ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.8, August 2008, pp.2018−2031
DOI: 10.3724/SP.J.1001.2008.02018

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

数据空间技术研究*

李玉坤, 孟小峰+, 张相於

(中国人民大学 信息学院,北京 100872)

Research on Dataspace

LI Yu-Kun, MENG Xiao-Feng+, ZHANG Xiang-Yu

(School of Information, Renmin University of China, Beijing 100872, China)

+ Corresponding author: E-mail: xfmeng@ruc.edu.cn

Li YK, Meng XF, Zhang XY. Research on dataspace. *Journal of Software*, 2008,19(8):2018−2031.
http://www.jos.org.cn/1000-9825/19/2018.htm

Abstract: This paper introduces the concept and characters of dataspace, and presents a framework for dataspace integration and management system. Based on the framework, this paper further summarizes research works on data model, integration, query, update, storage, index, evolution and systems of dataspace. Challenges and future work on dataspace research are analyzed.

Key words: DBMS; dataspace; dataspace management

摘 要: 阐述了数据空间概念及其特性,提出了数据空间集成与管理系统框架.在此基础上,进一步从数据模型、数据集成、数据查询、数据更新、存储索引、数据演化和系统实现几个方面对数据空间研究工作进行了总结分析.讨论了数据空间研究面临的挑战和未来的研究工作.

关键词: 数据库管理系统;数据空间;数据空间管理
中图法分类号: TP311 **文献标识码**: A

数据库技术在过去 30 年里为推动企业数据管理的发展做出了巨大贡献.但是,随着数字化技术和互联网的发展,Web 日益成为一个巨大的信息共享平台,数据管理呈现出新的特点[1].一个特点是海量.全球的数据量在以指数级的速度迅猛增长,据保守估计,目前每年全球至少产生 15 亿 TB 的新数据.另一个特点是共享.互联网和通信设备的普及使人们能够很容易地实现数据的共享,数据库之间也因此建立起越来越密切的联系.还有一个特点是多样化.今天人们所面临的数据已不再是关系模型下纯粹的结构化数据,图片、音频、视频、文档等非结构化数据大量涌入到人们的应用中来.

新的数据特点使数据管理技术面临挑战,传统的 DBMS 在这些挑战面前显得无能为力.为大家所熟悉的 Oracle,DB2 和 SQL Server 等商业关系数据库已经广泛应用于各行各业.在很多人看来,似乎一切都是如此完美,所有的数据管理问题都会在这里得到答案.然而,事实并非如此.进入 21 世纪,我们忽然发现管理着世界上最大、

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z155 (国家高技术研究发展计划(863)); the National Basic Research Program of China under Grant No.2003CB317000 (国家重点基础研究发展计划(973); the Program for New Century Excellent Talents in University of China under Grant No.NCET-04-0051 (新世纪优秀人才支持计划)

Received 2008-01-20; Accepted 2008-05-19

最丰富的数据集合,而且主要为个人服务的 Google,MSN,Yahoo 都不使用传统的 RDBMS,而是另辟蹊径去寻找能够更好地满足数据管理需要的方法.随着企业内部非结构化数据信息越来越多,企业数据管理问题会更加突出.个人信息管理同样面临这一问题,由于个人能够支配的时间有限,以及缺乏数据管理技术的支持,个人数据量的剧增导致个人信息管理效率下降,人们将大量的时间耗费在信息的收集和查找方面.

数据管理面临的挑战促使我们去寻求一种新的数据管理技术——数据空间(dataspace).

1 数据空间概述

与传统的 RDBMS 和数据集成技术相比,数据空间具有鲜明的特点.本节结合相关研究工作,对数据空间基本概念及其特性进行分析和归纳.

1.1 数据空间基本概念

数据空间是与主体相关的数据及其关系的集合[2],数据空间中的所有数据对于主体来说都是可以控制的[3].主体相关性和可控性是数据空间中数据项的基本属性,一般所说的数据空间实际上是指主体数据空间,与之相对的是公共数据空间.图 1 显示了主体数据空间和公共数据空间的关系.主体数据空间是公共数据空间的一个子集,随着主体需求的不断变化,数据项不断从公共数据空间纳入到主体数据空间中.一般所说的数据空间都是指主体数据空间.

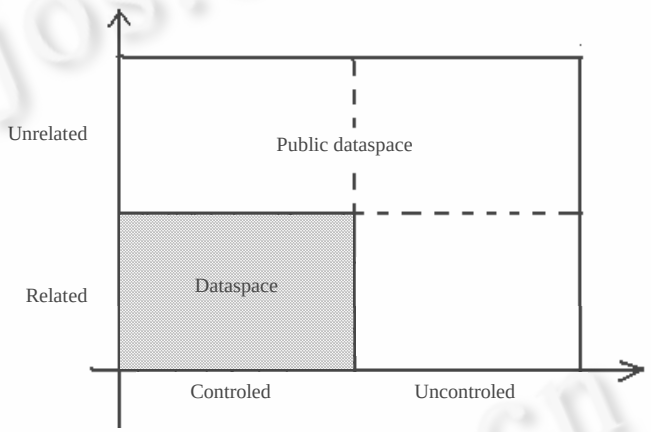

Fig.1 Dataspace and public dataspace

图 1 数据空间和公共数据空间

主体、数据集、服务是数据空间的 3 个要素.主体是指数据空间的所有者,可以是一个人或一个群组,也可以是一个企业.数据集是与主体相关的所有可控数据的集合,其中既包括对象,也包括对象之间的关系.主体通过服务对数据空间进行管理,例如数据分类、查询、更新、索引等,都需要通过数据空间提供的服务完成.表 1 显示了数据空间与关系数据库系统的主要区别.与传统数据管理技术相比,数据空间在数据模型、数据操作、数据对象、数据关系以及构建成本上都有明显的不同[2]:

Table 1 Comparison of dataspace and RDBMS

表 1 数据空间与 RDBMS 的比较

	Dataspace	RDBMS
Data model	Graph model, schema-later	Relation model, schema-first
Service quality	Best-Effort	Exact, complete
Data type	Multi-Source, heterogenous	Single-Source, fixed type
Data relation	Complex, dynamic, evolution	Simple, stable
Construction approach	Pay-as-You-Go	Pay-before-You-Go

(1) 数据模型.传统关系数据库基于关系模型,数据关联是基于关系表的.数据空间的逻辑模型是一个图,这一点与网状模型类似.但有一点不同,无论是关系模型、层次模型、面向对象模型,还是网状模型,都是一种模式优先(schema-first)的逻辑结构,即数据库依赖于严格的数据模式.而数据空间的一个重要特点是从数据到模式(from-data-to-schema),它并不依赖严格的数据模式,数据模式可以是松散的、滞后的.数据模式是在数据的基础上,根据主体需求逐步演化出来的.

(2) 数据操作.由于传统的数据管理技术具有模式优先的特性,使得数据操作基于严格的数据操纵语言,操

作结果是准确的、完整的.而在数据空间中,没有严格的数据模式,数据关系是根据主体需要逐步建立的,因此数据操作具有 Best-effort 的特性,即查询或搜索结果不一定是最优的,可能是次优的、近似的.

(3) 数据类型.数据空间的数据来自多个不同的数据源,数据格式多种多样.一个数据空间中可能包含关系表、文本、电子邮件、图像、音频、视频等多种异质的数据.而在传统的关系数据库中,数据源单一,数据格式就是关系表,支持的数据类型也是有限的预定义的数据类型.

(4) 数据关联.数据空间中数据关联是基于对象的,即任何对象之间都可以建立关联,只要这种关联对数据空间主体是有用的.因此,数据对象之间关联是复杂的、动态的、演化的.而传统的数据管理技术,数据关联建立在表一级,这种关联往往是稳定的,而且类型也相对单一.

