面向共享的地学数据语义标签提取与推荐方法

王敬悦[1] 王卷乐[2,3] 韩保民[1] 张 敏[2,4] 李 凯[5]
（1.山东理工大学建筑工程学院，山东淄博 255049；2.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；3.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；4.中国科学院大学，北京 100049；5.中国矿业大学（北京）地球科学与测绘工程学院，北京 100083）

摘要：地球科学数据具有丰富的语义信息，这为探索地学奥妙带来广阔空间的同时，也为数据共享带来了挑战。缺少语义的认知和关联使得用户难以从复杂、海量、多源、异构的地球科学数据中发现符合自身需求的数据。以中国科学院《地球大数据科学工程》的共享数据集为研究对象，对数据文本进行分词和标签提取，实现数据形式的统一，以达到多源异构数据的规范化管理。研究表明，通过对数据集实现标签提取、标签推荐以及知识图谱的构建，可促进海量地球科学数据的管理和精准服务。研究结果可为更多的科学数据共享平台提供借鉴。

关键词：地球科学；数据共享；语义标签；相似推荐；知识图谱；元数据
DOI：10.3772/j.issn.1674-1544.2023.02.010 **CSTR：**15994.14.issn.1674-1544.2023.02.010
中图分类号：P208 **文献标识码：**A

Semantic Tag Extraction and Recommendation Method for Geoscience Data Sharing

WANG Jingyue[1], WANG Juanle[2,3], HAN Baomin[1], ZHANG Min[2,4], LI Kai[5]
(1.Shandong University of Technology, School of Civil and Architectural Engineering, Zibo 255049; 2.State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101; 3.Collaborative Innovation Center for Development and Utilization of Geographic Information Resources in Jiangsu Province, Nanjing 210023; 4. University of Chinese Academy of Sciences, College of resource and Environment, Beijing 100049; 5. College of Geoscience and Surveying Engineering, China University of Mining & Technology (Beijing), Beijing 100083)

Abstract: Geoscience data has rich semantic information, which not only brings vast space for geoscience exploration, but also challenges for data sharing. The lack of semantic cognition and association makes

作者简介：王敬悦（1997—）女，山东理工大学硕士生，研究方向为土木水利；**王卷乐**（1976—），男，中国科学院地理科学与资源研究所研究员，博士生导师，研究方向为科学数据共享、地理信息系统与遥感应用（通信作者）；**韩保民**（1969—），男，山东理工大学建筑工程学院教授，研究生导师，研究方向为GNSS精密定位；**张敏**（1994—）女，中国科学院地理科学与资源研究所博士生，研究方向为灾害文本挖掘与知识图谱；**李凯**（1998—）男，中国科学院地理科学与资源研究所硕士生，研究方向为大地测量学与测量工程。

基金项目：国家重点研发计划项目“地球表层系统科学数据挖掘与知识发现关键技术与应用”（2022YFF0711600）；中国科学院战略性先导科技专项（A类）“地球大数据资源分类与多维标签管理系统研发”（XDA19090200，XDA19040501）。
收稿时间：2022 年 11 月 26 日。

it difficult for users to find data that meets their own needs from complex, massive, multi-source and heterogeneous scientific geo-data. This paper took the shared dataset of the Big Earth Data Science Project of the Chinese Academy of Sciences as the research object. Through word segmentation and label extraction of data text, the data form was unified to achieve standardized management of multi-source heterogeneous data. The research practice showed that the management and precise service of massive earth science data could be promoted through the realization of label extraction, label recommendation and the construction of knowledge graph on the data set. This study result is expected to provide reference for more scientific data sharing platforms.

Keywords: Earth science, data sharing, semantic tags, similarity recommendation, knowledge graph, metadata

0　引言

地球系统是由大气圈、水圈（含冰冻圈）、生物圈、岩石圈和人类圈所构成的自然社会综合体，是一个开放的复杂系统。在地球各圈层交互作用和人类活动的不断影响下，产生了覆盖面广、类型丰富、变化快速、海量且开发潜力巨大的科学数据。地球大数据是一种典型的科学大数据，具有空间属性，数据来源涵盖空间对地观测数据，包括陆地、海洋、大气及其他人类活动相关的数据[1]。地球科学领域的科学数据管理、挖掘、共享与应用是当前全球研究的前沿和热点。

我国高度重视地球科学数据的共享和管理。2002年启动科学数据共享工程，在多个地球科学领域建立了科学数据平台。2005年，国家科技基础条件平台中心成立，推动了多个地学数据中心建设[2]。2008年，地学领域的国家重点基础研究计划发展项目（973计划）开展数据汇交[3-4]。2013年，科技基础性工作专项项目开展数据汇交[5]。2018年，中国科学院“地球大数据科学工程”启动[6]，在资源、环境、生物、生态等多个领域开展数据汇交和共享。2019年，多个地学领域的国家科学数据中心成立[7]，并于2020年开始承接国家重点研发计划项目数据汇交。然而，通过汇交实现的科学数据集成管理通常是以原研究项目（课题）为单位组织的。这种组织方式便于检查和归档，但很难整合、集成和提供精准服务。因为缺乏必要的数据语义和类目关联，很难对解除项目（课题）组织后的分散数据进行集成，无法按照特定主题应用需求聚合抽取数据资源。这种情况也加剧了科学数据资源在汇交共享中的破碎化，无法在整体上给出地球科学数据资源领域分布全景视图。因此，非常有必要开展地学科学数据的集成和关联研究，以为其数据高效管理提供支持。

