文章编号: 1006-2475(2019)11-0018-05

基于数据虚拟化技术的多来源数据集成方法

张子晔[1],刘玉龙[1],呼 北[2]

(1. 中国电子科技集团公司第十五研究所系统八部,北京 100083; 2. 中国电子科技集团公司第十五研究所系统一部,北京 100083)

摘要: 司法业务数据存储没有统一的格式标准,各机关在进行数据查询访问时存在数据孤岛现象。为解决数据访问之间的异构性,本文提出一种基于数据虚拟化的多来源司法数据集成方法,通过数据虚拟化技术建立元数据映射关系,利用中间件构成数据交换中心,实现多机关多类型司法数据集成。利用改进的 K-means 聚类算法对虚拟对象元数据进行聚簇,缩短数据访问时间,提高司法数据查询效率。本文方法可以忽略数据存储异构性的影响,实现各司法机关无障碍数据访问通道。

关键词: 数据虚拟化; 中间件; 元数据; 改进的 K-means 聚类算法

中图分类号: TP391 文献标识码: A doi: 10.3969/j.issn.1006-2475.2019.11.004

Multi-source Data Integration Method Based on Data Virtualization Technology

ZHANG Zi-ye[1], LIU Yu-long[1], HU Bei[2]

(1. The Eigth System Department of the 15th Research Institute of China Electronics Technology Group Corporation, Beijing 100083, China; 2. The First System Department of the 15th Research Institute of China Electronics Technology Group Corporation, Beijing 100083, China)

Abstract: There is no uniform format standard for judicial business data storage, and there exists data islands in each organization's data query and access. In order to solve the heterogeneity between data access, this paper proposes a multi-source judicial data integration method based on data virtualization, which establishes metadata mapping relationship through data virtualization technology, and uses middleware to form a data exchange center to realize multi-organ and multi-type judicial data integration. The improved K-means clustering algorithm is used to cluster virtual object metadata, shorten data access time and improve judicial data query efficiency. The proposed method can ignore the influence of data storage heterogeneity and realize accessible data access channels of various judicial organs.

Key words: data virtualization; middleware; metadata; improved K-means clustering algorithm

0 引 言

随着国家网络强国战略、“互联网+”行动计划、大数据战略、新一代人工智能战略等重大决策部署,将互联网技术引入到各司法机关成为一项重要的司法建设内容。

但各司法机关独立存在,其数据没有统一的存储标准,存在数据存储不规范的情况。由于其数据的独立性,造成数据无法在各司法机关之间进行流转,产生了数据孤岛现象[1]。

传统的数据查询方法只能对固定格式的数据库进行访问,在多源异构的数据集成中存在问题。针对此种现象,国外为解决跨领域之间信息共享问题而制定的信息交换框架,实现了国家信息交换模型体系结构组成和基于 NIEM 的信息交换实施过程[2]。

但在此方面,中国起步较晚。早期的数据集成方法主要是数据仓库和中间件集成,但单纯的 2 种集成方式,需要通过 ETL 从物理空间集成数据并进行存储,占用了大量的物理空间。而新兴技术多为结合云平台实现多源数据共享,例如湖南依托“互联网+”技术构建部门“点对点”的信息共享查控执行联通机制,实现了对案件各流程、各节点的监督和有效控制[2]。广东基于云计算技术从基础设施层、平台服务层、应用服务层实现了司法信息资源共享平台的建

收稿日期: 2019-03-30; 修回日期: 2019-04-14

基金项目: 国家重点研发计划资助项目(2018YFC0831202)

作者简介: 张子晔(1995-),女,天津人,硕士研究生,研究方向: 计算机软件开发及数据集成,E-mail: 18810934029@163.com; 刘玉龙(1981-),男,研究员级高级工程师,硕士,研究方向: 大型信息系统架构设计和项目管理,E-mail: lyl_nci@126.com; 呼北(1994-),男,硕士研究生,研究方向: 计算机应用技术,E-mail: hubei004@sohu.com。

设方案[4]。但涉及司法数据，云平台存在一定的安全性问题并需要较强的技术支持，对多源异构司法数据集成带来不便[5]。

为此，本文提出一种基于数据虚拟化的数据集成方法，利用虚拟数据空间对多来源数据进行流转，满足司法数据集成的要求。

1　数据集成技术

数据集成技术将多来源数据库数据整合在一起，实现对不同来源数据库数据的无障碍查询与修改。通过数据集成，用户不仅可以在不了解各机关数据存储区别的前提下，实现多源异构数据的流转，还能极大地提高获取数据的效率。主要的数据集成技术有以下 3 种：