(5) 构建方式.传统数据库管理系统的构建往往是一步到位的,即通过分析相应的需求,设计出数据库模式,并在较长时间内保持稳定,这是一种 pay-before-you-go 的集成方式.而数据空间的构建是一种 pay-as-you-go 的集成方式,这是一种基于用户需要的演化集成方式,只有当用户认为必要的时候,才会将对象保存到数据空间中,才会在对象之间建立关系.相对于传统的集成系统而言,这种数据管理方式前期成本比较低,也更为实用.

数据空间的特性导致其数据计算也不同于传统的数据管理技术.演化是数据空间的一种重要的计算形式,指的是数据空间系统会随着时间以及应用的变化而不断自我进化.演化的目的是更好地满足主体的需求.Pay-as-you-go 特性降低了构建数据空间的前期代价,但由于缺乏模式对数据关系的刻画,只能提供低水平的服务,所以数据空间需要通过演化进一步强化数据关系,提高操作效率,更好地满足用户需求.

1.2 数据空间与数据集成

数据集成的提出是为了解决分散、异构数据的共享与管理问题.数据空间本质上也是针对这一问题,但在集成对象、集成方式等方面,又与传统的数据集成技术不同.图 2 对目前的数据管理技术进行了总结分类[2,4].横坐标 x 表示系统对语义的要求,随着 x 的增大,对语义的要求逐步降低,对数据模式的依赖程度也逐步降低;纵坐标 y 表示系统对更新操作的支持,随着 y 的增大,系统对数据更新的支持逐步降低.

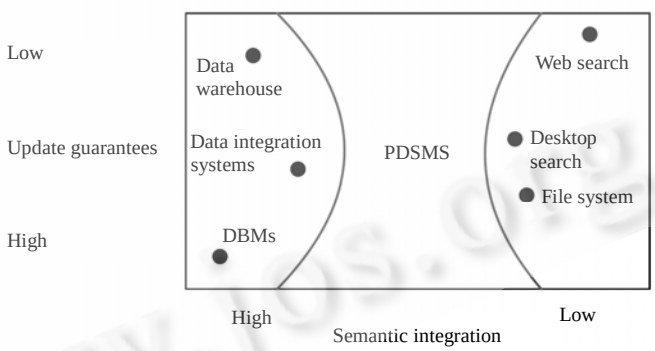

Fig.2 A space of data management solutions

图2 数据管理策略分布

由图 2 可以看出,传统的 DBMS 是一种模式优先的数据管理方式,有严格的数据模式,通过事务处理和并发控制支持数据更新和数据一致性,数据操作基于严格的数据操纵语言,需要详细了解数据模式及其关联.这类系统目前主要应用于数据模式相对稳定且能提前确定的领域.

对于桌面搜索引擎、文件系统和 Web 搜索引擎,不需要严格的数据模式,也不能集成复杂的语义信息.查询操作只能通过简单的关键字匹配实现,操作非常简单,同时,语义信息的缺乏使得数据查询能力比较弱.在数据管理方面,对于用户来说,对公共 Web 数据空间不必进行任何管理,对文件系统,仅需要进行目录划分、文件备份等简单的管理操作;对于桌面搜索引擎,也需要借助其提供的服务进行参数设置,建立索引.

传统的数据集成系统需要有固定的数据模式,然后将不同数据源的信息通过统一接口集成起来,从而实现信息的共享.与 RDBMS 相似,传统集成系统也是一种 Schema-first 的数据管理技术,但是对数据更新的支持程度

低于 RDBMS,数据仓库对数据更新的支持程度则更低一些.

数据空间管理系统(DSMS)希望在两种极端的数据管理技术的中间区域建立一种新的数据管理模式,使得对数据模式的依赖程度不像传统 RDBMS 那么强,也不像桌面搜索引擎那样没有数据模式.用户操作简单,并且能够实现比文件系统和桌面搜索引擎更有效的查询功能.从多数据源方面考虑,数据空间与集成系统相似,但两者仍然有很大的不同.表 2 对数据空间和传统的集成系统进行了比较.

Table 2 Comparison of dataspace and data integration system

表2 数据空间与集成系统的比较

	Dataspace system	Integration system
Data model	Graph model, loose schema	Relation model, schema-first
Data object	Multi-Source of various domain	Multi-Source of single domain
Data storage	Distribute, co-existence	Centralized, integrated
Construction approach	Pay-as-You-Go	Pay-Before-You-Go

由表 2 可以看出,数据空间逻辑上是一张图,没有严格的数据模式,数据操作针对多数据源,数据分布存储,其构建是一个 pay-as-you-go 的过程.而传统的集成系统一般基于关系模型,具有严格的数据模式,虽然是多数据源,但大都来自同一个领域,其构建是一种 pay-before-you-go 的方式.因此,数据空间既不同于 RDBMS、文件系统、桌面搜索引擎,也不同于传统的集成系统.

1.3 数据空间的不确定性问题

异构多数据源特性不可避免地为数据空间带来不确定性.这种不确定性可以分为 3 个层次(如图 3 所示),即数据本身的不确定性、模式匹配的不确定性和查询处理的不确定性[5].

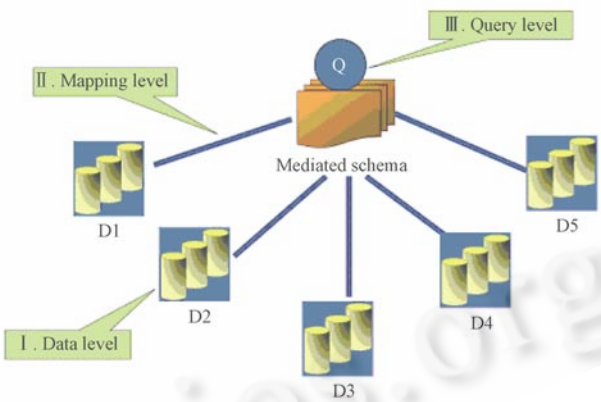

Fig.3 Uncertainty of dataspace

图 3 数据空间中的不确定性问题

首先,数据本身是不确定的.数据空间面对的是多种多样的数据,有些数据本身就具有不确定性,例如,通过信息抽取(information extraction)技术以自动方式从文本或者半结构化的数据源中抽取的数据.由于抽取技术所限或数据源本身的原因,这些数据通常是不准确的.还有一些数据是从 Web 数据源中抽取的.因此,数据空间无法保证所抽取数据的可靠性和实时性.

其次,模式匹配(也称为语义映射)是不确定的.不同数据源对应不同的数据模式,需要与全局模式建立语义映射.这种语义映射具有不确定性,例如,全局模式中的地址字段可能对应局部模式的家庭地址,也可能对应邮箱地址,需要根据具体数据的特点来确定.文献[5]对模式匹配中不确定性的表示、计算、应用进行了系统的讨论.通过建立局部模式和全局模式的语义映射,一方面可以将局部数据源的数据按照全局模式的要求集成到数据空间中,另一方面也可以将用户提交到数据空间上的查询转换为具体数据源上的查询.

第三,查询的不确定性.数据空间的查询通常都是以关键字的方式提交的.这种查询方式不同于传统的结构化查询,语义表达能力比较弱,其本身存在着不确定因素:一是关键字表达的查询内容不确定,用户很难通过关

键字清楚地表达自己的真实意图,系统通常需要将关键字查询转化为一些可能的结构化查询,提交到具体的数据源,这一转化过程也是不确定的;二是查询结果也不确定,关键字查询返回的结果可能很多,究竟哪些结果是用户真正想要的,需要系统对查询结果给出评价.

