科技资源元数据的关联与推荐方法

宋 佳[1,4] 高少华[2] 杨 杰[1,3] 诸云强[1,4,5]

（1.中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室，北京 100101；
2.武汉大学资源与环境科学学院，湖北武汉 430079；3.中国科学院大学，北京 100049；
4.江苏省地理信息资源开发与利用协同创新中心，江苏南京 210023；5.白洋淀流域生态保护与京津冀可持续发展协同创新中心，河北保定 071002）

摘要：大数据背景下，科技资源发现和推荐的关键是建立海量、多类型科技资源间的关联，并对其进行相关度排序。在深入研究科技基础性工作专项科技资源核心元数据的基础上，选择科技资源的内容特征、资源地点和资源时间为关联要素。然后结合专家打分和层次分析法，提出了科技资源元数据语义相关度算法，建立了科技资源间的关联。进一步按照相关度计算结果对科技资源进行排序，并将相关度高的科技资源优先推荐给用户。最后以科技基础性工作专项项目汇交的科技资源元数据为例，开展了科技资源元数据关联与推荐的实践。本研究提出的方法为促进海量科技资源的精准发现、智能推荐与共享应用提供了借鉴。

关键词：科技资源；元数据；语义关联；语义相关度

中图分类号：G203 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2017.05.005

Association and Recommendation Method for Metadata of Scientific and Technical Resources

SONG Jia[1,4], GAO Shaohua[2], YANG Jie[1,3], ZHU Yunqiang[1,4,5]

(1.State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101;2.School of Resource and Environment Science, Wuhan University, Wuhan, Hubei 430079;3.University of Chinese Academy of Sciences, Beijing 100049;4.Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, Jiangsu 210023;5.Collaborative Innovation Centre for Baiyangdian Basin Ecological Protection and Jingjinji Regional Sustainable Development, Hebei University, Baoding, Hebei, 071002)

Abstract: In the context of big data, efficient discovery and recommendation for scientific and technical data resources is to build the association between these data resources and then sort them by relevancy. Based on the investigation of core metadata of National Special Program on Basic Works for Science and Technology of China, this study chooses the content, location and temporal information of the data resources as association

作者简介：宋佳（1980—），男，博士，中国科学院地理科学与资源研究所助理研究员，研究方向：地球信息科学（通讯作者）；**高少华**（1993—），女，武汉大学资源与环境科学学院硕士研究生，研究方向：地图学与地理信息系统；**杨杰**（1990—），男，中国科学院地理科学与资源研究所硕士研究生，研究方向：地学模型数据匹配方法；**诸云强**（1977—），男，博士，中国科学院地理科学与资源研究所研究员，主要研究方向：地学数据本体与共享、资源环境信息系统。

基金项目：科技基础性工作专项项目“科技基础性工作数据资料集成与规范化整编”（2013FY110900）；国家自然科学基金重点项目“网络文本蕴含信息理解与知识图构建”（41631177）。

收稿时间：2017年7月31日。

factors. Then, a semantic relevance algorithm is proposed based on the method of expert scoring and analytic hierarchy process, and the semantic association between these data resources is achieved in this study. These data resources are able to be sorted in terms of the semantic relevance, and the data resources with high relevance value can be recommended to the users. The proposed method is validated in the application case of data archiving and sharing for the projects of National Special Program on Basic Works for Science and Technology of China, and it has great significance in promoting the accurate discovery, intelligent recommendation and sharing for scientific data.

Keywords: scientific and technical resources, metadata, semantic association, semantic relevance

1　引言

科技资源包括科学数据、图集、志书/典籍、标本资源、标准规范、论文专著或研究报告等。在大数据背景下，各类科技资源实体的数量以前所未有的速度增长，如何有效地将这些实体进行关联并在检索过程中为用户推荐最相关的科技资源，已经成为一个迫切需要解决的科学问题。

元数据的提出为科技资源之间的关联提供了必要条件。所谓元数据，就是关于数据的数据。王国复等[1]认为，元数据是对数据资源的规范化描述，它是按照一定标准（即元数据标准），从数据资源中抽象出相应的特征属性，组成的一个特征元素几何（即元数据元素）。元数据不仅方便用户使用数据资源，而且随着网络技术的发展和数字化资源的猛增，元数据在数据共享、资源的发现以及知识管理方面的作用越来越明显，越来越为人们所重视[2]。虽然目前众多项目、系统和平台在建设过程中纷纷涉及并颁布共享相关元数据标准[3]。但是，由于科技资源类型较多，其中涉及的学科领域也很多。所以，在数据描述等方面普遍存在语义异构现象，以及关键词在表达搜索意图时的局限性，使得以关键词匹配为主的检索方法在效率和结果质量上不再满足用户需求[4]。

