卓越科学家数据语义关联与搜索发现研究*

金国栋 范炜
(四川大学信息管理技术系，成都 610664)

摘要：以人物数据为中心，探索卓越科学家语义描述和关联发现的技术实现路径。利用Sesame对数据进行存储管理，发布语义数据，实现围绕卓越科学家的搜索发现。从学科领域、奖项、组织机构、科研成果、地理位置等多个维度挖掘卓越科学家的关联信息，实现了卓越科学家数据的关键词检索，对外提供Web Service，通过人物中心节点图和人物地图可视化地显示卓越科学家数据的语义关联。

关键词：卓越科学家；语义关联；搜索发现
中图分类号：G254
DOI：10.3772/j.issn.1673—2286.2014.04.004

1 引言

当今时代，科学呈现快速发展与学科交叉融合态势。紧跟科学发展前沿和实践科学发展观需要强有力的信息资源基础与服务手段。以学术人物为中心的资源关联角度，能够有效定位专家和参考权威成果，是信息资源管理支撑科学研究与社会管理事务的重要研究课题。刘俊婉（2010）[1]通过ISI Highly Cited（高被引科学家）数据库完成对“杰出科学家”的识别，将“杰出科学家”限定为“ISI高被引科学家”的代名词。另外，学术荣誉称号也是科学家科研活动的重要评价指标。汪士（2013）[2]将中国科学院院士作为我国杰出科学家的典型群体。

所谓“卓越科学家”，即专业领域顶尖专家和学术精英，他们通常由不同科研领域内的顶尖专家和学术精英构成，具有权威话语权，是学术共同体的领袖。卓越科学家数据是以卓越科学家为中心的相关数据的聚合，现阶段网络中卓越科学家数据相对分散，缺乏专门针对人物数据的发布平台，阻碍了共享和交换。

语义网的目标是建立机器可读可理解的数据网络（Web of Data），在此基础上实现语义推理。开放数据在公共信息服务领域有大量的应用，在遵循一定协议和规则的前提下，能够有效实现数据的互通与共享。这为围绕卓越科学家的关联发现提供了技术手段和数据基础。

本文从人物数据关联角度，探索语义描述、存储、发布与搜索的资源应用，以期促进卓越科学家的关联发现。

2 相关研究

（1）人物描述

通用人物描述主要有Brickley D等人提出的FOAF（The friend of a friend Project）[3]、Google和Yahoo!发布的Schema.org[4]、DBpedia[5]的人物描述方案以及W3C规范中的vCard[6]。四者的对比情况见表1。

表1四种描述方案中，FOAF通过描述文档之间的关联构建社交网络；Schema.org可以帮助搜索引擎更好地理解网页内容；DBpedia实现了对人物的百科全书式地描述；vCard则主要用于进行个人信息的交换。

在特定人物描述模型的构建方面，其中比较典型的有诺贝尔奖（Nobel Prize）获奖者模型。诺贝尔奖官方网站[7]定义了诺贝尔奖获奖者模型，描述了获奖者的姓名、性别、出生时间及地点、死亡时间及地点、获奖学科、获奖年份、获奖原因、颁奖时所在机构以及研究领域等信息，突出表现获奖者在所属领域内的主要学术成就。

* 本研究得到四川大学中央高校基础科研业务经费项目“关联数据集描述与发现服务研究”（编号：skq201204）和四川大学大学生创新创业训练计划项目“可视化语义搜索引擎——以卓越科学家搜索为例”（编号：20130564）资助。

表1 人物资源描述方案比较

描述方案	描述词汇	说明
FOAF	foaf:name，foaf:account，foaf:img，foaf：weblog，foaf:homepage，foaf:publications等。	主要描述用户Web主页中通常包含的内容，将人与互联网联系起来。
Schema.org	name，image，url，email，follows，knows，sibling，spouse，telephone等。	对网页中的人物实体进行标注，引入了对亲人的描述。
DBpedia	name，birthDate，award，hometown，institution，education，knownFor，nationality，parent，wife等。	对人物的描述比较充分，描述了人物的姓名、出生等基本信息，组织机构等社会活动信息以及家人等人物关系。
vCard	v:title，v:email，v:photo，v:agent,v:tel，v:postcode，v:org等。	又称电子名片，是一种简单的交换个人信息的方法，关注人物的联络信息、地点信息和机构信息。

以上人物描述形式中，通用描述方案强调描述的广度，能够适用于大范围的人群，特定描述模型强调描述的深度，着重体现某一类人群的突出特点。因此，在考虑人物描述的通用基础之上，突出表现卓越科学家的学术特点，丰富化和精准化卓越科学家的描述。

（2）卓越科学家数据分布情况

DBpedia从Wikipedia的页面中抽取多语种的结构化数据，与Freebase、GeoNames等其他数据集相连接，共描述了超过198,000个人物[8]，包括物理、化学、计算机等众多学科领域的卓越科学家。诺贝尔官方网站存储了所有诺贝尔奖获奖者的信息，截至2013年共有876位获奖者。另外，其他著名奖项，如数学学科的菲尔兹奖（Fields Medal，国际杰出数学发现奖）、计算机学科的图灵奖（A.M. Turing Award）等也不同程度地存储了获奖者的姓名、出生、死亡、科研机构、教育背景和学术成就等信息，截至2013年，菲尔兹奖共有52位获奖者，图灵奖共有60位获奖者。

