面向科学数据的搜索引擎voovle

李成赞，沈志宏，黎建辉
中国科学院计算机网络信息中心，北京 100190

摘 要： 科学研究产生的科学数据资源日益激增，仅在中国科学院科学数据库“十一五”信息化建设中整合全院科学数据库形成的共享数据量就多达148TB。为了充分发挥这些共享科学数据的价值，结合科学数据的专业特点，提供一套跨库统一检索、不同领域科学资源关联及发现的搜索工具，具有重要的实用价值。本文阐述面向科学数据搜索引擎voovle提出的背景、面临的问题，重点介绍了voovle的功能与实现，详述元数据模型等几项关键技术。最后通过案例介绍了voovle目前的应用情况、存在的不足以及下一步的发展方向等。

关键词： 科学数据；搜素引擎；元数据；语义；关联数据

voovle: A Scientific Data-Oriented Search Engine

Li Chengzan, Shen Zhihong, Li Jianhui
Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China

Abstract: With the development of scientific research, scientific data has been exploding increasingly. In the days when scientific research activities and data consuming across multiple disciplines become more and more common, it is necessary to provide a unified and cross-database search engine for open data and metadata. In the first part of this paper, the authors show the background and the possible problems of the scientific data-oriented search engine. Then, this paper presents voovle as a scientific data-oriented search engine and focuses on the functions and implementation. In section three, some key technologies of voovle are discussed in detail. Finally, current application situation, existing shortcomings and the future direction of voovle are briefly described.

Keywords: Scientific database; Search engine; Metadata; Semantic; Linked data

1. 背景

科学数据是各类社会科技活动所产生的原始性、基础性的数据，以及按照不同需求系统加工的数据集和相关信息[1]。科学数据是科技活动的结晶，是科技发展与创新的基石，是国家安全和社会经济发展的重要保障[2]。

许多发达国家建立了一批全球性的数据共享工程，如美国的“分布式最活跃数据档案中心群”、世界气象组织的“全球电信数据交换系统”和国际减灾协会组织的“国际灾害信息资源网络”等工程[3]。

我国自2002年起开始了国家科技基础条件平台建设的试点工作[1]，以整合离散的海量科学数据资源，将各部门、各单位所积累的科学数据资源纳入到国家科学数据共享统一框架，形成跨部门、跨学科、多层次的国家科学数据共享服务体系，使海量的科学数据资源的潜在价值得以充分发挥与增值[4-5]。中国科学院（以下简称“中科院”）“十一五”数据应用环境启动实施以来，按照“统筹规划、整合集成、公开共享、服务科研”的原则，已有62家研究所共同参与了科学数据资源建设[6]。截至2010年底，科学数据库已整合共享数据148TB，形成了538个数据库。其中包括化学、材料、空间、天文、遥感、人地系统、动物、微生物等8个主题库，聚变、青海湖、冰雪冻土、生态功能区划等4个专题库，化合物和植物物种2个参考型库，以及土壤、海洋、地球化学等37个专业库。形成了由专业库、主题库、专题库和参考型库组成的资源整合框架[7]。

随着科学数据开放共享进程的加快，科学数据开放的内容、广度和深度都有了很大程度的提升，但是同时也依然面临着诸多现实问题亟待解决。

1）缺乏对分布开放的科学数据资源进行有效组织、提供高效利用的措施

科学数据之间的内在关联、学科之间的交叉性的客观存在越来越被人们所认知，而当前数据应用环境中的共享科学数据资源虽然数据量颇为庞大，但是各个数据库物理上相互独立，各自为战，没有很好的将科学数据逻辑上应有的关联性表达出来。要充分发挥这些科学数据库的优势和价值，就必须将这些分布的科学数据库的数据资源有机整合和关联起来，形成一个逻辑上的整体。