为了解决这个问题，面向语义的数据关联的研究应运而生，其主要是通过数据语义，建立数据间的关联。目前的研究方法大致有两类：一是通过建立相应的领域本体进行推理，实现数据间的关联。如王东旭[5]、侯志伟[6]、孙凯[7]分别研究了地学领域的本体，并构建了地学数据的空间本体、时间本体以及形态本体，并将其应用于对地理空间数据的语义检索和发现中，取得了较好的检索结果。但是，科技资源涉及多个学科及多个领域，采用基于本体的语义推理方法需要构建面向多学科、多领域的完整的知识概念体系，这将导致面向科技资源的本体构建过程变得极为复杂而难以完全实现。二是通过定量地描述元数据之间的相关度来建立数据之间的关联。该方法具有构建过程简单且适用于多学科、多领域、多来源科技资源的特点。现在已有很多学者在此方面展开研究。例如，在对地理空间数据的研究上，诸云强[8，11]通过考虑数据主题、分类、空间拓扑、时间拓扑、空间精度、时间粒度、数据类型、数据格式等8个基本特征提出了地理空间元数据的多尺度和定量的关联方法，并通过计算相似度实现数据的推荐。赵红伟[9-10]根据地理空间数据在空间、时间、内容上的语义关系，提出了地理空间数据本质特征语义相关度计算模型，并利用RDF设计了地理空间元数据关联模型。通过计算元数据之间的语义相关度构建了地理空间元数据关联网络，从而有效地支持了地理空间语义关联检索与推荐等，提高了检索的查准率。针对极地科学数据，罗侃[12]建立了极地科学元数据关联指标体系，实现极地科学数据的关联查询应用。由以上研究可以看出，通过建立元数据之间的语义关联，可以丰富检索结果，使用户更容易得到所需要的数据，并且基于元数据语义关联的方法还可以实现数据的推荐，即在基于元数据语义相关度计算的基础上，将相关度高的数据排在

前面，优先推荐给用户。

目前，针对计算元数据间的语义相关度，进行关联与推荐的研究，主要以空间数据为代表。对于科技资源来说，空间数据只是其中的一部分资源，而对于其他类型科技资源的关联与推荐鲜有研究。本文以科技基础性工作专项项目产生的科学数据、图集、志书/典籍、标本资源、标准物质、论文专著和研究报告等科技资源为研究对象，通过元数据语义相关度的计算探讨科技资源的关联与推荐。

2　科技基础性工作专项科技资源核心元数据规范

作为科技创新活动的要素，科技资源是一切科技活动的核心。科技资源涵盖的学科领域众多，具有资源类型种类繁多、结构差异较大的特点。而元数据则是对各种科技资源的外部形式和内部形态的详细描述，为了能够对不同类型的信息资源进行描述和处理，不同领域的专业人员研究并定制了用于各个领域和各种场合的元数据标准[13]。其中，在国际上常用的元数据标准包括描述网络信息资源的都柏林核心元数据、描述国家数字地理空间数据的术语及其定义集合的地理空间元数据内容标准、地理信息元数据（ISO 19115）、地理信息服务（ISO 19119）等。除此之外，还有各学科领域的相关元数据标准，如生态科学数据元数据、气象数据核心元数据等。

现有元数据标准主要适应各学科领域的特定资源，而科技资源包含数据、志书/典籍、标本资源、标准物质等，难以直接采用现有的元数据标准。所以，在综合考虑各类科技资源共性特征的基础上，提出了科技基础性工作专项科技资源核心元数据规范，如表1所示，包括19个核心元数据项。其中，必选项有16项，可选项有3项。该规范已经应用在科技基础性工作专项项目的科技资源汇交工作中。

通过对表1所述科技基础性专项科技资源的核心元数据项进行选择，下文选择了表达科技资源核心特征的三要素，即资源内容（中文名称、关键词、资源类型、资源描述摘要、资源学科分类）、资源地点和资源时间，建立了科技资源之间的关联，并计算相应的语义相关度，为用户进行推荐。

