第2卷 第5期 2020年10月
Vol.2 No.5 Oct.2020

ISSN 2096-742X
CN 10-1649/TP

文献DOI：
10.11871/jfdc.issn.
2096-742X.2020.
05.002

文献PID：
21.86101.2/jfdc.
2096-742X.2020.
05.002

页码：13-22

开放科学标识码
（OSID）

基于领域本体的科技资源聚类方法研究

葛胤池[1]，张辉[1]，宋文燕[2*]，王轩[1]

1. 北京航空航天大学计算机学院，北京 100191
2. 北京航空航天大学经济管理学院，北京 100191

摘　要：【目的】针对科技资源分散、异构的特点，采用聚类的方法将分散、相关、相似的科技资源集成为多类型组合的资源池，以提高发现资源和利用资源的效率。本文提出一种基于领域本体的高维科技资源聚类方法。【方法】本方法构建了科技资源领域本体树和概念语义关系矩阵，并对其使用主成分分析（PCA）方法进行降维处理以构建科技资源向量空间，最终对科技资源向量空间应用K均值聚类算法得到聚类结果。与传统方法相比，本方法更适合于处理多源异构的科技资源数据。【结果】选取某国家生物种质资源库的资源数据作为科技资源集合，利用本方法得到了合理的聚类结果。【结论】本文提出的科技资源聚类方法具有三个特点：一是利用本体概念语义关系降维处理，有效降低了计算复杂度；二是较好地保留了重要的科技资源特征信息；三是生成的科技资源向量空间与聚类簇比较准确。本方法在一定程度上解决了多源异构科技资源数据的特征表示难、聚类效果差等问题。

关键词：科技资源；异构；领域本体；语义关系；聚类

Scientific and Technology Resource Clustering Based on Domain Ontology

Ge Yinchi[1]，Zhang Hui[1]，Song Wenyan[2*]，Wang Xuan[1]

1. Beihang University, School of computer Science and Engineering, Beijing 100191, China
2. Beihang University, School of economics and management, Beijing 100191, China

Abstract: **[Objective]** Clustering can gather scattered, heterogeneous but related, similar scientific and technology resources into a multi-type resource pool, which makes resource discovery and utilization more efficient. This paper proposes a clustering method for massive high-dimensional scientific and technology resources based on domain ontology. **[Methods]** This method constructs the ontology tree and concept semantic relationship matrix in the field of science and technology resources, and uses the Principal Component Analysis (PCA) method to reduce the dimensions to construct the vector space, to which the K-means clustering method is applied eventually to obtain the clustering result. Compared with the traditional methods, this method has a stronger processing capacity for multi-source heterogeneous scientific and technology resource data. **[Results]** In this paper, the rational clustering results can be obtained by the proposed method on a certain biological germplasm resource library test. **[Conclusions]** In general, the clustering method of scientific and technology

基金项目：国家重点研发计划“分布式科技资源体系及服务评价技术研究”(2017YFB1400200)；国家自然科学基金面上项目(71971012)；国家科技重大专项（2017-Ⅰ-0011-0012）

*通讯作者：宋文燕 (songwenyan@buaa.edu.cn)

resources proposed in this paper has three characteristics: first, the use of ontology concept semantic relations to reduce the dimensionality, which effectively reduces the computational complexity; second, better maintenance of important scientific and technology resource feature information; and third, more accurate resource vector space and clustering results. The proposed clustering method solves the difficult problems in feature representation and poor clustering effect of multi-source heterogeneous scientific and technology resource data to a certain extent.

Keywords: scientific and technology resources; heterogeneity; domain ontology; semantic relationship; clustering

引　言

科学技术是第一生产力，而科技资源是推动科学研究和技术创新的重要组成部分。随着科技服务等新型服务业的兴起，对科技资源进行科学有效的管理提出了更高的要求[1]。科技资源有多种分类方法，从资源形态上包含人力资源、生物种质资源、科研仪器设备、科学数据等几大类，而按教育部学科门类划分，可分为 13 个学科门类 110 个一级学科。科技资源具有地理分布、结构复杂、特征众多、数量巨大且变化频繁的特点，并根据科技资源全生命周期管理的需要随时会有新增、变化、消耗及销除等情况。科技资源集成是将分散的、相关的、相似的科技资源整合为有信息组织形态的一体，提高科技资源共享效率，促进协同创新和提高企业竞争力的方法。[2] 科技资源集成围绕完成某一项目或任务集成所需要的科技资源成套组合，形成科技资源池，提供整体解决方案。科技资源的分散性、异构性会阻碍其有效的集成与共享，因此我们需要对科技资源进行聚类以提高集成效率。