DBpedia和诺贝尔官方网站均在人物的描述中引入了语义信息，并提供开放的数据接口，便于共享和交换。其中DBpedia的数据能以N-Triples或Turtle格式整体下载，也能通过SPARQL语句进行查询，诺贝尔官方网站提供了REST API和SPARQL Endpoint两种数据获取方式。菲尔兹奖和图灵奖等其他著名奖项的人物数据无开放的数据接口，需要手工搜集。

（3）找寻与发现手段

找寻卓越科学家相关信息，一般通过人物搜索和学术搜索两种途径。在人物搜索方面，微软的人立方[9]是一个典型例子。人立方通过对网页内容进行自然语言处理，抓取出网页中的人名，构建人物数据库，形成人与人之间的关联，供用户检索和浏览。但由于数据来源于普通网页内容的抓取，数据内容的质量较低，结构化程度不高，且学术性相关信息少。在学术搜索方面，谷歌学术搜索[10]和微软学术搜索[11]的应用十分广泛。谷歌学术搜索收录了论文、图书、科技报告、文摘等多种学术资源，内容涵盖了自然、人文、社会等多种学科，同时支持中英文等多种语言的检索[12]，能让用户像使用通用搜索引擎一样地使用学术搜索引擎，降低了检索难度，很重要的一点是研究者可以利用谷歌学术搜索的“被引用次数”来查看某一篇文章的被引文献，从而可以追踪不同研究者基于同一研究主题的相互引用关系[13]。但是谷歌学术搜索并未直观地展示研究者之间的关系，且没有开放数据调用的API。微软学术搜索能可视化地展示研究者之间的合著和引文关系，可以查看以某研究者为中心形成的合著和引文关系节点图。

3 卓越科学家数据的语义模型

（1） 人物关联顶层设计

从通用的人物模型出发，提取出与卓越科学家相关的基本描述属性，并融合科研描述属性，将卓越科学家与地理位置、组织机构、科研成果、学科、奖项等类进行关联，构建卓越科学家的人物关联模型。

人既具有生物性，又具有社会性和精神性。社会性主要包括人的社会活动和社会关系等，精神性主要包括人的精神状态、心理活动和思维活动等。社会性和精神性最大程度地体现了卓越科学家与其他人物群体的不同，因此，本文从卓越科学家的社会性和精神性两个层面切入，将

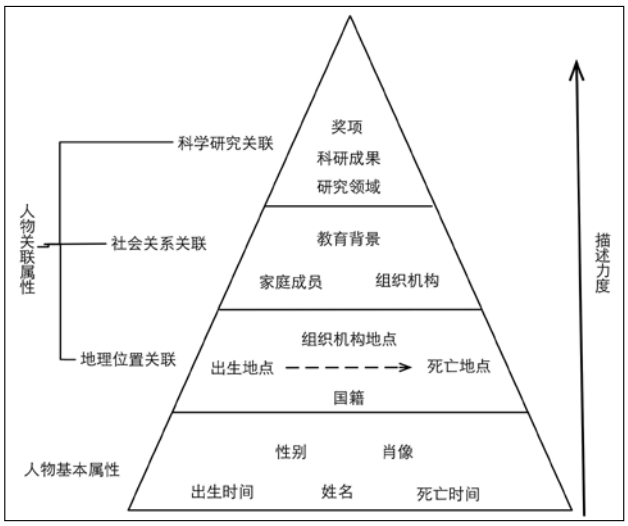

图1 卓越科学家描述的金字塔模型

其描述属性分为基本属性和关联属性。描述模型见图1。

图1基本属性主要描述卓越科学家的基本特征，包括姓名、性别、肖像、出生时间和死亡时间等，通过基本属性的描述，可以在大体上形成一个人物的形象；关联属性从地理位置、社会关系和科学研究的角度，描述了卓越科学家之间丰富的关联信息，其中地理位置信息描述了人物从出生到死亡经历的重要的地理位置变化，社会关系从家庭、教育和工作三个角度关联了人物的家庭成员、教育背景和相关组织机构，科学研究主要描述人物的研究领域、重要研究成果和所获著名奖项。上述两大类的四种属性对卓越科学家的描述力度呈递增关系，基本属性的描述力度最小，科学研究关联属性对卓越科学家的描述力度最大，即最能体现卓越科学家群体的特点。

（2）人物关联定义

基于人物的属性，人物之间的关联可以相应地分为直接关联和基于中间关联层的推理关联两种。直接关联即两个人物之间通过属性直接产生联系，如配偶关系；基于中间关联层的推理关联指两个人物之间需要借助中间层进行两次或以上的直接关联，才能产生联系，以人物A与人物B的校友关系为例，$School(X,S)$表示人物X是学校S的学生，$Alumna(A,B)$表示人物A和人物B是校友，则$(\exists S)[School(A,S) \wedge School(B,S)] \rightarrow Alumna(A,B)$，该关系需要通过学校作为中间层经过两次关联推理得到。