3　科技资源元数据语义相关度计算

语义相关度的计算不仅包括传统字面匹配的相关度，还包括体现语义层次上概念间关系的计算。例如，用户需要与“北京市”有关的科技资源，而像“京津冀地区”的科技资源在字面匹配相关度上与“北京市”不相关，但在空间概念上相关。即两种资源的字面匹配相关度很小，但空间概念上的语义相关度很大。因此，在基于元数据的语义关联时，需要综合考虑词汇层面之间的相关度，以及词汇在空间关系、时间关系等其他语义关系上的相关度。

本文通过计算元数据在资源内容、资源地点和资源时间上的相关度，采用层次分析法，确定不同层次下各因素在本层次中的影响程度，即权重值，最终计算出不同科技资源对象之间的语义相关度。

3.1　内容相关度

资源内容包括中文名称、关键词、资源类型、资源描述摘要和资源学科分类。为了建立元数据的关联，前三者需要计算两两科技资源的字面匹配相关度，后者属于资源所属的学科，需要计算类与类之间的相关度。因此，资源内容相关度计算如式（1）所示。

$$S_1 = W_{11}S_{11} + W_{12}S_{12} + W_{13}S_{13} + W_{14}S_{14} + W_{15}S_{15} \quad (1)$$

在式（1）中，S_1 表示内容相关度；S_{11}、S_{12}、S_{13} 和 S_{14} 分别表示中文名称、关键词、资源类型和资源描述摘要的内容字面匹配相关度；S_{15} 表示资源学科的类别相关度；W_{11}、W_{12}、W_{13}、W_{14} 和 W_{15} 表示各因素的权重值。

内容字面匹配相关度计算首先要对文本进行分词，可借助IkAnalyzer分词软件[14]实现分词，两两对比计算所包含的相同词语的个数所占比例，分别得到中文名称、关键词、资源类型和资

表 1　科技基础性工作专项科技资源核心元数据规范

描述项	内容	类型	是否必选
资源标识	由系统自动产生，唯一标识资源	字符型	是
中文名称	反映资源内容及主要特征的中文名称	字符型	是
英文名称	中文名称对应的英文名称	字符型	否
资源学科分类	资源所属的学科分类	字符型	是
关键词	反映数据集的主要内容及特征	字符型	是
资源类型	如科学数据、图集、志书/典籍、标本资源、标准物质、论文专著和研究报告等	字符型	是
资源格式	资源存储格式	字符型	是
资源时间	资源内容的时间点或时间范围	日期型	是
资源地点	资源内容表述的地理位置	字符型	是
资源描述摘要	资源内容、特征等的简要描述	字符型	是
共享方式	完全开放共享、协议共享、暂不共享	字符型	是
最新修订时间	资源的最新更新时间	日期型	是
资源质量描述	对资源质量相关属性的描述	字符型	是
在线链接地址	在线获取或访问资源的网络地址	字符型	否
缩略图	反映资源概貌、内容或特征的图片	图片型	是
来源项目	支持资源产生的项目信息	字符型	是
资源负责方	产生资源或拥有资源处置权并对资源质量负责的个人及单位信息	字符型	是
资源管理方	保藏、管理和对外提供资源服务的单位及联系人信息	字符型	否
元数据管理信息	项目数据汇交联络人相关信息	字符型	是

源描述摘要在字面上的匹配相关度。

类别相关度计算采用文献[15]提供的方法。本文采用的学科分类参照国家标准《学科分类与代码GB/T 13745−2008》。具体做法是：假设需要计算相关度的两个类别在分类树上的节点分别为 X 和 Y，找到距离 X 和 Y 最近的父节点 P，可根据式（2）计算类别相关度。

$$S_{15}=\frac{2\times N(P)}{N(X)+N(Y)+2\times N(P)} \qquad (2)$$

在式（2）中，$N(X)$、$N(Y)$分别表示 X 和 Y 到 P 的距离，$N(P)$表示 P 到根节点的距离。

若存在两个科技资源实体属于多个类别，分别计算每个类别的相关度，并取最大值作为这两个科技资源实体最终的类别相关度。

3.2　地点相关度

资源地点是资源内容所表述的地理位置。资源地点相关度的计算首先需要根据该地理位置，得到其地理坐标后，将其表达为一个空间几何对象；然后根据空间几何对象之间的空间关系，计算对象之间的空间语义相关度。

