属性维度（横向）和数据规模维度（纵向）划分数据粒度，为用户建立合理的数据组织。

● 数据安全需求：包括安全需求接口以及对需求的理解。根据数据提供者的数据安全特性，结合数据提供者对数据的隐私保护需求，建立数据安全需求规则，实现安全规则的建立、分解、合并优化等，以实现数据资源安全的复杂应用需求。

（2）面向数据使用者的功能

针对数据使用者对数据开放的应用需求，数据自治开放应用平台在数据自治开放数据资源管理系统的基础上，基于数据驱动的软件开发思想，建立一套面向数据自治开放应用系统的整体解决方案。

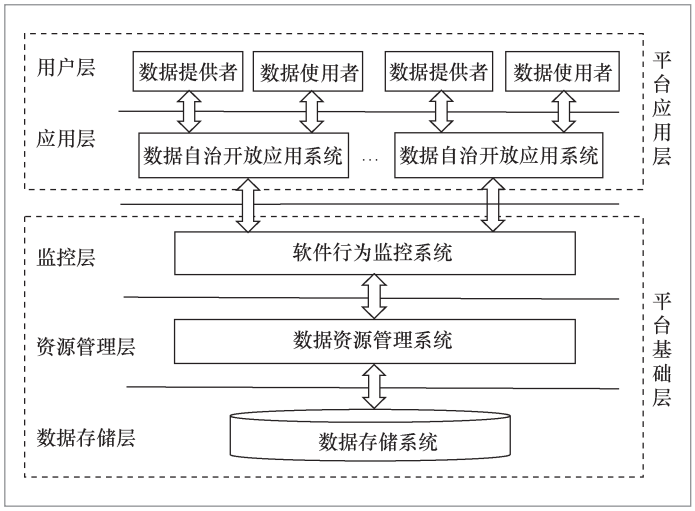

图1 数据自治开放平台层次架构

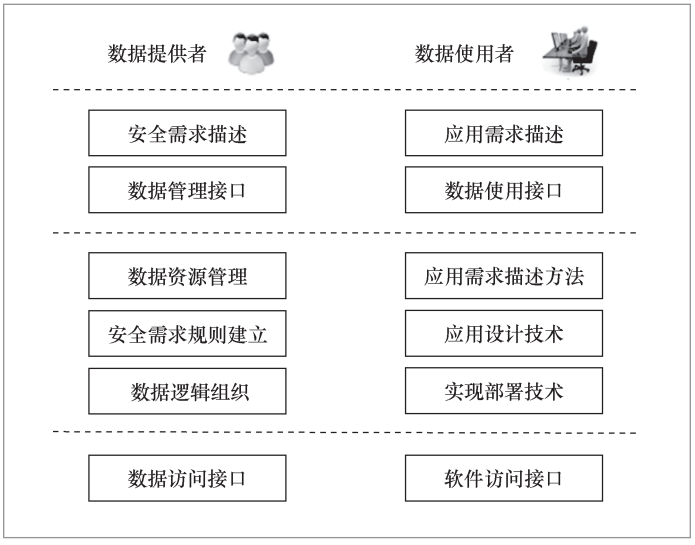

图2 数据自治开放应用平台的功能组成

● 数据自治开放应用需求描述方法。需求是整个数据自治开放应用系统最关键的输入，由业务需求、用户需求、功能需求、数据需求等组成。基于常规软件需求描述内容，针对数据自治开放应用需求存在模糊性、不确定性、变化性和主观性的特点，提出面向数据自治开放应用环境的N-元素描述法，体现数据自治开放应

用的业务流程、数据、功能、性能及安全约束等要素，从使用前景和范围文档来记录业务需求。

● 数据自治开放应用设计技术。结合数据自治开放应用需求，制定数据自治开放应用系统设计中需要遵循的原则、部署应用的规划。基于接口隔离原则、依赖反转原则等，给出数据自治开放应用的基本设计流程、组织结构、模块划分、功能分配、接口设计、运行设计和数据结构设计以及应用中系统功能模块之间层次结构与调用关系的设计原则。

● 数据自治开放应用的实现策略。基于数据自治开放应用分析及设计结果，给出数据自治开放应用的实现部署策略。

2.3 数据自治开放应用平台的对外接口描述

数据自治开放应用平台对外需要与数据资源管理系统和软件行为监控系统进行交互。根据图2所示的数据自治开放平台架构，其对外接口主要有以下两种类型。

● 软件访问接口：为数据自治开放应用系统提供了与软件行为监控系统之间的交互接口，即实现数据自治开放应用的各项业务功能的前端程序开发接口。数据自治开放系统的软件访问接口可采用OpenDL SDK的数据开放应用软件开发工具包进行设计与实现。

