以数据空间理念建立关系发现应用——NSTL智能检索平台的实践*

王莉，梁冰，白海燕
（中国科学技术信息研究所，北京 100038）

摘要：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。文章以NSTL智能检索平台的实践为背景，提出采用数据空间理念构建知识关联网络，实现pay-as-you-go模式的关系发现应用。

关键词：关系发现；智能检索；知识关联网络；数据空间；pay-as-you-go
中图分类号：TP391.3
DOI：10.3772/j.issn.1673—2286.2014.06.008

1 引言

NSTL智能检索平台是国家十二五科技支撑计划“面向外文科技文献信息的知识组织体系建设与应用示范”建设内容之一，重点解决知识抽取、自动标引、知识网络构建、关联数据、多维索引、自动聚类、个性化服务、自然语言检索、双语查询等关键技术在复杂应用下融会贯通的问题，通过超级科技词表（以下简称STKOS）的支撑，深入揭示NSTL海量科技文献信息资源，建立知识网络，为科技创新主体提供高效便捷的知识服务和科研文献信息支撑环境。该课题属于工程建设项目，所有的技术成果最终都集中反映在智能检索平台中，除了通过对数据精细化处理提升现有的检索功能之外，一个非常重要的任务就是挖掘新的应用，带给用户更丰富的、智能化的使用体验。“关系发现”就是这样一个应用，采用数据空间理念构建知识关联网络，实现pay-as-you-go模式的关系发现，引导用户参与知识探索。

2 数据空间理论

数据空间（Dataspace）是一种数据管理理念，由Michael Franklin[1]等人在2005年SIGMOD大会中首次提出，其核心思想是“淡化模式，凸显数据”。概括地说，一个数据空间是与主体相关的数据及其关系的集合。其中，主体即数据空间的所有者，可以是一个人、一个机构组织，甚至是某一个应用；数据集是由与主体相关的各种异构的、多来源数据资源及其关系构成；数据空间中的所有数据对于主体来说都是可控的。当开始构建这一数据集合时，不需要定义严格的数据模式（schema-later），数据模式可以是松散的、不完整、不精确的。随着用户的使用，数据模式以一种pay-as-you-go的方式根据主体需求逐步演化[2]。

在传统数据库管理系统中，数据库模式是事先设计好的，明确定义数据之间的关联关系，这种关联往往是稳定的，而且类型也相对单一，这种方式称为pay-before-you-go。pay-as-you-go，也称为“现收现付”，与之相对应，是一种滞后构建方式，即只有当用户认为必要时，才建立数据之间的关联。pay-as-you-go特性降低了构建数据空间的前期代价，与此同时，由于缺乏模式对数据关系的刻画，只能提供近优的、次优的搜索服务（best-effort），但是，这一状况会随着演化逐步得到改善。演化以及对人在数据管理中主体作用的关注，是数据空间理念与传统数据管理的本质区别。

* 本研究得到国家十二五科技支撑计划课题“信息资源自动处理、智能检索与STKOS应用服务集成”（编号：2011BHA10B05）资助。

数据空间作为一种新的数据管理方法，并不局限于某一种具体的算法和技术，其理论方法研究包括概念、模型、查询、索引、更新、演化等多个方面，相关应用则集中在个人数据管理、企业数据管理、科学数据管理、Web数据管理等领域。对于文献信息智能检索这样一个应用场景而言，构建知识关联网络是一个复杂的工程，一步到位、完整地建立所有知识节点之间的关联关系是不现实的。笔者认为，采用pay-as-you-go模式构建关系发现应用非常合适，它可能不是一个最优的服务，但是通过人的参与可以步步逼近最合适的答案，更能从这一探索过程中萌发更多思想。