同时，由于数据来源于不同领域的不同学科，数据呈现出了较多的异构性。从科学数据库应用上看，存储格式和领域模型的异构性急需关注。首先，存储格式异构性体现在结构化、半结构化以及非结构化上。结构化的形式多以关系型数据库为主，而关系型数据库又因采用不同诸如Oracle、MySQL、SQL Server数据库管理系统而导致数据库模型各异。其次，在领域模型上，由于学科的差异，不同应用领域中的数据模型有着先天的差异性。而且即便在同一个领域，不同建库单位所采用的概念模型、物理模型也存在着千差万别[8]。所以，要在物理上完全整合异构科学数据资源，就会受到上述数据异构性问题的困扰，除此之外，还包括因为保密性要求高低不同带来的安全权限限制问题、物理上集中之后的数据更新维护问题、物理集成和应用需要专业人员参与等诸多因素制约，几乎不可行。

因此为了把分布在不同领域、不同单位、不同系统中的异构科学数据整合起来，实现高度集成的数据查询功能及增值服务，需要制定一种伸缩性良好的元数据模型和中间件的普适性方案，对数据库中的数据进行规范化处理，对集中的元数据信息进行统一管理，实现对不同系统中的异构数据集进行方便集成。这样的数据整合，对原始数据库管理系统的改造很小，同时可实现用户透明地访问原始系统的所有数据资源。

2）缺乏统一的科学数据检索入口

目前面对种类繁多的科学数据库，尚没有一个统一的科学数据检索入口。用户如果要对科学数据进行检索和获取，首先需要知道所要查询的科学数据属于哪个单位，甚至需要知道这些科学数据在哪个网站上，大多用户遇到这类问题往往无所适从。如果能为用户提供一套面向科学数据查询的单一入口的检索方案，简

化科学数据检索过程，将对科学数据的共享和最大利用发挥积极的作用。

3）传统搜索引擎不能满足科学数据库中的数据检索的需要

传统的搜索引擎并不适合于科学数据库中的数据检索。传统搜索引擎技术普遍采用关键字匹配方式进行检索，忽视语义关联以及专业领域特性，从而在检索中出现系统查全率、查准率不高，经常出现错检、漏检等情况。检索到的数据缺少语义和上下文语境支持[1]，而且结果往往只局限在库的层面上，很少获取到数据层面的结果[8]。而科学研究跨学科、跨领域的重要性日益凸显，不同学科、不同领域数据间的关联关系也更加突出。目前还没有一个既能够支持科学数据跨学科、跨领域、支持关联发现的广度搜索需求，又能够满足单一学科、单一专业领域的深度检索需求的科学数据搜索引擎。

总之，随着数据应用环境建设的不断完善，数据存量和数据增量直线递增。如何设计一套科学数据搜索引擎，准确获取用户所需专业知识，最大限度发挥科学数据的作用，已经成为科学数据共享进程中必须解决的一个问题[1]。可以说，随着学科关联、交叉与融合的进一步深入，如何结合科学数据的专业特点，针对科学数据资源在开放元数据资源集成的基础上提供跨库统一检索功能及入口，提供不同领域科学资源的深度和广度检索，提供关联及发现支持，将具有深远的现实意义。

