计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20131103
51(2):290-301,2014

一种新的智慧城市数据共享和融合框架——SCLDF

陈真勇[1] 徐州川[2] 李清广[2] 吕卫锋[1] 熊 璋[2]
[1](软件开发环境国家重点实验室(北京航空航天大学) 北京 100191)
[2](先进计算机应用技术教育部工程研究中心(北京航空航天大学) 北京 100191)
(chzhyong@buaa.edu.cn)

A Novel Framework of Data Sharing and Fusion in Smart City—SCLDF

Chen Zhenyong[1], Xu Zhouchuan[2], Li Qingguang[2], Lü Weifeng[1], and Xiong Zhang[2]
[1](*State Key Laboratory of Software Development Environment* (*Beihang University*), *Beijing* 100191)
[2](*Engineering Research Center for Advanced Computer Application Technology* (*Beihang University*), *Ministry of Education*, *Beijing* 100191)

Abstract Smart city is a new concept and model of urban development, and it is the combination of urbanization development and the new generation of information technologies such as Internet of Things, cloud computing, mobile network and big data. With the explosive growth of the quantity of data in cities, how to share and fuse the massive, heterogeneous, multi-source data in smart city becomes a core issue which must be solved. In this paper, the characteristics and drawbacks of traditional data sharing and fusion technologies are firstly described and analyzed in details, and then some frameworks and thoughts such as semantic Web, data vitalization and Internet of data, which may resolve the data sharing and fusion problems in smart city, are introduced. Based on these studies, a new framework of data sharing and fusion in smart city—smart city linked data framework is proposed. Then the overall layered structure, advantages against the other relevant frameworks and technologies are described briefly. Next, the functions, technologies and challenges of each layer are described in details. The concept of data semantic annotation tag (DSAT) are proposed. And the technologies, methods and classifications of DSAT are described in details. At last, the relevant issues about data linked layer are analyzed in details.

Key words smart city; data fusion; semantic Web; linked data; Internet of data; data vitalization

摘 要 智慧城市是物联网、云计算、移动网络、大数据等为代表的信息技术与城市化发展相结合的产物.如何有效地实现对智慧城市中海量、异构、多源数据的数据共享和融合是智慧城市必须要解决的核心问题.首先分析了传统数据融合技术的特点,然后阐述了当前可以用来解决智慧城市大数据共享和融合问题的技术.在此基础上提出了一种新的智慧城市数据共享和融合框架——智慧城市数据互联框架(smart city linked data framework),并详细阐述了其总体架构以及架构中每层的作用、关键技术和需要解决的问题,最后详细阐述了数据语义标注标签和数据互联层相关问题.

关键词 智慧城市;数据融合;语义网;关联数据;数据互联网络;数据活化

中图法分类号 TP311

收稿日期:2013-07-29;**修回日期**:2013-10-30
基金项目:国家"八六三"高技术研究发展计划基金项目(2013AA01A601);国家自然科学基金项目(61170178)
通信作者:徐州川(xuzhouchuan@gmail.com)

随着物联网、云计算、移动网络、大数据等相关技术的发展,智慧城市中数据容量和类型的急剧增长,如何有效地管理、分析和整合这些大数据,从数据中提取出有用的信息并将信息转化为价值,成为众多互联网企业和学术界的研究重点和热点.

2012 年 IT 产业界和学术界对"大数据"的关注度不断提升,存储和处理大数据的技术得到空前的发展.研究者们对于如何存储、管理、分析和理解大数据提出了许多技术,大数据相关技术得到了极大的发展.Google 提出的大数据处理和存储技术 Map Reduce[1] 和 GFS(google file system)[2]、备受业界推崇的 Apache 旗下顶级项目 Hadoop、迅速崛起的 NoSQL 数据库技术以及 Hinton 教授提出的深度学习[3] 都成为当前的研究热点.

然而如何将多源分散的数据有机整合起来,有效地实现不同数据源的数据共享和融合还没有得到真正的解决.在智慧城市中数据的来源非常分散,如环境传感器数据、移动网络数据、互联网数据和各种信息系统数据,如何将这些分散的数据互联起来,实现数据共享和融合,提高数据利用率,是智慧城市中亟待解决的核心问题.

数据融合(data fusion)[4] 技术主要是指整合表示同一个现实世界对象的多个数据源和知识描述,形成统一的、准确的、有用的描述的过程,最早应用于军事领域中的遥感数据.传统的数据融合方法主要包括数据仓库、中间件和联邦数据库,这些技术主要用于解决企业多个异构数据集数据的共享和融合问题,建立在规模较小又不太分散的系统上.传统的数据共享技术主要包括语义标注和 Web API 技术.语义标注技术的标准主要包括 Microformat,RDFa,Microdata 等.然而语义标注技术具有使用范围较窄、描述能力有限的缺点.Web API 技术是当前数据共享应用采用最多的形式,缺点是开放接口不一致,返回的数据没有并联性,因而不能实现数据之间的互联.智慧城市中的数据具有海量、异构、多源的特点,因此解决智慧城市大数据的共享和融合问题需要提出新的数据共享和融合技术.

由 Berners-Lee 等人于 2001 年提出的语义网(semantic Web)[5] 概念是对 Web 3.0 的一种设想.互联数据(Linked Data)[6] 是语义网中的数据描述框架的实现,它是一种通过发布结构化数据使数据互联进而提高数据应用价值的框架.Linked Data 适用于分散、孤立、异构、海量的互联网数据,因此对智慧城市大数据的共享和融合具有指导意义.Xiong 等人[7] 提出了一种大规模数据分析范式——数据活化(data vitalization),其中智能活化单元思想对解决智慧城市数据共享和融合问题有指导和借鉴作用.Fan 等人[8] 提出了一种扩展物联网的思想——IOD(Internet of data),IOD 将数据类比为物联网中的实体,利用数据标签进行数据关联,是实现数据共享和融合的一种新的思路,对指导智慧城市数据共享和融合有积极的作用.

本文提出一种新的智慧城市数据共享和融合框架——智慧城市数据互联框架(smart city linked data framework, SCLDF).SCLDF 框架结合 Linked Data 技术、数据活化思想、IOD 思想,以及多媒体标注和描述、文档标注等多种技术,实现智慧城市数据共享和融合,同时还具有数据描述、访问寻址、语义标注、数据分类等多种功能.SCLDF 通过数据的关联关系实现数据共享和融合,不仅可以提高大数据的处理性能,加深数据语义处理的深度,还扩展了数据处理和分析的广度,使数据的分析和处理不再局限在少量孤立数据集内,而将大数据有机整合起来,极大地提高了数据的利用率.

1 相关技术

本节简要阐述 SCLDF 相关技术和思想,分析这些技术和思想的主要特点和优缺点.