1）联邦式数据库。

联邦式数据库是最早的数据集成方法，通过共享同一种数据模型，建立不同数据库之间的访问接口，形成统一数据整体，联邦式数据库分为紧耦合联邦数据库和松耦合联邦数据库 2 种模式。

2）数据仓库模式。

数据仓库是一种面向主体的数据集成方法，对于多来源不同结构的数据，需要进行一定的数据清洗和数据加工，形成相同的数据模式，才可加入到数据仓库中，形成大规模的数据整体。但数据仓库只提供数据的查询功能，并不能对其中的数据进行修改，有着一定的局限性[6]。数据库和数据仓库的对比如表 1 所示。

表 1　数据库和数据仓库对比表

区别	数据库	数据仓库
数据状态	当前实时数据	历史性、完整性、跨时间变化的数据
功能	日常基本操作	长期数据检索、决策规划
视图	单一视图	多维多样视图
规模	GB 到 TB	大于等于 TB
数据变化	支持增、删、改、查操作	可添加
处理量	事务吞吐量、小批次、高并发	查询吞吐量、大批量、高吞吐

3）中间件模式。

中间件位于数据源系统和应用程序之间，起到承上启下的作用，向下接收来自不同数据源的数据。中间件技术通过对多来源数据抽取、格式转换、元数据管理建模等方式向上为访问集成数据的应用提供统一数据模式和数据访问的通用接口[7]。

2　数据虚拟化技术

通过以上数据集成技术，逐步发展出以 Hadoop 为代表的分布式存储计算架构、以 HDFS（Hadoop Distributed File System）为代表的分布式文件系统、以 BigTable、DynamoDB 为代表的多种存储模型、大规模并行处理（Massively Parallel Processing，MPP）数据库等大数据处理技术[8]。但传统的数据集成技术占用大量的物理空间，存在成本高、效率低、占用空间大等问题。针对这些问题，需要构造一个虚拟空间，为此提出了数据虚拟化技术。

数据虚拟化技术可以通过对多源异构数据的逻辑虚拟化，构造一个虚拟的数据空间，使用者通过统一的访问方式，获取到多来源的信息数据，实现多来源数据的统一集成[9-10]。通过数据虚拟化，使用者不需要学习多种不同数据库的数据处理方法，大大方便了不同司法机关之间的数据流通，提高了司法业务的处理效率[11]。

3　改进的司法数据集成方法

由于传统的数据集成方法占用大量的物理空间与设备[12]，当需要访问多来源、多类型数据时造成了极大的不便[13]，为此本文在基于数据虚拟化技术的基础上，对不同机关案件信息数据进行集成，利用中间件实现临时数据存储，设计一种新型司法数据集成方法。根据各层具有的不同功能，将该架构分为 3 层，如图 1 所示，分别为数据源层、数据集成层和应用层。