2. 科学数据搜索引擎voovle

针对上述诸多问题，本文开发了面向科学数据、可实现跨库关联检索的voovle检索系统。

1）voovle的功能

voovle具备在服务器端自动创建、更新科学数据索引的功能，能够帮助用户通过简单的界面操作、快速检索和定位数据，浏览详细的数据信息。

voovle具备自动抓取通过VisualDB（中科院计算机网络信息中心研发的一套可视化关系数据管理发布平台）[9]部署的数据库系统中的元数据信息，并进行

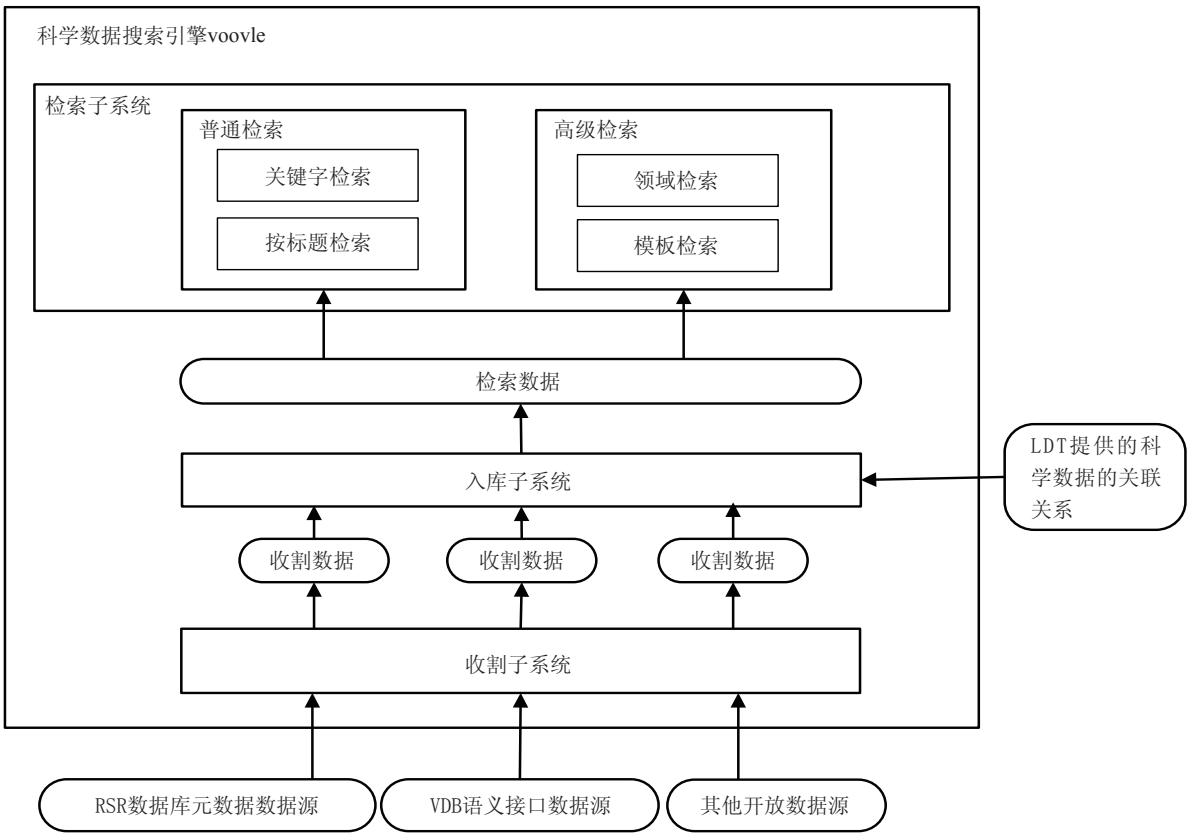

图1 科学数据搜索引擎voovle的整体结构图

基于语义的半结构化集中存储的功能。在此基础上提取索引信息，提供多关键字的统一精确检索和模糊检索、按数据集的领域检索、基于规则的关联发现和语义检索等功能。最终为用户提供一个科学数据资源的集中检索和统一定位的入口，来支持用户快速的进行深度和广度数据检索，同时通过浏览检索到的元数据信息，可以通过页面连接定位到各数据库系统的具体数据上。为了避免某些站点因服务停止等原因造成无法访问的问题，voovle同时提供了保存数据快照的功能。

2）voovle的整体结构

科学数据搜索引擎系统voovle的整体结构如图1所示。该搜索引擎主要由检索子系统、入库子系统、收割子系统三部分组成。voovle主要通过检索子系统对外提供服务。而元数据的统一收割和集中入库，则分别由收割子系统和入库子系统实现。三个子系统相互独立，保证了系统的稳定运行。

（1）收割子系统

收割子系统负责从站点注册中心获取注册的站点列表，然后通过注册站点对外开放的接口逐一对注册站点的元数据进行抓取的工作，同时将这些元数据及该站点内数据资源的关联关系等统一封装成RDF格式，提供给入库子系统进行加工和统一入库。收割子系统支持定时自动收割和手动收割两种方式。值得说明的是，通过VisualDB部署的站点可以通过中科院计算机网络信息中心提供的数据收割中间件wod直接对外提供元数据生成服务。

（2）入库子系统

入库子系统对收割子系统收割到的元数据信息进行整合、索引，然后通过关联发现工具(LDT)对跨库数据资源间的关联规则进行设定与加工，最终将加工处理后的RDF（Resource Description Framework）信息进行统一入库操作，为检索子系统提供可供检索的数据支持。