1.1 Linked Data

文献[5]提出的语义网思想是对未来互联网的一种设想,通过在互联网上提供统一数据描述实现数据网络(Web of data).W3C 对语义网的描述是可以让人们在网络上创建和存储数据、构建词汇库、为处理数据设定规则,从而形成数据网络,其最终目标是使计算机利用数据网络产生更多有效的、支持可信交互的系统,提供更加有用的网络应用[9].

Linked Data[6] 则是语义网思想的一个轻量级实现,它鼓励人们将已有数据发布成 Linked Data 推荐形式,并将数据互联起来,在网络上形成标准形式的、可以管理和使用的大规模分散数据,从而在互联网上形成一张超大规模的、相互连接的数据网络.Linked Data 中数据发布者就像构建普通网站一样发布数据,每个发布者自身负责数据的维护、数据之间的连接和数据的访问控制等问题.简而言之,Linked Data 是一种结构化描述数据和使数据互联的数据描述和发布规范.

尽管 Linked Data 的发展非常迅速,但关于 Linked

Data 的应用还十分有限,究其原因,Linked Data 技术研究面临着巨大的挑战:首先,Linked Data 所实现的数据连接方式以及相关应用接口和交互模型还没有被广泛接受;其次,Linked Data 技术如何实现自动、有效的数据连接还没有得到真正解决;此外,Linked Data 技术中的本体(连接模型)自动映射还没有得到解决,解决连接模型映射和数据连接是 Linked Data 面临的主要挑战,Linked Data 的数据连接采用互联网中的文档连接,当数据更新时如何有效维护连接、减少死连接产生也是 Linked Data 要解决的主要问题之一;最后,如何保证数据真实性和安全性还有待进一步研究.

尽管如此,Linked Data 技术仍在不断发展,基于 Linked Data 的应用也会越来越多. Linked Data 为数据共享和融合的研究和工程发展作出了巨大贡献.

1.2 数据活化

文献[7]针对大规模数据集分析问题提出的数据活化思想是一种适应复杂数据分析应用、高灵活性的自适应数据集分析思想. 数据活化将数据划分为一组智能活化单元,活化单元是具有智能计算能力的数据集合,具有获取环境信息、分析信息、互相通讯和单元进化等智能行为. 随着用户需求的变化,活化单元可以自发进行进化以适应新的查询请求. 数据活化是解决多源数据集分析问题的思想,具有灵活、智能、适应能力强的优点,但仍存在以下技术问题亟待解决:

1) 实现活化单元的智能行为. 数据活化要求活化单元具有较高的智能行为,根据接收到的数据查询主动适应新的查询,如何实现活化单元的智能行为是数据活化的主要难题.

2) 活化单元的抽象. 活化单元在数据活化中处于核心地位,对于底层存储要能够合理抽象数据和计算能力,对于上层用户要提供良好的用户接口. 活化单元的设计如何满足复杂的异构数据源和多变的用户接口是数据活化的另一主要挑战.

3) 活化单元的协调和管理. 正如 GFS 一样,数据活化中的 Master 单元负责协调和调度,如何有效协调和管理不同的数据单元是数据活化中的关键问题.

1.3 Internet of Data

文献[8]基于物联网的思想在数据领域提出了 Internet of Data(IOD). IOD 思想将数据比作物联网中的实体,数据虚拟标签(virtual tag)比作物联网中的 RFID 标签,使用信息隐藏技术在数据内嵌入虚拟标签,通过标签内存储的元数据实现对数据的认证、分析、可视化、存储和管理. IOD 中的数据实体指分散的、不便于管理和索引的非结构化数据文件,主要指多媒体文件. 利用多媒体文件数据存在的冗余性,采用信息隐藏方法将虚拟标签嵌入多媒体文件内,从而实现对数据生命周期的记录、管理及数据互联.

IOD 的数据实体要求数据能够利用信息隐藏技术嵌入信息,因此数据实体要有一定的信息冗余,具有这样性质的数据主要是多媒体文件,如图片、音频、视频等. IOD 通过虚拟标签对数据实体进行标识、标注和互联,记录数据元数据、数据日志和数据关联,实现多媒体数据文件的语义标注、生命周期管理、内容共享和融合. IOD 技术存在的主要问题如下:

1) 嵌入方式的限制. 信息隐藏方式不仅限制了数据类型,其本身也存在许多问题,如内容失真、算法复杂度和安全性及嵌入容量限制等.

2) 存储内容的限制. 虚拟标签记录的主要是数据日志和数据之间的关系,而数据日志和数据关系仅仅是数据互联的一部分. 因此限制了 IOD 的应用范围.

3) 对每一个数据实体都要设置一个 Unique ID,因此需要中心服务器制定一套复杂的标识系统对数据进行注册和寻址等操作.

2 智慧城市数据互联框架

本节首先阐述智慧城市数据互联框架的主要结构,然后详细描述框架中每层的作用、技术和相关问题.

2.1 概 述

本文在 Linked Data、数据活化和 IOD 等技术基础上提出了智慧城市数据互联框架 SCLDF. SCLDF 是一种智慧城市大数据的数据共享和融合框架,它用于解决智慧城市中大数据的共享和融合问题. SCLDF 采用分散自治架构,通过数据图模型描述数据之间的关联关系,从而形成数据网络以实现数据共享和融合. SCLDF 框架如图 1 所示,框架自下而上分为 4 层:

1) 数据存储层(data storage layer). 数据存储层是各种异构数据源存储形式的抽象. 智慧城市中数据源存储形式有很多种,如关系型数据库、半结构化文档、非结构化文档、多媒体数据等等,因此数据存储层具有海量、异构和分散的特点.

2) 数据转换层(data transformation layer). 为了实现数据共享和融合,数据转换层将底层不同

存储形式的数据转换为统一的图模型描述,为数据的共享和融合提供统一的数据描述. Linked Data 采用 RDF 描述数据,数据通过 RDF 形成数据图模型并相互关联. 本文借鉴 Linked Data 技术采用 RDF 对数据进行描述.

3) 数据互联层(data linked layer). 统一的数据描述通过数据互联形成数据网络,即为数据互联层. 数据互联层是 SCLDF 实现数据共享和融合的核心和基础,其作用是形成数据网络,自动维护数据关联,为数据共享和融合应用提供互联数据基础. 图 1 中数据互联层互联数据网络中的每个节点代表智慧城市中不同的数据集,如各种信息系统数据、环境采集数据等等,具有自动变化、自动维护数据关联关系等智能行为.

4) 数据共享层(data sharing layer). 数据共享层是利用数据网络实现数据共享和融合接口和应用的实现层. 数据共享层是真正为用户提供数据共享和融合接口、服务和应用的实现层.