物以类聚，聚类就是指将数据划分成有意义或有用的组（簇）的方法，组内元素尽量相似，不同组的元素尽量不相似。在科技资源集成中，对科技资源进行聚类形成资源包后，后续可以通过一定的规则进行集成组合，为不同用户提供个性化的科技服务。在聚类时，需要对科技资源集合进行数据预处理，选取数据的属性和维度。由于科技资源的分散性和异构性，其维度即属性数目将会非常高，容易陷入维度灾难。传统的聚类算法应对高维数据时往往表现较差。为此先对高维数据进行降维处理再进行聚类[3]。目前常见的降维方法有：主成分分析（PCA）[4-5]、Kohonen 自组织特征映射（SOFM）[6] 以及多维缩放（MDS）等算法。此类算法一般需要预先确定数据的维数以及属性信息，无法适用于科技资源多源异构、变化频繁的情况。同时对于高维度、海量的数据使用此类算法时需要大量的计算资源。

为此本文设计了一种基于领域本体的科技资源信息降维和聚类方法。该算法相比传统聚类分析的方法有以下优点：（1）利用领域本体语义关系以适应多源异构的科技资源数据；（2）适用于海量高维科技资源数据。

1　背景

1.1　科技资源整合与共享的意义

在信息时代，科技资源作为一种重要的信息资源和战略资源，对一个国家的科技发展和进步具有非常重要的意义。我国经过长时间的科技创新发展，已经产生了大量的科技资源，而这些科技资源既是我国科技创新的重要成果也是支撑我国新一轮科技创新活动的重要保证。能否充分有效地利用这些科技资源，对于我国的科技创新与发展而言至关重要[7]。

世界各国特别是欧美等发达国家都在积极推动科技资源整合与共享工作，来促进科技创新与经济发展。如美国通过立法与专项资金支持的方式来积极推动科技资源共享，建设数据共享平台；欧盟也通过建设覆盖整个欧洲地区的科技资源共享平台的方式推动科技资源共享[8]。

近年来，随着国家对科技资源的重视以及投入的增加，我国已经在科技资源整合与共享方面取得了明显的成效。但从总体上看，我国科技资源的整

合及共享服务体系依然处于初级阶段，与发达国家相比仍然存在较大差距。为此，我们必须不断加大科技资源共享力度，解决当前科技资源遇到的问题，从而推动创新型国家建设[9]。

1.2 科技资源的特征

科技资源作为国家战略资源，具有稀缺性和增值性资源普遍共有的特性，同时科技资源还具有地域分布、差异性、异构型等特点。科技资源的特征主要包括：

（1）稀缺性

科技资源作为国家科技创新发展与进步的重要资源，相对于科技资源日益增长的需求而言总是稀缺的。科技资源的稀缺性主要体现在两方面：科技资源总产出相对不足；科技资源利用率较低。

（2）分布的差异性

科技资源的分布受到区域的经济与科技发展状况的影响，不同地区的性质各异、层次不同、各具特色的经济发展模式与科技发展政策会导致区域科技资源分布的差异性[10]。另外，各种差异性受到地域差别的影响，不同的地域具有不同的特色资源。

（3）增值性

科技资源能为科技活动提供支持，同时科技活动对科技资源进行深层次的挖掘与使用，可以实现价值的转换与增加。通过科技资源开发，既可以转化为新的科技价值，还可以转化为社会价值、经济价值等。

（4）异构性

科技资源包含范围比较广，包括了人力资源、生物种质资源、科研仪器设备、科学数据等。其中人才、仪器、信息等资源结构性质各异，使用方式和评价指标等也有极大的区别。如何将这些分散的、多样的、异构的科技资源与海量个性化的需求相匹配，是提高科技资源服务质量的关键问题。

1.3 科技资源共享面临的挑战

在不断推动科技创新发展过程中，通过政府财政专项资助和科研计划等方式的支持，我国的科技资源越来越丰富，科技资源建设取得了较大的发展。但是我国的科技服务与共享体系依然不够完善，导致科技资源并没有得到充分的利用[11-12]。我国在科技资源共享过程中存在科技资源建设重复多但同时利用率低，科技文献资源质量不高，科技资源管理人才队伍建设不足等问题[13]。

同时在科技资源服务建设过程中，因为科技资源的异构型和分布上的差异性等，科技资源及其信息往往具有分散、封闭、异构和孤立等特点。如何将分散在不同地理位置、不同部门的，具有不同属性的异构异质异种科技资源匹配多用户的个性化需求，成为推动科技资源共享与服务的一个关键问题。例如，科研工作者想要完成某项科研课题，为此需要获取特定种类的资源，则可以通过检索科技资源集成产生的多源资源服务包来确定可用的资源范围，并结合具体业务需求、预算、地理位置等进行组合和筛选实现效益最优的个性化科技服务。