对于具体的空间几何对象，其类型包括点状对象、线状对象和面状对象。鉴于科技资源的复杂性，为了实现统一表达，均将其所描述的地理位置映射到同一尺度下的空间面状对象上。

空间关系包括空间拓扑关系、空间度量关系和空间方位关系。对于空间语义相关度的计算，空间方位关系对空间语义相关度计算的影响不大，空间度量关系与具体的拓扑关系相关，而空间拓扑关系较为复杂，因此，如何根据实体之间的拓扑关系计算资源地点相关度是解决问题的关键。对于面状实体之间的拓扑关系，本文将考虑空间拓扑关系即相等、相接、相交、包含、被包含和相离 6 种关系，如表 2 所示。假设用户需要的科技资源地理位置映射到的空间实体为X，关联的科技资源的地理位置映射到的空间实体为Y，则根据X和Y的空间拓扑关系在计算资源地点相关度 S_2 时，可以分为以下几种情形。

（1）当X和Y的拓扑关系为相等或是Y被包含于X时，S_2=1；

（2）当X和Y的拓扑关系为相接时，可根据相交边界的长度以及Y的边界长度进行计算，具

体的可由式（3）求得：

$$S_2=\frac{L(X\quad Y)}{L(Y)} \tag{3}$$

在式（3）中，$L(Y)$表示实体Y的边界长度，$L(X\quad Y)$表示实体X与实体Y两个实体相接的边界长度。

（3）当X和Y的拓扑关系为相交或是X包含于Y时，具体的可由式（4）得到：

$$S_2=\frac{A(X\quad Y)}{A(Y)} \tag{4}$$

在式（4）中，$A(Y)$表示实体Y的面积，$A(X\quad Y)$表示实体X与实体Y两个实体相交的面积。

（4）当X和Y的拓扑关系为相离时，可根据X和Y的空间距离进行计算，具体的可由式（5）得到：

$$S_2=\frac{1}{D(X,Y)} \tag{5}$$

在式（5）中，$D(X,Y)$表示实体X和实体Y之间的空间距离。

3.3　时间相关度

资源时间指的是资源内容的时间点或时间范围。对于采用时间点描述的资源时间，例如科技资源中标本的采集、制备时间，为统一时间描述，需要将其转化为时间段。对于所有的时间段，均统一为以天为最小时间单位，然后采用Allen[16]提出的时间区间代数理论，根据时间段之间的13种拓扑关系如表3[9]所示，计算资源时间相关度S_3。

根据时间拓扑关系，假设用户需要的科技资源的时间段为X，推荐的科技资源的时间段为Y，计算资源时间相关度S_3，可将其分为以下几种情况：

当时间拓扑关系为相交、包含等的时间段时，可依据X和Y重叠的时间长度，由式（6）得到：

$$S_3=\frac{L(X\quad Y)}{L(Y)} \tag{6}$$

在式（6）中，$L(Y)$表示时间段Y的时间长度，$L(X\quad Y)$表示时间段X与时间段Y重叠部分的时间长度。

当时间拓扑关系为相接、相离的时间段时，可依据X和Y的时间距离，由式（7）得到；

$$S_3=\frac{1}{D(X,Y)} \tag{7}$$

在式（7）中，$D(X,Y)$表示时间段X和时间段Y之间的时间距离。

3.4　语义相关度

根据对资源内容相关度、资源地点相关度和

表2　面状实体—面状实体拓扑关系

编号	拓扑关系	图示
1	相等	
2	相接	
3	相交	
4	包含	
5	被包含	
6	相离	

表3　时间段—时间段拓扑关系

编号	中文	英文	图示
1	相等	Equals	
2	包含	Contains	
	在…期	During	
3	结束于	Finishs	
	以…结束	FinishedBy	
4	开始	Starts	
	以…开始	StartedBy	
5	相交	Overlaps	
	被相交	OverlappedBy	
6	相接	Meets	
	被相接	MetBy	
7	早于	Before	
	晚于	After	