● 数据访问接口：通过OpenDL运行库与数据资源管理系统交互，支持Web及各种移动设备的访问，供外部软件使用数据，实现对物理存储的数据的管理，如新增、更新、冻结等。

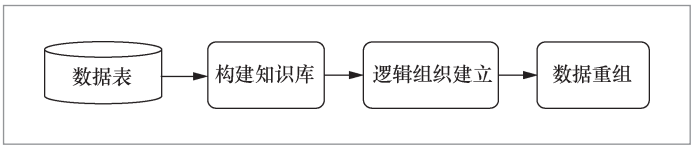

图 3 基于知识库的数据逻辑组织技术路线

3 关键技术实现

3.1 数据逻辑组织

在数据自治开放应用环境中，数据提供者与数据使用者往往并不相同。因此，首先需要将数据提供者提供的数据资源进行整合，设计面向使用者的数据逻辑组织。由于在传统应用中，数据的产生者和数据使用者为同一主体，因此其数据逻辑组织大都建立在应用需求已知的情况下。数据库设计由应用驱动，首先进行概念模型设计，进而进行逻辑表组织和数据表创建。而在面向大数据的数据自治开放应用中，数据的使用者和数据使用任务都存在较大的不可预知性和可变性，因此，采用传统的数据组织方式并不可行。

与传统的数据库关系模型对比，本体模型能够实现与现实世界的直接对应，同时保留具体数据之间的关联关系，能更准确地与自然语言相对应，从而呼应可变的上层应用需求[6]。为此，建立基于本体知识库的数据划分和逻辑组织。

如图3所示，基于本体知识库的数据逻辑组织技术，首先从原始多源关系型数据集中提取候选的本体概念以及概念的属性和层次结构，然后采用自顶向下的方法先对每个概念进行选择，判断概念是否符合本体模型以及概念的属性；然后为当前概念添加新的子概念和属性，并为属性添加值域约束和类型约束；最后对本体模型中每一个概念进行词义扩展，建立同义词库和反义词库。领域本体模型完成后，根据模型重新划分数据。首先去除本体概念中不存在的属性和表，然后根据本体属性的

约束提取满足条件的数据实例，最后打破原有的数据表之间的业务关联关系，根据本体概念之间的关系建立数据表之间新的联系，从而得到面向数据自治开放的新的数据逻辑组织。

3.2 安全需求描述与校验

安全需求面向数据提供者。数据提供者的数据资源开放首先需要保障数据的安全性，而如何描述和表达数据资源的安全性需求成为待解决的首要问题。由于安全性需求存在多样性的特点，以医疗数据为例，患者医生的个人信息属于个人隐私，直接存储在数据库中，需要针对具体数据制定安全规则，防止隐私泄露；医生诊断习惯、院内具有优势的诊疗方案等为数据分析的结果，涉及医生及医院的利益，需要针对用户的数据行为制定安全规则；而诸如医生手术成功率、医院经营状况等医院内部数据，只允许内部访问，需要根据用户的访问级别制定安全规则。针对不同行业，安全性需求既有共性，又有个性；既有显性需求，又有隐性需求。因此如何准确地表达、描述安全需求，成了迫切需要解决的问题。

安全性需求由数据提供者指定，若要求用户以系统指定形式制定规则，对用户而言要求太高。半自动化方式（通过界面引导用户添加对数据和行为的描述）相对而言对用户要求较低，但由于数据本身和数据的访问行为多样性，系统无法穷举用户的数据访问行为，导致局限性太大。一般用户对自然语言都能熟练掌握，若安全规则的生成对数据提供者的要求仅是以自然语言方式提供描述，无疑会大大降低用户的学习成本，从而提升系统的友好度。另外，安全规则的生成需要与具体的数据、用户的数据访问行为以及用户的访问级别相关联，如何设计一种结构能够在保存以上信息的同时，准确地实现从自然语言向该结构的转换，成为当前迫切需要解决的问题。本文提出的安全需求接口如下所示。