3 知识关联网络

知识关联网络是由知识节点及其之间错综复杂的关系构成的网状结构。从文献中构建知识关联网络主要基于文献的外部特征和内容特征完成。外部特征包括题名、作者、机构、出处、参考文献等信息，十分清晰。内容特征则相对复杂，既包括分类号、关键词和主题词这类相对明确的、描述文献主旨的特征点；也包括从文本信息中识别的知识对象及知识对象之间存在的隐含逻辑关系，更准确地说，是一个知识发现的过程。NSTL经过十余年的积累，拥有海量科技文献信息资源，可以预见，基于这样海量的文献构建的知识网络将是一个规模巨大的复杂网络。以数据空间理念指导知识网络的构建，虽然淡化模式，但并不是抛弃模式，本文提出利用本体对网络结构进行合理规划，有机组织并存储数据。这一技术方案能有效地降低网络的复杂性，是提高操作效率、支撑上层应用的关键。

3.1 以本体为知识关联模型

NSTL知识关联模型引入本体技术，通过构建NSTL核心本体规范知识关联方式，降低知识关联网络的管理和维护成本。NSTL核心本体包括资源（Resource）、代理（Agent）、事件（Event）和主题（Subject）四大部分，涉及31个核心类（见表1）。

其中，资源部分分为合集（Collection）和文献（Document）两大类。合集是一组相关文档集合，包括期刊（Journal）和丛书（Series）。文献进一步分为单册（Item）、单篇（Single Document）和片段（Document Part）。单册是指将一组文档集结成册，单独出版发行，分为期刊中的一期（Issue）、会议录（Proceeding）和单本成册的文集汇编（Compilation）。单篇包括图书（Book）、文章（Article）、科技报告、学位论文、专利、标准和科技要览。片段是从单篇中析出的知识单元，主要包括章节（Chapter）、图（Image）和表（Table）。NSTL对资源组织本体的研究起步较早，在参考MarcOnt Ontology、Bibliographic Ontology

表1 NSTL本体核心类

序号	类名	标识
1	资源	Resource
2	合集	Collection
3	期刊	Journal
4	丛书	Series
5	文献	Document
6	单册	Item
7	卷期	Issue
8	会议录	Proceeding
9	文集汇编	Compilation
10	单篇	Single_Document
11	图书	Book
12	论文	Article
13	学位论文	Thesis
14	专利	Patent
15	标准	Standard
16	计量规程	Verification_Regulation
17	科技要览	Overview
18	片段	Document_Part
19	章节	Chapter
20	图	Image
21	表	Table
22	代理	Agent
23	个人	Person
24	团体	Group
25	机构	Organization
26	主题	Subject
27	范畴	Category
28	概念	Concept
29	事件	Event
30	会议	Proceeding
31	项目	Project

Specification等书目本体，结合FRBR、OAI-ORE、DCMI等概念模型和标准规范的基础上，以连续出版物为主体构建了本体模型，并进行了小规模的关联数据实验[3]。本课题对文献资源的组织充分吸纳了前期研究成果。同时，代理部分参考FOAF本体，用来描述与资源和事件相关的人（Person）、机构（Organization）或组织（Group）；事件部分参考EVENT本体，用来描述会议和项目；主题部分参考SKOS本体，用来描述资源相关主题以及STKOS。

图1展示了由本体结构生成的基本关联框架。

3.2 抽取元数据构建关联网络

利用本体结构从元数据中提取核心类与属性，这一过程既是本体的实例化过程，同时也是关联网络的持久化过程，需要重点考虑关联网络的存储方案。毋庸置疑，RDF模型是描述关联关系的最佳选择，任何复杂关系都可以分解为多个简单的二元关系；然而RDF在实际应用中仍然存在很多问题，其中最典型的就是，如何对大规模RDF数据进行高效存储、查询进而推理。笔者认为，RDF模型能够简单而清晰地刻画出数据之间的关系，然而在实际构建关联网络时，是否使用RDF数据格式并不重要。Franklin[1]在数据空间的早期研究中即提出用带标签的图刻画数据空间，Blunschi和Dittrich等人[4]也提出将数据从物理设备、表示形态中释放出来，采用一种逻辑的、图的形式刻画，并给出了一种统一资源视图的概念和形式化表示方法，以此为基础设计实现了iMeMex原型系统。图形数据库是一种“合理的知识保存和描述的方式”[5]，基于这一思路展开研究与实验，最终确定以Neo4j为主，辅以Solr实现NSTL知识关联网络的存储。

