“辽宁科技大学学术著作出版基金”和
“辽宁科技大学人才项目”资助

数据空间中基于语义的实体搜索

杨丹 著

東北大學出版社
Northeastern University Press

“辽宁科技大学学术著作出版基金”和“辽宁科技大学人才项目”资助

数据空间中基于语义的实体搜索

杨　丹　著

东北大学出版社
·沈　阳·

图书在版编目（CIP）数据

数据空间中基于语义的实体搜索 / 杨丹著. — 沈阳：东北大学出版社，2019.10
ISBN 978-7-5517-2304-6

Ⅰ. ①数… Ⅱ. ①杨… Ⅲ. ①数据管理－研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2019)第 226867 号

出 版 者：东北大学出版社
地址：沈阳市和平区文化路三号巷 11 号
邮编：110819
电话：024－83683655(总编室)　83687331(营销部)
传真：024－83687332(总编室)　83680180(营销部)
网址：http://www.neupress.com
E-mail: neuph@neupress.com
印 刷 者：沈阳航空发动机研究所印刷厂
发 行 者：东北大学出版社
幅面尺寸：170mm×240mm
印　　张：9.5
字　　数：171 千字
出版时间：2019 年 10 月第 1 版
印刷时间：2019 年 10 月第 1 次印刷
组稿编辑：曲　直
责任编辑：孙德海
责任校对：刘乃义
封面设计：潘正一

ISBN 978-7-5517-2304-6　　定　价：52.00 元

前　言

随着大量有价值的、公开的各种数据资源的不断出现，如作者文献库DBLP、Citeseer等，电影数据库IMDB，维基百科、Freebase、DBpedia等知识库，社交媒体数据（微博、微信等），这些数据资源构成了异构的、松散结构的、丰富的数据空间。这些异构数据阻碍了数据空间中各种资源间的交互以及数据集成、资源搜索等进一步应用。由此产生了一种很简单、有效地对数据空间数据建模和搜索的方法，即以实体作为基本数据单位来组织这些数据，以实体搜索的方式来对这些异构数据进行信息检索。目前，学术界和工业界都已经开始进行以实体为中心的研究和应用，旨在原始的异构数据之上建立更有通用性的数据应用价值。语义实体搜索越来越具有普遍性和重要性，已经成为用户获取信息的重要方式。

本书是著者近十年科研成果的集合，围绕数据空间中基于语义的实体搜索关键技术展开，全书共分7章。

第1章阐述背景及意义，并介绍数据空间的概念、特性和国内外研究现状。

第2章主要介绍一种以实体为中心（entity-centric）的数据模型。针对数据空间中实体的异构性、实体间存在着丰富的语义关联关系，提出了一种以实体为数据单位、分层的图数据模型lgDM，由实体关联数据图G_D和实体关联模式图G_S组成。lgDM能够描述异构的实体类、实体及属性值，并能够描述实体类间、实体间丰富、复杂的关联关系。研究了图模型lgDM的权重设置方法、建立索引的方法和模型所具有的查询能力。实验结果表明了所提出的数据模型lgDM在描述丰富语义关联关系方面的有效性。

第3章主要介绍数据空间中基于聚类的实体关联关系挖掘算法CFRQ4A。提出了四阶段的实体关联关系构建模型，并且在实体关联关系构建的整个生命周期中引入了关联关系约束验证来确保关联关系的正确性。提出了由实体聚类、候选实体对过滤、关联关系归纳和推理、关联强度量化四步骤组成的基于

聚类的实体关联关系挖掘算法 CFRQ4A，用较少的手工来逐步地发现实体关联关系。实验结果表明了所提出的 CFRQ4A 算法的准确性和有效性。

第 4 章主要介绍数据空间中基于时间的集合式实体识别（collective ER）算法 T-CER。提出了包括预处理、blocking、表象聚类和时间约束检查四步骤的集合式实体识别算法 T-CER，解决了数据空间中具有时间信息的集合式实体识别问题。针对数据空间中实体随时间演化的特性，在表象聚类步骤提出基于演化的实体识别聚类算法 TE-Clustering，在相似度度量方法中引入属性演化系数和关系演化系数来捕捉时间演化对相似度的影响。并且给出基于识别顺序依赖图 G_{depend} 来解决集合式实体识别的识别顺序问题的方法。大量实验结果表明了所提出的 T-CER 算法和 TE-Clustering 算法的准确性与有效性。