语义标签为数据资源的集成关联管理与高效发现提供了一种方法支持。标签是冗长的文本信息的简洁凝练形式，也是文本核心语义的表现形式[8]。数据集本身的元数据描述信息是由数据提供者原始创建的。这些信息受到数据集提供者的专业背景和个人喜好的影响而表现各异，尤其是部分专业数据集的描述信息非常薄弱。在后期的数据集成管理阶段，由于知识产权限制或专业人力缺乏等原因，很难直接对这些数据集的元数据进行补充和修改。通过增加数据集语义标签，可以丰富其数据描述内容，从而促进分散、异构数据的集成管理。主要体现在以下3个方面：一是大量增加的规范化标签信息不仅使原数据集的内容表现更加直观，而且为信息检索和归类提供更加简单的处理模式[9]。将标签作为文本信息的语义索引，将搜索词与相关的标签进行匹配，可加快检索速度、缩减查询计算的存储空间开销[10]。二是标签之间丰富的语义信息可以促进数据的有效归类和相互关联。不同的数据集之间，通过同样的标签可以更加容易地相互关联起来，进而形成关联网络，促进数据管理者对所集成数据的整体管理。三是数据的标签可以进一步和用户的画像特征进行关联，有助于针对不同用户实现个性化精准数据推荐。

语义标签在地学数据管理方面的结合还比

较少见，但其在信息管理领域已经有很好的技术基础和应用实践。在标签提取方面，陈伟鹤等[11]在提取语义标签时，选择文本中具有高度共现关系的高频词或短语作为标签，从而将关键词提取问题转化为公共子串匹配问题。罗燕等[12]利用齐普夫定律和词频统计相结合的方法和TF−IDF（Term Frequency−Inverse Document Frequency）词项的语义表示方法，对文本中的词语进行权重排序，并进行文本语义标签的提取。李鹏等[13]利用PageRank算法，对带有标注信息的网页数据中的图节点进行权重计算，并根据节点权重值的大小进行语义标签提取。夏天[14]基于TextRank的算法思想，构建候选语义标签图，引入覆盖影响力、位置影响力和拼读影响力等因素，计算词语之间的影响力概率转移矩阵，通过迭代计算分值大小提取候选关键词，挑选分值大的前n个关键词作为语义标签。在精准推荐方面，早期的标签推荐方法[15]被称为基于条件概率的标签推荐方法，主要是利用标签之间出现的条件概率来建立标签相关性。近年来，为提升标签推荐的效果，采用深度学习的方式来建立标签相关性[16−17]。标签推荐的主要任务是捕获标签之间的相关性特性。Song等[18]以文档词频为依据，对文档中每个标签进行重要度排序并进行标签推荐。孙传明等[19]对基于项目的协同过滤算法进行改进，并引入项目标签信息，计算项目间的相似度。魏玲等[20]将用户个人信息、交易信息、偏好信息和行为信息作为标签进行用户画像，计算用户间的相似度。董立岩等[21]结合项目—用户评分矩阵、项目属性矩阵和项目权重，计算用户间的相似度。吴佳婧等[22]使用相似性将项目属性差异作为权重，以此进行信息推荐。为标签引入推荐功能不仅能提升平台用户的体验，提高生成标签的数量和质量，还能建立更高效的信息检索系统[23]。

然而，由于地球科学数据自身的复杂、海量、多源、异构特点，如何借助语义标签技术加强本领域的数据管理和共享服务是一个现实的需求和难题。本文拟引入标签管理的方法，用于地球科学大数据的管理和共享。以中国科学院发起的“地球大数据科学工程”为对象，在自然语言处理技术支持下，通过文本处理建立标签，实现数据之间的主动关联，并建立数据图谱网络，促进地球大数据的管理与共享服务。

1 数据与方法

1.1 数据源

“地球大数据科学工程”是中国科学院于2018年发起的一个数据工程。该工程整合共享的数据资源不仅包括空间对地观测数据，还包括陆地、海洋、大气及与人类活动相关的数据，统称地球大数据[23]。地球大数据汇聚来自于9个项目的数据，涵盖对地观测数据、地理资源数据、生态系统地面观测数据、生态多样性与生态安全数据、海洋数据、三级区域集成数据、“一带一路”区域集成数据、联合国可持续发展目标实现支撑数据等8个主要维度[24]。本文研究的数据源是地球大数据前期汇总的1 169条数据集的元数据信息，由中国科学院空天信息创新研究院提供。

在文本分词过程中为提高分词精度，引入了学科领域术语。这些术语均来自超级科技词表（STKOS）。为了提高标签推荐的准确度，爬取中国知网中地学领域相关的文章摘要，增加语料信息。