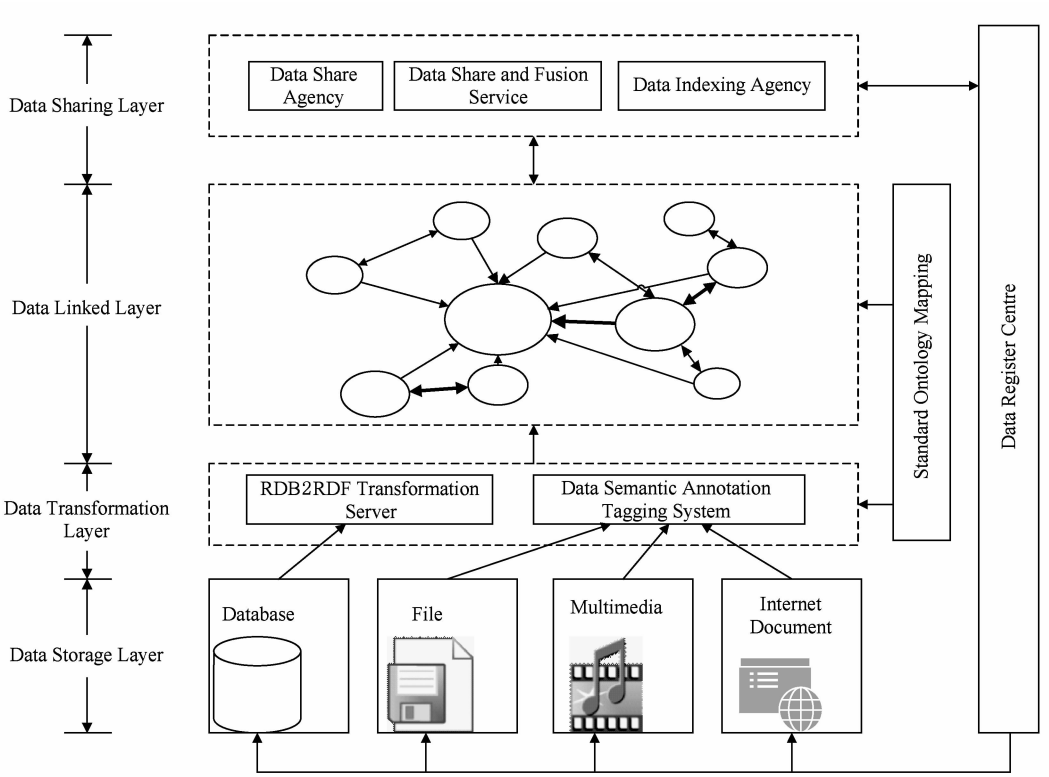


Fig. 1 Architecture of smart city linked data framework.

图 1 智慧城市数据互联框架结构图

SCLDF 中还包括了标准本体映射和数据注册中心. 标准本体映射用于解决多源数据采用不同描述词汇产生的数据描述问题,而数据注册中心则可以解决数据真实性和安全性问题.

SCLDF 是一个分层、灵活、轻量级的智慧城市数据共享和融合框架,结合了传统数据融合、Linked Data、IOD 和数据活化等技术. 相对于其他数据共享和融合技术,SCLDF 在多方面进行了改进,具有以下几个优点:

1) 分层的设计框架使 SCLDF 具有高可扩展性和灵活性,能够适应大规模、异构、多源、分散的数据共享和融合;

2) 采用分散自治的框架,能够真正解决海量数据的共享和融合问题,减少了数据共享和融合的成本,是一种轻量级的数据共享和融合框架;

3) 可以描述多种类型的数据,提出了数据语义标注标签技术,增强非结构化数据与结构化数据之间的互联;

4) 采用结构化和语义化的数据描述方案,为大数据的其他分析技术提供了更深入的数据描述模型,为智慧城市处理和分析数据提供了良好的数据基础.

2.2 数据存储层

数据存储层包含的数据类型有以下 3 种:

1) 结构化数据. 主要指关系型数据库数据. 尽管关系型数据用传统的数据仓库技术可以解决多数据源之间的融合问题,但无法实现多数据源的数据关联,其次数据仓库是对多个数据库中的数据有规划地、人为干涉地设计多源数据库分析技术,而智慧

城市中的结构化数据库规模庞大并且分布非常分散,数据仓库无法应用于这样的应用场景.

2) 半结构化数据.半结构化数据主要指结构变化非常大的数据,如 XML、HTML、各类报表等,因而无法直接使用数据库进行存储.一般的做法是将半结构化数据转化为结构化数据,或者转化为统一描述的图模型或树模型.

3) 非结构化数据.传感器收集的数据大部分都是非结构化数据,因此智慧城市中存在大量的非结构化数据.非结构化数据主要包括非结构化文档、图像、音频和视频.对于非结构化数据,一般采用标注的方法将信息转化为结构化数据,从而实现对数据的管理、处理和分析.由于计算机视觉、机器学习和人工智能等领域的发展,基于多媒体内容分析的应用越来越广泛,如基于内容的图像检索技术等.将标注技术、基于内容的分析技术和语义结合起来,是多媒体数据分析领域的发展趋势.

数据存储层数据存储形式主要有两种:一种是存储在各类数据库中的结构化数据;另一种是以文件形式存储的半结构化或非结构化数据.

2.3 数据转换层

由 2.2 节可知,在智慧城市中,数据的底层存储形式和种类非常繁多,如何将这些多源的、分散的、异构的数据转换为一种统一的描述,并有效地解决数据更新、数据转换效率、数据转换质量是数据转换层的主要任务.

数据转换层针对不同的数据类型,主要采用两种技术:对结构化数据(主要是关系型数据)采用 RDB 到 RDF 映射技术(RDB2RDF);对于非结构化数据(主要是多媒体数据)采用数据语义标注标签对数据进行 RDF 标注.数据语义标注标签是本文根据多媒体标注技术、多媒体语义描述技术和 IOD 中的虚拟标签提出的一种语义化标注方法.

RDB2RDF 技术是将关系型数据映射到图模型数据的方法.文献[10]对 RDB 到 RDF 映射的研究进行了分类.根据创建映射的方式,RDB2RDF 分为自动创建映射和领域语义驱动创建映射.Berners-Lee 在文献[11]中提出了一些自动映射规则,Virtuoso RDF View[12]和 D2RQ[13]均采用了这种思想.自动映射的方式简单,不需要复杂的领域语义,但是产生的本体不是领域相关的标准本体,增加了数据互联的不一致性.领域语义驱动创建映射的方式采用领域相关的本体知识实现 RDB2RDF.这种方式需要复杂领域本体知识,但是映射后的数据是标准本体数据,增强了数据的一致性,实现数据之间的互联.根据映射的实现方式,RDB2RDF 分为静态映射和动态映射.静态映射采用数据仓库中的提取、转换和加载(extract, transform, load, ETL)实现,Byrne 在文献[14]中使用批处理创建 RDB 到 RDF 仓库的映射规则.ETL 方法具有良好的性能,允许在数据上附加处理和分析以及逻辑运算,但不能及时反应数据的变化.动态映射只有在收到查询请求时才动态地产生 RDB2RDF 映射,虽然可以及时反应数据变化,但是查询性能比静态映射差.