资源时间相关度的计算，元数据语义相关度S可由式（8）得到：

$$S = W_1S_1 + W_2S_2 + W_3S_3 \quad (8)$$

在式（8）中，S表示元数据语义相关度；S_1、S_2、S_3分别表示上文计算得到的资源内容相关度、资源地点相关度和资源时间相关度；W_1，W_2，W_3表示各相关度的权重值。

元数据语义相关度的计算还需确定各因素的权重值大小。权重的确定可以由层次分析法获得。层次分析法（AHP）是由美国运筹学家Saaty[17]提出来的，其原理简单，且数学推理严格，具有很广泛的应用。根据层次分析法，首先对各种因素按照其影响程度分级分类，本文将影响元数据语义相关度的因素分为两级，分别为一级因素（资源内容、资源地点和资源时间）和二级因素（中文名称、关键词、资源类型、资源描述摘要和资源学科分类）；然后按照层次构造判断矩阵；最后通过相关计算得到每个因素的权重值。具体方法如图1所示。

（1）构造判断矩阵

判断矩阵表示在同一层次下不同因素对上一级某因素的重要程度，将因素之间的相对重要性用数值表示，构成矩阵形式。因此，对于构造的判断矩阵A=$(a_{ij})_{n\times n}$，其中a_{ij}为因素i与因素j重要性比较结果，a_{ij}的取值一般为1，2，…，9以及它们的倒数。1表示因素i与因素j同等重要，3表示因素i比因素j稍微重要，5表示因素i比因素j明显重要，7表示因素i比因素j强烈重要，9表示因素i比因素j极端重要，而2、4、6、8分别有3、5、7、9相应的类似含义，只是程度稍小，可由专家打分得到。

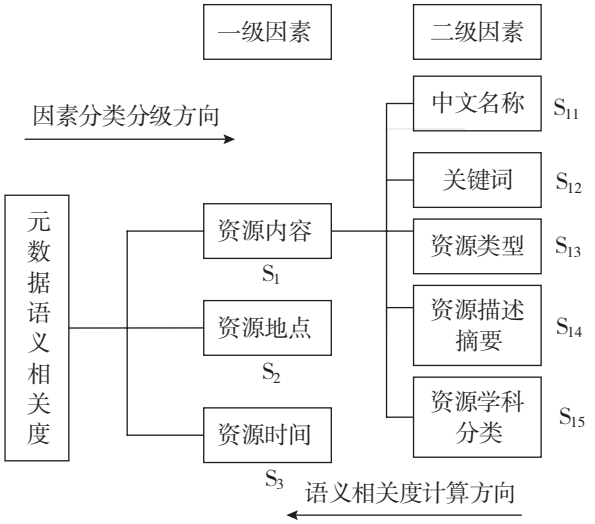

图1　元数据语义相关度计算

（2）一致性检验

在利用判断矩阵计算各因素的权重前，首先要对判断矩阵进行一致性检验。判断矩阵是否满足一致性检验，关系到后续由判断矩阵得到的权向量是否能真实反映各因素之间的客观权重。一致性检验主要由“一致性比例（CR）”来确定，其计算方法如式（9）：

$$CR = \frac{CI}{RI} \quad (9)$$

其中，CI为一致性指标，RI为平均随机一致性指标。当CR<0.10时，则认为判断矩阵一致性是可以接受的，否则需要对判断矩阵进行适当的改正。一致性指标CI的计算方法如式（10）：

$$CI = \frac{\lambda_{max} - n}{n-1} \quad (10)$$

在式（10）中，λ_{max}为判断矩阵的最大特征值；n为判断矩阵的阶数。

平均随机一致性指标RI是通过多次重复进行随机判断矩阵特征值的计算后取算术平均值得到的，其值可根据判断矩阵的阶数，经查表得到。

（3）计算权重值

计算同一层次下各种因素对上一层次中相关联因素的影响程度，即权重值，归结为计算判断矩阵的特征向量的问题。具体的计算方法有和积法、特征向量法和最小二乘法等，为了计算简便，本文将采用近似计算和积法。

根据图1，对影响元数据语义相关度计算的各级因素，通过组织相关专家进行打分，得到各级因素的判断矩阵如表4、表5所示。经过计算，表4和表5两个判断矩阵均满足一致性检验。因此，利用这两个判断矩阵得到的权向量可以真实反映各因素之间的客观权重。