1.2 方法

本文对数据集的元数据信息进行分词和权重处理获取数据标签。使用python−bs4爬取大量论文摘要进行模型训练，对数据标签进行相似度计算并进行可视化以及知识图谱构建。本实验流程见图1。

1.2.1 标签提取方法

为在海量数据中快速定位用户关注的内容，需要对标签进行管理。本实验使用的文本数据是地球大数据的元数据描述文本，包括编号、名称、类型、简介、学科分类等数据元素信息。由于这些文本数据缺乏词边界，在提取标签之前要进行分词和停用词过滤。经过分词处理后，文本被转化为一个词条集合。该集合包含了一些与关键词无关的项，如虚词、无意义的字序列片段或

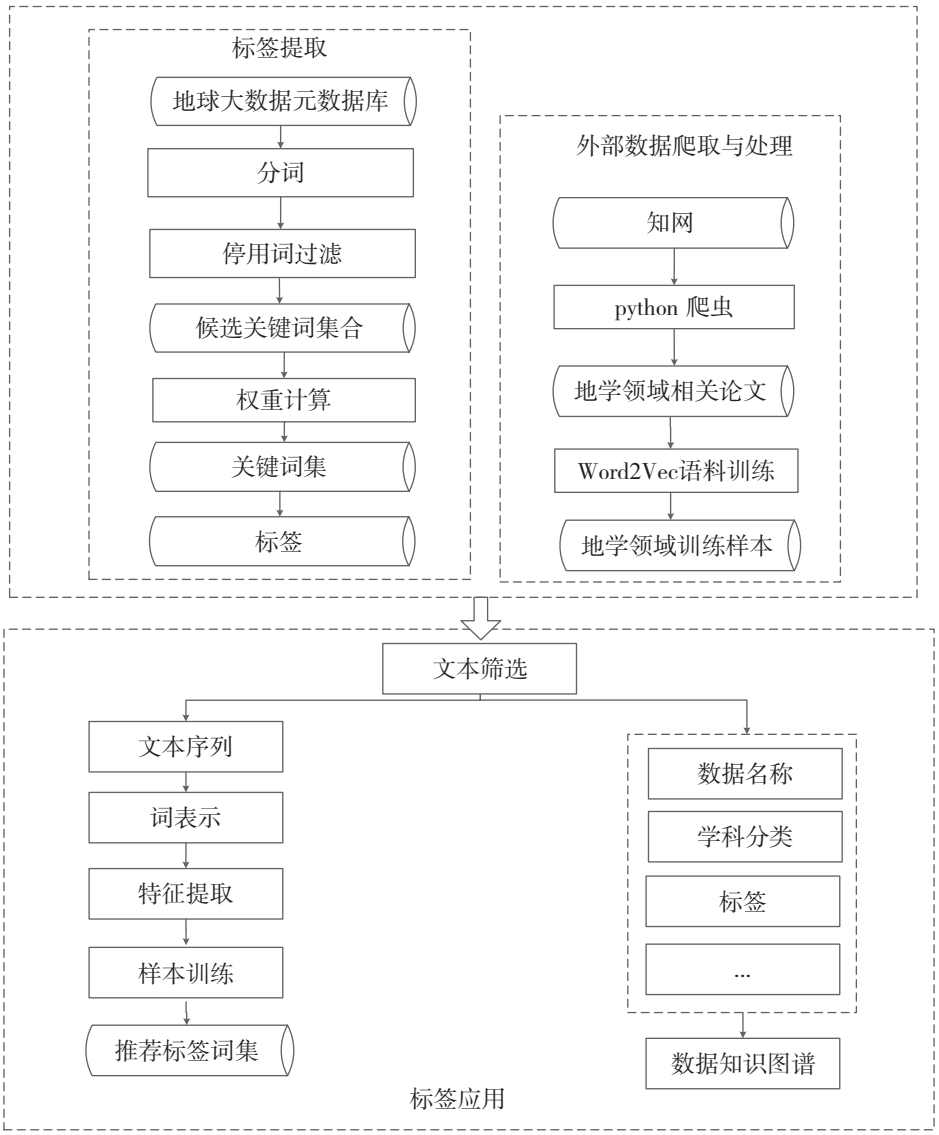

图 1　实验流程

者一些符号，在处理过程中加入哈尔滨工业大学的停用词词库以剔除这些无关项。为选择精度相对较高的分词模型，本实验使用Jieba分词和中国科学院分词系统进行对比，通过精确度、召回率和F1 值的大小来确定两种分词工具的准确性。选取精度较高的分词工具进行分词处理。同时，获取STKOS中 60 多万个天文、地理、生物、社会等领域的专业术语，将其加入相应分词工具的用户自定义词典来进一步提高分词的精度。

对分词后的文本数据使用TextRank算法来计算各个词或词组在文本中的权重。其主要思想是：将分词后的每个词作为节点，并为节点之间的边引入了权值，通过计算节点权重并进行倒序排列，得到权重最高的T个单词作为标签。