许多学者对多源异构资源的集成进行了研究。在科技资源领域中，于阳对江苏省科技资源信息使用 Hadoop 大数据平台实现了不同来源的科技数据合并与存储[14]；李宗俊等提出了利用科技资源池作为虚拟化容器进行资源集成的方法[15]；宫萍等提出通过建立统一的适用于多源异构科技资源的元数据格式规范来构建基于语义本体的科技资源集成建模框架[2]。此外，针对类似应用情景的其他领域资源集成的研究中，汤华茂通过构建制造资源的分布式语义描述模型实现了异构制造资源的虚拟化描述[16]；程臻利用本体建模的方式用统一描述框架描述了异构云制造资源并建立起虚拟资源本体层次模型[17]。

以上研究虽然大都能够实现异构资源的整合，但并未提出实现个性化资源服务的完整方法。为此本文提出了一个适用于多源异构科技资源聚类的方法，通过聚类的方式将海量的、异构异质的科技资源有效地集成起来，以形成科技资源（服务池）供后续的检索和优化配置。

2 基于领域本体的科技资源聚类方法

来自不同领域的科技资源往往具有不同的描述方法及元数据标准，这便是其异构性所在。为了尽可能多和完整地保留各个领域科技资源信息的完整性，本方法对相关科技资源领域构建概念领域本体树，并将每一个科技资源的元数据信息根据概念集合进行向量化表示。据此得到的科技资源向量虽保有足够的信息，但向量的维数随着领域范围的扩大、异构性的增强而逐渐增多，容易陷入维度灾难。因此本文在进行聚类前，使用 PCA 方法对稀疏语义关系矩阵进行降维处理以得到属性较少的科技资源向量，以避免聚类出现效率低、效果差的问题。

本方法主要分为三部分，如图 1 所示：构建领域本体树及语义关系，以计算不同概念之间语义距离；根据语义距离构建科技资源向量空间；对科技资源向量进行聚类。

2.1 领域本体语义关系定义与构建

领域本体是对特定领域之中概念及其相互之间关系进行形式化表达的领域知识库，可以在宏观上反映出领域知识的梗概全貌，并可以为特定领域信息的检索、分类提供有力的支持 [18]。

在本文中，定义本体结构 $G=(V,E)$，其中为概念集合，每个概念作为树中的一个节点；为概念间的语义关系，作为边集。本体中的概念间关系可分为上下位关系和相关概念的其他关系，其中上下位关系构成了本体的树形结构，称为本体的层次树，相关概念的其他关系构成本体结构中的非上下位关系 [19]。如图 2。