本文基于人物之间的中间关联层进行推理关联，结合已建立的人物类，在人物关联模型的中间关联层中融入了地理位置、组织机构、学科、奖项、科研成果等五个中间类。其中，人物的地理位置信息包括出生地点、死亡地点、受教育地点、国籍等；组织机构信息包括所在教育机构和研究机构的信息；奖项信息主要描述人物在所处研究领域获得的著名国际奖项；学科信息主要为人物研究领域的相关信息；科研成果信息主要描述人物的重要科研成就。对这些信息进行描述，一方面有助于加强对人物进行多维度的揭示，如地理位置信息有助于对人物群体按地点进行关联分析，学科、奖项和科研成果有助于了解人物的研究领域，发掘人物之间在科学研究方面的合作关系。另一方面，可以挖掘出人物之间

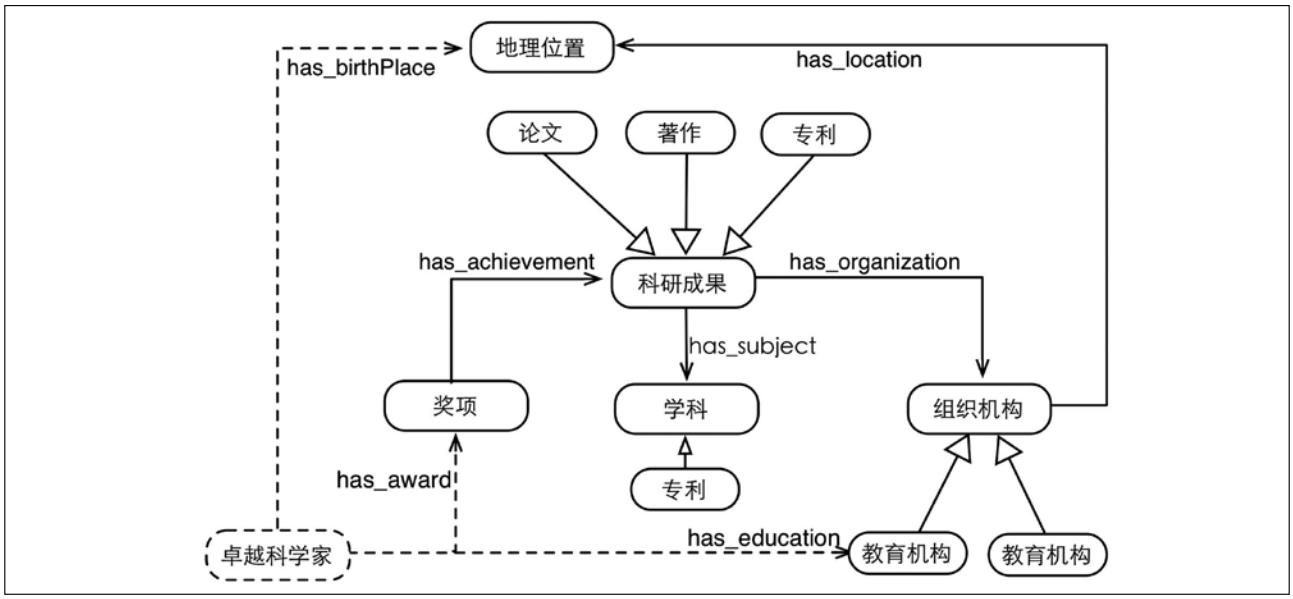

图2 人物中间关联层模型

众多的隐性关联。人物中间关联层模型如图2所示。

图2共有12个类，其中科研成果类派生出论文、著作和专利三个子类，组织机构类派生出教育机构和科研机构两个子类。卓越科学家类通过人物的出生地点、教育背景和所获奖项分别与地理位置、教育机构和奖项类产生关联，图中以虚线表示；组织机构通过所在地与地理位置类关联，科研成果类根据论文、著作、专利等的所属机构与科研机构关联，同时根据其所属学科与学科类关联，奖项类通过获奖原因与科研成果类关联，图中以实线表示。

基于以上分析，人物之间的关系分为直接和间接两类，共有七种。直接关系为家庭成员关系，卓越科学家之间的家庭成员一般有科研合作关系或处于相同科研领域；间接关系为相同的出生或死亡地点（相同地理位置）、校友（相同教育机构）、同事（相同科研机构）、共同研究领域（相同领域）、奖项共享（相同奖项）以及科研合作（相同科研成果）等六种。

人物之间形成的关系网络图为有向图G=（P，R），其中P表示图中的节点，即人物，R表示图中的边，即人物间的关系。则人物P_i与P_j之间的上述七种关系可以对应分别表示为r_{ij}^k。若P_i与P_j之间的第k种关系存在，则r_{ij}^k为1，否则为0。根据不同关系所基于的属性的描述力度不同，为上述关系设置不同的权重值，分别为0.2，0.05，0.075，0.1，0.15，0.2，0.225。因此，人物P_i与P_j之间的属性关联值可定义为：

$$r_{ij}=\sum_{k=1}^{7}a_k r_{ij}^k \quad (1)$$

其中，a_k为第k种关系所占权重。

4 关联发现设计与实现

本文设计的关联发现系统主要由卓越科学家数据采集、语义存储管理、语义发布与搜索等功能模块组成，提供关键词检索、基于图的人物关系可视化以及Web Service调用。系统技术架构如图3所示。