W3C 一直致力于 RDB2RDF 的研究,在 2012 年推出了 RDB2RDF 标准语言 R2RML[15]和 RDB2RDF 直接映射标准[16],R2RML 是可定制的 RDB2RDF 标准,可以将数据库中的属性映射到领域本体,因此可以利用 R2RML 进行 RDB2RDF 工程.一些发布工具(如 Virtuoso RDF)[17]已经开始支持 R2RML 标准.

数据语义标注标签技术将在第 3 节详细介绍.

2.4 数据互联层

经过数据转换层转换的数据提供了数据 RDF 描述,通过 RDF 数据图模型形成了数据互联层.与 Linked Data 相似却不相同,数据互联层使用了 Linked Data 的技术,但对其进行了改进.

与 Linked Data 一样,数据互联层数据遵循以下规则:使用 HTTP URI 作为唯一标识数据的名字;使用 HTTP 协议作为获取数据的协议;获取和发布数据采用标准的形式,如 RDF/XML, N3, Turtle 等格式的数据、SPARQL 查询语言;数据之间存在关联,从而可以获得与之相关的数据[6].

RDF, RDFS, OWL, SPAQL 是数据互联层描述和查询数据最基本的语言.

RDF[18]是一种描述对象("资源")之间关系的数据模型,并为这种数据模型提供简单的语义.RDF 允许用户自定义连接类型(即本体)[19].本体提供了对给定领域的一种共识,这种共识对于消除术语差别是必要的.

RDFS[20]和 OWL[21]是用于定义本体的语言.W3C 在 2012 年推出的 OWL 新标准 OWL2[21]是一个更丰富的词汇描述语言,可以刻画诸如类间关系、类的基数、相等、属性的特征和枚举类等.OWL2 是对 RDFS 和 OWL 的极大扩展.

SPARQL[22]是 W3C 推荐的 RDF 标准查询语言.SPARQL 提供和 SQL 类似的查询接口,可以方便地查询和更新 RDF 数据.

图 1 所示的数据互联层中的节点代表智慧城市中不同的数据集,而节点之间的箭头表示数据集之间 RDF 连接,线的粗细代表连接数据的多少. 通过数据集之间的 RDF 连接和本体映射,不同数据集形成了统一的数据描述,数据集间形成了互联网络,为数据共享和融合提供了统一的数据互联和描述基础.

虽然数据互联层和 Linked Data 采用了相同的描述和查询技术,但是为了解决 Linked Data 中存在的问题,数据互联层在以下几方面做了改进.

如 1.1 节所述,由于 Linked Data 允许数据发布者自己定义本体,而不强制数据和已有的本体之间产生映射,导致数据描述产生数据本体不一致. 为了解决数据互联问题,需要不同的语汇表映射到目标模式中,解决不同数据源的本体映射,将多源本体描述数据融合起来. 为了解决词汇表映射问题,SCLDF 引入了标准本体映射. 标准本体映射不仅要定义标准本体,而且要兼容本地本体和其他标准本体,如 FOAF[23],SIOC[24],SKOS[25],DOAP,vCard[26],Dublin Core[27],OAI-ORF 等,将本地本体和其他标准本体映射到标准本体,以解决词汇表映射问题. 在第 4 节中将详细描述标准本体映射.

产生数据冲突的一个主要原因是数据的真实性问题,在 Linked Data 上每个组织或个人都可以发布数据,这就造成了数据真实性问题. 而智慧城市中的数据发布者一般来说是固定的城市管理或服务机构,因此 SCLDF 中提出了数据注册中心策略. 发布者在发布数据之前必须在数据注册中心注册数据资源,以证实数据的真实性. 数据使用者在使用数据时可根据数据注册中心提供的数据判断数据真实性. 在 Linked Data 上推荐解决数据真实性问题的方法是通过启发式算法将数据按照真实性进行排序,而在智慧城市中的数据发布者是特定的,因此解决数据真实性的有效方式是采用注册的方式.

采用 RDF 描述数据的一个主要问题是数据关联关系的自动维护. 数据之间存在的关联和冲突不可能全部人为解决,因此必须有一种自动处理数据关系的机制,实现数据关系的自动化. 由于 SCLDF 是分散自治的结构,采用数据活化中的智能单元概念为每个自治的数据集赋予智能关系计算能力可作为解决数据关系自动化的一种方式.

综上所述,数据互联层采用 RDF,RDFS 和 OWL 等标准对数据进行描述和关联,采用标准化本体映射解决本体映射问题,采用数据注册中心解决数据真实性问题,采用智能关系计算能力自动处理数据关联关系,每个数据集节点都具有独立、自治、互联和智能的特点.

2.5 数据共享层

数据共享层是实现数据共享和融合的关键层,其作用是利用数据互联层形成的数据网络为最终用户提供数据共享接口、服务和应用. 数据共享层要解决的问题主要是如何利用数据互联层形成的数据网络提供有效并且友好的数据共享接口、服务和应用. 下面介绍几种设想的接口、服务和应用:

1) 数据浏览器应用. 正如互联网中的文档一样,数据互联层的数据是相互连接的,通过抓取数据互联层的互联数据,为用户提供友好的数据查询、浏览及搜索服务和接口,以便用户使用数据提供更高层的数据应用.

2) 数据搜索引擎应用. 是数据共享层非常重要的应用,将数据网络作为一个整体为用户提供数据服务. 数据搜索引擎可以为普通用户和开发者提供接口,让用户通过特定的方式搜索数据,开发者通过数据搜索引擎提供更高层的数据服务.

3) 特定领域应用. 开发面向特定领域的应用,如智能交通应用、智能医疗应用、智能 GIS 应用以及各种互联数据应用等等.

4) 数据共享接口和服务. 为用户和开发者提供数据标注、数据互联等服务接口,使经过授权的用户可以发布数据和提供相关应用.

3 数据语义标注标签

数据语义标注标签(data semantic annotation tag, DSAT)是本文针对半结构化数据和非结构化数据提出的一种语义标注方法,对无法存储在关系型数据库的半结构化数据和非结构化数据采用标注方法进行数据共享和融合. 以下将数据语义标注标签简称为数据标签(data tag).

数据标签是对文献[8]中虚拟标签的改进. 虚拟标签和 RFID 标签作用一样,起到标识、寻址访问和记录部分元数据的作用. 数据标签具有相同的作用,但具有更广泛的意义.