不确定性为数据空间技术研究带来了挑战,首先是不确定性的表示问题,其次是基于不确定性的计算问题.对数据空间模型、查询、更新、演化、索引和存储技术的研究,都必须要考虑到数据空间的这一特性.

2 数据空间相关技术研究

作为一种新型数据管理技术,数据空间的研究仍处于起步阶段.其本质也是解决数据集成与管理问题,只是集成的对象不再是一个固定的领域,而是与主体相关的所有数据.这一节首先提出一个数据空间系统框架,然后从数据模型、集成更新、数据查询、存储索引、数据演化、系统实现几个方面,对数据空间的相关工作及研究成果进行总结分析.

2.1 数据空间系统框架

作为一种数据管理技术,数据空间同样面临数据模型、查询、更新、存储、索引等传统问题,但由于数据特点不同,使得这些问题的解决不同于传统数据库,同时,数据空间还面临数据演化等新的研究问题.图 4 所示为本文提出的数据空间系统框架,主要包括以下几部分:数据集成引擎、数据空间引擎、数据演化引擎和数据输出引擎.

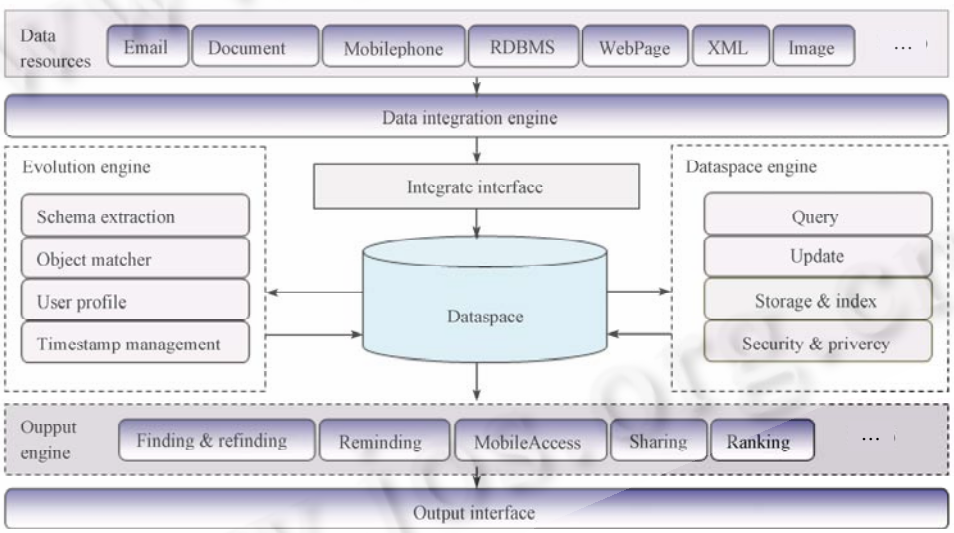

Fig.4 A framework for dataspace integration and management

图 4 数据空间集成与管理框架

(1) 数据集成引擎(data integration engine):负责管理数据源,包括监控数据空间内部和外部的变化,进行数据集成和数据更新.数据抽取是数据集成的第一步,通过包装器(wrapper)实现,每个包装器对应不同的数据源,负责对特定格式的数据对象进行信息抽取和标识,并通过模式匹配确定数据对象在数据空间中是否已经存在,完成数据集成操作.例如,当需要集成一篇 PDF 格式的论文时,相应的 Wrapper 会自动将论文题目、作者、摘要、参考文献等信息抽取出来,用这些信息标识该数据对象.目前在 IR(information retrieval)领域有大量关于信息抽取[6−8]、分类[9]以及模式匹配方面的研究工作可以借鉴.

(2) 数据空间引擎(dataspace engine):负责数据的存储、索引、访问、查询、安全等,是数据空间系统框架的核心部分.目前研究工作主要围绕数据空间概念[10]、模型[11,12]、索引[13]技术.这一部分将是未来数据空间研究工作的重点.

(3) 数据演化引擎(evolution engine):包括数据模式的抽取、数据关系的发现、数据重要性及相关性的自适应计算、索引的自动优化等,其目的是使数据空间的操作更加高效.主体的需求是数据空间演化的动力,因此,演化引擎也包括主体特性的提取以及形式化表示.

(4) 数据输出引擎(output engine):建立数据空间的目的是实现数据的高效访问.由于数据的异构性和访问方式的多样性(可以通过手机等移动设备访问个人数据空间),查询结果的输出和展示方式也与传统的数据库技术不同,查询结果往往包括不同类型的数据.如何通过数据共享技术、高效的数据查询算法和排序算法提高输出效率,如何进行界面设计等,也是数据空间研究所面临的问题.

2.2 数据空间模型

数据模型包括数据结构、数据操作和数据一致性,是数据管理技术的基础和核心.数据空间模型需要能够概括数据空间的特点,提供高效的数据服务.数据库研究者对于数据空间模型从基本概念、数据特点、数据表示等方面进行了研究.

(1) 基于关系模型的数据统一表示.在这种方法中,人们试图基于关系模型面向多种数据源建立一个全局的数据视图,实现对数据的统一访问.全局视图独立于各数据源.Information Manifold[14],SEMEX[15],Haystack[16]和 MyLifeBit[17]都是采用这种方式.这种方式可以利用关系数据库中成熟的技术,但是由于关系数据库自身对模式的严格依赖以及对关键字查询支持的局限性,使得这种数据空间模型不能很好地适应数据空间管理的需要.

(2) 以 XML 为代表的半结构化数据描述.XML 是最重要的半结构化数据描述方法之一,也可以用来实现 schema-later 的集成方式.因此,借助 XML 刻画数据空间也是很自然的方法.由于 XML 提出是针对异构数据的交换,因此,尽管 XML 可以用来描述数据,但是还很难成为一种刻画数据的逻辑模型[18],而且如果不结合其他技术,基于树形结构的 XML 在对基于图的操作实现方面也存在不足.因此,人们更多地将 XML 作为一种实现的方法,而不是用它来刻画数据空间的逻辑模型.

(3) iDM 数据空间模型.这是 Dittrich 等人提出的数据空间模型[12].该模型的主要特点包括:用图刻画数据空间;提出了一种统一资源视图的概念和形式化表示方法,能够实现对各种数据类型(如文档、目录、关系表、XML 文档、数据流等)的统一表示;突破了数据对象和文件系统的边界,将对象内部数据和外部数据统一表示,例如,一个 XML 文件可以表示为一个叶子节点,也可以作为中间节点,继续细化表示其内部的数据信息;设计实现了一个查询语言 iQL,并实现了一个原型系统 iMemex[4].

此外,Franklin[2]提出用带标签的图刻画数据空间,数据空间底层核心应当支持多种数据模型,如关系模型、XML 数据等;Zhuge[19]提出了资源空间模型 RSM(resource space model)的概念和理论体系,基于 RSM 对信息资源进行有效的分类和管理.数据空间管理的是与主体有关的所有数据信息,数据分类也是其面临的主要问题之一,如果利用 RSM 刻画数据对象之间的这种分类关系,就可以使图的复杂性降低,从而提高操作效率.此外,在数据表示、数据约束、数据操作方面,仍有许多问题需要进一步研究.

(1) 数据表示.时间特性是数据空间的重要特性,也是数据空间演化和关联查询的重要依据.在个人数据空间中往往需要通过时间关联搜索某个数据对象,例如查询在海南开会期间收到的邮件.数据空间不但要保存对象当前的状态,而且有时还需要记录其历史信息和变化轨迹.如何有效地表示数据空间对象的时间属性,仍是一个需要研究的问题.此外,数据空间不但包括文本、图片、关系表等简单数据,而且包括任务、服务、知识等复杂数据,因此,数据空间模型需要能够刻画这些复杂的数据信息.