（3）检索子系统

检索子系统是voovle的核心部分，基于入库子系统处理整合的RDF数据基础上，对外提供跨库统一检索服务，支持基于规则的推理和关联发现。检索子系统提供普通检索和高级检索两种功能。

①普通检索主要为检索目标较为模糊的用户提供检索服务，包括关键字搜索和按标题搜索两个功能。

②高级检索为熟悉科学数据且检索目标明确的用户提供检索服务，包括领域检索和模板检索。

领域检索功能介绍：只提供特定领域的科学数据资源的检索服务，屏蔽其他领域的检索结果，方便用户快速定位相关资源。

模板检索功能介绍：提供针对资源某一个具体属性的检索服务（不同于关键字检索功能的是，关键字检索功能只能对资源全部属性进行统一检索），方便用户更精准的定位数据。

3. 关键技术

1）元数据模型

元数据是关于数据的数据，在数据共享过程中元数据具有数据发现、数据获取、数据管理与交换等功能。元数据为各种形态的数字化信息单元和资源集合提供规范、描述方法和检索工具。为分布的、由多种数字化资源有机构成的信息体系提供整合的工具与纽带[10]。为各种信息的集成提供支持，为集中检索提供保障。当前国内外数据共享平台多以元数据为核心[11]。

voovle中提出了一套元数据模型。将科学数据资源的元数据信息在物理上进行集成和统一存储。各注册站点都按照该元数据模型对收割子系统提供接口。该元数据模型由两部分组成：一部分是在参考都柏林核心（Dublin Core）元数据[12]基础上提出的通用元数据元素，这一部分元数据元素比较固定；另一部分是将各站点设置的共享索引数据项作为元数据模型的特殊构成元素，这部分元数据元素具有更多的自主性和灵活性，各建库单位可以根据领域科学数据模型自行进行选择和设定，不同单位不同科学数据模型的共享索引（元）数据项各不相同。这套元数据模型即兼顾了元数据的通用性，又考虑到了不同领域科学数据的特殊性和扩展性。其中通用元数据元素部分示例如表1所示，由索引数据元数据元素如表2所示（以中国脊椎

表1 通用元数据元素示例

术语名称	含义	来源
title	标题	DC元数据元素集
keyword	关键词	DC元数据元素集
abstract	摘要	DC元数据元素集
created	创建日期	DC元数据元素集
source	数据来源	DC元数据元素集
url	url	DC元数据元素集
creator	创建者	DC元数据元素集
classification	数据分类	DC元数据元素集
type	类型	DC元数据元素集
relation	关联	DC元数据元素集
organization	单位	DC元数据元素集
contact	联系人信息	DC元数据元素集

表2 索引（元）数据示例

名称	含义	来源
cname	中文名	站点索引数据项
genus	拉丁属名	站点索引数据项
cgenus	中文属名	站点索引数据项
family	拉丁科名	站点索引数据项
cfamily	中文科名	站点索引数据项

动物分类代码数据库中的斑头雁分类代码记录为例）。

2）RDF

RDF(Resource Description Framework)是W3C提出的一种元数据描述模型，RDF标准设计的目的是描述Web上的资源、内容和关系[12]。RDF现在已经成为了用于描述一般信息、资源和关系的标准，将这些Web信息赋予确定的语义。同时，通过RDF还能描述概念实体以及实体之间的关系。总之，RDF提供了分布web环境中对各种信息资源的一种统一的语义描述方式[13]。在RDF模型中，现实世界被描绘并解释成一系列资源(resource)组成的集合，每个RDF语句可以表示为一个三元组<subject，predicate，object>，它表示subject对应资源的predicate属性的值是object对应资源或值。在RDF基础上，RDF Schema通过一系列有确定语义的词汇(例如“: class”)来描述概念层次的语义，在其中可以确切定义类概念和属性概念的语义。我们采用RDF来描述网格资源信息内容，就可以为它们赋予确定的语义，从而实现数据资源的统一描述[12]。