整个构建过程包括三个关键步骤：第一，建立元数据格式到本体框架的映射规则，支持元数据的抽取；第二，建立本体框架到数据模型的映射规则，支持元数据的存储；第三，以本体框架为枢纽，采用Java编程，Neo4j内嵌式调用方式（Java 1.7.0_51、Solr4.5.4、Neo4j 1.9.3）完成实例数据的抽取和转换。以下重点讨论元数据格式到本体框架，以及本体框架到存储模型两大映射中的核心问题。

1）从元数据格式到本体框架的映射

关联网络来源数据分为文献数据和STKOS两大部分，NSTL现有文献数据采用Oracle数据库存储，STKOS数据则需要通过XML交换格式从超级科技词表发布平台定期收割而来，因此，需要分别完成Oracle表结构和XML结构到本体的转换。

对于关系数据库和本体之间的映射，相关研究较多，并且提出了一些通用的、经过正确性证明的转换规则。本文参考文献[6]给出的本体元素与关系数据库模式各元素之间的对应关系，结合NSTL文献资源元数据存储结构（见图2所示），形成以下基本映射原则：

首先，期刊具有层级结构，分为母体、卷期和论文三级存储，可以分别映射为本体结构中资源类下的合

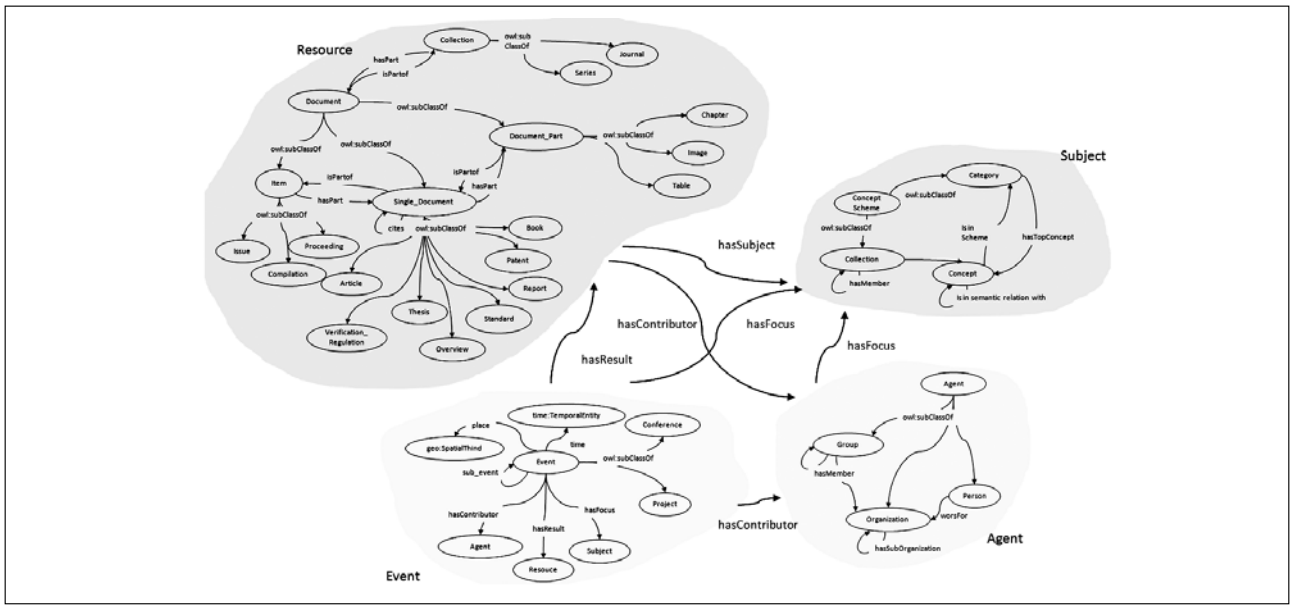

图1 NSTL知识关联模型基本框架

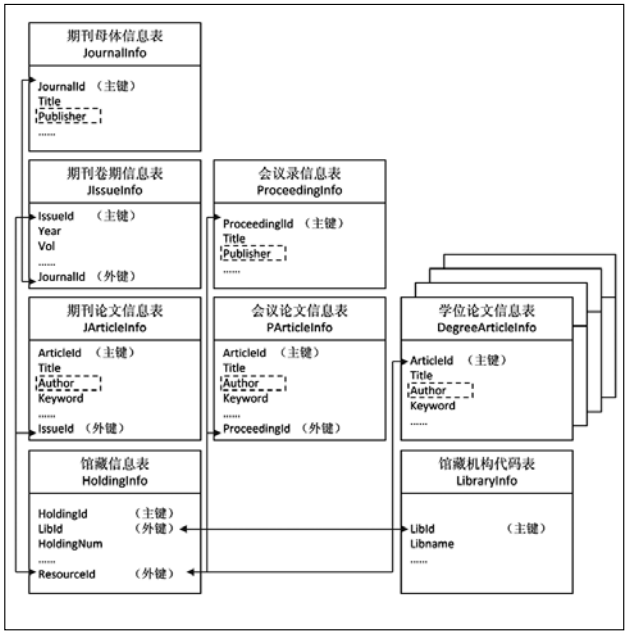

图2 不同类型的文献数据在NSTL Oracle仓储中的存储结构