首先,数据标签采用语义化描述,使数据标签的应用范围更广泛. 其次,数据标签的形式有多种,包括嵌入数据内的数据标签、嵌入文件头和文件尾的数据标签以及存储在数据库或文件中的独立数据标签. 此外,数据标签使用的技术包括数字水印、多媒体标

自动标注、多媒体描述、文档标注等,对多媒体和文档都可以进行标注.

按照数据标签的存储形式,数据标签可分为嵌入式数据标签(internal data tag)和独立数据语义标注标签(external data tag);按照数据类型进行分类,数据标签可分为多媒体文件数据标签(multimedia data tag)、文档数据标签(document file data tag)、Web内容数据标签(Web content data tag),如图2所示:

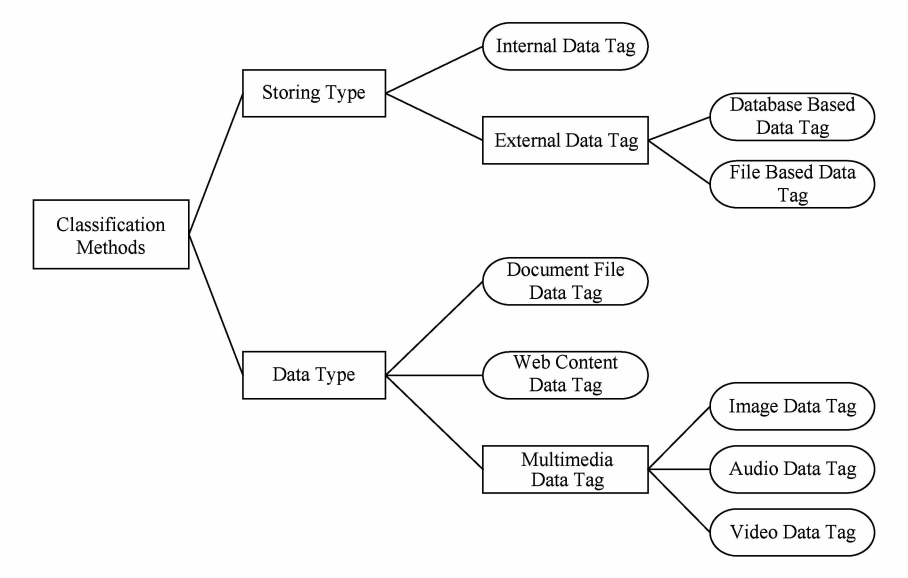


Fig. 2 Data semantic annotation tag classifications.

图2 数据语义标注标签分类

3.1 嵌入式数据语义标注标签

将数据标签嵌入数据中的数据标签称为嵌入式数据标签.为了实现数据自身和其他SCLDF中数据的关联,在数据标签中存储的是数据唯一标识符URI和基本元数据;数据之间的关联关系则存储在SCLDF数据互联层形成的数据网络中,因此数据通过数据标签和其他数据产生关联.其嵌入过程如图3所示:

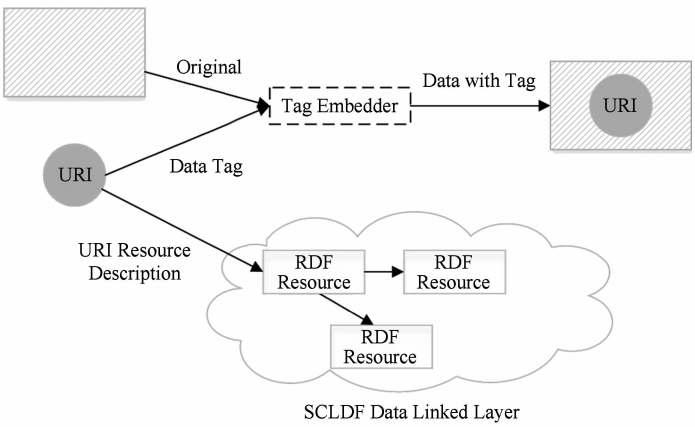


Fig. 3 Embedding of internal data tag.

图3 嵌入式数据语义标注标签嵌入过程

当通过带有数据标签的数据查询相关数据时,需要从数据中提取出数据标签,通过数据标签记录的URI和SCLDF数据互联层产生关联,从而实现从数据共享和融合,如图4所示.

嵌入式数据标签将数据标签和数据自身存储在一起,嵌入到数据内容或数据文件中,实现数据标签和数据紧密结合.嵌入式数据标签的嵌入方法如表1所示.主要包括数字水印技术、文件格式和文件元数据.

采用数字水印方法将数据标签嵌入数据内容中要求数据有数据冗余.数字水印方法的数据标签是对IOD虚拟标签的一种实现,主要用于多媒体文件,具有以下优点:1)安全性,由于信息直接嵌入在载体数据中,嵌入过程加密,保证了信息的安全性;

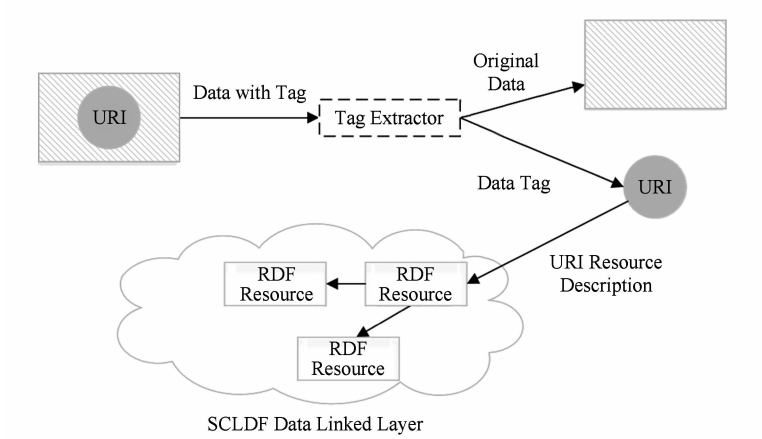


Fig. 4 Extraction of internal data tag.

图 4 嵌入式数据语义标注标签提取过程

2)隐蔽性,水印信息不容易被察觉;3)与载体紧密性,由于嵌入载体数据内,使嵌入的信息不易丢失,便于管理.

Table 1 Internal Data Tag

表 1 嵌入式数据标签

Location	Requirements	Methods
Data Content	Data Redundancy	Watermarking, RDFa, Microformat
Metadata	File System Support	Embed in File Attributes
File Header/Tail	File Format Dependency	Embed Data Tag in File Header/Tail

关于 Web 文档的标注技术应用已经非常广泛,Web 的语义标注语言如 RDFa[28],Microfomat[29] 和 HTML Microdata 等都是对 Web 文档进行标注的标准或事实标准,这些结构化的描述语言将语义信息以 Markup 的方式写入 Web 文档中,从而实现语义标注.

将数据标签嵌入数据的元数据中要求数据必须能够存储元数据. 文件系统为了方便文件的管理,为文件提供了元数据描述,即文件属性. 一种嵌入数据标签的方式是扩展文件系统中的文件属性,将必要的信息嵌入在文件属性中,从而实现数据之间的关联.