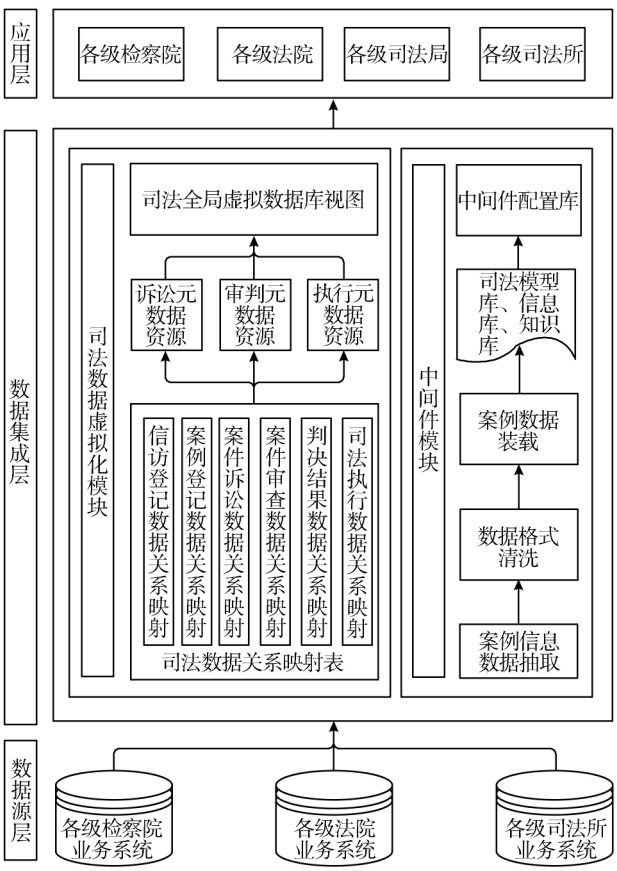

图 1　司法数据虚拟化集成架构

对于用户，各机关单位查询所需司法数据流程如图 2 所示。具体步骤如下:

1) 应用层需要发送数据查询请求。

2) 进入数据集成层，根据司法全局虚拟化视图按映射关系查找对应元数据，其中包括案件的登记、诉讼、审判、执行等司法信息。

3) 依据数据的使用频率决定是否利用中间件模块对指定司法数据进行查询存储。

4) 根据元数据信息在数据集成层查找到从数据源中整合的司法数据信息，最后返回查询结果。

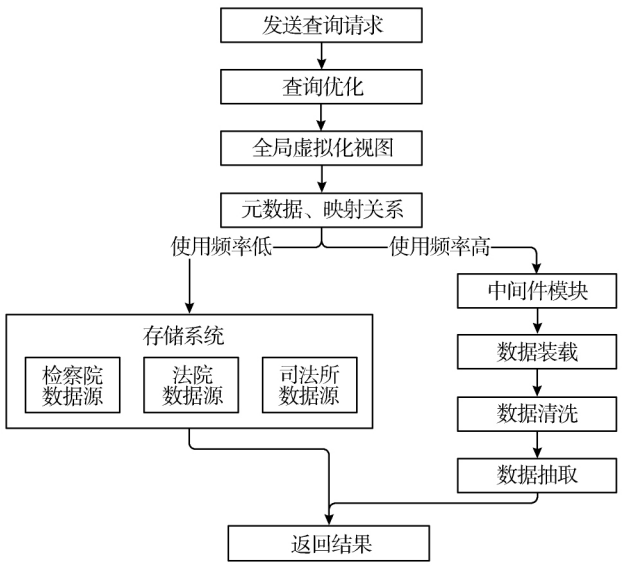

图 2 司法数据查询流程

3.1 数据源层

数据源层是司法数据的各种来源，是不同数据类型的数据系统集合，它是所有数据的来源，包括各级检察院业务系统、各级法院业务系统和各级司法所业务系统的数据。由于各司法机关的数据库系统没有统一的标准，所以数据的类型多种多样，其中包含主流的数据库系统，例如 My SQL、SQL Server、Oracle 等多种数据库。司法数据的内容也各有不同，例如司法机关所包含的司法鉴定人员、司法鉴定机构、法律援助机构、法律援助机构人员等各种数据内容。

3.2 数据集成层

数据集成层是数据架构的重要部分，用户无需了解数据结构，可以按照需求访问数据，它分为司法数据虚拟化模块和中间件模块 2 个部分。

3.2.1 司法数据虚拟化模块

数据虚拟化模块通过对数据资源的逻辑虚拟化，实现数据的集成管理并提供统一的访问接口，以便为各种数据消费需求提供跨数据源整合的数据服务。此模块并不直接存储数据的物理信息，而是首先建立多种数据格式的映射规则，再通过元数据信息对多机关数据进行管理。通过司法数据关系映射表建立信访登记数据、案件登记数据、案件诉讼数据等关系映射，再提取诉讼、审判和执行的元数据资源，用于建立司法全局虚拟数据库视图。所以，在数据虚拟化模块中，元数据建模以及搜索效率极大地影响了此种架构的使用效果。

3.2.2 中间件模块

由于数据虚拟化中数据存取时间较长，当数据的使用频率较高时，临时开辟空间用于中间件模块。其对多源异构数据经过案例信息数据抽取、格式清洗、案例数据装载，并匹配相应司法模型库、信息库和知识库，最终经过一系列处理，建立其间的映射关系，最后将其存放到中间件配置库中。定时对中间件模块所使用空间进行更新释放，使其数据不断更新流转。中间件模块处理流程如图 3 所示。