4.1 数据采集与预处理

（1）数据范围

为了更好地获取数据和展示关联，以学科作为数

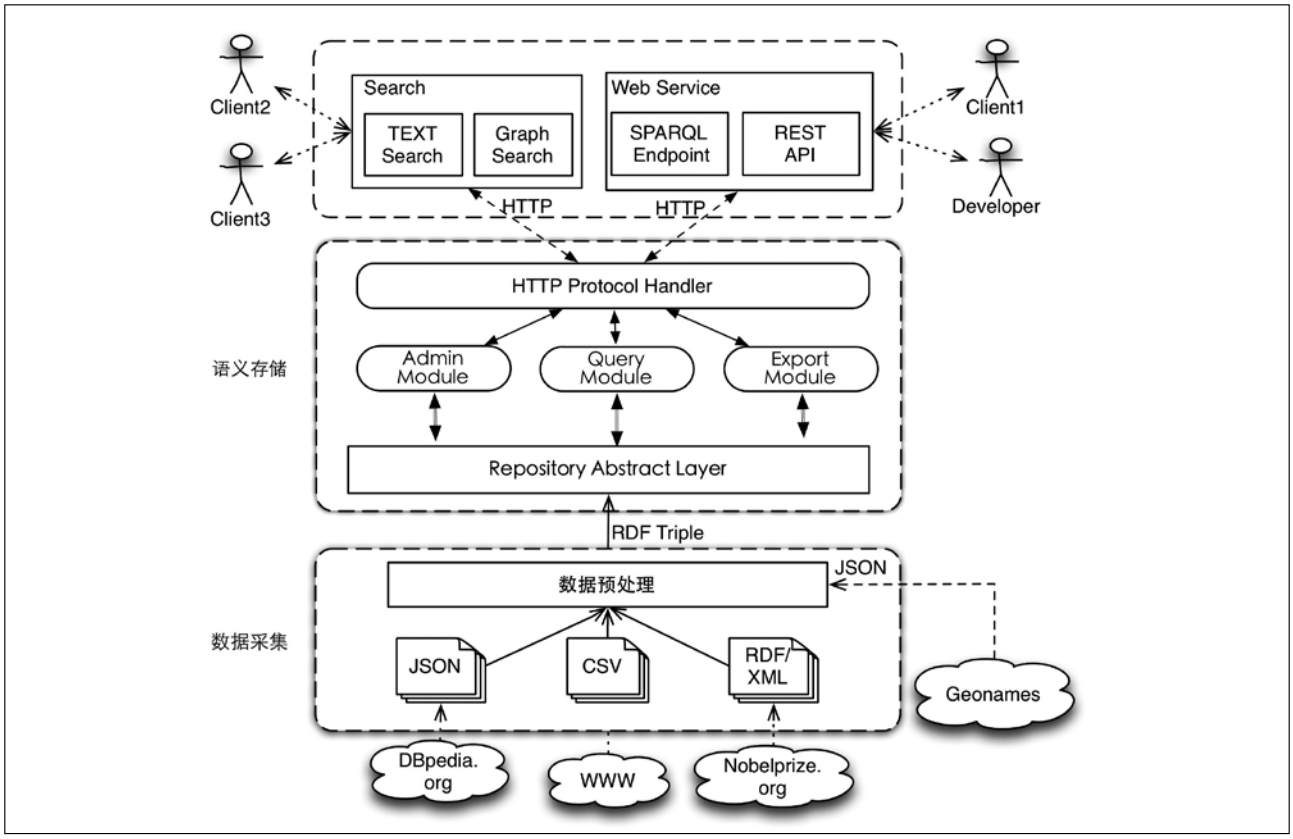

图3 系统技术架构图

表2 人物数据范围列表

奖项名称	描述方案	获奖人数
诺贝尔物理学奖	物理学	196
诺贝尔化学奖	化学	166
诺贝尔生理或医学奖	生物医学	204
菲尔茨奖	数学	52
沃尔夫数学奖	数学	54
图灵奖	计算机科学	60
IEEE荣誉奖章	电气电子工程学	92

据选择的切入点，选取物理、化学、生物医学、数学、电气电子工程学等自然科学领域的卓越科学家作为研究对象。获得学科领域内著名奖项的科学家在一定程度上可以作为该学科领域内卓越科学家的代表。因此，本文选取获得诺贝尔科学奖、菲尔兹奖、沃尔夫数学奖（Wolf Prize in Mathematics）、图灵奖和IEEE荣誉奖章等自然科学领域著名奖项的797位卓越科学家构成实验数据集，其中部分科学家获得两项及以上奖项。

如表2所示，实验数据集中每个人物形成一条记录，每条记录设置firstName、lastName、birthDate、deathDate、birthCity、deathCity、gender、education、award、familyMember、img、nationality等12个描述字段，分别描述卓越科学家的名、姓、出生日期、死亡日期、出生地点、死亡地点、性别、教育背景、所获奖项、家庭成员、肖像、国籍等信息。

（2）数据采集说明

实验数据集通过以下方式得到：

• 诺贝尔奖官方网站的开放数据描述了获奖者的姓名、出生、死亡、获奖时所在机构、获奖学科、获奖时间和获奖原因等信息，可以通过两种方式访问，一种是通过REST API，返回CSV或JSON文件，另一种是通过SPARQL Endpoint查询返回RDF格式的结果。本文采用第二种方式获得诺贝尔科学奖的全部人物数据，导出为RDF/XML格式存储于本地。