安全需求接口={input, output}。

安全需求接口的input：用户的自然语言查询输入。

安全需求接口的output：{安全规则名，规则涉及的对象，规则对应的操作，访问控制权限}。

规则名：规则的统一命名。

规则涉及的对象：安全需求中涉及的数据对象。

规则对应的操作：该安全需求要求该数据对象进行何种操作。

访问控制权限：该安全需求要求该数据对象进行上述操作的权限，如受限的或禁止的等。

如**图4**所示，该关键技术首先分析数据库逻辑结构，构造实体类、属性以及实体与实体及属性之间的关系，生成知识库。另外，接受用户安全规则的自然语言描述，并将该安全规则分解为数据访问行为描述以及该数据访问行为的访问控制处理。数据访问行为及其访问控制均采用Key-Value的二元组形式化方式描述，用户可以

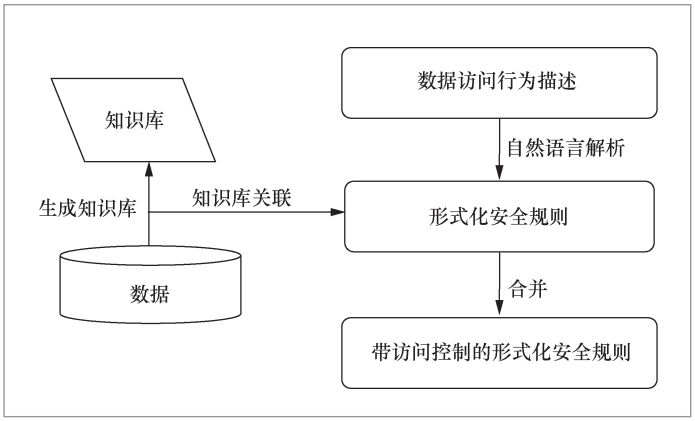

图 4 安全需求接口技术路线

通过浏览器查看，并可直接通过该图形化界面对其进行修改和完善，最终进行预保存。在获得用户的查询意图（语句）后，即可与安全规则进行匹配，对违反安全规则定义的数据访问需求进行屏蔽和提示。

3.3 分析需求描述与转换

分析需求面向数据使用者。数据自治开放平台在确保数据安全性的同时，需提供一个用户友好的访问接口，以便用户便捷地查询分析相关数据。在传统数据库管理系统中，用户通过标准的数据库查询语言或交互式的可视化查询界面描述所要进行的查询分析操作。然而，上述方式需要用户具有数据库查询语言基础，并且对数据库中表的结构较为熟悉，能自行确定连接条件等。这对于数据提供者和数据使用者并不相同的数据开放而言，较为困难[7,8]。除此以外，逻辑较为复杂的查询分析很难或者无法用这种可视化的查询界面表达出来。特别是对面向大数据的开放来说，对数据挖掘算法的有效支持是其必备的一个功能，但通常来说普通用户并不能精准地表述自己数据挖掘的意图，更不可能透彻理解数据挖掘的相关算法。自然语言是人类最为熟悉的交流工具，设计一个支持自然语言的查询分析接口，将极大地提高数据的易用性，这种方便用户的查询方式将把数据迅速推广普及至各个行业。

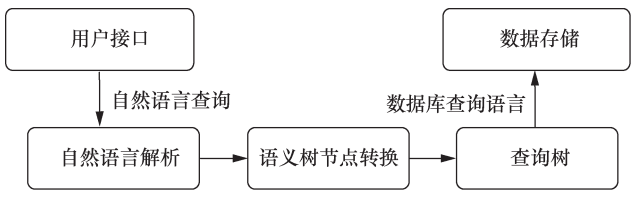

图 5 自然语言查询接口技术路线

本文提出的自然语言查询分析接口接收用户的自然语言查询需求，并将其转化为数据管理层支持的查询语言，这代表查询接口接收的语言是自然语言的一个子集，同样，数据库本身内容的确定性以及查询句型的有限性和不带感情色彩等特点，使得语义的歧义性、模糊性问题在查询接口中可得到有效处理。

自然语言查询接口={input，output}。

查询接口的input：用户的自然语言查询输入。

查询接口的output：对应的数据库查询语言。

如**图5**所示，自然语言查询以用户自然语言为输入，首先通过文本解析器将自然语言转化为语义树；进而根据语义相似度及映射规则将树节点归类，并与自治开放数据中的底层数据对象建立对应关系。在此基础上，根据设计的查询树结构以及节点类型调整语义树至基本查询树，选取Top$k$棵基本查询树与用户进行交互，从而返回一棵最正确的基本查询树，再将基本查询树进行模糊节点扩充，生成一个层次清晰、关系明确的查询树结构并返回。最后，通过查询转换将查询树转换为SQL或其他数据库可执行语句，传入下层接口执行。