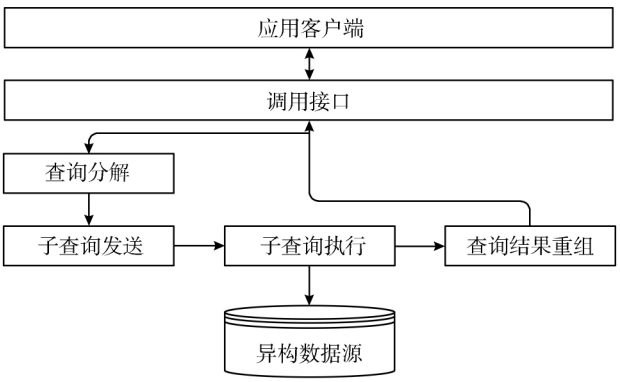

图 3 中间件模块

例如，在集成司法审判信息时，其后台数据库存储结构如图 4 所示，分为案件表、当事人表、审委会表、庭审记录表等，对于不同数据表，一次性设置属性。

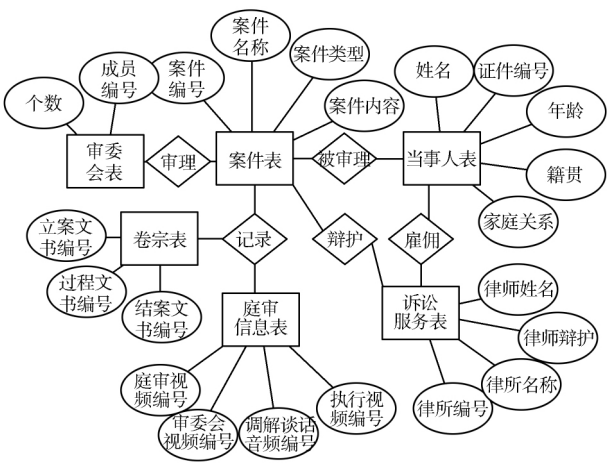

图 4 司法审判数据库 ER 模型

3.3 应用层

应用层是面向各司法机关，如各级检察院、法院、

司法局和司法所，提供不同业务的数据访问功能。用户通过应用层将数据请求传达到数据集成层，在数据集成层获取到所需数据时，再返回应用层以供用户使用。在应用层上，可以开发不同的数据访问接口和访问工具，给用户提供更加方便的数据查询访问方法。

4　虚拟对象元数据分簇算法

在数据集成架构中，数据集成层是数据汇集的中心，各司法机关为获取有效数据信息，需要依靠数据查询方法。由图 2 司法数据查询流程可看出，对于数据虚拟化系统，元数据信息是获取数据的关键。各种不同的物理数据源分布在各个司法机关的数据库中，数据虚拟化系统中并不存在真的物理数据，只是保存了物理数据对应的元数据信息[10]。元数据集成替代数据集成能够避免大量数据的移动和存储，有效降低数据集成的成本[14]。通过对元数据的相关查询，可以获取到用户需要的数据信息，所以元数据对数据虚拟化系统的查询效率有着重要的影响[15]，为此本文采取元数据分簇算法进一步优化数据虚拟化系统的查询效率[16]，提高数据集成方法的可行性。

4.1　问题分析

在数据集成框架中，当查询某一案件审判执行相关信息时，首先根据被诉讼人的身份信息从检察院获取案件诉讼批号，再依次根据批号在法院查找对应审判信息，并获取司法部或所在司法局对此案件的司法执行信息。司法数据查询过程如图 5 所示。

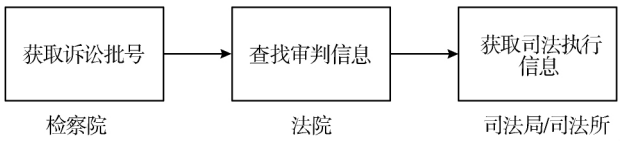

图 5　司法数据查询过程

当数据庞大时，这种数据查询方法构成一种树形结构，需要根据查询指令要求分层进行查询。在判断数据集成方法是否合理时，查询数据的效率高低可以作为一项衡量内容[17]。当数据量或者数据存储结构复杂多样时，树形结构会变得十分庞大，会大大增加数据的查询时间。为此需要设计一种扁平化的元数据组织结构，缩短数据查询的路径，从而提高搜索效率[18]。