• 通过DBpedia的SPARQL Endpoint构造SPARQL语句，查询所有获奖者的姓名、出生时间及地点、死亡时间及地点、性别、机构、肖像、国籍等信息，以JSON格式返回，获奖者信息按奖项进行分类，每个奖项以一份JSON文件的形式单独存储于本地，共采集到七个奖项共600余条人物信息。

• 调用GeoNames的Web Service API，获得人物相关地区的经纬度及行政区域划分数据，返回JSON文件。

• 人工辅助采集。卓越科学家的部分信息，如科研成果和部分地理位置信息等需要通过搜索引擎在WWW中人工采集，采集结果以CSV格式存储于本地，其中科研成果信息选取能代表科学家获奖原因的被引率最高的一篇论文、一本著作或一项专利，得到824条记录。

（3）数据预处理

由于采集到的原始数据来源多样，格式不统一，在进行存储之前，有必要对其进行预处理。数据预处理分以下三步进行：

• 完整性检查。对数据完整性的检查分为两个方面：第一是否采集了数据范围内的所有人物的信息；第二是每个人物的每个字段是否都有描述信息。本文的实际采集情况是第一种全部采集完整，而第二种存在部分不完整。

• 一致性检查。对数据中的日期、组织机构名称、空值等进行一致性的检查。原始数据中日期格式有“MM-DD-YYYY”、“YYYYMMDD”等多种，均转化为“YYYY-MM-DD”格式；原始数据中组织机构名称有简称与全称两种形式，均转化为简称；原始数据字段中的空值有“NULL”、“None”以及空字符串等多种形式，为了方便处理，本文中将日期字段的空值设为“1111-11-11”，其余字段的空值设为“NULL”。

• 数据合成。不同采集来源的原始数据之间存在大量的重合，为了进一步减少数据集的冗余，优化系统的查询结果，本文对不同来源同一对象的描述信息进行合成，按照六个类分别存储为6份本地JSON文件。

4.2 语义数据存储与发布

（1）语义描述

按照关联数据发布的流程[14]，采集得到的JSON数据需要添加语义描述，转化为语义数据，再进行存储和发布。添加语义描述的步骤如下：

• 设定http://www.excellentscientists.org为语义数据的基础URI；

• 选择词汇表。为增强数据的互操作性，在JSON数据的描述字段的基础上融入了FOAF、RDF、RDFS、OWL等的部分描述词汇；

• 添加内部链接和外部链接。添加本地文件中类之间的关系链接，以及与GeoNames、DBpedia和Nobelprize官网等的链接。

通过上述步骤将JSON文件转化为RDF文件，以Turtle格式存储。以对居里夫人（Marie Curie）及其所获奖项和科研成果的描述为例，如图4所示。图中（1）描述了居里夫人的姓名、出生时间及地点、死亡时间及地点、获得奖项、家庭成员等人物信息，利用owl:sameAs与foaf:page链接至DBpedia，利用scientist:birthCity与scientist:deathCity链接至Geonames，并与奖项（prize）、科研成果（achievement）、组织机构（organization）等类形成关联。（2）对1903年的诺贝尔物理学奖进行了描述，包括获奖时间、获奖者、获奖原因等，通过prize:title链接至诺贝尔官网，同时与科研成果类形成关联。（3）中描述了科研成果信息，包括类别、作者、领域以及相关组织机构等，与学科类和组织机构类关联。

（2）语义数据存储

常见RDF文件的存储管理方案有Jena[15]、Sesame[16]和4Store[17]等，其中，Sesame最早作为On-To-Knowledge项目的一部分，由荷兰公司Aduna开发，后推出开源版本。本文在比较之后，选择Sesame作为存储方案，主要出于以下三个方面的考虑：Sesame由Java语言编写实现，具有良好的跨平台性；在RDF数据的导入和查询的速度方面，Sesame有不错的表现[18]；Sesame除了可以作为Java类库本地调用以外，还可以利用其内嵌的HTTP Server封装为一个独立的系统，通过客户端程序远程调用，能够满足本地存储管理与远程查询的需求。

本文主要调用Sesame中的RDF Model API、Rio API和Repository API创建RDF存储查询系统，并建立spoc、posc、cosp三种索引，以提高检索效率。首先利用Model API和Repository API创建一个本地存储库，添加索引，然后利用RepositoryConnection接口连接本地存储库，导入上文中转化的RDF文件，再利用Query Engines实现SPARQL语句查询模块，最后调用Rio API将查询返回结果封装为JSON、XML及RDF格式等，导出查询结果。

（3）语义数据发布

系统在存储的基础上实现了SPARQL Endpoint。在Endpoint中，用户可以输入SPARQL语言进行查询，查询语句通过HTTP协议传送至语义存储模块，该模块执行查询操作后将结果返回Endpoint，最后将结果以HTML形式显示在浏览器页面上。查询结果可以导出为XML、JSON以及常见RDF格式文件。以对居里夫人的获奖情况和研究领域的检索为例，设计检索语句如下：

```
SELECT ?field ?prizeTitle ?prizeYear
WHERE { ?id data:firstname ‘Marie’; data:lastname ‘Cruie’;
        scientist:prize ?prize; ?prize prize:title ?prizeTitle ;
        ?prize prize:year ?prizeYear; ?prize:achiv ?achievement ;
        ?achievement achievement:field ?field . }
```