集、单册和单篇。同样，会议录映射为资源类下的单册，会议论文映射为资源类下的单篇；其他文献类型则直接映射为单篇下的对应子类。相应地，二维表的列大多可以直接映射到对应类的属性。

其次，馆藏机构代码表可以映射为本体结构中代理类下的机构，同时，二维表的列转换为类的属性。

第三，期刊和会议录单独设计馆藏信息表，用于描述资源和馆藏单位之间多对多的联系，首先确保资源和收藏单位分别映射为本体中的资源类和代理类，继而为这两个类分别建立馆藏属性，通过这两个属性保持类之间的关联。对于学位论文、科技报告等单篇著录文献，馆藏信息作为字段出现在文献基本信息表中，同样，资源类中应该保留该属性，馆藏单位对应的代理类也需要建立相应的关联属性。

第四，基础数据没有对作者和机构进行规范，相关信息存储在文献基本信息表中，需要提取出来分别映射为代理类下的个人和机构。

第五，文献基本信息表中关于主题的信息涉及关键词、标引词、范畴、分类号等字段，直接转换为资源类的属性，不需要映射为主题类。

XML格式到本体结构的映射完成的是STKOS到主题类的转换。STKOS词表总体设计了来源术语、基础术语、基础概念、规范概念、范畴类、来源词表、范畴体系等7大类元素，智能检索平台中主要使用的是基础术语、基础概念、规范概念和范畴类，图3展示了这几部分之间的基本关联。目前STKOS尚未全部完成，并没有建立正式的开放发布服务，这里构建关联网络所需的基础数据从STKOS协同工作平台中收割而来，以描述基础术语的termRecords.xml文件为核心，没有复杂的嵌套结构。词表的映射暂时不需要和主题之外的其他类建立关联，映射关系非常简单，即使后期交换格式发生变化，程序调整也很容易。