(2) 数据约束.数据空间本质上有 schema-later 的特性,但并不是 no-schema,否则就会大大降低数据空间的查询能力.这样,数据空间的数据约束就应当是一种弹性的、不严格的方式.如何形式化这种数据约束方式以及如何量化表示这种数据不一致性,需要进一步研究.

(3) 数据操作.由于数据空间中的各个数据源分布在不同位置,并且可能各自有一套数据操作机制,数据空间要求主体能够随时随地便捷地对数据进行访问,这使得数据查询和更新操作更加复杂.

从以上分析可以看出,虽然对数据空间模型的研究已经有一些工作,但是还有很多问题需要解决.有些概念

也需要进一步澄清,例如,主体作为数据空间的要素之一,在数据空间模型中的具体地位和作用是什么,如何将主体的特性、地位等数据信息形式化等.

2.3 数据空间集成与更新

数据空间的构造有两个途径:一个是数据空间集成,通过集成将新的数据对象保存到数据空间;另一个是数据更新.图 5 所示为个人数据空间集成框架,包括两部分,一部分是数据集成引擎,负责新数据对象的集成和原数据对象的更新,它包括包装代理、数据对象识别、相关性评估几个部分;另一部分是数据监控引擎,负责监控数据空间内部和外部的变化,以支持自动的数据集成.

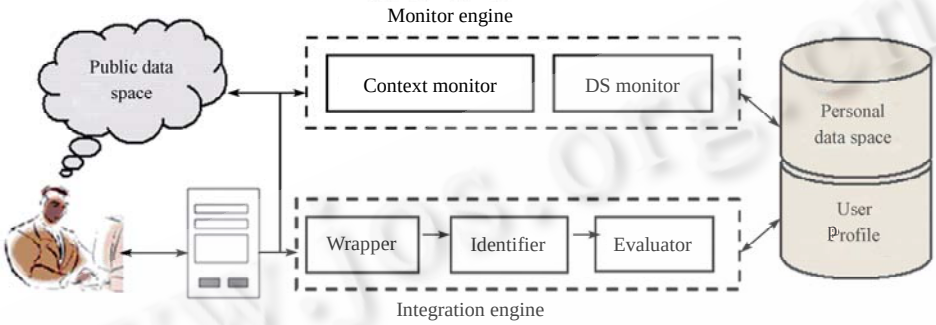

Fig.5 A platform for personal dataspace integration

图 5 个人数据空间集成框架

(1) 数据包装代理(wrapper).由于数据空间包括多个不同类型的数据源,因此,不同的数据格式需要不同的数据抽取技术.其主要任务是从特定数据项中抽取特征信息并按照数据空间要求进行形式化表示.

(2) 数据对象识别(identifier).该模块包括两项内容:一是将待处理的对象与数据空间现有数据模式比较,确定其对应的数据模式.这一步需要借鉴模式匹配和实体识别[20,21]的相关工作.二是将待处理的对象与数据空间现有对象进行匹配,确定该数据对象在数据空间中是否存在,以确定下一步需要执行的任务.

(3) 数据相关性评估(evaluator).在保存之前通过评价机制(evaluator)对数据对象与主体的相关性进行评估,如果数据对象与主体相关,则将该对象集成到数据空间,否则放弃该数据对象.

数据空间监控引擎包括两部分.一部分监控数据空间外部数据源的变化(context monitor),例如在桌面建立了新的文档、接受了新的邮件等;另一部分监控数据空间内部数据的变化(DS monitor),例如对某个文档进行了修改.监控引擎监控到数据变动后,通知集成引擎完成数据集成.

关于数据集成[22]已有很多研究工作,涉及体系架构[23,24]、数据仓库[25]、企业数据集成[26]、数据共享[27]等多个方向.这些数据集成技术的一个共性是基于模式优先的传统数据管理技术,在很多应用场景下,这种数据管理技术不能很好地适应面向主体数据集成的要求.与传统的数据集成相比,数据空间面对的是邮件、文档等非结构化、语义结构复杂、分布存储的数据.数据空间集成不仅包括数据对象的集成,而且包括数据关系的集成.Dong[15]提出基于数据空间中的数据关联,实现面向用户任务的多数据源数据集成.Jens Dittrich[4]提出基于包装器自动提取数据对象之间的语义关系进行自动集成.桌面搜索引擎也是一种数据集成方式,自动监控桌面数据的变化并尽可能快速地对索引进行更新.虽然目前针对数据空间集成提出了一些策略和方法.但仍需要进一步对数据空间集成模型、算法进行研究,以提高集成效率和数据质量.分布存储和数据共存的特性要求数据空间系统能够自动高效地发现其内部和外部数据的变化,数据空间集成的前提要求判断数据与主体的相关性,这些都是数据空间集成面临的研究问题.

数据更新是改变数据空间的另一种方式.在传统的关系数据库中,数据更新通过数据操纵语言实现.主体对需要进行的更新是明确知道的.而且关系数据库有完备的机制保证数据更新的并发操作以及数据一致性.数据空间则不同,用户对数据的更新可以通过其他工具独立完成,例如,利用 Outlook 接收新邮件、通过 Word 对文档内容进行修改.这就需要数据空间能够自动检测数据对象的变化并完成更新操作.此外,由于数据空间中的数据

分布存储,而且用户对数据的改变具有随意性,使得数据空间数据一致性问题更加突出.数据空间的一致性问题主要来自数据更新、多数据版本、数据扩展等方面.

(1) 数据更新引起的一致性问题.在个人数据空间中,一般是单用户使用,数据的并发问题并不突出.但是在企业数据空间中,可能涉及到多个用户对同一个数据项的修改,这时就要用到并发控制、事务处理等技术来保证事务的一致性.由于数据空间管理的数据来自多个不同的数据源,因此,其事务处理机制与传统的关系数据库会有所不同.例如,数据更新请求需要重写并分发到不同的数据源,数据源自身的问题或者网络原因都可能导致事务无法正常完成,因此,需要针对数据空间特点研究相应的事务处理策略.

(2) 数据空间中的版本控制.版本控制是数据空间中的一个非常重要的问题,许多数据信息,如档案资料、个人文档往往需要保留多个版本,一些个人信息也需要保留不同时期的版本,如联系方式、工作单位信息等.需要根据数据空间的特点来研究适应的版本控制策略.

(3) 数据扩展.数据空间中的数据是高度异构的,但是数据操作要求有相对一致的格式,这时就需要按照全局模式对不同数据源的数据进行一定程度上的扩展,使其满足操作需要.再有,如果两个数据源在各自的局部环境中使用正常,但是到了数据空间这个全局环境中则可能会遇到问题.例如,一个电话号码在局部环境中可能是正确的,但是在数据空间中就可能与其他电话号码重复,这时就需要进行扩展,加上相应的区域信息.此外,用户提交的数据操作可能不被某些数据源支持,这时也需要数据空间扩展来提供.

由于分布存储、多数据源等特性,数据空间更新需要转换为对各数据源的操作.数据源自身对数据更新的支持以及所采用的安全策略,都会对数据空间的更新操作带来影响.此外,用户经常需要随时随地进行数据空间的更新,例如,通过手机更新通信录、随时随地访问并修改某一篇技术文档,这些都会带来数据一致性和版本控制问题.