RDF文档使用XML编写。使用XML技术为基石建立数据集成平台，成为数据集成技术发展的趋势。通过使用XML，RDF信息可以轻易地在使用不同操作系统和应用语言的计算机之间进行交换，而XML本身的扩展性可以方便的支持不同领域的科学数据结构。

voovle中的入库子系统最终将不同建库单位的科学数据按照元数据模型进行收割并以RDF的格式在TDB物理库中进行重新组织、统一整合和集中存储。以中国土壤数据库中的土壤普查农田肥力表中的红土记录为例，其具体RDF格式如下：

```
<rdf:RDF>
    <rdf:Description
    rdf:about="http://www.soil.csdb.cn/wod/resource/fertility_80/topsoilNutrient/20422>
    <!—通用元数据项-->
    <dc:title>红土</dc:title>
    <dc:keyword>农田,耕层,土壤肥力,土壤普查</dc: keyword >
    <dc: creator >中国科学院南京土壤研究所</dc: creator >
    <!—索引元数据项-->
    <!—基本数据类型-->
    <topsoilNutrient:profile_type>A11—C型</topsoilNutrient:profile_type>
    <topsoilNutrient:landuse>旱地</topsoilNutrient:landuse><topsoilNutrient:om>1.12</topsoilNutrient:om>
    <topsoilNutrient:soil_type_name>红土</topsoilNutrient:soil_type_name>
    <topsoilNutrient:agrochem_property>
    据49个农化样分析：有机质含量1.12%，全氮0.067%，速效磷6ppm，速效钾123ppm。有效微量元素硼0.64ppm，钼0.17ppm，锰5ppm，锌0.8ppm，铜0.6ppm，铁4ppm。
```

</topsoilNutrient:agrochem_property>

<!—引用类型-->

<topsoilNutrient:soilSubclassSubclass_id rdf:resource="http://www.soil.csdb.cn/wod/resource/fertility_80/soilSubclass/70"/>

<!—集合类型-->

<topsoilNutrient:sublocation2sublocation_id rdf:resource="http://www.soil.csdb.cn/wod/resource/fertility_80/sublocation2/587"/>

<topsoilNutrient:sublocation2sublocation_id rdf:resource="http://www.soil.csdb.cn/wod/resource/fertility_80/sublocation2/589"/>

</rdf:Description>

</rdf:RDF>

3）开发工具选择

voovle采用Jena作为查询、语义解析和推理的工具。选择Jena的原因主要包括：Jena是一套表示和处理半结构化数据的Java开源工具包，它包含RDFS的应用程序接口API（Application Programming Interface），支持读写RDF数据，并支持数据表达、解析、查询和推理，从而支持语义相关应用开发[14]。Jena提供了ARQ查询引擎，支持通过SPARQL查询语言对本体模型进行查询。SPARQL查询是面向数据的，SPARQL语言本身不支持任何推理功能[15]。Jena的推理子系统（Inference Subsystem）允许将一些推理引擎或推理机引入到Jena中，从而获得本体数据中隐含的信息。推理机制支持RDFS语言，可以执行从实例（instance）到类（class）的推理。Jena提供基于规则的推理机，包括RDF推理机等，在基于规则的推理机中包含了一般的推理功能，同时也支持用户自定义推理规则[16-17]。

4. 应用情况

目前科学数据搜索引擎voovle已投入运行，为用户提供科学数据统一检索的服务。目前voovle中共收录了包括青海湖联合科研基地数据库、中国黑土生态数据库、中国土壤数据库、东北植物与生境数据库、中国湖泊科学数据库等在内的37家建库单位的124个数据集中的元数据资源。此外，voovle还收录了CERN野外台站、e-CarbonScience数据集元数据信息，元数据记录数达到了5,646,706条。下面以“斑头雁为例”简单介绍一下voovle的使用流程。

首先登录科学数据搜索引擎voovle的检索主页面，如图2所示。用户可以选择进行精确查询或者模糊查询，也可以对检索的范围进行选择，如图3所示进行数据集选择。检索设置完成之后只需要输入检索关键字，比如“斑头雁”，点击检索即可。