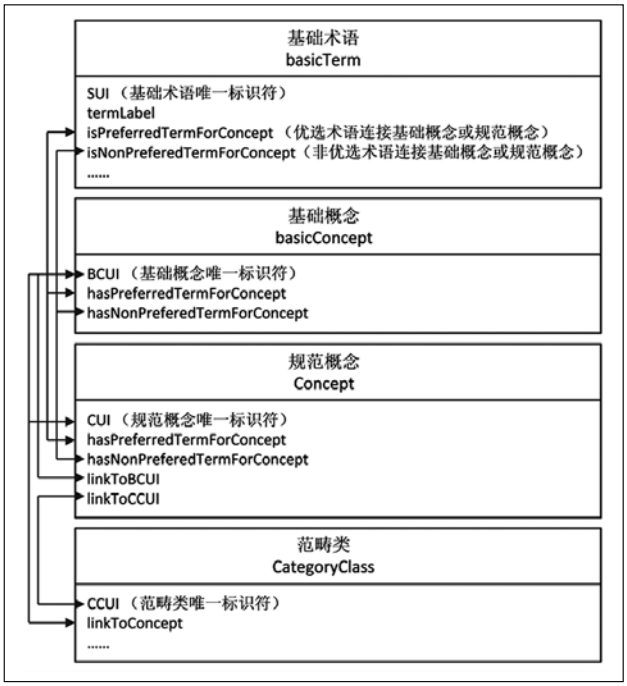

图3 STKOS核心元数据格式之间的关联

2）从本体框架到数据模型的映射

本文采用Neo4j和Solr相结合的存储方案。Neo4j[7]是目前比较主流的一款图形数据库开源软件，以图论为理论基础，采用节点和边的形式刻画实体及实体之间的关系，善于处理具有复杂连接关系、低结构化的数据。自2003年起，Neo4j已经被作为24/7的产品使用。在讨论Neo4j的高性能时，主要指的是它的查询能力，即读性能表现。Neo4j能够很好地支持大规模的关联数据查询，但是首要解决的技术问题是如何建立复杂的网络图，这是一个数据写操作问题。课题在前期技术方案选型阶段对Neo4j进行了内嵌式数据入库性能测试，测试分别对单属性节点和双属性节点进行无条件入库和关系去重入库，实验表明，节点属性的数量和入库操作的复杂度对入库速度的影响较大。

知识关联网络中，每一个知识点都包含了丰富的

表2 单属性节点和双属性节点在关系去重条件下的入库表现

程序实现方式	属性个数	节点数量（对）	入库时间（毫秒/1000对节点）	说明
嵌入式关系去重	1	10亿	100	随着数据量的增大，速度逐渐变慢。在导入3亿数据时，速度大概在1s左右。
嵌入式关系去重	2	5亿	150	随着数据量的增大，速度逐渐变慢。在导入3千万数据时，速度大概在10s左右。

注：以上测试环境配置为普通PC电脑（CPU i7-3770 @ 3.40GHz 四核，内存8 GB，硬盘1 TB / 7200 转/分），操作系统Windows 7 旗舰版 64位 SP1（DirectX 11），Neo4j1.9.3。写速度与文件系统的查找时间和硬件有很大关系，上表中的绝对数据只是起参考作用，重点关注的是不同实验结果之间的对比。

属性，在Neo4j工程化实现中，为每一个节点保留各种细致属性是不可行的，因此引入Solr全文搜索引擎，实现对节点及属性的全存储和全索引（包括去重操作），Neo4j中只保留节点及其关键属性。

Solr是基于Lucene的搜索服务器，以Document为原始对象对资源进行存储索引。每个Document由若干Field组成，而每个Field表示索引资源的一个属性。Field是一个关联的键值对，取值往往切分为若干Term，Term是索引的最小概念。Solr1.3版本引入了多核（MultiCore）概念，每个Solr core由它自己的配置文件和索引数据组成。多核配置实现了一个Solr实例管理多个索引文件。课题在将NSTL本体框架映射到Solr数据模型时，分别建立了文献资源和超级科技词表两套索引，用两个core进行管理。