1.2.2　相似计算方法

有了丰富的语义标签后，仍然要解决相关性排序的问题才能支持不同数据集之间的高效匹配。而将标签向量化则是相关性计算和排序的基础。Word2Vec是一款利用深度学习来训练生成词向量工具。该工具可以将一个个单词转化为向量表示。通过测量词向量之间的距离，可以判断词语间的关联关系。词间距离越近，说明这些词的共现频率越高，相关性就越强，而相关性强的词就会被优先推荐。本实验采用Word2Vec方法进

行标签推荐，首先将这些词进行向量化表示，通过向量与向量之间的距离来度量它们之间的相似度，从而发现这些词间存在的潜在关系。

Word2Vec模型主要由词表示层、特征提取器以及多标签分类器3个部分组成。模型输入通常是一个长度为n的文本序列$w_1, w_2, \cdots, w_n$，其中每个w_i代表一个单词。w_i是一个长度为V_{text}的One−Hot向量，V_{text}为文本词表的大小。将原始的单词One−Hot向量通过词表示层转化，变成对应的词向量$e_1, e_2, \cdots, e_n \in R^{D_{embedding}}$，其中词向量的维度用Dembedding表示。随后在文本特征提取器中，输入转化后的词向量，提取有效的文本表达形式。最后，将这些文本输入到一个多标签分类器中，得到k个标签的伪概率t_1, t_2, ⋯, $t_k \in [0, 1]$，其中k为标签词表的大小，模型的损失函数为交叉熵函数[25]。

其概率公式为：

$$P(W) = P(w_1) \times P(w_2 \mid w_1) \times \ldots \times P\left(w_N \mid w_1^{N-1}\right) = \prod_{i=1}^{N} P\left(w_i \mid w_1^{i-1}\right) \quad (1)$$

Word2Vec模型包括连续词袋模型（Continuous Bag−of−word Model，CBOW）和Skip−gram模型。本文采用Skip−gram模型。该模型是通过输入的一个中心词来预测周围词。Skip−gram模型的目标函数为：

$$\psi = \sum \log\left\{P\left[Context\left(w_t\right) \mid w_i\right]\right\} \quad (2)$$

使用相似度计算为每个数据的标签推荐最相关的5个标签，以便读者找到自己最感兴趣的数据。

1.2.3 知识图谱方法

丰富的语义标签以及不同标签之间的相关性可以支撑构建知识图谱。知识图谱可以“实体—关系—属性”的方式展示地球科学领域数据资源的总体情况。如通过图谱可以发现哪一地学分支学科的数据密集，而哪一地学分支学科数据较为稀缺；哪一地理区域的数据更为集中，而哪一地理区域的数据则有空白；哪些研究热点数据发展较快，并发现这些热点之间的可能联系等。有了这个整体认识，可以促进地球大数据管理的长期规划和顶层设计。在本文研究方案中，使用图形数据库Neo4j对数据集中每个数据之间的实体关系进行连接，以便清晰地展示各数据与属性之间的关系。对数据产品名称、简介、学科分类、发布时间、数据标签以及发布机构建立图谱关系，将数据通过要素连接，使地球科学数据形成一个关系群体。当以数据标签为数据查询的切入点时，与该标签相关的数据可被准确筛选过滤出来。通过标签的关联关系，可以寻找到与该标签有关的所有数据，并将这些数据内容以可视关系网络的形式进行展示。

三元组是知识图谱的一种通用表示方式，即$G=(E,R,S)$，其中$E=\left\{e_1, e_2, \ldots, e_{|E|}\right\}$是知识库中的关系集合，$|R|$是包含的不同关系，$S \subseteq E \times R \times E$代表知识库中的三元组集合。三元组的基本形式主要包括实体1、关系、实体2和概念、属性、属性值等。将整理好的数据信息以三元组形式（<产品名称—属于—学科分类><产品名称—属于—标签><产品名称—发布于—机构><产品名称—时间—发布年>）保存于数据库中，这些三元组本质上就是数据网络中的知识关联。

2 实验结果

2.1 标签提取对比

首先从1 169个“地球大数据科学工程”数据集元数据文本中提取出能够反映数据内容和数据特点的词汇作为标签。使用分词工具对该数据集语料文本进行分隔处理，根据词语在文本中的重要性赋予其权重，选择权重值高的词语作为数据标签。图2为地球大数据语料文本。

对于部分数据集文本，采用jieba分词和中国科学院分词系统（NLPIR）进行分词对比，并采用准确率、召回率和F1值对两种分词方法进行精度评价，选出精度较高的分词工具，进而对文本进行整体的分词处理。