4 典型项目应用

4.1 医疗数据自治开放应用

近年来，医院信息化建设的深入促进了医疗大数据的形成。医疗数据的开放可以改变医院信息系统中医疗数据的封闭现状，实现医生对患者的跟踪诊疗，提高医治过程的透明化，为医院的临床、科研及管理运营等提供重要的数据获取途径，有助

于提升医疗数据的再利用价值，提高医疗服务的质量和水平。

本文在自主研发的数据自治开放模型、应用技术与方法基础上，选取了上海市某三甲医院作为数据开放的应用示范，实现在自治开放环境下临床医疗数据的开放及深度开发利用。该医院作为全国知名大型综合性医院，一方面，接收来自全国各地的初诊和复诊患者，医生需要获取患者的前期诊疗数据；另一方面，为了方便患者的后续治疗，又需要把患者的治疗数据开放给其他医院。因此，医院面临数据开放的典型性问题，即如何最大限度地从第三方获得有效的数据，如何在保证数据安全性的前提下为第三方开放数据。

4.2 医院数据自治开放应用系统

针对医院的业务需求，医院的医疗数据自治开放系统体系框架如图6所示。

图6的系统体系框架包括数据提供方、数据使用方和分级诊疗数据自治开放应用系统三方。数据提供方主要为社区医疗服务中心、二级医院和三级甲等医院等，患者可能在这些机构做过治疗或化验；数据使用方主要为政府相关部门、第三方医疗研究机构或个人等，他们需要获取相关诊疗数据作为患者后续治疗、医学研究或政府决策等用途；分级诊疗数据自治开放应用系统部署在试点医院，它主要由数据整合、数据分析模型库、数据浏览和展示、数据申请和审核以及使用等多个功能模块组成。其中，数据整合主要完成来自社区医疗服务中心、二级医院和三级甲等医院多级来源数据的整合和数据划分。数据分析模型库根据预收集的数据分析需求，如疾病分期、分型预测、诊疗方案学习等，构建一系列分析模型，并在后期进行模型的累积和演化更新。数据逻辑组织展示和浏览主要从元数据、时间、来源等多维度对数据进行展示，供使用者进行筛选。数据提供方可根据用户的权限和其筛选数据的安全规则进行使用审核。最后，向符合安全规则的数据访问需求提供相应数据供用户使用。

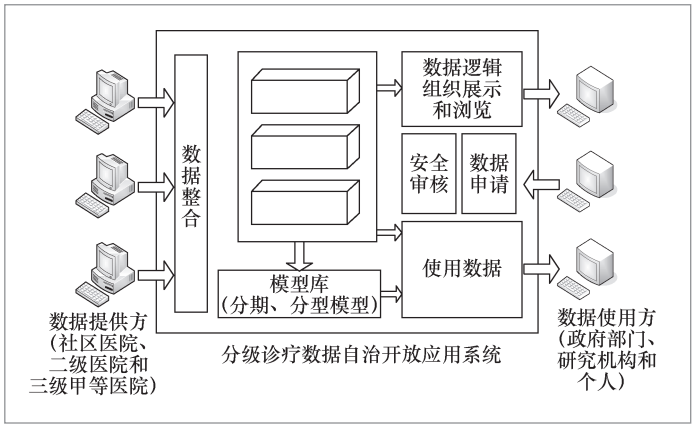

图 6 系统体系框架

在上述分级诊疗数据自治开放应用系统中，首先需要对医院不同信息化系统（如医院信息系统、影像归档和通信系统、电子病历系统、实验室信息系统等）进行数据整合，抽取患者信息、医生信息、医院信息、诊断信息、检查检验信息、治疗信息的结构化与非结构化数据，建立全方位数据资源体系。

以国内某三甲医院为例，经过相同患者识别、信息整合，最终的表有400多张，再将与临床无关的表去掉，还有60多张数据表。不论是数据提供者还是数据使用者，都并非数据库专业用户，直接在原始表结构上进行安全规则定义和查询需求定义均十分困难，因此建立基于本体的数据逻辑组织十分必要。在图6的系统体系框架中，数据逻辑展示和浏览模块采用本文提出的基于本体的数据逻辑组织将极大地方便数据提供者和数据使用者，是应用系统的有力支撑。同时，在建立本体的过程中，对字段进行合并重组，将大大减少数据量。以甲状腺数据集为例，重组后的数据