还可以根据特定的文件格式,将数据嵌入到文件头或文件尾中. DIG35[30] 是一种描述图像元数据的语言,文献[30]对图像各种文件格式 DIG35 数据的 Internal Model 嵌入方法进行了详细的描述,这些方法均可以用于数据标签的嵌入. 这种方式的优点是简单易用,方便管理,缺点是依赖于不同的文件格式,格式转换时容易丢失信息,相比数字水印方法安全性较差.

嵌入式数据标签将 URI 和数据紧耦合,具有方便数据管理、信息不易丢失的优点,实现了数据和其语义标注的紧密关联.

3.2 独立数据语义标注标签

独立数据标签的标签部分和数据部分相分离,标签存储在数据库或独立文件中,通过记录数据和独立标签的关联信息实现数据和其他数据的互联.

独立标签的适用范围较广,因为其对数据形式没有限制,既可以是各种各样的文件,也可以是存储在数据库的数据或其他形式的数据.

采用数据库存储标签的独立数据标签和传统的文档标注技术、多媒体标注技术相似,都是通过数据库管理系统对数据的标注信息进行管理. 这种方式对于数据转移较少的情况较为适用,如互联网资源和本地文档管理、较少变动的数据库数据等.

采用文件存储标签的独立数据标签是将数据的结构化描述存储在结构化文档中的标注方法,和以数据库存储为基础的方式不同,这种方式的数据标签通过将数据和标签文件存储在相同目录中实现标签和数据的关联,可以随数据进行转移,但易出现丢失标签或失去关联的情况.

3.3 自动数据语义标注标签

与多媒体标注和文档标注技术一样,手工方式的数据语义标注标签存在很大的局限性,人的主观性和语汇的多义性、同义性都会导致描述的不一致. 因此,自动标注技术得到了研究者的关注,许多研究者纷纷提出不同的自动标注的方法[31-32]及自动语义标注的方法[33-34].

很多研究者在语义网和 Linked Data 领域提出了不同的自动标注框架,这些框架都局限在特定

数据的语义标注上,如文献[35]提出语义网中图像语义标注框架,采用SVM、图像分割、特征提取等技术实现图像自动语义标注;文献[36]提出了一种针对网络非结构化数据和半结构化数据的语义标注框架,采用贝叶斯网络实现自动语义标注.

利用这些已有的特定领域的自动标注技术实现数据语义描述,通过数据语义标注标签进行数据关联,是实现自动数据语义标注标签的有效方式.

4 SCLDF数据互联层

SCLDF数据互联层是SCLDF实现数据共享和融合的核心层,为数据共享层的接口、应用和服务提供数据基础.SCLDF数据互联层为了提供统一和互联的数据网络,需具备以下3个功能:

1) 数据关联的自动维护,如数据关联关系的自动发现、更新和删除;

2) 解决多本体描述的映射问题;

3) 解决数据冲突问题.

为了实现数据关联的自动维护,数据服务器必须具备一些关联关系的智能计算能力.与数据活化中的智能单元功能相似,数据服务器通过感知数据变化、数据关系变化等环境信息自动维护数据关系.

为了解决多本体描述问题,需要制定智慧城市标准化本体和标准映射机制.解决数据冲突问题需要解决数据互联层的数据融合问题.

4.1 数据关系自动维护

自动维护数据关联关系包括3个方面:数据关系自动发现、更新和删除.

关联数据关系是数据发布者在发布数据时必须执行的过程,但是发布者无法知道所有的数据关系,因此需要在发布数据后自动发现数据关系,实现数据网络逐渐融合,形成紧密关联的网络.通过数据搜索引擎和数据浏览器提供的数据关系服务和应用自动关系发现代理可以实现数据关系自动发现.数据服务器上的自动关系发现代理通过周期查询或订阅的方式向数据搜索引擎请求数据关系,搜索引擎通过查询将可能的存在关系的数据返回给代理,代理通过计算自动关联数据实现数据关系自动发现过程.

数据关联关系自动更新和删除是维护数据关联的重要过程.有效解决数据关联关系自动更新和删除的方法主要有两种:文献[37]提出的数据源发布Linked Data更新日志,采用嵌套更新集合方式组织更新日志,其他数据源通过访问更新日志来更新数据关联关系;PTSW(ping the semantic Web)是一个存储RDF文档的仓库,RDF文档的创建或更新会通知PTSW,其他可能连接此RDF文档的数据源通过访问PTSW的API获取这些更新信息,从而实现数据关联自动更新,这是一种中心注册的方式.这两种方式各有优缺点,中心注册方式只需要订阅注册中心,但容易产生单点失效;而发布数据更新日志需要订阅所有关联数据源,但不会产生单点失效.在SCLDF中,由于智慧城市数据源本身的特点,采用中心注册方式比较适用.智慧城市中建立一个更新注册中心,数据发布者在数据集服务器上使用数据关系维护代理程序,通过周期访问或订阅方式访问更新注册中心来更新数据关联关系.更新注册中心通过一些有效手段,如分布式和多复制备份就可以有效防止单点失效.

4.2 标准本体映射与数据融合

由于不同的组织发布数据时采用不同的本体描述数据,如采用RDB2RDF技术自动生成的自定义本地本体、或映射到不同领域不同标准的本体.为了实现不同数据源真正的互联,需要将不同的本体定义映射到统一本体描述上.此外,不同的数据源描述同一实体时数据可能产生冲突,解决数据冲突是实现SCLDF数据融合的关键.

构建标准化本体是一个复杂而漫长的过程,Linked Data中没有制定标准化本体,而是采用映射机制实现不同数据本体之间映射,解决本体不一致问题.SCLDF中的标准本体映射分为两部分:一部分是制定标准本体;另一部分是实现本体映射机制.为了加强数据描述的一致性,数据或采用SCLDF标准本体,或映射到标准本体.

关于制定标准本体的研究虽然有很多,但却没有突破,这也是Linked Data为何放弃标准化本体的原因.从文献[38]对标准化本体工程进行的综述不难看出,标准化本体主要集中在顶级本体的标准化,因为细化本体过于复杂而无法统一.但是在智慧城市中,由于数据来源都非常明确,如交通、通信、教育、医疗等服务部门,这些部门制定相关领域的标准本体通过注册即可成为标准化本体的一部分,而其他数据要成为SCLDF的一部分必须采用映射机制与注册本体进行本体映射.

本体映射是实现异构数据通用性的重要研究领域,文献[39]对本体映射问题进行了综述.W3C推荐标准RIF[40]是一组规则交换标准语言,提供传递映射和局部映射,能很好地描述本体映射.