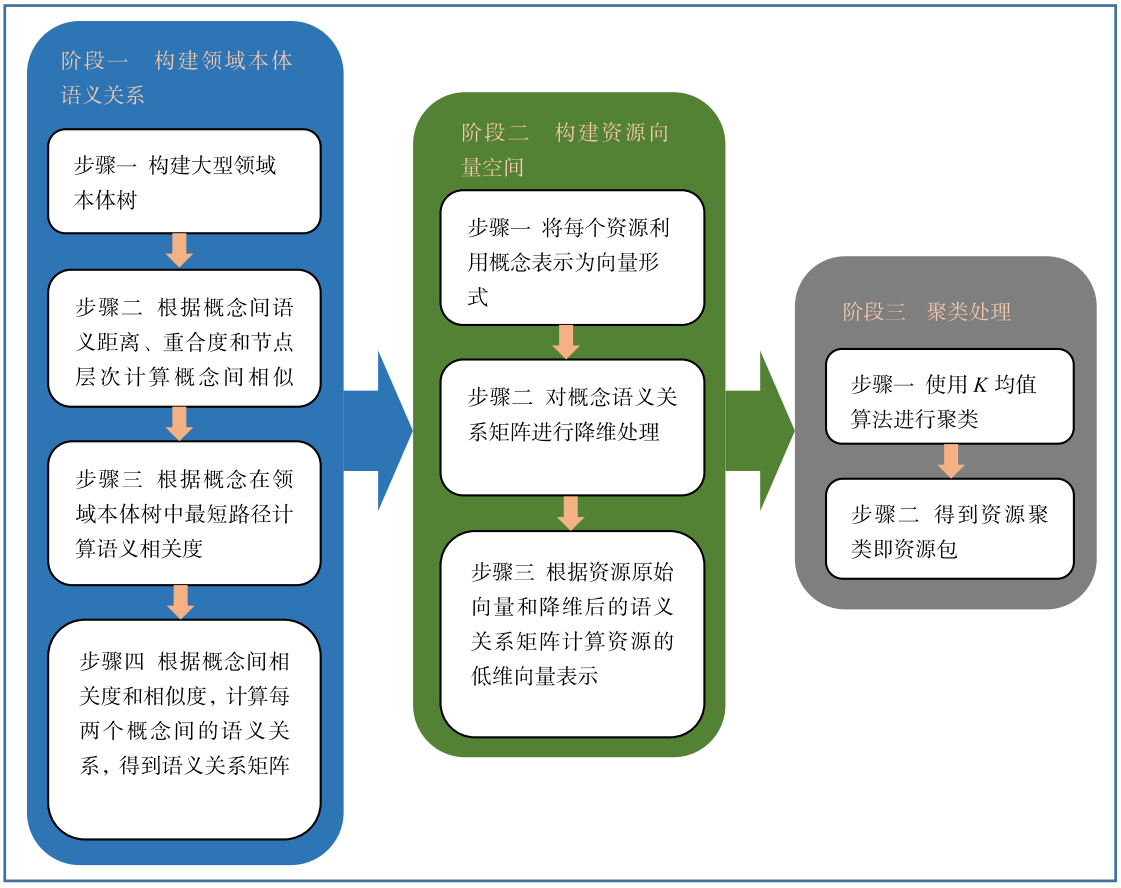

图 1 方法流程

Fig.1 Method flow

本体结构中，概念间的语义关系包含概念间语义相似度和概念间语义相关度。概念间语义相似度主要度量了本体中的上下位关系，概念间的相关度主要度量本体中概念间特有的关系[20]。

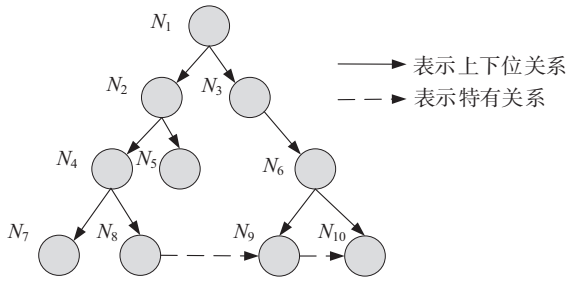

图 2 本体领域树示意图

Fig.2 An ontology domain tree

2.1.1 领域本体语义关系定义

本小节对领域本体结构中语义关系的相关概念进行定义。

定义一：语义距离。设 N_i、N_j 为本体领域树中任意两个概念节点，语义距离 $d(N_i,N_j)$ 表示从 N_i 到 N_j 所经过的路径长度。

定义二：语义重合度。设 R 为本体层次树的根。$NS(N_i)$ 是从 N_i 出发，向上直到根 R 所经过的概念节点集合。N_i、N_j 语义重合度表示为：

$$C(N_i,N_j)=\frac{NS(N_i)\cap NS(N_j)}{NS(N_i)\cup NS(N_j)} \tag{1}$$

定义三：节点层次。Level(N_i) 表示节点 N_i 在本体领域树中所处的层次。

定义四：概念节点 N_i、N_j 的相似度。定义为：

$$\mathrm{sim}(N_i,N_j)=\frac{1}{d(N_i,N_j)+1}\times C(N_i,N_j)\times\frac{1}{\left|\mathrm{Level}(N_i)-\mathrm{Level}(N_j)+1\right|} \tag{2}$$

定义五：概念节点 N_i、N_j 的语义相关度。定义为：

$$\mathrm{rel}(N_i,N_j)=\begin{cases}1 & N_i\equiv N_j\\ \dfrac{\lambda}{\mathrm{Shortest}P(N_i,N_j)+\lambda} & N_i\neq N_j\end{cases} \tag{3}$$

其中 Shortest$P(N_i,N_j)$ 为 N_i、N_j 间的最短距离。语义相关度主要用于表示领域本体树中具有非上下关系的节点间的相关程度。

2.1.2 领域本体语义关系构建

领域本体结构中概念间的语义关系应包含两种关系：本体中的上下位关系和本体中定义的其他关系。因此定义概念间语义关系 $R(N_i,N_j)$ 为：

$$R(N_i,N_j)=\mathrm{sim}(N_i,N_j)+\mathrm{rel}(N_i,N_j)-\mathrm{sim}(N_i,N_j)\times\mathrm{rel}(N_i,N_j) \tag{4}$$

由式 (4) 可以得到本体概念集合中的所有概念间语义关系，并可以表示出本体的语义关系矩阵 $S_{m\times m}$：

$$S=\begin{bmatrix}s_{11} & \cdots & s_{1m}\\ \vdots & \ddots & \vdots\\ s_{m1} & \cdots & s_{mm}\end{bmatrix} \tag{5}$$

其中 $s_{ij}(0\le i,j\le m)$ 表示概念 N_i 与 N_j 之间的语义关系，即 $s_{ij}=R(N_i,N_j)$。易知 S 为对称矩阵，即 $s_{ij}=s_{ji}$。

2.2 科技资源向量空间表示

每一个科技资源都可以根据领域本体树中的概念集合唯一表示为词袋（Bag of words, BOW）向量形式，即：

$$W_i=[w_{11}\quad\cdots\quad w_{1m}]^{\mathrm{T}} \tag{6}$$

其中，$0\le i\le k$，k 为科技资源集合中的资源总数。

对于异构异质异种的科技资源集合，由于概念领域的较大差异，其向量表示会呈现出极其稀疏的特性。且构建的领域本体树概念数目越大、覆盖领域越广，这种现象也越严重，将会增加分析和计算的难度和成本[21]。