基于表4和表5判断矩阵的结果，利用近似

计算和积法，得到各因素的权重值，如表 6 所示。

4　实例分析：“森林土壤剖面调查数据”元数据的关联与推荐实验

本文以资源中文名称为“2009—2010 年中国森林土壤剖面调查数据”的元数据为被关联对象，利用提出的科技资源元数据关联方法计算其他资源对象与被关联对象之间的语义相关度。按第 3 节方法和表 6 权重进行计算，得到语义相关度排序后前 15 条的计算结果，如表 7 所示。

表 7 列出的资源在相关性上随着相关度的减小而减弱，并且用本文方法计算得到的相关度与资源的实际相关程度有较好的吻合。被关联资源“2009—2010 年中国森林土壤剖面调查数据”所属学科为农学，资源类型为数据，资源时间为 2009 年 1 月 1 日至 2010 年 12 月 31 日，资

表 4　资源内容判断矩阵

S_1	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}
S_{11}	1	3	9	8	6
S_{12}	1/3	1	7	7	5
S_{13}	1/9	1/7	1	1/4	1/7
S_{14}	1/8	1/7	4	1	5
S_{15}	1/6	1/5	7	1/5	1

λ_{max}=5.3610，*CI*=0.0902，*CR*=0.0806<0.10，满足一致性检验

表 5　元数据语义相关度判断矩阵

S	S_1	S_2	S_3
S_1	1	4	2
S_2	1/4	1	1/3
S_3	1/2	3	1

λ_{max}=3.0154，*CI*=0.0077，*CR*=0.013<0.10，满足一致性检验

表 6　各因素权重值

二级因素	权重	一级因素	权重
中文名称	0.4799	资源内容	0.5573
关键词	0.2770	资源地点	0.1224
资源类型	0.0308	资源时间	0.3203
资源描述摘要	0.1202		
资源学科分类	0.0922		

表 7　“2009—2010 年中国森林土壤剖面调查数据”相关的元数据及相关度计算结果

中文名称	资源学科分类	空间关系	时间关系	资源类型	相关度
2009—2010年中国森林土壤标本资源	农学	相等	相等	数据、图集	0.9323
2009—2010年中国森林土壤指标检测数据	农学	相等	相等	数据	0.8725
2009—2010年森林土壤调查技术规程	农学	相等	相等	其他	0.8652
2009—2010年中国所有森林土壤土类和三大阶梯主要山系垂直带谱土类的标本图集	农学	相等	相等	图集	0.7371
2008—2013年中国东部土系调查土壤剖面数据	地球科学	被包含	包含	数据	0.6000
土壤生态样方调查与考察数据采集与处理规范	地球科学	相交	在……期	标准规范	0.5412
俄罗斯—蒙古土壤剖面数据集（2008—2010年）	地球科学	相接	包含	数据	0.5199
中俄蒙考察区森林野外样方调查数据集（2008、2010年）	地球科学	相交	相交	数据	0.4670
东北森林植物种质资源现状评估报告	生物学	相等	包含	研究报告	0.4346
土壤科学调查技术规程（2016年）	地球科学	被包含	晚于	标准规范	0.4027
2008—2013年中国东部土系225个整段模式标本数据	地球科学	被包含	包含	标本资源	0.3977
《中国土壤系统分类检索》第四版	地球科学	相等	包含	志书	0.3977
秦巴山区土壤背景数据	农学	被包含	早于	数据	0.3514
中国土地覆被数据库（1992、1993、2000、2001、2005、2006、2009、2014）	地球科学	相等	相交	数据	0.3019
全国地区土壤中有机氯农药及其他持久性有机污染物调查数据	环境科学技术及资源科学技术	相等	早于	数据	0.2894

源地点为中国。由表 7 可以看出，前 2 条资源在资源学科分类、资源时间、资源地点、资源类型的特征上与被关联资源完全相同，因此综合后的语义相关度最高。而第 3 条和第 4 条资源在资源学科分类、资源时间、资源地点的特征上与被关联资源完全相同，而在资源类型上与被关联资源不同，因此综合后的语义相关度较前 2 条略低。从表 7 中还可以看出，因为考虑了资源的学科特征，除了被关联资源所属的农学外，地球科学、生物学等其他相关学科的资源也可以被关联起来。同理，因为考虑了资源的类型特征，除数据类型以外的其他资源类型，如志书、图集等也可以被关联起来。这些不同学科、不同类型的科技资源不仅丰富了关联与推荐的结果，而且可以作为原有科技资源的一种补充，使用户从多个方面充分获得所需科技资源的相关信息。