2.4 数据查询处理

传统的数据查询技术主要有两类:一类是基于关键字的查询,另一类是基于数据模式的结构化查询.分布存储、多数据源和模式松散的特点使得数据空间的查询不同于传统的数据查询.由于缺乏语义信息,关键字查询的能力和效率比较低,依赖数据模式的结构化查询在 Schema-later 的数据空间中也不适用.因此,在数据空间中,需要将关键字查询和结构化查询结合起来,支持更加复杂、灵活的查询需求.

数据空间应当支持用户查询数据空间的任意数据.多数据源特性要求数据空间支持查询转换;对于具有严格数据模式的数据源,要求数据空间能够进行结构化查询;数据空间还需要支持元数据查询,即除了返回用户查询结果以外,有时还需要返回用户结果所来自的数据源以及如何计算得到的.因此,数据空间的查询比传统的方式更为复杂.涉及查询优化、查询转换、查询接口等多方面.

基于图的查询优化技术.数据空间可以看作一个以数据对象为节点,以数据关系为边的图,基于图的数据搜索技术可以应用到数据空间查询.这方面的工作有很多,涉及图索引结构[28,29]、图的相似性度量[30]、图搜索优化[31]等多个方面.关联查询是数据空间中重要的查询技术.图中每一条链上的节点都可以认为具有相关性,但关联程度会因节点之间的距离或权重而不同.这就为基于数据关联的查询带来了不确定性.数据空间中查询成本的构成主要有两种,一种是数据源分布造成的数据传输成本,另一种是查询计算成本.通过对图的优化可以降低计算成本.虽然目前已有一些基于图的查询方法,但是,如何将这些方法应用到数据空间中,仍是需要进一步研究的问题.由于数据源分布造成的传输成本,需要采用数据缓存或多副本的方式加以解决.

通过查询转换实现多数据源的查询.由于数据空间中的数据多样性,用户需要查询的数据可能分布在不同的数据源上,所以,对数据空间的查询将不可避免地被转换为对多个数据源的查询.当前有大量关于集成环境下查询转换的工作[32,33],还有一些工作描述了查询转换复杂度和数据模式转换语言之间的相互依赖关系[34],但是,这些工作的基础都是基于模式的集成系统,对于数据空间这种松耦合的情况有借鉴意义,但并不完全适用,需要根据数据空间的特点对算法进行调整.

查询接口.查询接口包括查询方式和查询语言.数据空间应当支持结构化查询、关键字查询、导航、浏览等多种查询方式.弱化模式的特点使关键字查询和导航查询成为重要的查询方法.查询技术的实现基于数据空

间模型和具体的数据组织方式,不同的数据组织方式依赖于不同的技术.如果数据空间通过 RDBMS 或 XML 实现,就要借鉴基于关系数据库的关键字查询[35,36]和基于 XML 的关键字查询[37]的相关工作.传统 RDBMS 下的 SQL 查询语言是结构化查询语言,在 schema-later 的数据空间中并不适用,但是可以借鉴其语法形式.iDM[12]提出了一种数据查询语言——iQL,基于 iQL 可以执行简单的关键字查询,也能根据其提供的语法进行语义查询.

查询结果排序和显示.不确定性使得数据空间中的查询大部分是 Best-effort 查询,即查询结果可能不是最优的.查询结果的展示需要满足以下 3 点:一是能够以相对统一的方式显示不同类型的数据;二是能够进行合理的排序;三是支持用户在查询结果的基础上进一步执行更精确的查询.目前有很多排序算法,例如,如何对 XML 进行查询排序[38],以及如何在集成的环境下寻找正确的答案[39].关于 Web 搜索引擎结果网页的排序算法也有很多工作,这些工作都可以借鉴到数据空间中.此外,由于数据空间是与主体对应的,因此,主体的行为特性也是影响查询结果的重要因素[40].

数据查询是数据空间方面重要的研究问题.尽管目前有很多查询方法和排序算法,但是专门针对数据空间这种异构数据查询的研究工作还比较少.分布存储的数据特性使得数据空间查询代价的计算与传统的关系数据库不同,查询优化的策略也有区别.

2.5 数据存储和索引

数据的组织存储与索引的最终目的是服务于数据操作,由于在数据空间中数据查询与更新操作与传统数据库不同,因此数据空间需要采用与之相适应的数据存储与索引技术.分布式存储是数据空间的重要特征.图 6 显示了数据空间的存储结构,数据空间中的各种类型数据分布存储在不同的设备上,用户通过全局视图实现对所有数据的访问,每个数据对象在数据视图中有对应的描述信息,这些信息包括数据自身属性、与主体的关系、物理存放位置、访问情况等,这些数据信息的形成也具有 pay-as-you-go 的特性,根据主体的需求确定需要记录的信息.

从图 6 可以看出,数据空间中的数据主要包括两类:元数据和内容数据.元数据是指数据描述信息,例如数据项名称、数据项的基本属性、数据项之间的关联关系等;内容数据是指数据项的内容,如文档所对应的文件.元数据信息存放在数据空间的系统区域,以全局视图的形式存在.用户通过该数据视图对数据内容进行操作.用户提交查询后,查询处理引擎首先通过全局视图找到要查找的对象及其物理位置,然后将数据取出并返回.内容数据一般就存储在其原始的数据源中,或由主体指定存放在特定的位置.由于全局视图是数据操作的基础,因此对全局视图的存储与索引非常重要.为了提高数据操作的效率,元数据信息应当存放在读取效率最高的地方,例如用户访问最容易的个人计算机上,也可以在不同位置存放多个副本,这样可以大大提高数据操作的效率,但多版本数据的一致性保证会带来一些额外的开销.

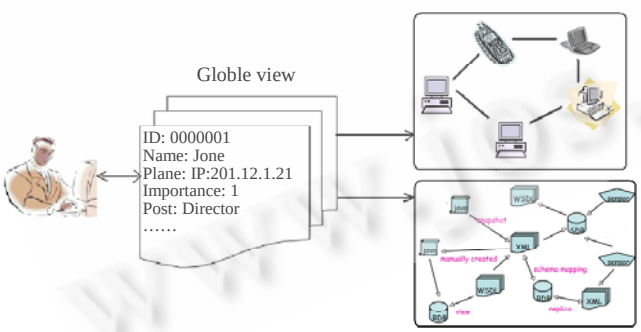

Fig.6 Data storage of dataspace

图 6 数据空间的数据存储

传统数据库以表为存储单位,采用基于行的数据存储技术,其优点是便于基于关系的查询与更新操作.这种数据组织方式对于数据空间并不适用,一方面,数据多样性造成表内容的稀疏,从而导致存储空间的浪费和数据读取开销的增大;另一方面,当新的数据项不能匹配已有数据模式的时候,需要对数据模式进行调整,这样会造成数据存储模式的调整和应用软件的修改,从而增加数据操作的成本.针对这种复杂多变的数据存储,有许多研究工作.1985 年,Copeland 和 Khoshafian[41]就提出了一种垂直的数据存储策略,用三元组进行数据的存储.三元组的内容分别是〈对象 ID,属性名,属性值〉.这种数据存储方式最大的优点就是能够适应数据模式频繁变化的需要,数据项属性的增加不再引起数据模式的变化.这种数据组织方式适合商务领域商品信息千差

万别的数据特点,在电子商务领域得到了成功的应用[42].由于主体信息的个性化和多样性,这种数据组织方法可以应用在个人数据空间中.类似的工作还有 XML 方面的,XML 既是一种数据描述方式,又是一种数据组织方式.其基于对象的数据组织方式与数据空间的要求也是匹配的,但是 XML 在复杂数据操作方面仍有一定的局限性.