为了降低高维数据的计算分析难度，本文采用主成分分析的方法对语义关系矩阵进行降维。主成分分析方法是将多个具有相关性的要素转化成几个不相关的综合指标的分析与统计方法，可以在保证主要信息少量丢失的前提下，对高维数据进行降维处理，把一些作用较低或不相关的指标省去，起到简化研究和提高计算效率的作用。

经过主成分分析后，本体概念集合将保留 n 个主要概念，且 $n<<m$。经过降维后的语义关系矩阵 S'

可以表示为：

$$S' = \begin{bmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nm} \end{bmatrix} \tag{7}$$

每个科技资源在降维后的语义关系矩阵 S' 下对应的向量形式表示为：

$$W_i' = S' \cdot W_i = [w_{11}' \quad \cdots \quad w_{1n}']^{\mathrm{T}} \tag{8}$$

2.3 科技资源聚类

聚类分析用于将数据划分成有意义或有用的组（簇）。在本文中，对科技资源进行聚类形成许多资源包，以供后续通过一定的规则进行集成组合，为不同用户提供个性化的科技服务，提高检索查询效率。

本文采用经典的 K 均值聚类算法。K 均值聚类算法可以描述为：首先选择 K 个初始质心，每个点被指派到最近的质心，而指派到一个质心的点集为一个簇。然后以每个簇的均值替换更新每个簇的质心，重复这个过程，直到簇不发生变化或质心不发生变化即收敛[22]。其时间复杂度为 $O(I \times K \times k \times n)$，其中 I 为收敛所需迭代次数，K 为聚类簇数，k 为点数，n 为属性数量。当 K 显著小于 k 时，K 均值算法的计算时间可视为与线性相关。算法流程如表 1 所示。

本文使用科技资源向量间的欧式距离度量点间距离，基于肘部法则（Elbow Method）来选择合适的 K 值。肘部法则是一种 K 均值聚类簇数的选择方法，它通过寻找畸变程度得到大幅改善的 K 值来确定聚类簇数。

表 1 聚类算法

Table 1 Clustering algorithm

Algorithm *K*-均值聚类算法**********
Input: 科技资源向量空间，*K*个初始质心
Output: 聚类结果
1. repeat
2. 将每个点指派到最近的质心，形成*K*个簇
3. 重新计算每个簇的质心
4. until 质心不发生变化

续表

3 实验分析

为了验证本文设计的聚类方法，使用“中国科技资源共享网”(https://www.escience.org.cn) 中的部分水生生物种质数据作为科技资源数据集进行实验。该数据集包含国家水生生物种质资源库提供的 3 606 个与水生生物相关的资源的名称、描述等资源元数据信息。

3.1 水生生物种质资源本体树构建

为了应用本文的方法，首先需要建立水生资源数据的领域本体树。本文设计的领域本体树主要针对水生资源数据标题、描述以及关键词中出现的词汇用手动建立。本体树含有 27 个概念及个体，如图 3 所示。

将此领域树通过 2.1 节描述的方法建立降维矩阵，将 27 个概念组成的高维向量转换为 3 个主要概念组成的向量。通过分析降维矩阵可知，对每个主要概念贡献最大的前 5 个主要概念如表 2 所示。

通过这些信息可知：主要概念 1 侧重于分子和细胞工具相关资源；主要概念 2 侧重于斑马鱼资源；主要概念 3 侧重于水生生物。

表 2 贡献前 5 的主要概念

Table 2 Top 5 main contributing concepts

概念	主要概念1	主要概念2	主要概念3
#0	抗体	突变体斑马鱼	特色水生动物
#1	质粒	野生型斑马鱼	藻类和原生动物
#2	细胞系	转基因斑马鱼	长江鱼类
#3	分子和细胞工具	斑马鱼	珍稀水生动物
#4	突变体斑马鱼	水产细菌	水生植物