第 5 章主要介绍数据空间中时间感知的查询时实体识别与数据融合框架 TQ-ER，在保证查询数据质量的同时，使实体识别适应用户查询的实时性要求。实验结果表明了 TQ-ER 的有效性和正确性。

第 6 章主要介绍数据空间中基于关联关系的关键字查询意图消歧算法。针对关键字查询存在的语义模糊性，利用实体类间、实体间的关联关系提出了包括关键字语义项映射、目标实体类识别和候选查询集生成三步骤的关键字查询意图消歧算法。实验结果表明了所提出的关键字查询意图消歧算法的准确性和有效性。

第 7 章主要介绍语义实体搜索原型系统 KeymanticES 的设计与实现。基于对数据空间中基于语义的实体搜索关键技术的研究成果，实现了基于语义的实体搜索原型系统 KeymanticES。来自学术领域的真实数据集上的实验结果表明了 KeymanticES 的有效性。

本书适合大学三年级以上的计算机专业本科生和研究生，以及具有类似背景的对数据集成和语义实体搜索等方向感兴趣的人士。语义实体搜索发展极其迅速，目前已成为研究的热点之一。著者自认为才疏学浅，加之时间和精力所限，书中不妥之处在所难免，承蒙诸位读者和专家不吝告之，将不胜感激。

杨　丹

2019 年 9 月

目 录

第1章 绪 论

1.1 研究背景和意义

随着数字化技术和互联网的发展，数据管理和计算模式呈现出如下新的特点。一是海量化。全球的数据量在以指数的趋势迅猛增长，目前每年全球至少产生15亿TB的新数据。二是多样化与异构。随着网络技术的发展和Web技术的日益成熟，Internet收集了海量的信息资源，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，大量的XML文档、文本等半结构化数据，图片、音频、视频、文档等非结构化数据大量地涌入到应用中。三是松散化。这些资源具有分布分散、结构松散，并且更新变化快等复杂特性。四是共享化。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。

随着信息技术的不断发展，计算机逐步成为人们日常工作和生活的必备品。同时，E-mail信息、工作文档文件、收集的参考资料、图片和视频等个人信息也在急剧膨胀，并且这些个人数据管理呈现出如下新的特点：数据量成倍增长，数据的更新日新月异；数据的形式趋向多样化，管理的目标包括结构化、非结构化和半结构化的数据，以及动态的音频、视频等流数据；数据间的语义关联性更强，而且这种关联更难被发现和提取。这些复杂的特性决定了无法用单一、传统的关系数据库系统来组织和管理新环境下的数据。桌面搜索工具虽然为用户管理个人数据资源提供了方便，但它主要是面向全文的搜索，得到的还是相对“独立的”无关联的资源，并没有打破资源自身的界限。

面对以不同形态存在且相互关联的多种资源信息的混合体，目前还没有一

个成熟的管理软件有效地管理它们，人们还是通过手工对它们进行分门别类的管理或基于桌面搜索管理，无法实现语义查询和进一步深入查询，更不能获得资源之间的关联关系，导致数据资源利用率不高。无论是传统的数据库技术还是面向全文的桌面搜索技术，均已无法满足这些异构多样数据管理的新要求。与新的数据特点相适应，人们对信息的管理能力和服务模式也提出了新的要求，传统的数据库管理系统在这些新的要求面前显得无能无力，不能满足这些复杂数据管理的新要求。数据空间[1]就是在这一背景下提出的新的概念和技术。数据空间是基于 pay-as-you-go 思想进行集成的一种数据组织形式，不依赖于严格的数据模式，并且能随着时间演化，在任意时候提供给用户尽最大努力的结果，能够满足上述数据特点的数据管理的要求。数据空间将是数据管理的又一新目标，代表了一种新的管理数据的理念。数据空间技术是数据库管理技术的进一步发展，该技术的发展与成熟将代表数据管理进入一个新的里程碑，数据空间的相关研究成果将为管理开放的数据资源提供良好的支持，达到提高资源利用率和工作效率的目的，具有广阔的前景。