由于数据空间是新的数据管理技术,因此,对于数据空间的组织存储及索引技术的研究并不多.Dong 在 SIGMOD2007 发表的论文[13]比较系统地研究了数据空间索引技术,将关键字索引和结构化索引技术结合起来,基于传统的倒排表索引方法,在索引中考虑了数据属性、数据项之间的关系、数据模式的层次结构以及数据模式中的同义词.目前关于非结构化数据的索引[43,44]、图的索引[45]、XML 数据的索引[46]都有一些研究工作.不同的索引技术适用于不同场景.

索引的更新是一个挑战性的问题,由于数据源对自己的数据项有控制权,数据项的变化可以通过数据源自身来完成,这样,数据空间并不能及时发现,这就存在索引的及时更新问题.此外,个人数据空间和企业数据空间的管理模式不同,存储策略和索引策略也会不同.

2.6 数据空间演化

数据空间演化的原因在于数据空间主体的发展和变化,一方面,主体不断将自己需要的数据集成到数据空间中,另一方面,数据空间原有的数据对主体的作用也会发生变化,这就需要数据空间通过演化来适应这种变化.数据空间的演化实质上就是通过自适应技术实现数据空间的自调优.目前直接针对数据空间演化的工作还比较少,但可以借鉴其他领域关于数据演化的研究工作[47,48].文献[47]研究了 Web 数据的动态特征,例如变动的频率、幅度,在数据空间中同样存在这样的问题.数据源及数据项的访问频率、变动特征是数据空间演化的重要依据.文献[48]主要研究了网络的演化问题.这些演化的概念和方法都可以借鉴到数据空间中.主体特性也是数据空间演化的重要依据,在个人数据空间中,主体行为特征的分析有助于进行数据存储和索引优化.目前在主体行为分析和个性化特征[49]提取方面有很多研究工作可以借鉴到数据空间中.

从演化内容上,数据空间演化可以分为数据演化、关系演化和演化计算.数据演化包括数据对象及其属性的增加、修改和删除.关系演化是指对数据空间对象关联关系的增加、修改和删除.例如,当我们将一篇论文保存到数据空间中时,如果这篇文章的作者已经存在于数据空间中,就需要在新的数据对象和已有数据对象之间建立关联.演化计算主要是指数据计算策略和方法的优化,当数据空间的数据及其关系发生变化以后,会引起数据存储、索引以及查询优化算法的改变,从而满足用户操作的需要.

从实现的角度考虑,数据演化分为手动和自动两种方式.手动方式是指当用户保存数据或意识到数据之间的关联时,通过手工方式实现,如文献[50]提出的一种用户建立数据关联(iTrail)的方法;自动方式是另一种演化方式,当发生数据更新时,数据空间系统自动根据预先设定的规则完成数据演化.这两种方式是互补的.自动演化方式需要一定的数据量和操作历史记录为基础.只有当数据空间的数据量达到一定程度才能进行自动演化.

从演化的时机考虑,数据空间演化可以分为基于时间的演化和基于事件的演化.基于时间的演化是指数据空间根据预先设定的任务,定时启动演化服务,例如,可以设定每天定时按照用户访问频率调整数据对象的重要性;基于事件的演化是指当特定的事件发生时,执行相关的演化服务,例如,将一个新的数据对象保存到数据空间时,数据空间自动标识该对象并进行数据强化.

数据空间演化和数据空间的数据量、数据质量都有关系.当数据空间的数据量很少时,数据之间的关联也比较简单,还无法总结出演化的规则,随着数据量的增大和数据关系复杂性的提高,才有了数据空间演化的基础和必要性.关于数据空间演化,目前还没有系统化的研究工作.数据空间演化的概念定义、演化模型是首先需要解决的问题.

2.7 数据空间原型系统

目前,数据空间技术的研究不仅包括概念、模型、查询、索引等基础理论和算法,而且系统实现及应用研究也日益引起关注,主要的研究工作还是围绕个人数据空间管理.人们常用的个人数据管理工具主要是文件系统和桌面搜索.特别是桌面搜索日益成为一个重要的研究与应用领域,桌面搜索本质上也可以看作一种数据集

成方式,通过建立全文索引,为用户提供与 Web 搜索相似的查询功能.结果排序是桌面搜索需要解决的重要问题.为了对结果进行有效排序,人们提出了一些算法,例如,根据文件访问顺序、文件名称相似性、文件存放位置等信息在数据对象之间建立连接,由此将 PageRank 算法引入到桌面搜索结果排序[51,52].但是,数据对象之间的这种连接仍是一种抽象的关系,不同于数据对象之间的具体关系,如文章与作者之间的关系、文章之间的引用关系.基于数据空间的个人数据管理系统不仅集成个人数据对象,也集成数据对象之间的关系.代表性的原型系统有 Dittrich 等人开发的 iMemex 和 Dong 等人开发的 Semex.表 3 对比分析了它们与文件系统、桌面搜索等个人数据管理工具的区别.

Table 3 Comparison of systems for personal data management

表 3 个人数据管理系统比较

	File system	Desktop search	iMemex	Semex
Model basis	Hierarchical structure	N/A	Graph	Graph
Schema model	Tree	N/A	iDM	Class-Based mediated schema
Relationship query	No	No	Yes	Yes
Keyword query	Yes	Yes	Yes	Yes
Data integration	N/A	Text-Based index	Wrapper-Based	Association-Based
Semantic integration	No	No	Yes	Yes
Index	N/A	Full-Text index	Full-Text index, vertical view	N/A
Dataspace-Oriented	No	No	Yes	Yes

从表 3 可以看出,与文件系统和桌面搜索相比,基于数据空间的 iMemex 和 Semex 原型系统具有以下特点:将个人数据空间视为图结构,并基于图结构刻画个人数据空间的数据及其关系;面向多个不同的数据源;通过在数据对象之间建立关联实现基于语义的数据集成;提供基于数据关系的查询.这样,用户就可以更加高效地管理个人数据空间.例如,利用桌面搜索引擎可以实现关键字查询,但很难实现如下的查询:论文 Indexing Dataspace 的作者最近两年所发表过的文章;参加 SIGMOD2007 期间个人写的关于数据空间的一篇文章;关于某个附件的一封邮件的发件人地址、电话等挡案信息,等等.用户要完成此类查询需要手工执行很多操作.数据空间可以较好地支持此类查询.通过构造个人数据空间,用户可以实现复杂的语义查询,实现随时随地对个人数据的快速访问,可以方便地备份个人重要数据,保持异地数据同步.通过构造群组数据空间,群组成员可以方便地进行信息的共享与交流.

在系统架构以及功能方面,这些原型系统各有特点.iMemex 基于数据空间模型 iDM,针对多个数据源,实现基于 wrapper 的数据集成,通过定义查询语言 iQL 可以方便地进行查询.Semex 系统对个人数据集成中参照协调[53]问题给予了关注.例如,同一个数据对象往往具有不同的表示形式;不同类别的数据对象之间往往会发生关联;同一对象不同版本的协调性.此外,还有其他一些个人数据空间管理系统,如 MyLifeBit,HayStack,SIS 等.MyLifeBit 侧重于对文本和多媒体数据的集成与管理;Haystack 侧重于通过分析用户个性化信息对个人数据进行标注;SIS 着眼于基于现有的文本资源,建立简洁、高效的查询接口.群组数据空间和企业数据空间的相关工作较少,相关的原型系统也较少.在数据空间的实现和应用技术研究方面,还有许多问题需要解决.传统数据库中的数据一般集中存放,软件也是非常庞大的集中式的服务器系统.数据空间管理系统应该采用什么样的系统架构仍然是一个需要研究的问题.