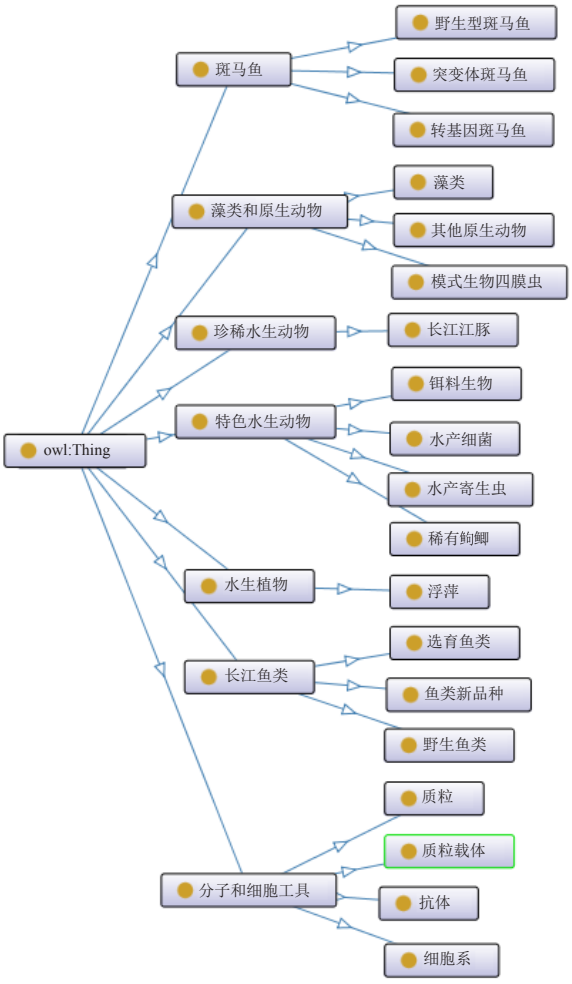

图 3 领域本体树

Fig.3 Domain ontology tree

3.2 水生生物种质资源向量表示

科技资源相关描述主要由标题与详细描述两部分文本组成。为了将描述转换为向量，首先分别把标题和详细描述分别转换为与概念同维度的 27 维向量。具体方法为：如概念出现在文本中则设置为 1，否则设置为 0。之后将标题向量、详细描述向量和关键词向量叠加形成科技资源数据描述向量。

以水生生物资源数据的一个资源为例，其标题为“工具质粒 (pT2(kop:Cre-UTRnos3, CMV:EGFP))”，描述为“由国家斑马鱼资源库收集、保藏，用于科学研究目的的工具质粒。DNA 资源，经由每年不少于一次转化、质粒提取、验证工作维护。资源常年以 DNA 样品方式保藏和分享。资源类型为工具质粒。”关键词为“斑马鱼；工具质粒；DNA”。转换后向量中非 0 值以及对应概念如表 3 所示。

其中向量下标为 9 的“质粒”取值为 3，因为概念出现在标题、描述和关键词中；下标为 17 的斑马鱼取值为 2，因为概念同时出现在描述和关键词中。

表 3 科技资源向量实例

Table 3 Technology resource vector example

向量下标	概念	向量值
9	质粒	3.0
17	斑马鱼	2.0

通过以上方法对水生生物种质资源数据中 3 606 个样例进行向量化，并通过 2.2 节描述的方法进行降维得到 3 维空间中的点集。

3.3 水生生物种质资源聚类

对 3.2 节降维后的结果使用 *K* 均值算法对科技资源进行聚类。根据肘部法则进行 *K* 值的选取，根据图 4 聚类簇数量与误差平方和（Sum of the Squared Error, SSE）关系图，本例中 *K* 值选取为 6，即聚类簇数为 6。聚类结果如图 5 和表 4 所示。

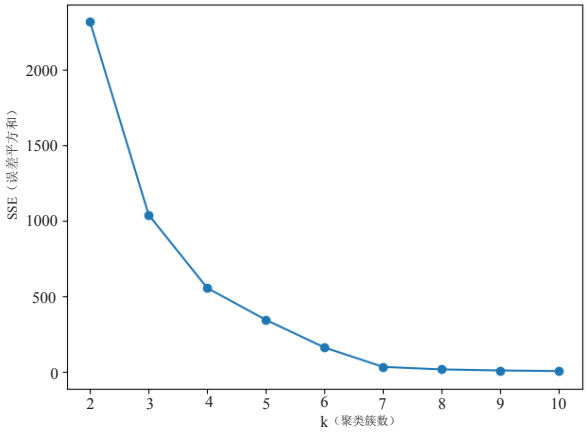

图 4 聚类簇数量误差平方和 (SSE) 关系图

Fig.4 Relationship between the number of clusters and SSE

由聚类结果可知，本文提出的聚类方法将 2 606 条水生资源数据聚成了 6 类，且聚类结果具有明显的语义意义，与数据提供机构给出的主题分类（图 6，来自“中国科技资源共享网”）能够较好地吻合，其准确率为 99.6%。相关科研工作者提出科技资源需求时，可以通过检索条件检索到相关的资源包并进行优化配置实现个性化服务。