1.2 数据空间概述

本节首先对数据空间的概念进行介绍；接着对数据空间的特性进行分析、归纳，并且与传统的数据库系统和数据集成系统进行比较。

1.2.1 数据空间的概念

数据空间(Dataspace)的概念最初由 M. Franklin、A. Halevy 和 D. Maier 几位学者于 2005 年在 SIGMOD Record 的论文 *From databases to dataspaces: a new abstraction for information management* 和 PODS 2006 的论文 *Principles of dataspace systems* 中提出。学者们根据当前数据与信息的增长对数据管理技术需求的发展情况，针对现有传统数据库技术的不足，提出了一种新的信息管理抽象方法，并系统地分析了数据空间技术的目标，以及构建数据空间的支撑平台所面临的挑战。在论文中给出了数据空间及其组件的一个例子(如图 1.1 所示)，将数据空间建模成一系列参与者(participants)和关系(relationships)。图中的大矩形框中表示了数据空间中各种异构类型(结构化、半结构化、非结构化)的数据资源(即参与者)及其彼此间丰富的关系。从图中可以看出，参与者可以是关系数据

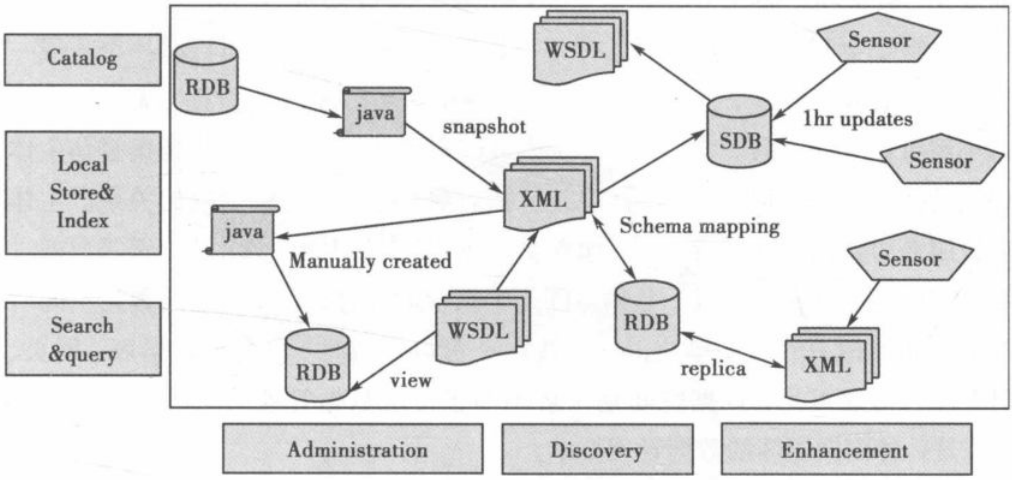

图 1.1 一个数据空间及其组件的例子

库、XML 资源库、文本数据库、Web 服务和软件包等，甚至是传感器。矩形框的外围左边和下边分别给出了数据空间的组件(模块)，包括目录服务组件、本地存储和索引组件、搜索和查询服务组件、管理组件、发现服务组件和提高组件，用来提供数据空间管理系统的各种管理和支持服务。2006 年，A. Halevy 等在荣获 VLDB 十年最佳论文奖的报告 *Data integration—the teenage years* 中对数据集成技术所面临的挑战性问题进行了分析，其中包括采用 pay-as-you-go 数据管理思想的数据空间技术。

数据空间是基于 pay-as-you-go 思想进行集成的一种数据组织形式。数据空间在本质上可以被看作对数据集成框架的下一步演化，但在集成对象、集成方式等方面与传统的数据集成技术不同。一个数据空间是由一系列相关的异构资源对象集和资源对象间的关联关系集组成的。提供 Web 级别的数据集成需要一个能为现实世界中任意关系提供建模的系统，并且能随着时间演化，在任意时候提供给用户尽最大努力的结果。从数据管理角度来说，数据空间是对新的数据特点的一种刻画，许多在数据管理和相关领域的研究问题都与数据空间相关，因此其主要研究问题包括数据模型、实体识别、模式匹配和模式映射、关键字查询、数据集成等。

数据空间是与主体相关的数据及其关系的集合，数据空间是与主体相对应的，数据空间中的所有数据对于主体来说都是可以控制的。主体相关性和可控性是数据空间中数据项的基本属性。数据空间分为主体数据空间和与之相对的公共数据空间。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。主体、数据集、服务是数据空间的三个要素。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；也就是说，一个人可以有一个数据空间，一个项目小组可以有一个数据空间，一个企业可以有它的数据空间。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，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。数据空间是数据项的集合，数据项是与数据空间所对应的实体相关的信息单位，一个数据项可以是邮件、文件、数据表、网页、PPT 等。由此可见，数据空间是一种不同于传统数据管理的新的数据管理理念，是一种面向主体的数据管理技术。