3 数据空间研究现状及挑战性问题

作为一种新型数据管理技术,数据空间日益引起研究界和工业界的关注.目前的研究工作主要从两个方向展开,一是针对数据空间概念、模型、查询、索引等理论与算法;二是以个人数据管理为应用背景,进行个人数据空间管理及应用技术的研究.对于数据更新、数据演化、企业数据空间的研究,目前相关工作还比较少.总体来说,数据空间技术的研究还处于起步阶段,许多问题有待进一步解决,主要的挑战来自以下方面:

(1) 数据空间的模型与数学基础.完备的数学模型是数据管理技术的基础,关系模型和关系代数使关系数据库有了坚实的理论基础,得到了巨大的发展.数据空间具有自己的特性,能否基于这些特性建立完备的数学模型是亟需解决的问题.

(2) 数据清洗与数据质量.数据优先、淡化模式的特点必然使数据空间的数据质量下降,造成数据操作效率低下.数据清洗、数据强化是提高数据质量的方法,但是这类操作的代价往往会比较大,需要研究高效的算法和适用的策略.

(3) 数据空间的不确定性.不确定性问题正逐渐成为热点研究问题.在数据空间中,从数据到模式、pay-as-you-go、best-effort、数据演化等基本特征使数据空间具有与生俱来的不确定性.不确定性数据的计算、不确定数据的表示等都为数据空间研究带来了挑战.

(4) 数据空间的个性化和多样性.主体的个性化使得数据空间具有个性化特征.个人数据空间和企业数据空间面对的问题有很大区别,计算方法也不同,例如采用的存储策略、索引方法等.即使针对个人数据空间,也因为主体的不同呈现出多样性.

(5) 数据空间管理系统(DSMS).数据空间需要什么样的管理系统,是像传统的关系数据库那样以集中的、庞大的系统软件形式呈现出来,还是以服务的方式分布在不同的机构和位置,这就涉及数据空间系统架构.此外数据空间的实现基于信息抽取、模式识别、人机交互、移动数据管理等众多领域的研究成果.

4 结 论

数据空间是一个新的研究领域.在数据模型、数据操作、数据索引方面已有一些相关工作,但还有许多关键问题没有解决,这为国内的数据库研究者提供了机遇,也提出了挑战.数据空间不仅为我们提出了一种新的数据管理方法,而且代表了一种新的对待数据的理念.数据空间技术的研究涉及一些深层的理论问题,研究成果可以直接应用于企业和个人数据管理.因此,数据空间研究具有重大的理论价值和应用价值.

References:

[1] Meng XF. From Database to Dataspace, From Enterprise to People. Annual Report of WAMDM Lab., School of Information, Renmin University of China, 2006. 2−7 (in Chinese). http://idke.ruc.edu.cn

[2] Franklin M, Halevy A, Maier D. From databases to dataspaces: A new abstraction for information management. SIGMOD Record, 2005,34(4):27−33.

[3] Jones W, Bruce H. A report on the NSF-sponsored workshop on personal information management. Seattle, 2005. http://pim.ischool.washington.edu/pim05home.htm

[4] Blunschi L, Dittrich JP, Girard OR, Karakashian SK, Salles MAV. A dataspace odyssey: The iMeMex personal dataspace management system. In: Proc. of the 3rd Conf. on Innovative Data Systems Research (CIDR 2007). 2007. 114−119. http://www.cidrdb.org/

[5] Dong X, Halevey A. Data integration with uncertainty. In: Proc. of the 33rd Int'l conf. on Very Large Data Bases (VLDB 2007). New York: ACM Press, 2007. 687−698.

[6] Zhao HK, Meng WY, Yu C. Automatic extraction of dynamic record sections from search engine result pages. In: Proc. of the 32nd Int'l Conf. on Very Large Data Bases (VLDB 2006). New York: ACM Press, 2006. 989−1000.

[7] Cohen WW, Ravikumar P, Fienberg S. A comparison of string distance metrics for name-matching tasks. In: Proc. of the IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03). 2003. 73−78.

[8] Minkov E, Cohen WW, Ng AY. Contextual search and name disambiguation in email using graphs. In: Proc. of the 29th Int'l conf. on Research and Development in Information Retrieval (SIGIR 2006). New York: ACM Press, 2006. 27−34.

[9] Dredze M, Lau TA, Kushmerick N. Automatically classifying emails into activities. In: Proc. of the 2006 Int'l Conf. on Intelligent User Interfaces (IUI 2006). New York: ACM Press, 2006. 70−77.

[10] Halevy AY, Franklin MJ, Maier D. Principles of dataspace systems. In: Proc. of the 32nd Int'l Conf. on Principles of Database Systems (PODS 2006). New York: ACM Press, 2006. 1−9.

[11] Freeman E, Gelernter D. Lifestreams: A storage model for personal data. SIGMOD Record, 1996,25(1):80−86.

[12] Dittrich JP, Antonio M, Salles MAV. iDM: A unified and versatile data model for personal dataspace management. In: Proc. of the 32nd Int'l conf. on Very Large Data Bases (VLDB 2006). New York: ACM Press, 2006. 367−378.

[13] Dong X, Halevy A. Indexing dataspaces. In: Proc. of the 27th Int'l Conf. on Management of Data (SIGMOD 2007). New York: ACM Press, 2007. 43−54.

[14] Levy A, Rajaraman A, Ordille J. Querying heterogeneous information sources using source descriptions. In: Proc. of the 22nd Int'l Conf. on Very Large Data Bases (VLDB 1996). San Fransisco: Morgan Kaufmann Publishers, 1996. 251−262.

[15] Dong X, Halevy A. A platform for personal information management and integration. In: Proc. of the 2nd Conf. on Innovative Data Systems Research (CIDR 2005). 2005. 119−130. http://www.cidrdb.org/

[16] Karger DR, Bakshi K, Huynh D, Quan D, Sinha V. Haystack: A customizable general-purpose information management tool for end users of semistructured data. In: Proc. of the 2nd Conf. on Innovative Data Systems Research (CIDR 2005). 2005. 13−26. http://www.cidrdb.org/

[17] Gemmell J, Bell G, Lueder R, Drucker SM, Wong C. MyLifeBits: Fulfilling the Memex vision. In: Proc. of the 10th ACM International Conference on Multimedia. New York :ACM, 2002. 235−238.

[18] Abiteboul S. On views and XML. In: Proc. of the 18th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems (PODS 1999). New York: ACM Press, 1999. 1−9.

[19] Zhuge H. Resource space model, its design method and applications. The Journal of Systems and Software, 2004,72(1):71−81.

[20] Singla P, Domingos P. Object identification with attribute-mediated dependences. In: Proc. of the 9th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005). LNCS 3721, Springer-Verlag, 2005. 297−308.

[21] Tejada S, Knoblock C, Minton S. Learning domain-independent string transformation weights for high accuracy object identification. In: Proc. of the 8th Int'l conf. on Knowledge Discovery and Data Mining (SIGKDD 2002). New York: ACM Press, 2002. 350−359.

[22] Halevy A, Rajaraman A, Ordille J. Data integration: The teenage years. In: Proc. of the 32nd Int'l Conf. on Very Large Data Bases (VLDB 2006). New York: ACM Press, 2006. 9−16.

[23] Naumann F, Leser U, and Freytag JC. Quality-Driven integration of heterogenous information systems. In: Proc. of the 25th Int'l Conf. on Very Large Data Bases (VLDB 1999). San Fransisco: Morgan Kaufmann Publishers, 1999. 447−458.

[24] Papakonstantinou Y, Garcia-Molina H, Widom J. Object exchange across heterogeneous information sources. In: Proc. of the 11th Int'l Conf. on Data Engineering (ICDE 1995). Dallas: IEEE Computer Society, 1995. 251−260.