数据融合问题是 SCLDF 数据互联层的核心问题之一,在 SCLDF 中的数据融合要求从自治的数据源中找到最可信的数据,从而解决数据冲突,实现数据融合.数据库管理系统领域的数据融合技术可以有效解决不同数据库中的数据冲突问题,而 SCLDF 中的数据源具有分散、自治的特点,因此需要结合新的数据融合技术解决数据冲突.融合数据互联层中多源数据的原型系统包括 DERI Pipes[41] 和 KnoFuss architecture[42],结合传统的数据库数据融合技术,可有效解决数据互联层的数据融合问题.

5 SCLDF 在智慧城市中的应用

智慧城市涉及的领域非常广,因此在智慧城市中有许多数据,例如交通数据、地理信息数据、天气数据、个人数据和职能部门或公司数据等等.

下面以交通数据、地理信息数据、天气数据、部门数据和个人信息数据为例说明 SCLDF 在智慧城市中可以发挥的重要作用.

交通数据包括道路数据、公共交通(公共汽车、轨道交通、出租车)数据以及实时道路情况图像数据等.天气数据包括各地实时天气各项指数数据,如天气、温度、湿度、空气污染情况等.地理信息数据则包括城市各种详细地理信息数据,如经纬度、道路、街道、建筑、单位、部门等信息,地理信息数据在智慧城市中具有重要的意义,因为许多智慧城市应用都是基于位置信息的服务.部门数据包括部门基本信息、职员公开信息、提供服务的基本信息等,如医院的数据包括医院基本信息、不同科室的专家信息、专家坐诊时间信息、病人病历信息等.个人信息数据包括社交信息、个人存储在特定应用中的数据(如日程安排表)、地理位置信息等.

这些数据都可以通过制定相关领域本体,将数据转化为结构化的 RDF 数据,发布在 SCLDF 框架中进而相互关联,形成巨大的数据网络,图 5 为上述数据的简单数据互联示意图.

如图 5 所示,地理信息数据、交通数据、天气数据、建筑物数据、部门和公司数据及个人信息数据通过 RDF 相互联系.如交通数据中包含的道路信息、街道信息、交通图像信息等,这些信息都和地理位置信息相关,因此通过 RDF 关联和地理信息数据相关联.与此信息类似,建筑信息的地理位置和地理信息数据关联,部门与公司的公开职员数据和个人信息关联,个人信息的地理位置信息和地理信息关联,天气信息和地理信息数据相关联,建筑信息数据和地理信息相关联等等.

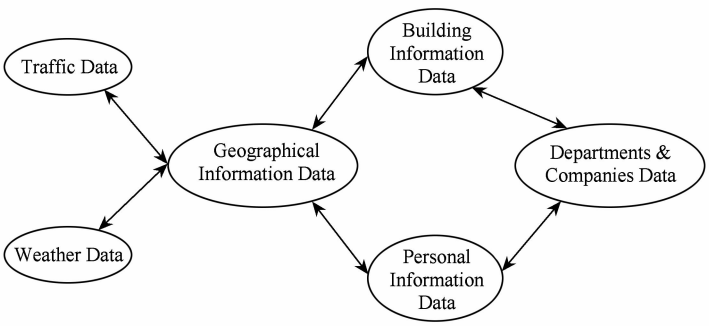


Fig. 5 An illustrate example application of SCLDF.

图 5 一个 SCLDF 应用例子

智慧城市中的智能应用就可以通过这些相互关联的数据为公众提供智能服务.例如,智慧医疗应用为公众提供专家预约服务,应用程序通过 SCLDF 搜集个人信息、医院信息、专家个人信息,通过推理和计算提出最匹配的预约信息;智慧交通应用为公众提供交通推荐服务,应用程序查询 SCLDF 中的交通数据、天气数据、个人位置信息数据和地理位置信息数据,综合这些关联信息为用户推理出合理的出行方案.

上述示例仅仅是在部分 SCLDF 共享和融合数据中的应用示例,随着数据源的增加,SCLDF 在智慧城市应用中将发挥更大作用.SCLDF 通过多源数据的共享和融合为智慧城市提供互联数据基础,从而产生更有价值的智能应用.

6 结 论

本文讨论了智慧城市数据共享和融合的基本问题,分析了不同解决方案的优缺点,基于 Linked Data, Internet of Data 和数据活化技术,提出了一种解决

智慧城市数据共享和融合问题的新方案——智慧城市数据互联框架 SCLDF.

本文论述了 SCLDF 的层次结构,详细讨论了每一层的作用、所需要解决的问题和使用的技术. SCLDF 数据转换层将数据存储层已有的关系型数据库数据、半结构化和非结构化数据标注内容转换为 RDF 描述,通过统一的描述为数据互联层的数据互联奠定数据描述基础;SCLDF 的核心层数据互联层是实现数据融合的核心层,通过改进 Linked Data 技术将所有数据关联起来实现数据共享和融合,连接原有的数据孤岛,形成数据网络;SCLDF 的上层是数据共享层,数据共享层为数据使用者提供了数据共享和融合接口、服务和应用,数据使用者通过数据共享层消费互联数据,实现数据价值的最大化.

本文对 SCLDF 进行了初步的阐述,为解决智慧城市数据共享和融合问题提供了一个新思路. 但 SCLDF 的研究还面临着很多挑战,需要更进一步的研究.

参 考 文 献

[1] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1): 107-113

[2] Ghemawat S, Gobioff H, Leung S T. The Google file system [C] //Proc of the 19th ACM Symp on Operating System Principles. New York: ACM, 2003: 29-43

[3] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554

[4] Gao Xiang, Wang Yong. Survey of multisensor information fusion [J]. Computer Measurement & Control, 2002, 10(11): 706-709 (in Chinese)
(高翔, 王勇. 数据融合技术综述[J]. 计算机测量与控制, 2002, 10(11): 706-709)

[5] Berners-Lee T, Hendler J, Lassila O. The semantic Web [J]. Scientific American, 2001, 284(5): 34-43

[6] Bizer C, Heath T, Berners-Lee T. Linked data-the story so far [J]. Int Journal on Semantic Web and Information Systems (IJSWIS), 2009, 5(3): 1-22

[7] Xiong Zhang, Luo Wuman, Chen Lei, et al. Data vitalization: A new paradigm for large-scale dataset analysis [C] //Proc of the 16th Int Conf on Parallel and Distributed Systems (ICPADS). Piscataway, NJ: IEEE, 2010: 251-258

[8] Fan Wei, Chen Zhenyong, Xiong Zhang, et al. The Internet of data: A new idea to extend the IOT in the digital world [J]. Frontiers of Computer Science, 2012, 6(6): 660-667

[9] W3C Consortium. Semantic Web [EB/OL]. (2013-07-28) [2013-07-22]. http://www.w3.org/standards/semanticweb/#w3c_overview