Neo4j采用节点、关系和属性灵活地表达复杂网络关系，数据结构非常简单。从本体框架到数据模型的映射，实质上是一个从领域模型到Neo4j节点空间的映射过程。文献[8]讨论了实现这一过程的一般方法，其主要步骤是：以Neo4j自动生成的参考节点作为图的起点，首先引入子参考节点，每一个实体类都用一个子参考节点表示，通过SUBREFERENCE关系挂接在参考节点上。领域对象实例，也就是应用域中的真实数据，对应每一个具体的节点，挂接在子参考节点上，用关系INSTANCE连接。最后将实例属性映射为节点属性。

按照图4描述的映射思想，本体框架中的类表现为Neo4j节点的类型，而Neo4j节点空间中每一个节点对应的是本体的实例，节点属性对应的是实例属性，从类的属性继承得到。为了满足工程化实施的需求，需要重点考虑类的属性到节点空间的映射问题。首先，采用“命名空间/节点类型/唯一标识符”规则建立节

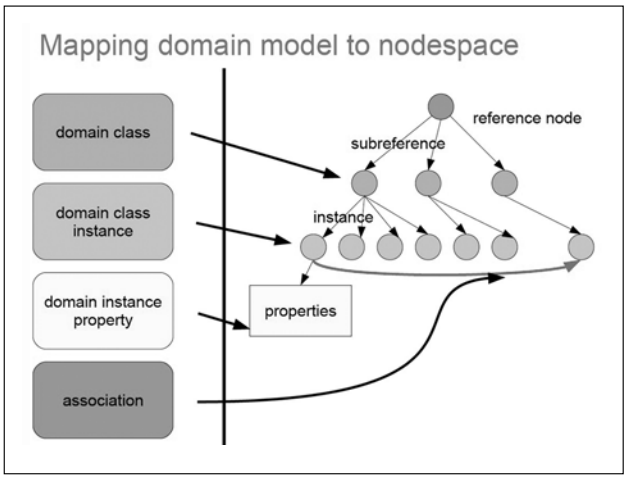

图4 Grails：从领域模型到节点空间的映射[9]

点统一命名机制，为每一个节点生成URI标识。需要注意，这一标识是在创建Solr索引时生成的，不同于Neo4j自动生成的节点id，而是作为节点的关键属性存在。节点通过URI与Solr索引数据之间建立关联，从而可以获取关于实例的详尽的属性信息。第二，本体框架中的类映射到节点实例层面时，作为节点的类型属性（type）保留下来。第三，本体框架中类之间的关联映射到实例层面时，主要是通过具体的属性建立起来的，在处理时，直接转换为节点间的关系，不需要作为节点属性保留。第四，为节点保留一个直观易懂的关键属性。这一属性主要用在可视化场景中，作为节点名称显示，因此强调的是“直观”而不是唯一标识，例如，对于一篇科技论文，DOI最具辨识度，可以直接作为唯一标识符使用，但是并不建议作为节点名称使用，采用论文标题则更直观易懂。经过以上处理，每个节点设计了3个属性：统一资源标识、资源类型和名称；定义了

inScheme（范畴表）、linktoCUI（链接到范畴类）、linktoCCUI（链接到规范概念）、broader（上位类）、narrower（下位类）、related（相关）、has_concept（包含概念）、usedFor（代替）、has_authorlist（有作者）、has_organization（有机构）、has_keyword（有关键词）、belong_to_Issue（属于卷期）、has_holdinglist（有馆藏）、belong_to_Journal（属于期刊）、belong_to_proceedings（属于会议）、has_citelist（有引文）等16种类型的节点关系，文献资源子空间和超级词表子空间通过has_concept关系连通。