其计算公式如式（3）～式（5）所示。

2002-2012年全球ENVISAT/ASAR波模式	该数据集为2002-2012年Envisat/ASAR波模式数据通过CWAVE_ENV算法反演得到的全球海况参数
2003-2015年CERN植物物候观测数据集	中国生态系统研究网络（Chinese Ecosystem Research Network，CERN）植物物候观测数据集，是
2005-2015年中国CERN40个台站气象辐	该数据集由分布在中国8个典型生态类型的40个气象观测站原位观测的气象辐射要素逐小时数据组
2002-2016年典型生态系统土壤含水量变	本数据集包含2002-2016年中国11个典型生态系统的综合观测场土壤含水量观测数据，定位观测
2001年亚热带典型常绿落叶阔叶混交林	常绿落叶阔叶混交林是常绿阔叶林和落叶阔叶林之间的自然过渡类型，是对气候波动较为敏感的
2016年环江喀斯特生态系统植被指数与	中国科学院环江喀斯特生态系统观测研究站（简称“环江站”）位于广西自治区的西北地区，是典
2000-2012年全国1km空间分辨率气温和	随着高分辨率卫星遥感技术的快速发展及其在生态学中的广泛应用，与时空分辨率相匹配的气象
1981-2010年中国逐月散射光合有效辐射	太阳辐射中的光合有效辐射（Photosynthetically Active Radiation，PAR）是植物光合作用的主要能
2005-2014年中国生态系统研究网络地下	地下水是一个地区重要的自然资源，地下水位数据可为研究地下水的长期变化提供重要的参考资
2005-2014年CERN野外台站气象观测场	土壤水分是影响陆地–大气边界层能量和物质传输的重要因子。土壤水分含量是中国生态系统研究
1995-2011年中国陆地生态系统土壤环境	土壤环境是地球环境的重要组成部分。目前土壤环境问题的关注重点在于土壤污染。我国土壤污
2005-2015年中国陆地生态系统光合有效辐	光合有效辐射在生态学、农学以及气候学等多个学科中都有重要的应用价值。该数据集涵盖了中国
2003-2005年中国典型生态系统收支分量数	长白山、鼎湖山、千烟洲、内蒙古、海北灌丛、海北湿地、当雄、禹城8个台站20003-2005年的GF
2003-2005年中国典型生态系统总辐射数据	长白山、鼎湖山、千烟洲、海北灌丛、西双版纳、禹城6个台站20003-2005年天空总辐射半小时观
2003-2005年中国典型生态系统平均温度数	长白山、鼎湖山、千烟洲、海北灌丛、西双版纳、禹城6个台站20003-2005年的平均温度半小时观
2003-2005年中国典型生态系统有效辐射数	长白山、鼎湖山、千烟洲、海北灌丛、西双版纳、禹城6个台站20003-2005年有效辐射半小时观测
2003-2005年中国典型生态系统降水量数据	长白山、鼎湖山、千烟洲、海北灌丛、西双版纳、禹城6个台站20003-2005年降水量半小时观测数
2003-2005年中国典型生态系统CO_2通量数	长白山、鼎湖山、千烟洲、海北灌丛、西双版纳、禹城6个台站20003-2005年CO2通量半小时观测
2003-2005年中国典型生态系统潜热通量数	长白山、鼎湖山、千烟洲、海北灌丛、西双版纳、禹城6个台站20003-2005年潜热通量半小时观测
2003-2005年中国典型生态系统显热通量数	长白山、鼎湖山、千烟洲、内蒙古、海北灌丛、海北湿地、当雄、禹城8个台站20003-2005年的显

图 2　语料文本

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (5)$$

式中，*TP*是指分词样本中正确分词的数量；*FP*是指分词样本中未能准确切分的数量；*FN*是指在分词样本里预测为不准确但实际却分词正确的数量。

通过表 1 中jieba分词和NLPIR的提取结果，可以看出jieba分词正确率为 65.54%，而中国科学院分词系统的正确率、召回率相对较低。在对地球大数据文本进行分词处理时，jieba分词情况效果更好，其正确率远高于NLPIR，且文本内容丢失情况也相对较少。

对分词后的词集进行权重计算，使用TextRank计算方法对每条数据进行关键词提取（图 3），显示关键词和该词在整个文本中的权重。其中，中国（1）、全球（0.747 863）、青藏高原（0.608 185）、资源（0.429 319）、环境（0.404 956）、国家（0.387 999）、土壤（0.384 693）、大气（0.378 582）、地表（0.375 199）植被（0.349 728）为权重最大的前 10 个词语；城镇化率（0.016 456）、水鸟（0.010 032）、栖息地（0.010 016）、野生植物（0.009 750）、大城市（0.009 741）、油气田（0.009 654）、植物引种（0.009 488）、工业用地（0.008 891）、栽培（0.007 322）、建设工程（0.007 173）为权重值最小的 10 个词语。由此可以看出，该数据集包含的数据关键词主要属于地球系统科学领域。

表 1　分词结果对比

名称	正确率	召回率	*F*1 值
中国科学院分词系统（NLPIR）	0.298 2	0.380 1	0.334 2
jieba分词	0.654 5	0.731 0	0.690 6