5　结论与展望

本文以多学科、多领域、多渠道、多类型的海量科技资源为研究对象，在综合考虑科技资源共性特征的基础上，提出科技基础性工作专项科技资源核心元数据规范，并选择了最能表达科技资源核心特征的几个要素，即科技资源内容（中文名称、关键词、资源类型、资源描述摘要、资源学科分类）、资源地点和资源时间作为关联项，提出面向科技资源的语义相关度算法。最后对语义相关度计算结果进行排序，优先将相关度高的科技资源推荐给用户。

（1）通过提出科技基础性工作专项科技资源核心元数据规范来降低关联和推荐的复杂性，以元数据作为科技资源关联的中介对象，经对元数据相关项之间语义相关度的计算，提取并定量地表达了其中隐含的语义信息，间接地建立了科技资源之间的语义关联方法。

（2）根据语义相关度对关联资源进行排序，定量地反映了资源之间的关联程度，为科技资源的精准发现、资源推荐和共享应用提供了方法支撑。实验结果表明，通过计算元数据之间的语义相关度对科技资源进行关联与推荐的方法，具有操作简单、构建方便的特点；通过计算语义相关度，可避免传统检索方法的局限性，推荐结果在一定程度上可以满足用户的不同需要；通过元数据建立科技资源之间的关联，具有较好的可扩展性。

（3）本文在层次分析法中确定的权重带有一定程度的主观性，后续研究可考虑引入机器学习的方法，通过训练样本确定权重大小，并在计算语义相关度时适当加入其他项进行计算。

参考文献

[1] 王国复, 涂勇, 王卷乐, 等. 科学数据共享中的元数据技术研究[J]. 中国科技资源导刊, 2008, 40(1): 30−36. DOI: 10.3772/j. issn. 1674−1544. 2008. 01. 006.

[2] 徐枫. 元数据技术及其在科学数据共享中的应用. 科学数据共享管理研究[J]. 北京: 中国科学技术出版社, 2002: 178−196.

[3] 黄如花, 邱春艳. 国内外科学数据元数据研究进展[J]. 图书与情报, 2014(6): 102−109.

[4] 侯志伟. 地学数据时间本体及其在语义检索中的应用: 以地质年代本体为例[D]. 北京: 中国科学院大学, 2016.

[5] 王东旭, 诸云强, 潘鹏, 等. 地理数据空间本体构建及其在数据检索中的应用[J]. 地球信息科学学报, 2016, 18(4): 443−452.DOI: 10.3724/SP. J. 1047. 2016. 00443.

[6] 侯志伟, 诸云强, 高星, 等. 时间本体及其在地学数据检索中的应用[J]. 地球信息科学学报, 2015, 17(4): 379−390.DOI: 10.3724/SP. J. 1047. 2015. 00379.

[7] 孙凯, 诸云强, 潘鹏, 等. 形态本体及其在地理空间数据发现中的应用研究[J]. 地球信息科学学报, 2016, 18(8): 1011−1021.DOI: 10.3724/SP. J. 1047. 2016. 01011.

[8] ZHU Y, ZHU A, SONG J, et al. Multidimensional and quantitative interlinking approach for Linked Geospatial Data [J]. International Journal of Digital Earth, 2017, 10(9): 1−21.DOI: 10.1080/17538947. 2016. 1266041.

[9] 赵红伟, 诸云强, 杨宏伟, 等. 地理空间数据本质特征语义相关度计算模型[J]. 地理研究, 2016, 35(1): 58−70.DOI: 10.11821/dlyj2016.01.006.

[10] 赵红伟, 诸云强, 侯志伟, 等. 地理空间元数据关联网络的构建[J]. 地理科学, 2016, 36(8): 1180−1189.DOI: 10.13249/j. cnki. sgs. 2016. 08. 008.