4.2　解决方法

本文采用的是一种改进的 K-means 算法，由于 JSON 数据格式便于在各种数据平台或数据库中读写，所以利用 JSON 数据格式封装数据源信息[19]，然后利用聚类算法将 JSON 文档分成不同簇，实现元数据分簇。这样在查询司法数据时，首先查找到第一层聚簇信息，再在其中进一步查找，缩短了数据查询路径。例如，当查询被诉讼人 w 的司法信息时，通过对 w 的 id 元数据信息聚类，可以直接查询到 w 各机关相关司法元数据信息，无需对每个司法机关进行查询，一定程度上提高了数据虚拟化查询效率。

4.3　改进的 K-means 元数据聚类算法

1) 规范 JSON 文档格式。

首先对元数据设定 JSON 文档格式，每个文档分为 5 个部分，id 代表元数据类型，source 代表元数据来源机关，position 代表司法数据的来源数据位置，reinformation 代表元数据相关信息，所有参数的大小为 1 ~ 10 的任意整数。例如一条司法所信息，id 为 JG_SFJG，source 为司法所，position 为人民调解，reinformation 为司法所相关信息，其他各个 JSON 文档数据多对应如上定义模式，为后面元数据聚类提供标准数据格式。

2) 确定最初 k 值和各中心点开始位置。

对于大规模数据查询，k 值的选取比较重要，如何选取 k 值，对噪声和离群值会产生较大的影响[20]，为此定义轮廓系数(Silhouette Coefficient)来确定初始 k 值：

$$s_i = \frac{b_i - a_i}{\max(a_i b_i)}$$

假设将所有数据集分为 k 类，a_i 是当前所选数据到其所在簇其他数据的平均距离，b_i 是当前所选数据到其他各簇中数据的平均距离的最小值[21]。当轮廓系数 $s_i = +1$ 时，说明当前对象的分簇正确，其与其他簇数据具有较大区别；当 $s_i = -1$ 时，说明分簇效果不好，当前数据与其本簇数据具有较大区别。

聚类中心个数的选择影响了整体的聚类效果，分别计算不同 k 值时 s_i 的情况，选取合适的初始 k 值。

3) 计算剩余节点距 k 中心点的位置，将结果赋值给距离最小的簇[22]，对于每个样本 x_i，将其标记为距离类别中心 a_i 的最近类别 j。剩余节点距 k 中心点的距离计算公式如下：

$$Dist_j = \arg\min_{i \le j \le k}\{\sqrt{\sum_{i=1}^{n}(x_i - a_j)^2}\}$$

在距离计算中，由 source 得出当前数据所属数据类别，计算其在整体 XML 文档中距中心点的位置。

4) 再次计算簇的平均值，设定新的中心点。

5) 重复步骤 3 和步骤 4，计算最小化平方误差 E，直至准则函数收敛。最小化平方误差 E 的计算公式如下：

$$E = \sum_{i=1}^{k}\sum_{x \in C_i} || x - \mu_i ||_2^2$$

4.4 实验结果

图 6 是选取 $k=6$ 时产生的聚类效果图，图(a) 中的圆点表示选定的聚类中心点，图(b) 是生成的聚类效果。

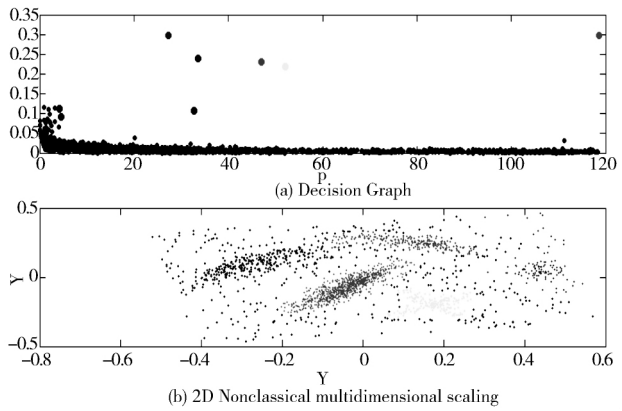

图 6 $k=6$ 时的聚类结果

利用对元数据的聚类算法，屏蔽了数据来源的异构性，对比加入聚类方法后对司法数据集成方法的性能影响，表 2 显示了不同 k 值下添加聚类算法前后的司法数据查询时间。

表 2 司法数据查询时间

k 值	查询时间/s	
	不聚类	聚类
$k=6$	0.5373	0.00753
$k=10$	0.7495	0.00926
$k=17$	1.6356	0.03739