2.2　相似度计算

在进行标签推荐时使用Word2Vec训练词向

全境 中国 样例
全球 海况 反演 海域 浮标 波高
植物 物候 种子 物候观测 环境 生态 木本 开花 中国 草本 成熟期
辐射 中国 气象 降雨量 风速 气压 温湿度 光合 风向 土壤 通量
含水量 土壤 水文学 专业 土壤学 中子 中国
混交林 习性 物种 气候 全球 群落
喀斯特 光谱 植被 投影 异质性 地物 西北地区 反射率
气象 插值 时空 降水 中国 栅格数据 气温 陆地 气候 空间数据 全球
辐射 生产力 植被 中国 陆地 散射光 冠层 光能 散射辐射
地下水位 地下水 中国 自动记录 人工 水资源 标准差
陆地 土壤 含水量 全国 土壤水分 地面 环境 农田 边界层 荒漠
土壤环境 元素 中国 陆地 监测数据 环境 表层 重金属 土壤学
辐射 光合 陆地 中国 生态 气候学 标定
太湖 江苏 水温 东湖 武汉 水色 电导率 水深 物理 湖泊 透明度
太湖 江苏 离子 东湖 武汉 溶解氧 水化学 湖泊 需氧量 化学 氮氨态
太湖 江苏 东湖 武汉 生物 湖泊 生物量 浓度
太湖 江苏 东湖 武汉 沉积物 湖泊 含水率 全磷
水温 胶州湾 大亚湾 水色 季度 盐度 三亚 物理 海湾 三亚湾 水深 透明度
大亚湾 胶州湾 三亚 需氧量 化学 水化学 海湾 三亚湾 季度
大亚湾 胶州湾 三亚 季度 海湾 生物 三亚湾
胶州湾 沉积物 含水率 三亚 海湾 三亚湾
沉降 中国 降水 大气 监测
沉降 中国 降水 监测 大气
沉降 中国 量化 文献 大气

图 3　提取的关键词

量，将每个标签转化为 300 维的词向量形式，以此计算词语间的余弦相似度。图 4 为标签向量化图。

以数据集中点击量较高的 2018 年全球重点区域植被叶绿素含量的数据为例，其标签包括植被、森林、草地、作物。为每个标签推荐与其相关的 5 个标签（表 2）。从表 2 可以看出，与植被最相关的标签分别是草地、林地、覆盖度、地物、草甸。其中，相关性最高的为草地，相似度为 0.547 7。这说明在地球科学领域研究中，植被常和草地、林地、地物、草甸结合在一起。

将该数据的标签和所有相关标签向量进行降维处理并投射到二维平面上，并将这些词进行可视化，词间余弦值越大说明两个词之间的关联度越高。从图 5 可直观地看出词和词之间的相关关系，与植被一词距离最近的就是降水量，其次是草地、林地、覆盖度等词语。将词语投射在二维平面上，可以更加直观地看出词间关系。

2.3 数据知识图谱建立

本文构建了一个数据知识图谱来显示数据组

国家 -0.66538703 -0.17268741 -1.6931654 -0.3916274 0.39734298 -1.1337138 -1.7234292 1.0913613 -2.0367901 -1.0446275 0.60993
11485 0.1388609 -1.1427922 -0.41179043 1.1685 -0.77843297 -2.1382058 1.4101357 0.49039707 -1.0711977 0.5729272 -0.7132094 (
8853 -1.023649 1.3064289 -0.53058743 -0.22881551 1.4924971 -1.8324097 -0.110120885 -1.4058546 -0.43256724 0.6905399 -0.9495
0.43123946 -0.95974517 0.40018493 -0.27334234 0.0051513915 0.5702276 -0.70809543 -1.8447366 0.47722048 0.05484027 -1.6407
经济发展 -1.7668408 2.1209154 0.41488886 -0.7248368 -0.71288776 -0.37138397 -1.301664 0.5998563 -0.45525163 -0.060096364 -1
7032 -0.40278542 -1.1776756 -1.1654335 0.7933078 -1.3679521 0.0405389 0.7766136 0.7671944 0.31117463 -1.6749855 2.0879142
7 0.524291 0.99041 0.35604352 -0.9519268 0.846221 -0.043387484 -0.12440735 1.7940898 1.3302987 -1.3270215 -1.7393669 0.2124
413931 -2.6510248 0.2383224 -0.4398663 -3.0857074 0.45679685 -1.4863366 0.87964255 -0.6628771 0.66733706 -0.77935463 0.241
断裂 2.6797607 -0.3667132 -0.58511096 0.4923944 1.098519 -0.30544078 0.26402026 0.35453588 0.7465995 0.030783646 -2.218606
41967 0.75807613 1.222541 -0.11840392 -1.2468214 -0.05238553 0.47794512 -0.7732405 -0.8418738 -1.2539976 2.1513076 -1.0237
8117299 -0.65284884 2.6051002 -0.19230227 -0.22669746 0.005165641 -1.4639708 0.0118569825 1.1148882 0.81120634 -0.7535969
.55970603 1.367746 -0.63529414 0.3481558 -1.5511597 -0.52066237 1.3274318 0.08166422 -0.55641526 -0.52169824 1.4203005 0.98
机制 1.1064421 0.7368279 1.1450739 -0.8763349 -0.5486763 1.0085185 -1.5418026 -0.96514654 -1.145781 0.6339334 1.6519299 2.82
08916 1.3497888 0.4462512 0.19350979 -0.026198747 0.5586513 1.1399106 -1.150106 -0.8229777 -1.6903011 -0.23077124 -0.37938
2 -1.015817 -0.5563732 0.6673214 0.3671404 -0.042794034 -0.5179284 0.09826592 -0.15700367 -0.45539466 0.6369232 -0.4259492
0.39850417 0.14976446 0.7196452 0.83230764 -0.3311423 1.0259949 -1.5619483 -0.1355602 -0.19737035 -0.31637046 0.4807998 1.9
人口 -1.8766605 -0.41987804 -0.023240015 -0.93210804 1.3285503 2.5704093 -0.7283807 -0.046815787 -1.4540163 0.93953884 0.31
261898 0.18473907 0.7901761 -1.2317293 0.23781118 -0.82639056 -0.2644876 0.413738 -0.49881178 0.98735243 0.9567543 0.29585
0.22122836 0.9367419 0.6813209 1.1158946 1.5126351 -0.13015015 -2.2363734 -1.1506084 0.56258726 0.34700936 -1.8771664 -1.46
6504 -0.22676496 -0.3259866 0.42817977 -1.1435752 0.5514343 0.09389014 -0.691323 -2.461649 -0.6999779 -1.523201 0.9723037 -
流体 0.16425991 -1.6801838 1.8086063 -0.4368717 -0.2605447 -1.0890154 1.9791187 0.38747412 -0.39762565 0.8869896 -0.8374114
8527274 -0.37569958 -0.42952448 0.66314346 1.092835 -1.1306527 -0.46605074 0.7479387 -1.4077353 -0.6654774 -0.73006004 0.8
45 -0.10572098 -0.46689445 -1.3275406 0.50348544 1.692785 -0.78596145 1.8446134 1.0699403 -0.7093415 1.7557445 1.9586354 0.
654 0.59397644 -0.7019586 1.4314827 0.026957506 2.9847443 0.9474208 0.07300913 -0.6395423 0.39135867 -0.5491106 0.8323731
文明 -1.588396 1.1280946 -1.1409094 0.027574938 -0.9587486 0.024072591 -0.16829392 0.059217405 -1.0081319 0.13606468 0.790
1.9423758 0.3342236 -1.0630218 0.5550418 -1.6708492 0.0016741363 -0.027842712 0.9867163 0.39978144 -0.17838371 -0.2988142