1.2.2 数据空间的特性

数据空间具有空间和时间特性。从空间上来说，数据空间的数据来自多个分布的自治的数据源；从时间上来说，数据空间中的数据也随着数据项的发展而不断变化，数据空间的大小是动态变化的，其中的数据是动态演化的，包含的信息量会不断增强，数据质量也会不断提高。与传统的数据管理技术类似，数据空间管理也面临数据模型及数据集成、查询与索引等各种技术的研究，但是由于数据特点不同，这些问题的解决不同于传统的数据库系统和数据集成系统。图 1.2 是不同的数据管理策略分布图，沿着语义集成度（横轴）和数据耦合度（数据间协调的松弛程度，纵轴）两个维度表示了已经存在的数据管理解决方案的分布。从图 1.2 可以看出：数据空间管理系统在两个维度上都处于中间的位置，在语义集成度上处于数据库管理系统、数据仓库系统、传统的数据集成系统与 Web 搜索系统和桌面搜索系统的中间；在数据间协调的松弛程度上处于数据库系统、传统的数据集成系统、桌面搜索系统与数据仓库系统和 Web 搜索系统的中间。因此，数据空间正好迎合了当今异构、复杂、多样化数据管理的新需求。下面分别将数据空间与传统的数据库系统和数据集成系统进行了比较。

1.2.2.1 与传统的数据库系统的比较

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与传统的数据库系统有明显的不同[1]，主要体现在以下五个方面。

① 数据模型。传统的关系数据库基于的是关系模型，数据关联是基于关系表的。数据空间的逻辑模型是一个图。数据库是模式优先（schema-first）的逻辑

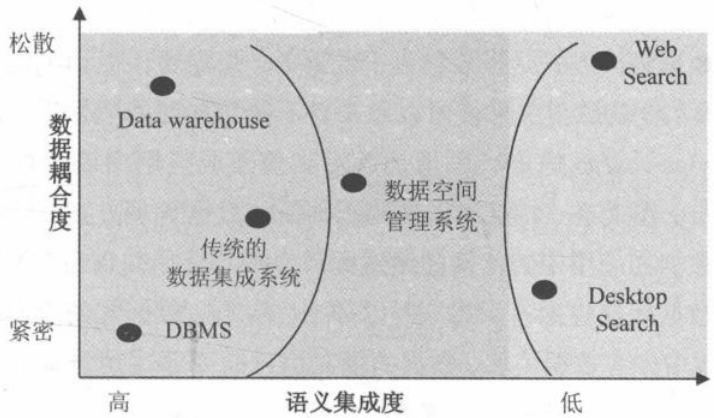

图 1.2 不同的数据管理策略分布

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。传统的数据管理技术具有完整的模式，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(如查询操作)具有尽最大努力的特性，查询结果可能是近似的、pay-as-you-go 的。

③ 数据类型。数据空间的数据来自多个不同的数据源，数据格式多样，如可能包含关系表、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，数据格式就是单一的关系表，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时才会将对象保存到数据空间中，才会在对象之间建立关系。这种数据管理方式因为比传统的集成系统的前期成本低，所以更为实用。

1.2.2.2 与传统的数据集成系统的比较

传统的数据集成方式是模式优先于数据的，只需要根据预先设计出的模式结构，通过模式间的映射关系就可以对来自不同数据源的数据进行集成。而数据空间的 pay-as-you-go 的集成思想是针对当前集成应用中以数据为中心的特征，数据优先于模式这一特点而提出的一种新的数据管理方案。目前半结构数据和无结构数据在应用中的比例已经达到了 80% 以上，并且还在不断增长。这意味着当前数据集成应用将面临一种以数据为中心、数据优先于模式的集成方式，即在集成中先有数据信息，数据的模式信息需要通过信息抽取和挖掘等方法在数据集成的过程中获得。传统数据集成方法显然已经无法适应新的应用需求。此外，当前数据集成中所要处理的数据信息具有更加明显的异构、海量、分布等特点，尤其是在数据的异构性方面已经不仅仅局限于模式上的异构，还包括类型上的异构。数据空间 pay-as-you-go 的集成方式中，将在用户认为必要时根据其需求抽取指定的数据信息和相应的结构化信息并在数据之间建立关联关系。这种集成思想不但能够提供实时而准确的数据信息，还能够提供对数据信息的统一高效的管理方法。图 1.3 给出了数据空间与传统数据集成系统在功能性和响应时间上的比较。从图中可知，数据空间技术即以数据为中心的 pay-as-you-go 思想的数据集成技术，对于推动数据库领域技术发展和为当前企业与个人的数据应用提供解决方案具有重要意义。