[11] ZHU Y, ZHU A, FENG M, et al. A similarity−based automatic data recommendation approach for

（下转第 103 页）

7 结论与展望

生物资源考察为基础数据的获取奠定了基础，对处于考察断层期的区域或物种等开展补充考察，对考察空白区域进行系统的考察，不仅能够摸清生物资源现状，也可以为资源的合理利用与开发等奠定基础。加强对环境脆弱区生物资源的考察，加强对个别濒危物种和具有较高开发利用价值物种的专项考察，对生物多样性和遗传多样性的保护以及对我国社会经济发展等具有重要的价值和深远的意义。

（1）近10年来，基础性工作专项中对陆生生物资源的科学考察与调查最为广泛深入，在全国及南北方等大尺度上的考察比较全面，但对蝎类资源以及生物入侵物种及现状的考察在时间上有着较长的空缺。

（2）在典型区域尺度上的生物资源考察次数最多，范围最广，但依然存在考察空白区域，空白区域主要集中在华东地区以及新疆大部分地区，且对动物资源的考察较为薄弱。

（3）近10年来，基础性工作专项中对大尺度的水生生物资源的调查非常全面，但缺乏对典型区域的水生生物资源调查，尤其是受污染严重的湖泊水系以及国家水利工程建设流域等。

(4) 目前，生物资源的科学考察与调查中对个别物种尤其是珍稀濒危物种的调查较少。

参考文献

[1] 中华人民共和国环境保护部[DB/OL].[2017−05−15]. http: //www.zhb.gov.cn/.

[2] 魏辅文, 聂永刚, 苗海霞, 等. 生物多样性丧失机制研究进展[J]. 科学通报, 2014(6): 430−437.

[3] 王昊, 吕植, 顾垒, 等. 基于Global Forest Watch观察2000—2013年间中国森林变化[J]. 生物多样性, 2015, 23(5): 575−582.

[4] 袁永锋, 李引娣, 张林林, 等. 黄河干流中上游水生生物资源调查研究[J]. 水生态学杂志, 2009(6): 15−19.

[5] 葛斌杰. 中国东海北部近陆岛屿植物资源科学考察[J]. 自然杂志, 2016, 38(2): 125−131.

[6] 孙鸿烈, 成升魁, 封志明.60年来的资源科学: 从自然资源综合考察到资源科学综合研究[J]. 自然资源学报, 2010, 25(9): 1414−1423.

[7] 杨光梅, 李文华, 闵庆文. 生态系统服务价值评估研究进展: 国外学者观点[J]. 生态学报, 2006, 26(1): 205−212.

[8] 武建勇, 薛达元, 赵富伟, 等. 中国生物多样性调查与保护研究进展[J]. 生态与农村环境学报, 2013, 29(2): 146−151.

[9] 国家地球系统科学数据共享服务平台[DB/OL]. [2017−05−25].http: //www.geodata.cn/index.html.

[10] 中国科技资源共享网[DB/OL].[2017−08−15].http: //www.escience.gov.cn/.

[11] 万本太, 徐海根, 丁晖, 等. 生物多样性综合评价方法研究[J]. 生物多样性, 2007, 15(1): 97−106.

[12] 彭羽, 卿凤婷, 米凯, 等. 生物多样性不同层次尺度效应及其耦合关系研究进展[J]. 生态学报, 2015, 35(2): 577−583.

（上接第44页）

geographicmodels[J].International Journal of Geographical Information Science, 2017, 31(7): 1403−1424. DOI: 10.1080/13658816. 2017. 1300805.

[12] 罗侃, 诸云强, 程文芳, 等. 极地科学数据关联方法及应用研究[J]. 极地研究, 2016, 28(3): 361−369.DOI: 10.13679/j. jdyj. 2016. 3. 361.

[13] 许鑫, 张悦. 非遗数字资源的元数据规范与应用研究[J]. 图书情报工作, 2014, 58(21): 13−20.DOI: 10. 13266/j. issn. 0252 – 3116.2014.21.002.

[14] IK−Analyzer.v[EB/OL].[2017−08−23].http: //code. google.com/p/ik−analyzer/.

[15] WU Z, PALMER M. Verb semantics and lexical selection[C]//32nd annual meeting of the association for computational linguistics. Las Cruces, New Mexico, Stroudsburg: Association for Computational Linguistics, 1994: 133−138.

[16] ALLEN J F. Maintaining knowledge about temporal intervals[J].Communications of the ACM, 1983, 26(11): 832−843. DOI: 10.1145/182.358434.

[17] SAATY T L. How to make a decision: the analytic hierarchy process[J]. European Journal of Operational Research, 1990, 48(1): 9–26.