图 4 标签向量化图

表 2 相关标签推荐表

标签	相似词	相似度
植被	草地	0.547 7
	林地	0.491 0
	覆盖度	0.461 5
	地物	0.458 6
	草甸	0.457 8
森林	保护区	0.532 6
	森林资源	0.442 3
	人工林	0.439 4
	阔叶林	0.397 0
	林地	0.391 2
草地	林地	0.685 8
	草原	0.661 5
	草场	0.642 3
	灌丛	0.631 8
	牧草地	0.609 6
作物	农作物	0.713 1
	小麦	0.692 9
	水稻	0.673 3
	大豆	0.544 9
	植物	0.510 2

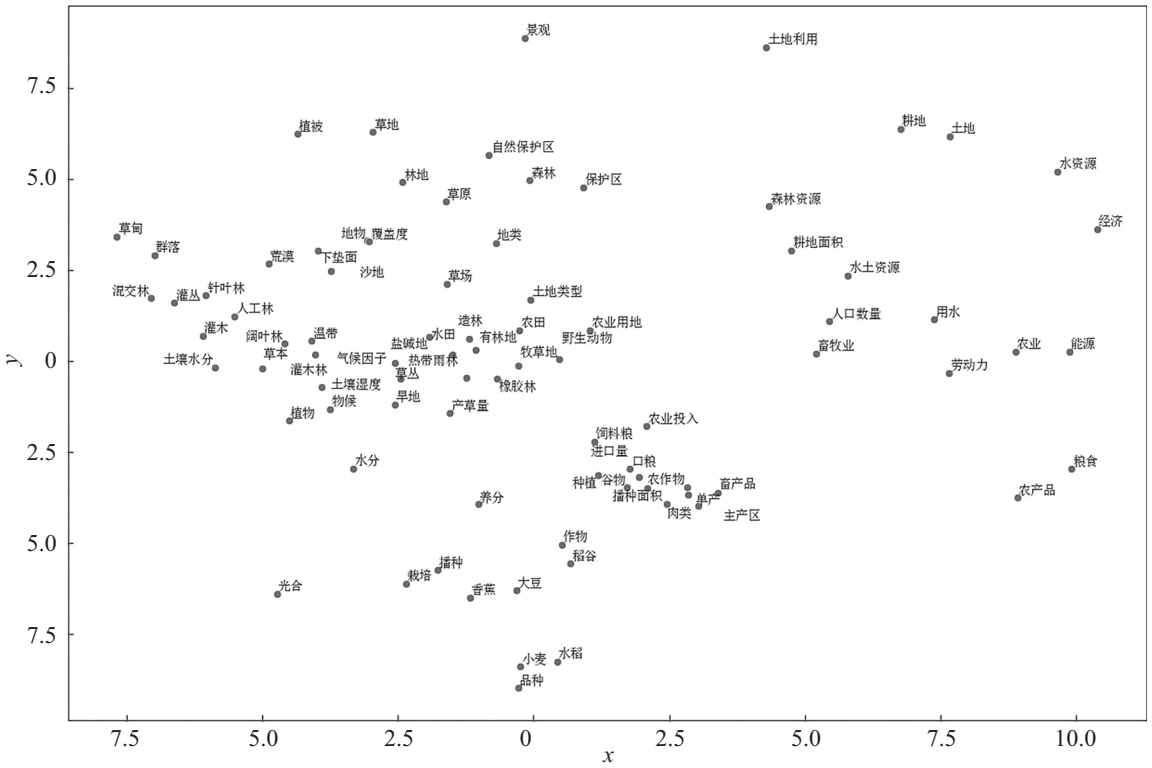

图 5　数据相关标签可视化

织结构之间的关系。图 6 显示了地球大数据数据组织形式。数据名称是唯一的标识符，集成了数据标题、数据简介、学科分类、数据标签、发表时间、发布机构、数据类型、通信邮件以及访问地址。

图 7 展示了个别标签与部分数据之间的关系。通过标签可以对那些看似没有关联的数据建立联系。如青藏高原地区土壤有机质数据（1979—1985 年）与全球重点区域总初级生产力季度产品（2017 年）数据，两个看似不相关的数据却都包含了对植物的研究，因此可以通过数据标签植物一词进行关联。当标签的数量不断增加时，数据间所体现出的关联性就越来越强。

地球大数据的数据知识图谱构建是利用数据组织模型，将数据信息和通过提取的标签信息以三元组的形式存储在Neo4j原生数据库中，并进行可视化展示。本试验结果显示，在图 8 中，包

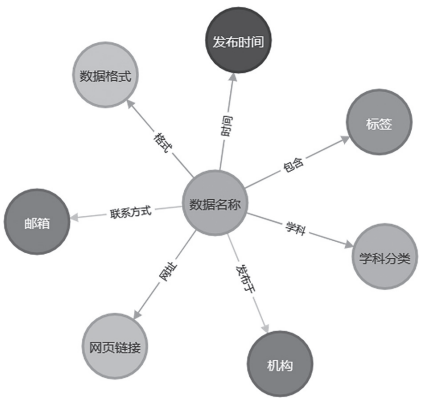