图5给出了一个概念和一篇期刊论文在节点空间中的定义，以及其关联关系的可视化展示。

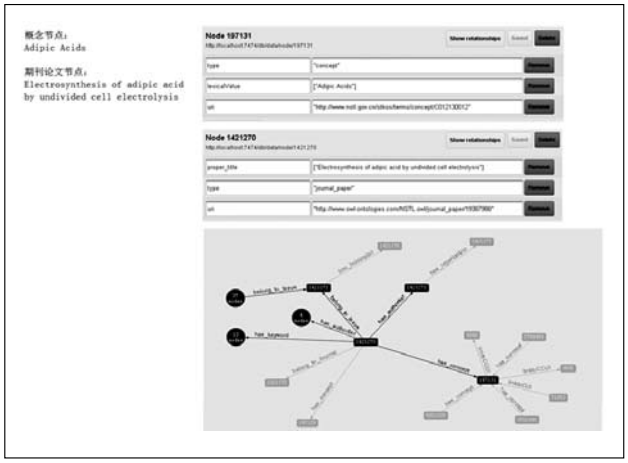

图5 期刊论文节点与概念节点的定义及关联图示（摘自Neo4j控制台）

课题先期从西文期刊和会议文献中抽取文献资源构建知识关联网络，生成的两个Solr索引文件共包括6,659,444个文档（numDocs）；在Neo4j中表现为6,879,092个节点，20,197,772个关系，占用3613MB磁盘空间。

4 pay-as-you-go模式的关系发现

NSTL智能检索平台实现的关系发现是一种交互性立体式的检索应用。首次检索采用常规的关键词匹配方式（辅以语言词典纠错、STKOS超级科技词表及历史检索输入提醒），检索结果将按照NSTL本体结构分面展示，这是关系发现的起点。发现过程采用Neo4j图遍历机制实现。简单地说，遍历一张图就是从图中任意节点开始，按照一定的规则，跟随节点的关系，访问关联的节点集合。Neo4j内部实现了最短路径、所有路径、所有简单路径、Dijkstra和A*等5种图算法，其遍历速度与图的规模大小没有关系，有效地解决了大批关联数据查询的问题。

关系发现应用中主要包括三个重要功能。一是以检索得到的文献集合为基础，采用图的形式展示不同知识点及知识点之间的关系，支持用户点击任意节点进行浏览，可以跟随知识点之间的链接扩展到更大的节点空间，突破检索结果集合的限制。二是专注于一篇文献，观察它在整个节点空间中与其他节点之间存在的关系，支持渐进式的探索。三是在一次或多次检索的结果列表中收集关注的实例（如作者、论文、机构），通过实例分析功能在整个节点空间中观察它们之间的关联关系，支持渐进式的探索。

关系发现应用具有交互性探索的鲜明特性，在人的主体作用下，可以发现那些隐藏的间接关系，也可称之为基于查询产生的关联痕迹。关系发现应用呈现出pay-as-you-go特征，但pay-as-you-go并不仅止于此，它是和演化联系在一起的，即，在必要的时机将这些关联痕迹固化下来，使原本松散的数据模式得以改进。

本文将演化作为下一步工作，主要从两个方面考虑：一是与个性化功能相结合，允许登录用户保存他认为有意义的关联路径，自定义起止节点间的关系，只对该用户可见并可用；二是与数据管理功能相结合，允许数据管理员建立数据之间的关联，对所有用户有效。其中，第二种方式虽然实现简单，但需要重点考虑触发机制问题，本体演化、数据清洗固然会引发关系变化，大众智慧（多数人认可的自定义关联）也可以作为一个触发点（至少可以加入推荐功能）考虑进去。

5 结语

文章以NSTL智能检索平台的实践为背景，提出采用数据空间理念构建知识关联网络，实现pay-as-you-go模式的关系发现应用，将检索过程变成一个对知识内容和知识点探索和挖掘的过程。

目前，NSTL智能检索平台还是一个原型系统，完成了主体功能的开发和测试；支撑应用的知识关联网络来源于NSTL生产平台中的西文期刊和会议文献，以及STKOS协同工作平台提供的超级科技词表数据。下一步工作主要包括：1）知识关联网络演化的设计与实现；2）可视化展示与交互功能的优化（设计更直观、更容易理解和交互的方式）；3）针对工程化实施的性能调优（采

用Neo4j HA集群模式提高读性能）。

关系发现应用基于图形技术帮助用户了解复杂纷乱的数据。在未来，随着本体进化、知识对象识别及抽取的广泛应用，结合正确的算法模型，关系发现将发挥巨大作用，引导检索真正走向智能。

参考文献

[1] FRANKLIN M, HALEVY A, MAIER D. From databases to dataspaces: a new abstraction for information management [EB/OL]. [2014-05-16]. http://www.sigmod.org/publications/sigmod-record/0512/sigmod-record.december2005.pdf#page=28.