[25] Calvanese D, Giacomo GD, Lenzerini M, Nardi D, Rosati R. Source integration in data warehousing. In: Proc. of the 9th Int'l Workshop on Database and Expert Systems Applications (DEXA'98). Dallas: IEEE Computer Society Press, 1998. 92−197.

[26] Halevy AY, Ashish N, Bitton D, Carey M, Draper D, Pollock J, Rosenthal A, Sikka V. Enterprise information integration: Successes, challenges and controversies. In: Proc. of the 25th Int'l Conf. on Management of Data (SIGMOD 2005). New York: ACM Press, 2005. 778−787.

[27] Ng WS, Ooi BC, Tan KL, Zhou AY. PeerDB: A P2P-based system for distributed data sharing. In: Proc. of the 19th Int'l Conf. on Data Engineering (ICDE 2003). Dallas: IEEE Computer Society, 2003. 633−644.

[28] Yan XF, Yu PS, Han JW. Graph Indexing: A frequent structure-based approach. In: Proc. of the 24th Int'l Conf. on Management of Data (SIGMOD 2004). New York: ACM Press, 2004. 335−346.

[29] He H, Singh AK. Closure-Tree: An index structure for graph queries. In: Proc. of the 22nd Int'l Conf. on Data Engineering (ICDE 2006). Dallas: IEEE Computer Society, 2006. 38.

[30] Holder L, Cook D, Djoko S. Substructure discovery in the subdue system. In: Proc. of the AAAI Workshop of Conf. on Knowledge Discovery in Databases. Menlo Park: AAAI Press, 1994. 169−180.

[31] Jiang HL,Wang HX, Yu PS, Zhou SG. GString: A novel approach for efficient search in graph databases. In: Proc. of the 23rd Int'l Conf. on Data Engineering (ICDE 2007). Dallas: IEEE Computer Society, 2007. 566−575.

[32] Halevy AY. Answering queries using views: A survey. VLDB Journal, 2001,10(4):270−294.

[33] Kolaitis P. Schema mappings, data exchange, and metadata management. In: Proc. of the 24th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems (PODS 2005). New York: ACM Press, 2005. 61−75.

[34] Lenzerini M. Data integration: A theoretical perspective. In: Proc. of the 21st ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems (PODS 2002). New York: ACM Press, 2002. 233−246.

[35] Hristidis V, Gravano L, Papakonstantinou Y. Efficient IR-style keyword search over relational databases. In: Proc. of the 29th Int'l Conf. on Very Large Data Bases (VLDB 2003). New York: ACM Press, 2003. 850−861.

[36] Bhalotia G, Hulgeri A, Nakhe C, Chakrabarti S, Sudarshan S. Keyword searching and browsing in databases using BANKS. In: Proc. of the 18th Int'l Conf. on Data Engineering (ICDE 2002). Dallas: IEEE Computer Society, 2002. 431−440.

[37] Shao F, Guo L, Botev C, Bhaskar A, Chettiar M, Yang F. Efficient keyword search over virtual XML views. In: Proc. of the 33rd Int'l Conf. on Very Large Data Bases (VLDB 2007). New York: ACM Press, 2007. 1057−1068.

[38] Guo L, Shao F, Botev C, Shanmugasundaram J. XRANK: Ranked keyword search over XML documents. In: Proc. of the 23rd Int'l Conf. on Management of Data (SIGMOD 2003). New York: ACM Process, 2003. 16−27.

[39] Levy AY, Rajaraman A, Ordille JJ. Querying heterogeneous information sources using source descriptions. In: Proc. of the 22nd Int'l Conf. on Very Large Data Bases (VLDB 1996). San Fransisco: Morgan Kaufmann Publishers, 1996. 251−262.

[40] Qiu F, Cho J. Automatic identification of user interest for personalized search. In: Proc. of the 15th Int'l World Wide Web Conf. (WWW2006). New York: ACM Press, 2006. 727−736.

[41] Copeland GP, Khoshafian S. A decomposition storage model. SIGMOD Record, 1985,14(4):268−279.

[42] Agrawal R, Somani A, Xu Y. Storage and querying of e-commerce data. In: Proc. of the 27th Int'l Conf. on Very Large Data Bases (VLDB 2001). San Fransisco: Morgan Kaufmann Publishers, 2001. 149−158.

[43] Bast H, Weber I. Type less, find more: Fast autocompletion search with a succinct index. In: Proc. of the 29th Int'l Conf. on Research and Development in Information Retrieval (SIGIR 2006). New York: ACM Press, 2006. 364−371.

[44] Cooper BF, Sample N, Franklin MJ, Hjaltason GR, Shadmon M. A fast index for semistructured data. In: Proc. of the 27th Int'l Conf. on Very Large Data Bases (VLDB 2001). San Fransisco: Morgan Kaufmann Publishers, 2001. 341−350.

[45] Chen Q, Lim A, Ong KW. *D*(*k*)-Index: An adaptive structural summary for graph-structured data. In: Proc. of the 23rd Int'l Conf. on Management of Data (SIGMOD 2003). New York: ACM Press, 2003. 134−144.

[46] Rao P, Moon B. PRIX: Indexing and querying XML using prufer sequences. In: Proc. of the 20th Int'l Conf. on Data Engineering (ICDE 2004). Dallas: IEEE Computer Society, 2004. 288−300.

[47] Ntoulas A, Cho J, Olston C. What's new on the Web? The evolution of the Web from a search engine perspective. In: Proc. of the 13th Int'l World Wide Web Conf. (WWW 2004). New York: ACM Press, 2004. 1−12..

[48] Marcel S, May RM, Bonhoeffer S. The evolution of network topology by selective removal. Journal of the Royal Society Interface, 2005,2(5):533−536.

[49] Song X, Tseng BL, Lin CY, Sun MT. Personalized recommendation driven by information flow. In: Proc. of the 29th Int'l Conf. on Research and Development in Information Retrieval (SIGIR 2006). New York: ACM Press, 2006. 509−516.

[50] Salles MAV, Dittrich J-P, Karakashian S.K, Girard OR, Blunschi L. iTrails: Pay-as-You-Go information integration in dataspaces. In: Proc. of the 33rd Int'l conf. on Very Large Data Bases (VLDB 2007). New York: ACM Press, 2007. 663−674.

[51] Chirita PA, Costache S, Nejdl W, Paiu R. Beagle^{++}: Semantically enhanced searching ranking on the desktop. In: Proc. of the 3rd Int'l Conf. on European Semantic Web Conf. (ESWC 2006). LNCS 4011, Springer-Verlag, 2006. 348−362.

[52] Chirita PA, Firan CS, Nejdl W. Pushing task relevant Web links down to the desktop. In: Proc. of the 8th ACM Int'l Workshop on Web Information and Data Management (WIDM 2006). New York: ACM Press, 2006. 59−66.

[53] Dong X, Halevy A, Madhavan J. Reference reconciliation in complex information spaces. In: Proc. of the 25th Int'l Conf. on Management of Data (SIGMOD 2005). New York: ACM Press, 2005. 85−96.

附中文参考文献:

[1] 孟小峰.从企业到个人,从数据库到数据空间.网络与移动数据管理实验室年报,中国人民大学信息学院,2006.2−7. http://idke.ruc.edu.cn

李玉坤(1969－),男,河北冀州人,博士生,高级工程师,主要研究领域为数据空间,个人信息管理.

孟小峰(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 数据集成,XML 数据库,移动数据管理.

张相於(1986－),男,硕士生,主要研究领域为数据空间.