根据表 2 的时间对比，在数据虚拟化层对元数据加入聚类算法，缩短了数据的查询时间。基于数据虚拟化技术的集成方法在不占用物理空间的基础上，利用元数据达到了高效的数据访问，提高了整体司法数据集成的性能和可靠性。

5 结束语

本文针对司法数据的处理与利用过程中存在数据孤岛的问题，提出了一种基于数据虚拟化的多来源数据整合架构，并利用元数据分簇算法提高数据查询的效率，对多源数据的整合具有基础性重要意义[23]。基于智慧司法工程的推进和司法办案需求，司法数据需要在司法机关之间流转共享。数据本身的属性是静态的，如果没有良好的数据共享机制的存在，即便是拥有海量的数据也很难产生相应的数据价值，多源数据集成实现了各司法机关对数据的有效查询工作，带动了整个智慧司法工作。

参考文献:

[1] 杨志农. 人民法院数据共享交换系统建设探索[J]. 网络安全技术与应用，2017(12):150-151.
[2] 戴剑伟，冯勤群，王刚. 美国国家信息交换模型原理分析[J]. 电子政务，2014(8):100-109.
[3] 鲁玉峰，李丹，王硕. 基于信息资源管理的制造企业数据集成规划的研究[J]. 智能制造，2016(10):26-30.
[4] 江国华，何盼盼. 数据共享与中国司法现代化[J]. 中国高校社会科学，2017(1):80-88.
[5] 马广惠，安小米，宋懿. 业务驱动的政府大数据平台数据治理[J]. 情报资料工作，2018，39(1):21-27.
[6] 王于丁，杨家海，徐聪，等. 云计算访问控制技术研究综述[J]. 软件学报，2015，26(5):1129-1150.
[7] 石峻峰，樊泽恒，武莉莉，等. 高校大数据集成管理研究[J]. 图书馆学研究，2014(21):47-50.
[8] 赵国锋，葛丹凤. 数据虚拟化研究综述[J]. 重庆邮电大学学报(自然科学版)，2016，28(4):494-502.
[9] 罗伟雄，时东晓，刘岚，等. 数据虚拟化平台的设计与实现[J]. 计算机应用，2017(A2):225-228.
[10] 葛丹凤. 多源异构的网络感知信息的元数据组织方法研究[D]. 重庆: 重庆邮电大学，2017.
[11] 王艳. 数字海洋多源异构数据管理优化研究[D]. 上海: 东华大学，2014.
[12] 丁祥郭. 消防信息系统数据集成的研究与应用[J]. 电信快报(网络与通信)，2018，560(2):24-28.
[13] 李杰玲，雷军程. 基于 XML 消除高校信息孤岛[J]. 怀化学院学报，2007(11):72-74.
[14] 冯勇，王明玉. 基于语义的轻量级数据集成方法[J]. 计算机工程与设计，2012，33(1):402-406.
[15] 耿玉水. 面向集团企业的数据集成模型构建方法研究[D]. 天津: 天津大学，2013.
[16] 王宗杰，侯贵法，王成耀，等. 基于元数据的分布异构数据集成研究[J]. 微计算机信息，2007，23(27):211-213.
[17] BACHTARZI C，BACHTARZI F，BENCHIKHA F. A model-driven approach for materialized views definition over heterogeneous databases[C]// 2015 1st International Conference on New Technologies of Information and Communication(NTIC). 2015:1-5.
[18] WANG J H，YU J X. Revisiting answering tree pattern queries using views[J]. ACM Transactions on Database Systems，2012，37(3):1-34.
[19] MUNSHI A A，MOHAMED A R I. Photovoltaic power pattern clustering based on conventional and swarm clustering methods[J]. Solar Energy，2016，124:39-56.
[20] 罗伟雄，刘岚，时东晓，等. 基于数据虚拟化技术的大数据资源中心建设[J]. 软件，2017(7):27-31.
[21] 丁遒劲，马袁燕，李勃慧. 多来源元数据集成中的组织管理框架研究[J]. 数字图书馆论坛，2017(12):60-64.
[22] 任建勋，杨栓. 数据虚拟化平台优化方案构想[J]. 河南水利与南水北调，2017，45(7):93-94.
[23] 缪谨励，陶留锋，谢飞，等. 基于虚拟数据库技术建立国土规划数据集成模型研究[J]. 地理信息世界，2016，23(4):31-36.