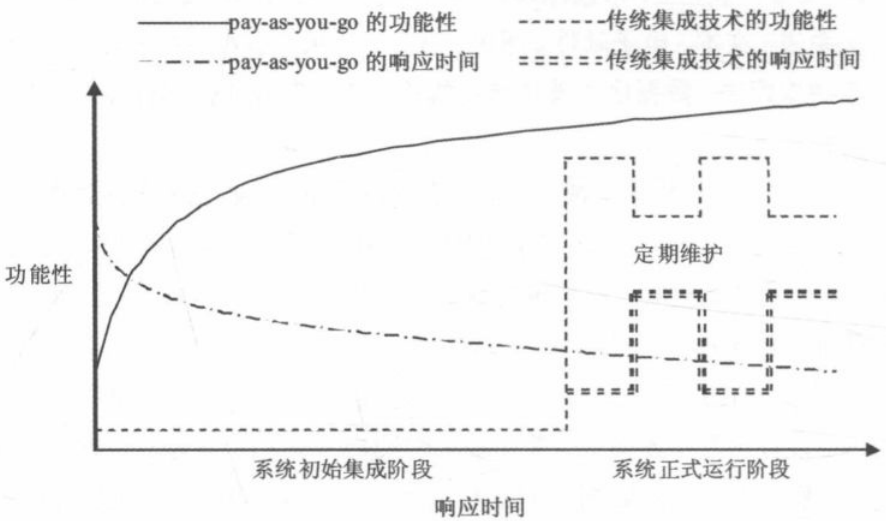

图 1.3 数据空间与传统数据集成系统的功能性和响应时间比较

数据空间在数据模型、数据对象、数据存储、创建方式等方面都与传统的

数据集成系统有明显的不同。主要体现在以下四个方面。

① 数据模型。传统的数据集成系统的数据模型是关系模型，而且是模式优先的，即基于数据的模式对数据进行集成。而数据空间的数据模型是图模型，数据模式可以是松散的。

② 数据对象。传统的数据集成系统的数据对象是来自同一个领域的多数据源，而数据空间中的数据对象来自多领域的多个数据源。

③ 数据存储。传统的数据集成系统中的数据存储方式是中心式的、集成的，而数据空间是分布式的、共存式的。

④ 创建方式。传统的数据集成系统是基于 pay-before-you-go 方式创建的，而数据空间是基于 pay-as-you-go 思想方式创建的。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。因此，传统的基于模式的数据集成技术对数据空间中资源的处理能力十分有限。

1.3 国内外研究现状

目前国内外数据空间的研究工作主要从两个方面展开：一是针对数据模型、查询、模式匹配、引用协调等理论与算法；二是以某种应用为背景原型系统及技术的研究。在数据空间的研究领域，已有的原型系统较多的是针对个人数据空间管理系统，目前国外典型的原型系统主要有 iMeMex 系统[2]和 SEMEX 系统[3]等。iMeMex 系统是瑞士的苏黎世理工学院实现的个人数据空间管理系统平台，该系统将物理与逻辑上独立的数据信息集成后以桌面的形式提供给用户，从而避免了用户烦琐管理底层数据。该系统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的统一表示。iMeMex 采用以元组形式组织数据的 iDM 数据模型[4]，提出一种统一资源视图的概念和数据之间关联关系的形式化描述方法，以实现对不同类型数据的统一表示。这种方法突破了数据对象与文件系统的边界，将对象内部数据和外部数据统一表示，在查询处理方面属于尽力而为的工作模式。iMeMex 系统中基于资源间的内在关联预先定义关联规则或关联轨迹(iTrail)构建关联图[5]，并采用基于分组压缩索引(简称 GCI)的查询处理技术实现关联图上的近邻查找。虽然在查询中考虑了语义信息，但 iTrail 规则或关联轨迹均需要被预先定义，因此查询结果的质