[2] 李玉坤,孟小峰,张相於.数据空间技术研究[J]. Journal of Software, 2008, 19(8): 2018-2031.

[3] 乔晓东,白海燕,梁冰.NSTL的关联数据构建与应用场景设想[J].数字图书馆论坛,2011(12):54-60.

[4] BLUNSCHI L, DITTRICH J-P, GIRARD O R, et al. A dataspace odyssey: the iMeMex personal dataspace management system [EB/OL]. [2014-05-16]. http://infosys.cs.uni-saarland.de/publications/BDG+07.pdf.

[5] 吴斌.图形数据库Neo4j的使用[EB/OL]. (2013-07-05) [2014-05-16]. http://blog.csdn.net/ub1010/article/details/9263325.

[6] 许卓明,黄永菁.从OWL本体到关系数据库模式的转换[J].河海大学学报(自然科学版),2006,34(1):95-99.

[7] neotechnology graphs are everywhere [EB/OL]. [2014-05-18]. http://www.neotechnology.com/.

[8] ARMBRUSTER S. Grails Goes Graph [EB/OL]. [2014-03-18]. http://download.irian.at/2012/f9u4mq2xdg/d2_f_1645_Stefan_Armbruster_grails_goes_graph.pdf.

作者简介

王莉，女，1974年生，硕士，中国科学技术信息研究所研究员，研究方向：数字图书馆。E-mail：wangli@istic.ac.cn。
梁冰，男，1974年生，博士，中国科学技术信息研究所高级工程师。
白海燕，女，1973年生，硕士，中国科学技术信息研究所研究馆员。

Relationship Discovery Based on the Concept of Dataspace - A Case Study of NSTL Intelligent Retrieval Platform

WANG Li, LIANG Bin, BAI HaiYan
（Institute of Scientific and Technical Information of China, Beijing 100038, China）

Abstract: Dataspace is a new data management concept with pay-as-you-go features. On the background of NSTL intelligent retrieval platform, this paper uses the concept of dataspace to establish an associated knowledge network, and develops relationship discovery functions of the pay-as-you-go model.

Keywords: Relationship discovery; Intelligent retrieval; Associated knowledge network; Dataspace; Pay-as-you-go

（收稿日期：2014-05-18）

以数据空间理念建立关系发现应用--NSTL智能检索平台的实践

作者： 王莉， 梁冰， 白海燕， WANG Li， LIANG Bin， BAI HaiYan
作者单位： 中国科学技术信息研究所,北京,100038
刊名： 数字图书馆论坛 ISTIC
英文刊名： Digital Library Forum
年，卷(期)： 2014(6)

参考文献(8条)

1. FRANKLIN M;HALEVY A;MAIER D From databases to dataspaces:a new abstraction for information management 2014

2. 李玉坤;孟小峰;张相於 数据空间技术研究 2008(08)

3. 乔晓东;白海燕;梁冰 NSTL的关联数据构建与应用场景设想 2011(12)

4. BLUNSCHI L;DITTRICH J-P;GIRARD O R A dataspace odyssey:the iMeMex personal dataspace management system 2014

5. 吴斌 图形数据库Neo4j的使用 2014

6. 许卓明;黄永菁 从OWL本体到关系数据库模式的转换 2006(01)

7. neotechnology graphs are everywhere 2014

8. ARMBRUSTER S Grails Goes Graph 2014