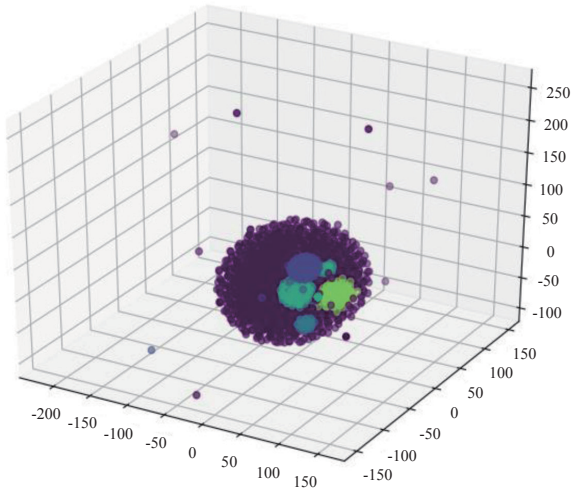

图 5 科技资源聚类向量空间

Fig.5 Technology resource clustering vector space

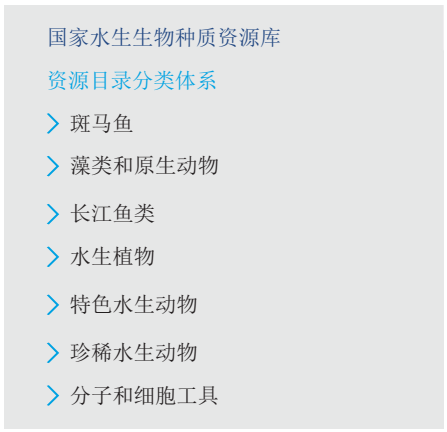

图 6 国家水生生物种质资源库主题分类

Fig.6 Theme classification of NABRC

综上，本文使用“中国科技资源共享网”（https://www.escience.org.cn）中水生生物种质资源验证了本文讨论的科技资源聚类方法，可以看到通过领域本体树对科技资源向量进行降维后依然保持了良好的原始数据特征，并取得了良好的聚类结果。这验证了本方法在高维数据集上应用的有效性。

表 4 科技资源聚类结果

Table 4 Technology resource clustering results

类别	1	2	3	4	5	6
数量	1772	1260	424	102	26	22
解释	藻类和原生动物	斑马鱼	模式生物四膜虫	水产细菌	长江鱼类	质粒载体、鱼类细胞系
错误	8	1	1	1	1	2

4 结论与展望

本文提出了一种基于领域本体概念树的科技资源向量化方法，给出了本体概念语义关系矩阵的构造方法和向量空间的构造方法，并利用该向量空间进行了聚类处理分析，目前在一般规模的真实数据集上得到了较好的聚类结果。证明通过本体概念语义关系降维后的向量在简化计算的同时，依然可以保留足够的科技资源特征信息。本方法具有针对多源异构的高维科技资源数据的处理能力，为领域广、数量大的异构科技资源集合进行聚类分析和个性化服务共享提供了技术支持。

通过本方法产生合适的科技资源聚类后，可以在由每个聚类中心组成的新向量集合中进行查询条件最近邻检索，并对检索结果对应的聚类包通过一定的规则进行选择集成组合，可以达到为不同用户提供个性化资源服务的目的。如何对资源包进行选择和最优化组合配置将是科技资源集成的另一个研究重点。

本文提出的异构数据聚类方法依赖于基于领域本体的数据预处理过程，目前对大规模构建领域本体仍然是一个困难的工作。本文未来将研究利用深度学习和知识图谱的技术自动构建大规模的本体领域网络以适应海量的、覆盖众多领域的科技资源数据集。

本文的另一个未来工作方向是改进聚类方法的灵活性和计算效率，以具备较强的领域适应性和规模适应性。

致　谢

感谢国家水生生物种质资源库于中国科技资源共享网公开发布共享的科学数据资源。

利益冲突声明

所有作者声明不存在利益冲突关系。

参考文献

[1] 赵启阳,张辉,王志强.科技资源元数据标准研究的现状分析与新的视角[J].标准科学,2019(03):12-17.

[2] 宫萍,王理,张辉,魏思远,王馨.基于语义本体的科技资源集成建模研究[J].标准科学,2019(03):36-40.

[3] Lee J A, Verleysen M. Unsupervised dimensionality reduction: overview and recent advances[C].The 2010 International Joint Conference on Neural Networks (IJCNN). IEEE, 2010: 1-8.

[4] Jolliffe I T. Principal components in regression analysis [M].Principal component analysis. Springer, New York, NY, 1986: 129-155.

[5] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis[J]. Journal of computational and graphical statistics, 2006, 15(2): 265-286.

[6] Kohonen T. Self-organized formation of topologically correct feature maps[J]. Biological cybernetics, 1982, 43(1): 59-69.

[7] 王志强,杨青海.科技资源管理标准体系研究[J].标准科学,2019(03):6-11.

[8] 李小平,徐汉川.科技资源及服务集成与优化[J].中国基础科学,2019,21(06):41-43+60+64.

[9] 赵男.浅析如何做好科技资源共享[J].科技风,2020(16):243.

[10] 杨子江.科技资源内涵与外延探讨[J].科技管理研究,2007(02):213-216.

[11] 国家科技基础条件平台建设战略研究组.国家科技基础条件平台建设战略研究报告[M]. 北京:科学技术文献出版社. 2006.

[12] 张渝英,董诚,王运红.科技资源共享研究框架体系的探讨[J].现代科学仪器,2007(05):3-9.