系统还设计了REST API访问方法，方便开发者获取系统中的人物数据，结果以XML格式返回。上述检

```
id:6 foaf:firstname 'Marie' ;
     foaf:lastname 'Curie' ;
     foaf:gender 'female' ;
     foaf:img 'www.nobelprize.org/nobel_prizes/physics/laureates/
1903/marie-curie_thumb.jpg' ;
     foaf:page <http://dbpedia.org/page/Marie_Curie> ;
     rdf:type foaf:Person ;
     rdfs:label Marie Curie, nee Sklodowska ;
     scientist:birthDate '1867-11-07'^^xsd:date ;
     scientist:birthCity <http://sws.geonames.org/756135> ;
     scientist:birthCityName 'Warsaw' ;
     scientist:deathDate '1934-07-04'^^xsd:date ;
     scientist:deathCity <http://sws.geonames.org/6443046> ;
     scientist:deathCityName 'Sallanches' ;
     scientist:nationality <http://www.geonames.org/798544/poland> ;
     scientist:award prize:nobelPhysics_1903 ;
     scientist:award prize:nobelChemistry_1911 ;
     scientist:education organization:Sorbonne_University ;
     scientist:daughter id:190 ;
     scientist:husband id:5 ;
     scientist:soninlaw id:189 ;
     owl:sameAs <http://dbpedia.org/resource/Marie_Curie> ;
```

（1）人物语义描述

```
prize:nobelPhysics_1903
     prize:year '1903'^^xsd:date ;
     prize:title <http://data.nobelprize.org/resource/
category/Physics> ;
     prize:owner id:4 ;
     prize:owner id:5 ;
     prize:owner id:6 ;
     prize:achiv achivement:Sur_une_nouvelle_substance
_fortement_radio-active_contenue_dans_la_pechblende .
```

（2）奖项语义描述

```
achivement:Sur_une_nouvelle_substance_fortement
_radio-active_contenue_dans_la_pechblende
    rdfs:label 'On the invisible rays emitted
by phosphorescent bodies' ;
    rdfs:type achivement:paper ;
    achivement:author id:4 ;
    achivement:author id:5 ;
    achivement:author id:6 ;
    achivement:year '1896' ;
    achivement:field field:nuclear_physics ;
    achivement:org organization:University_of_Paris .
```

（3）科研成果语义描述

图4 语义描述示例 (Marie Curie)

索语句可以转化为以下的REST请求：

http://localhost:8080/openrdf-sesame/repositories/scientist?query=select+?field+?prizeTitle+?prizeYear+where+%7B?id+data:firstname+'Marie'…etc…

4.3 基于关系的搜索发现

（1）关系计算

人物之间的相关度是对相邻人物之间亲疏程度的直接描述，对其进行定量计算有助于我们更好地分析人物之间的关系。

人物网络图中，相邻人物P_i与P_j之间的相关度大小不仅同P_i与P_j间的属性关联有关，还与P_i、P_j的共同相关人物，即朋友的朋友关联有关。计算思路如下：

- 找出图中P_i与P_j之间所有无重复节点的路径l_{ij}，并计算每条路径中除去P_i、P_j的节点个数m，记为该路径的度($m \geq 0$)；
- 计算所有度为m的l_{ij}的长度，即路径所有相邻人物P_i与P_j间关系大小r_{ij}的乘积，并求出每个m值下的最大值$\left|l_{ij}^{m}\right|$；
- 综合所有$\left|l_{ij}^{m}\right|$的值。路径长度越大，则在该路径上两端的人物之间的关系越弱，故设置递减系数β，为0到1之间的小数，利用指数函数对$\left|l_{ij}^{m}\right|$的值进行递减，本文取0.8。

主要计算公式如下：

$$Relevance(P_i,P_j)=\sum \beta^{m}\left|l_{ij}^{m}\right| \quad (2)$$

基于Sesame存储的人物语义数据，构建人物之间七种关系的查询语句，找出每个人物的相邻人物，即人物网络中的相邻节点，将其转化为邻接表。计算每两个人物之间的属性关联值，将结果存储于$N \times N$的相关度矩阵中（N为人物数目，实验数据集为797）。以居里夫人（Marie Curie）为例，与其相关度值最高的8位科学家的计算结果如表3所示，其中Pierre Curie、Irène Joliot-Curie、Frédéric Joliot与居里夫人均有家庭成员关系，Antoine Becquerel、Pierre Curie与居里夫人之间存在科研合作关系。

（2）关键词检索

用户可以输入人名关键词对人物进行检索。关键词检索模块将用户输入的文本信息包装成相应的SPARQL查询语句，通过HTTP协议对Sesame模块进行远程检索，Sesame模块将查询结果以JSON格式返回，关键词检索模块再对JSON结果数据进行解析，并依据上文计算的相关度值，对结果进行排名，相关度越高则排名越靠前，最后以文本形式返回给用户。以对居里夫人（Marie Curie）的检索为例，检索结果如图5所示，共检索到43位相关人物，排名前两位的为相关度值最高的Pierre Curie和次之的Irène Joliot-Curie。

（3）基于Graph的关联发现

为更好地展示关联发现的结果，在基于Graph的关联发现模块中将检索出的底层关联数据转化为JSON格式，调用d3.js[19]类库，以动态的中心节点图的形式展示人物关联。