[10] Sahoo S S, Halb W, Hellmann S, et al. A survey of current approaches for mapping of relational databases to rdf [OL]. (2009-01-08) [2013-07-22]. http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf

[11] Berners-Lee T. Relational databases on the Semantic Web [EB/OL]. (1999-08-27)[2013-07-22]. http://www.w3.org/DesignIssues/RDB-RDF.html

[12] Erling O, Mikhailov I. RDF support in the virtuoso DBMS [M] //Networked Knowledge-Networked Media. Berlin: Springer, 2009: 7-24

[13] Bizer C, Cyganiak R. D2rq-lessons learned [C/OL] //Proc of W3C Workshop on RDF Access to Relational Databases. (2007-09-08) [2013-07-22]. http://www.w3.org/2007/03/RdfRDB/papers/d2rq-positionpaper/

[14] Byrne K. Having triplets-holding cultural data as rdf [C] //Proc of the ECDL 2008 Workshop on Information Access to Cultural Heritage. Aarhus: ILPS, 2008

[15] Das S, Sundara S, Cyganiak R. R2RML: RDB to RDF mapping language [EB/OL]. (2012-09-27) [2013-07-22]. http://www.w3.org/TR/2012/REC-r2rml-20120927/

[16] W3C Consortium. A direct mapping of relational data to RDF [EB/OL]. (2012-09-27)[2013-07-22]. http://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/

[17] Virtuoso Open-source. Virtuoso R2RML support [EB/OL]. (2009) [2013-07-22]. http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtR2RML

[18] Klyne G, Carroll J. Resource description framework (RDF): Concepts and abstract syntax [EB/OL]. (2004-02-10)[2013-07-22]. http://www.w3.org/TR/rdf-concepts/

[19] Antoniou G. A Semantic Web Primer [M]. Cambridge: MIT Press, 2004

[20] Brickley D, Guha R. RDF vocabulary description language 1.0: RDF schema [EB/OL]. (2004-02-10) [2013-07-22]. http://www.w3.org/TR/rdf-schema/

[21] W3C OWL Working Group. OWL 2 web ontology language document overview (second edition)[EB/OL]. (2012-12-11) [2013-07-22] http://www.w3.org/TR/owl2-overview/

[22] The W3C SPARQL Working Group. SPARQL 1.1 overview [EB/OL]. (2013-03-21)[2013-07-22]. http://www.w3.org/TR/sparql11-overview/

[23] FOAF Project. The friend of a friend (FOAF) project [EB/OL]. (2010-10-17) [2013-07-22]. http://wiki.foaf-project.org/w/Main_Page

[24] Science Foundation Ireland. SIOC core ontology specification [EB/OL]. (2010-03-25)[2013-07-22]. http://rdfs.org/sioc/spec/

[25] Miles A, Bechhofer S. SKOS simple knowledge organization system reference [EB/OL]. (2009-08-18) [2013-07-22]. http://www.w3.org/TR/skos-reference/

[26] IETF. VCard format specification [EB/OL]. (2011-08) [2013-07-22]. http://tools.ietf.org/html/rfc6350

[27] Dublin Core Metadata Initiative. The Dublin core metadata element set [EB/OL]. (2013-06-17)[2013-07-22]. http://dublincore.org/

[28] Adida B, Herman I, Sporny M, et al. RDFa 1.1 primer [EB/OL]. (2012-06-07)[2013-07-22]. http://www.w3.org/TR/xhtml-rdfa-primer/

[29] Microformats Community. Introduction to microformats [EB/OL]. (2013-04-17)[2013-07-22]. http://microformats.org/wiki/introduction

[30] Digital Imaging Group. DIG35 specification version 1.1 [OL]. (2001-04-16)[2013-07-22]. http://www.bgbm.org/tdwg/acc/Documents/DIG35-v1.1WD-010416.pdf

[31] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models [C] //Proc of the 26th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2003: 119-126

[32] Kang F, Jin R. Automatic Image Annotation [M]. East Lansing: Michigan State University, 2007

[33] Srikanth M, Varner J, Bowden M, et al. Exploiting ontologies for automatic image annotation [C] //Proc of the 28th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2005: 552-558

[34] Carneiro G, Chan A B, Moreno P J, et al. Supervised learning of semantic classes for image annotation and retrieval [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(3): 394-410

[35] Khan L. Standards for image annotation using Semantic Web [J]. Computer Standards & Interfaces, 2007, 29(2): 196-204

[36] Rajput Q, Haider S. BNOSA: A Bayesian network and ontology based semantic annotation framework [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2011, 9(2): 99-112

[37] Auer S, Dietzold S, Lehmann J, et al. Triplify: Light-weight linked data publication from relational databases [C] //Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 621-630

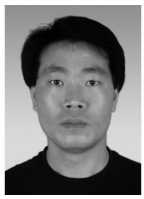
[38] Li Jing. Current trends on the ontology technology standardization [J]. New Technology of Library and Information Service, 2007, 2(8): 12-17 (in Chinese)
(李景. 本体技术标准化综述研究[J]. 现代图书情报技术, 2007, 2(8): 12-17)

[39] Kalfoglou Y, Schorlemmer M. Ontology mapping: The state of the art [J]. The Knowledge Engineering Review, 2003, 18(1): 1-31

[40] Kifer M, Boley H. RIF Overview [EB/OL]. 2nd ed. (2013-02-05)[2013-07-22]. http://www.w3.org/TR/rif-overview/

[41] Le-Phuoc D, Polleres A, Tummarello G, et al. DERI pipes: Visual tool for wiring web data sources [EB/OL]. (2008)[2013-07-22]. http://www.cs.vu.nl/~pmika/swc-2008/DERI%20Pipes-DERI%20Pipes.pdf

[42] Nikolov A, Uren V, Motta E, et al. Integration of semantically annotated data by the KnoFuss architecture [M] //Knowledge Engineering: Practice and Patterns. Berlin: Springer, 2008: 265-274

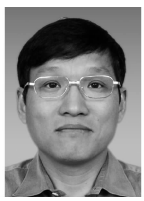


Chen Zhenyong, born in 1974. Associate professor of the School of Computer Science and Engineering of Beihang University, China. His current research interests include smart city, data vitalization and information hiding.

Xu Zhouchuan, born in 1989. Master candidate in the School of Computer Science and Engineering of Beihang University. His current research interests include semantic Web and linked open data.

Li Qingguang, born in 1987. Master candidate in the School of Computer Science and Engineering of Beihang University. His current research interests include information hiding and smart city.

Lü Weifeng, born in 1972. Professor in the School of Computer Science and Engineering of Beihang University. Member of China Computer Federation. His main research interests include intelligent transport system and smart city.



Xiong Zhang, born in 1956. Professor at the School of Computer Science and Engineering, Beihang University, China. His main research interests include the Smart City, data vitalization and large-scale computer application systems.