总量减小到30%，在重组后的数据集上进行查询执行，将显著提升查询执行性能[9]。进一步，对于数据使用者（以医生为例）来说，在构造数据分析模型时，将不可避免地从开放数据中获取数据。考虑简单查询意图“查询低钾的甲亢用户”，该语句的具体执行涉及数据库中的三表连接，且存在两种连接方式。若让用户直接书写结构化查询语句或书写基于本体的查询语句，则十分困难。本文提出的基于自然语言的分析需求与转换关键技术将帮助数据使用者输入自然语言语句提出查询请求，即可方便地获取所需数据构建模型，数据提供者定义安全需求的情况与此类似。以“普通用户不能查询医生的用药习惯”这一安全规则为例，很明显，该规则定义待保护的隐私数据为“用药习惯”，即医生根据诊断结果经常开的药。将该自然语言描述的待保护数据转换为描述数据访问的安全规则，并进一步对规则进行分解与合并，通过对数据使用者的数据访问行为进行审核判断，即可防止隐私数据的访问和泄露。可见，应用本文提出的关键技术可极大地提升数据自治开放平台的性能和可用性。

5 结束语

本文提出了针对数据自治开放上层应用的系统解决方案。在现有的大数据相关研究中，大多数工作集中于大数据的存储管理技术，然而针对大数据提供者和大数据使用者，如何设计方便、有效的查询分析需求接口、完全需求接口，如何提供清晰、有效的数据逻辑组织，以方便管理和使用数据，相关的研究还较少。本文对上述关键问题和关键技术展开研究，并提供了相应的解决方案。最后，将本文提出的应用平台与技术，在医疗领域的数据自治开放示范工程中进行验证，验证了本文方法的有效性。

参考文献:

[1] MANYIKA B J, CHUI M. Big data: the next frontier for innovation, competition, and productivity[R]. McKinsey Global Institute, 2015.

[2] 李姣, 郭海红, 郭珉江, 等. 美英政府开放健康医疗数据的主题分布与开放程度量化研究[J]. 图书情报工作, 2015, 59(20): 132-137.
LI J, GUO H H, GUO M J, et al. A quantitative analysis on the topic distribution and openness degree of health and medical data in the open government practice of the US and UK[J]. Library and Information Service, 2015, 59(20): 132-137.

[3] YANG K, JIA X H, REN K, et al. DAC-MACS: effective data access control for multi-authority cloud storage systems[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(11): 1790-1801.

[4] YANG K, JIA XH, REN K. Secure and verifiable policy update outsourcing for big data access control in the cloud[J]. IEEE Transactions on Parallel and Distributed Systems, 2015, 26(12): 3461-3470.

[5] COLLIAT G. OLAP, relational, and multidimensional database systems[J]. ACM SIGMOD Record, 1996, 25(3): 64-69.

[6] DEHAINSALA H, PIERRA G, BELLATRECHE L. OntoDB: an ontology-based database for data intensive applications[C]// Advances in Databases: Concepts, Systems and Applications, International Conference on Database Systems for Advanced Applications, April 9-12, 2007, Bangkok, Thailand. Heidelberg: Springer, 2007: 497-508.

[7] SAHA D, FLORATOU A, SANKARANARAYANAN K, et al.

ATHENA: an ontology-driven system for natural language querying over relational data stores[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 1209-1220.

[8] LI F, JAGADISH H V. Constructing an interactive natural language interface for relational databases[J]. Proceedings of the VLDB Endowment, 2014, 8(1): 73-84.

[9] HAO Q, LI Y, WANG L M. An ontology-based data organization method[C]// International Conference on Advanced Cloud & Big Data, August 13-16, 2017, Shanghai, China. Piscataway: IEEE Press, 2017: 135-140.

作者简介

陈德华(1976-),男,东华大学计算机科学与技术学院副教授,主要研究方向为数据库与数据仓库、大数据与深度学习。

潘乔(1977-),男,东华大学计算机科学与技术学院副教授,主要研究方向为数据库与数据仓库、大数据与深度学习。

王梅(1980-),女,东华大学计算机科学与技术学院教授,主要研究方向为数据库、多媒体。

乐嘉锦(1951-),男,东华大学计算机科学与技术学院教授、博士生导师,主要研究方向为数据库与智能信息处理、大数据与智慧医疗。

收稿日期: 2018-01-26

基金项目: 上海市科技发展基金资助项目(No. 16JC1400802)

Foundation Item: Shanghai Science and Technology Development Fund (No. 16JC1400802)