动态的中心节点图展示了以某人物为中心关联发现的结果，在关键词检索结果中点击View Graph查看

表3 关系计算结果示例

No.	科学家	相关度值	研究领域	主要关联
1	Pierre Curie	0.940	nuclear physics	夫妻关系，共享奖项（1903年诺贝尔物理学奖），共享科研成果，相同研究领域
2	Irène Joliot-Curie	0.890	nuclear chemistry	母女关系，相同研究领域
3	Frédéric Joliot	0.860	nuclear chemistry	亲人，相同研究领域
4	Antoine Becquerel	0.820	nuclear physics	共享奖项（1903年诺贝尔物理学奖），相同研究领域
5	George Hevesy	0.799	nuclear chemistry	相同研究领域
6	Otto Hahn	0.788	nuclear chemistry	相同研究领域
7	Frederick Soddy	0.760	nuclear chemistry	相同研究领域
8	Isidor Rabi	0.686	nuclear physics	相同研究领域

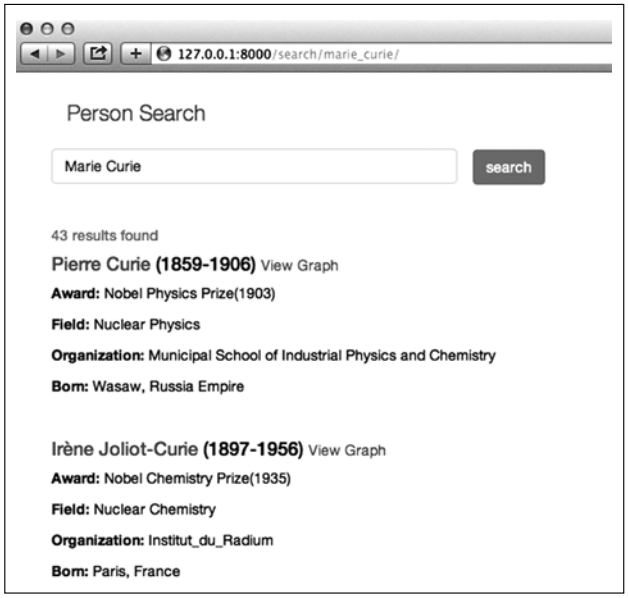

图5 基础检索示例 (Marie Curie)


```
{
  "nodes": [
  {"name":"Marie Curie", "img":"http://www.nobelprize.org/
nobel_prizes/physics/laureates/1903/marie-curie_thumb.jpg",
"group":0 },
  {"name":"Antoine Becquerel", "img":"http://www.nobelprize
.org/nobel_prizes/physics/laureates/1903/becquerel_thumb.jpg",
"group":1 },
  {"name":"Pierre Curie", "img":"http://www.nobelprize.org/
nobel_prizes/physics/laureates/1903/pierre-curie_thumb.jpg",
"group":1 },
  {"name":"Gabriel Lippmann", "img":"http://www.nobelprize
.org/nobel_prizes/physics/laureates/1908/lippmann_thumb.jpg",
"group":1 },
  {"name":"Jean Leray", "img":"http://www.math.sciences.univ
-nantes.fr/sites/default/files/layer1.jpg", "group":4 }
  ],
  "links": [
  {"source": 1, "target": 0, "value": 8.2 },
  {"source": 2, "target": 0, "value": 9.4 },
  {"source": 3, "target": 0, "value": 3.2 },
  {"source": 4, "target": 0, "value": 3.2}
  ]
}
```


图6 中心节点图JSON数据片段

相应人物的中心节点图。图中人物节点间连线的粗细对应人物之间的相关度值。以居里夫人（Marie Curie）为例，JSON数据片段如图6所示，图中nodes中存储了人物的姓名、肖像和学科信息，links中存储了人物之间的关联信息。检索结果如图7所示，图中不同学科的人物名称以不同的颜色标注。从图中可以看出，物理、化学、生物医学和数学领域的43位科学家与居里夫人（Marie Curie）形成了关联。