在检索结果的列表显示页面，用户可以点击资源标题或者“详细信息”链接，查看资源的元数据信

图2 科学数据搜索引擎voovle检索主界面—输入检索关键字

图3 科学数据搜索引擎voovle检索主界面—选择检索数据集

图4 科学数据搜索引擎voovle检索结果列表—查看元数据/查看原始数据

图5 科学数据搜索引擎voovle检索结果细览

图6 数据应用环境门户提供的科学数据检索入口

息和原始数据信息（当服务站点无法连接时会显示原始数据的快照信息）。同时用户也可以通过“查看原始数据”直接链接到数据资源的原始网站进行查看。其中检索结果的列表显示页面如图4所示。

在详细结果详细页面，用户可以直接查看数据的元数据信息和原始数据快照信息。同时在详细结果显示页面的下方，还将显示该数据资源的管理数据，这些数据可以是同数据集的关联数据，也可能是通过设定的推理规则推理出来的跨库关联数据资源。具体如图5所示。

另外中科院数据应用环境门户网站同样提供了科学数据搜索引擎的检索入口，调用页面如图6所示。

5. 总结

voovle科学数据搜索引擎为科研人员提供了面向数据应用的科学数据统一检索入口，支持科研人员快速检索、关联和发现感兴趣的数据。通过应用和总结，也逐步发现了voovle中需要进一步完善的地方。其一是改进改进索引的分词算法，提升分词效果；其二是提出一套有效查询结果的去重和排序方法；其三是增强基于领域的特色检索服务；其四是需要开展数据挖掘服务。另外需要增加用户搜索过程的信息跟踪，如记录用户点击哪些结果、对哪些关键词进行更改、查询关键词与所处地理位置的关系等信息，并对这些信息进行存储和分析，有助于基于元数据的检索、关联与发现的改进和用户行为等分析。因此，随着Linked Data、Web of Data的发展，下一步需开展voovle 2.0的研发，形成基于语义、更加专业化的科学数据搜索引擎，支持更便捷、更适宜的科学数据检索服务。

参考文献:

[1] 李丽亚, 宋扬. 基于Ontology的科学数据共享检索体系解析[J]. 情报理论与实践. 2009(5): 81.

[2] 姚松涛. E—Science环境下科学数据的整合与共享.现代情报. 2009(5): 128–130.

[3] 徐枫. 科学数据共享标准体系框架[J]. 中国基础科学, 2003(1): 44.

[4] http://www.china.com.cn/chinese/PI–c/239334.htm.

[5] http://www.amadata.net.cn/gxgc.aspx.

[6] http://www.risn.org.cn/Norm/Flfg/ShowInfo.aspx?ID=11545.

[7] http://www.csdb.cn/prohtml/0.news.news/pages/3105.html.

[8] 沈志宏, 吴开超. voovle: 面向科学数据的搜索引擎的设计与实现[J]. 科学数据库与信息技术论文集(十). 201007: 240–246.

[9] http://vdb.csdb.cn.

[10] 胡德华. 元数据对搜索引擎的支持力度研究. 中国卫生信息技术交流大会. 20041101: 90–92.

[11] 诸云强, 刘润达. 分布式地球系统科学数据共享平台研究. 计算机工程与应用. 2009(1): 245–248.

[12] http://dublincore.org/.

[13] 任磊, 谭跃生. 基于RDF元数据的网格资源统一描述方法. 内蒙古科技大学学报. 2009(2): 148–151.

[14] http://www.w3.org/TR/REC–rdf–syntax.

[15] http://www.w3.org/TR/rdf–sparql–query/.

[16] http://jena.sourceforge.net/.

[17] http://imarine.blog.163.com/blog/static/513801832008613744292 72/.

收稿时间：2011年6月21日

作者信息

李成赞

中国科学院计算机网络信息中心科学数据中心，软件工程师。主要研究方向为科学数据管理与发布、关联与集成。

沈志宏

中国科学院计算机网络信息中心，科学数据中心高级工程师。主要研究方向为数据库应用。

黎建辉

中国科学院计算机网络信息中心，博士，正高级工程师，硕士生导师。主要研究方向为大规模科学数据管理共享与应用研究。