[13] 孔德洋.我国科技资源共享问题探讨[J].中国科技资源导刊,2008,40(06):51-56.

[14] 于阳.科技资源信息相关集成方法研究[J].江苏科技信息,2020,37(11):25-28.

[15] 李宗俊,陈文杰.区域科技服务资源集成与关联研究[J].中国科技资源导刊,2019,51(06):1-5+58.

[16] 汤华茂,郭钢.云制造资源虚拟化描述模型及集成化智能服务模式研究[J].中国机械工程,2016,27(16):2172-2178.

[17] 程臻. 云制造服务平台关键技术研究[D].哈尔滨工业大学,2016.

[18] 唐琳,郭崇慧,陈静锋, 等.基于中文学术文献的领域本体概念层次关系抽取研究[J].情报学报,2020,39(4):387-398.

[19] 甘健侯,姜跃,夏幼明.本体方法及其应用[M].北京:科学出版社,2011.

[20] 郝文宁,冯波,陈刚,靳大尉,赵水宁.基于领域本体的文档向量空间模型构建[J].计算机应用研究,2013,30(03):764-767.

[21] 孙荣, 刘宗田, 廖涛, 等. 应用本体对特征向量降维研究[J]. 计算机工程与设计, 2010 (17): 3864-3867.

[22] Tan P N, Steinbach M, Kumar V. 数据挖掘导论[M]. 北京: 机械工业出版社, 2019.

收稿日期：2020 年 7 月 17 日

葛胤池，北京航空航天大学计算机学院，博士研究生，主要研究方向为数据挖掘、知识图谱、自然语言处理。
本文中负责方法设计、实验和主要论文撰写。

Ge Yinchi is a PhD student in School of computer Science and Engineering, Beihang University. His research areas are data mining, knowledge graph and NLP.
In this paper, he is responsible for the method design, experiment conduction and main paper writing.

E-mail:geyinchi@buaa.edu.cn

张辉，北京航空航天大学计算机学院，博士，教授，博士生导师，国家科技资源共享服务工程技术研究中心副主任，负责科技资源的整合、管理与共享服务的技术研发工作，发表相关学术论文 70 余篇，获得专利 6 项。目前主要研究领域为互联网信息检索、大数据管理与挖掘、知识发现与管理。本文中对文章总体框架进行指导。

Zhang Hui is a professor and the doctoral tutor at the School of Computer Science and Engineering of Beihang University. He also serves as the deputy director of the National Science and Technology Resource Sharing Service Engineering Research Center. His main research areas are Internet information retrieval, big data management and mining, knowledge discovery and management.

In this paper, he is responsible for the overall framework guidance of the article.

E-mail: hzhang@buaa.edu.cn

宋文燕，北航长聘副教授、博士生导师，上海交通大学工学博士，德国慕尼黑工业大学 (Technische Universität München) 博士后。长期从事复杂产品 / 服务系统、大规模个性化定制、模块化协同开发、可持续运营等理论及应用研究。已发表学术论文近 60 篇，其中 50 多篇发表在 IEEE T. Reliab., Int. J. Prod. Res., CIRP Ann.- Manuf. Techn. 等国际 SCI/SSCI 期刊上，在国际知名出版社 Springer 独立出版英文学术专著 1 部，在机械工业出版社等出版著作、教材 2 部，授权 / 公开国家发明专利多项，入选“北航青年拔尖人才支持计划”。本文中对文章整体框架进行指导。

Song Wenyan, Ph.D., is an associate professor and doctoral tutor of Beihang University. His research areas are engaged in theoretical and applied research on complex products/service systems, large-scale personalized customization, modular collaborative development, sustainable operation, and etc.

In this paper, he is responsible for the overall framework guidance of the article.

E-mail: songwenyan@buaa.edu.cn

王轩，北京航空航天大学计算机学院，硕士研究生，主要研究方向为知识图谱，自然语言处理，大数据分析。

本文中完成研究背景部分论文撰写。

Wang Xuan is a graduate student in School of computer Science and Engineering, Beihang University. His research areas are knowledge graph, natural language processing, and big data analysis.

In this paper, he is responsible for the background research of the paper.

E-mail: 837909408@qq.com

引文格式：葛胤池, 张辉, 宋文燕, 王轩. 基于领域本体的科技资源聚类方法研究[J].数据与计算发展前沿, 2020,2(5): 13-22.DOI:10.11871/jfdc.issn.2096-742X.2020.05.002.PID:21.86101.2/jfdc.2096-742X.2020.05.002.

Ge Yinchi, Zhang Hui, Song Wenyan, Wang Xuan. Scientific and Technology Resource Clustering Based on Domain Ontology [J].Frontiers of Data & Computing, 2020,2(5): 13-22.DOI:10.11871/jfdc.issn.2096-742X.2020.05.002.PID:21.86101.2/jfdc.2096-742X.2020.05.002.