（4）基于地图的关联发现

地图从地理位置信息的角度展示了人物之间的聚

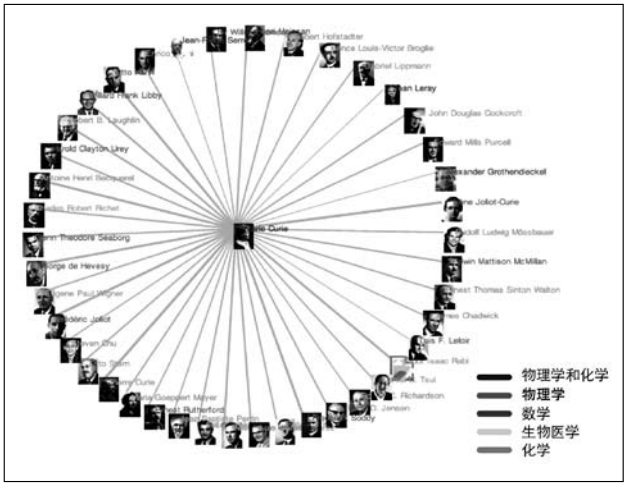

图7 中心节点图示例 (Marie Curie)


```
{
  "items":[
  {
    "label": "Pierre Curie",
    "birthCity": "Paris, France",
    "imageURL": "http://www.nobelprize.org/nobel_prizes/
physics/laureates/1990/taylor_thumb.jpg",
    "almLoc": "49.962908, 2.350044"
  },
  {
    "label": "Jean Perrin",
    "birthCity": "Lille, France",
    "imageURL": "http://www.nobelprize.org/nobel_prizes/
physics/laureates/1926/perrin_thumb.jpg",
    "almLoc": "50.633131, 3.057976"
  },
  {
    "label": "Rene Thom",
    "birthCity": "Montbeliard, France",
    "imageURL": "http://upload.wikimeia.org/wikipedia/
common/d/de/Rene_Thom.jpeg",
    "almLoc": "47.513491, 6.797619"
  }
  ]
}
```


图8 人物地图JSON数据片段

图9 人物地图示例 (France)

集关系，在基于地图的关联发现模块中，以地理位置为查询关键词构建SPARQL语句，利用已采集的人物数据和地理位置数据，关联查询与某地理位置相关人物的姓名、出生地点、头像和出生地点经纬度信息，返回JSON数据，再利用可视化工具Exhibit[20]绘制人物地图，可在关键词检索结果中点击View Map查看。以出生地为法国（France）的卓越科学家为例，JSON数据片段如图8所示，items中label字段存储人物姓名，birthCity字段存储出生地点，imageURL字段存储头像的URL，almLoc字段存储经纬度信息。绘制出人物地图如图9所示，地图中共聚集了37位卓越科学家。

5 总结与展望

本文从学科领域、奖项、组织机构、科研成果、地理位置等多个维度挖掘卓越科学家的关联信息，构建了人物关联模型，利用Sesame存储人物语义数据，对外提供关键词检索和Web Service，并融入可视化技术展示关联发现图，促进卓越科学家之间的关联发现。由于实验中数据集较小，选取的人物关系数量较少，影响了人物之间关联的发掘和关联度的计算。因此，在后续研究中将增强系统数据的开放性，鼓励用户贡献数据，并进一步增加关系维度，实现更加丰富且准确的关联发现。

参考文献

[1] 刘俊婉.杰出科学家论文影响力的社会年龄分析[J].情报学报,2010,29(1):121-127.

[2] 汪士.中外杰出科学家行政任职差异及其影响[J].科技进步与对策,2013,30(6):134-138.

[3] FOAF [EB/OL]. [2013-12-08]. http://www.foaf-project.org.

[4] Schema.org [EB/OL]. [2013-12-08]. http://schema.org.

[5] DBpedia [EB/OL]. [2013-12-08]. http://dbpedia.org.

[6] vCard [EB/OL]. [2013-12-08]. http://www.w3.org/Submission/2010/SUBM-vcard-rdf-20100120.

[7] 诺贝尔官方网站[EB/OL]. [2013-12-08]. http://nobelprize.org.

[8] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia - A crystallization point for the Web of Data [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 154-165.

[9] 人立方[EB/OL]. [2013-12-08]. http://renlifang.msra.cn.

[10] 谷歌学术搜索[EB/OL]. [2013-12-08]. http://scholar.google.com.

[11] 微软学术搜索[EB/OL]. [2013-12-08]. http://academic.research.microsoft.com.

[12] 陈国华,汤庸,彭泽武,等.基于学术社区的学术搜索引擎设计[J].计算机科学,2011,38(8):171-175.

[13] NORUZI A. Google Scholar: The new generation of citation indexes [J]. Libri, 2005, 55(4): 170-180.

[14] HEATH T, BIZER C. Linked data: Evolving the web into a global data space [J]. Synthesis lectures on the semantic web: theory and technology, 2011, 1(1): 1-136.

[15] Jena [EB/OL]. [2013-12-08]. http://jena.apache.org.

[16] Sesame [EB/OL]. [2013-12-08]. http://www.openrdf.org.

[17] 4Store [EB/OL]. [2013-12-08]. http://4store.org.

[18] HASLHOFER B, MOMENI R E, SCHANDL B, et al. Europeana RDF store report [J]. 2011.

[19] d3.js [EB/OL]. [2013-12-08]. http://d3js.org.

[20] Exhibit [EB/OL]. [2013-12-08]. http://www.simile-widgets.org/exhibit/.

作者简介

金国栋，男，四川大学公共管理学院信息管理技术系本科生。

范炜，男，1981年生，管理学博士，四川大学公共管理学院信息管理技术系讲师，研究方向：信息组织与信息检索。通讯作者，E-mail: fanwscu@163.com。

Semantic Association, Searching and Discovering for Excellent Scientists Data

JIN GuoDong FAN Wei

（Department of Information Management Technology, School of Public Administration, Sichuan University, Chengdu 610064, China）

Abstract: Centered on person data, this paper explores a technical route of semantic description and linkage discovery for excellent scientists. Based on a semantic data model of excellent scientists, we add semantic annotations, and manage the semantic data with Sesame. Then, we discover the linkage of excellent scientists by their discipline areas, awards, organizations, scientific research achievements and geographical locations. Finally, We provide web services through keywords retrieval and map their semantic associations.

Keywords: Excellent scientists; Semantic association; Searching and discovering

（收稿日期：2014-04-01）