图 6　数据组成结构

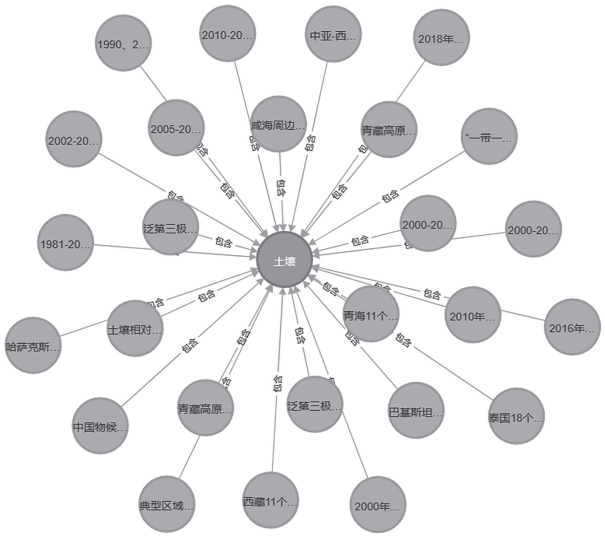

图 7　数据关联

括 3 902 个节点和 10 758 条边。节点包括数据标题、数据简介、学科分类、数据标签、发表时间、发布机构、数据类型、通信邮件以及访问地址。关系类型包括时间、格式、机构等。图 8 中深浅不同颜色的节点代表不同类别的实体，如数据集名称、数据标签、学科分类等数据要素，且要素间通过属性关系相互连接。当所有数据的重要信息都集成在图谱中时，就可以看到总体数据间存在复杂多样的关联性。

3　讨论

3.1　标签构建

本实验应用了NLPIR和jieba分词进行对比分析。实验表明，在对地球科学领域文本进行分词处理时，jieba分词的效果要比NLPIR的效果更优。这一对比认识可以帮助数据管理者选择更优的分词工具建立数据标签。从标签构建的实际效果来看，原有数据集共有 1 485 项关键词，且原数据集中的关键词由数据创建人提供，关键词质量也各有不同。很多关键词有着各种各样的问题，如有的字符过长，有的描述主观性较强。通过本文研究增加的标签数量达到 2 215 个。不仅增加了数据语义标签的数量，而且通过TextRank方法实现了标签的权重排序，显著提升了数据被检索发现的效率。此外，结合STKOS等规范术语库，提高了术语的语义规范化水平，也为未来加强不同学科之间的数据互操作提供语义规范化支持。总体来看，本实验对地球大数据文本描述信息进行标签提取，可以使原本数据内容各异的地球大数据有了统一形式的标签，数据管理者可通过标签进行数据管理查询和分类。这也证实了语义标签提取方法的可行性和在地球科学等领域的扩展性。但本实验中使用的是在无监督环境下的标签提取，这种采用文本数据原有的语义分布

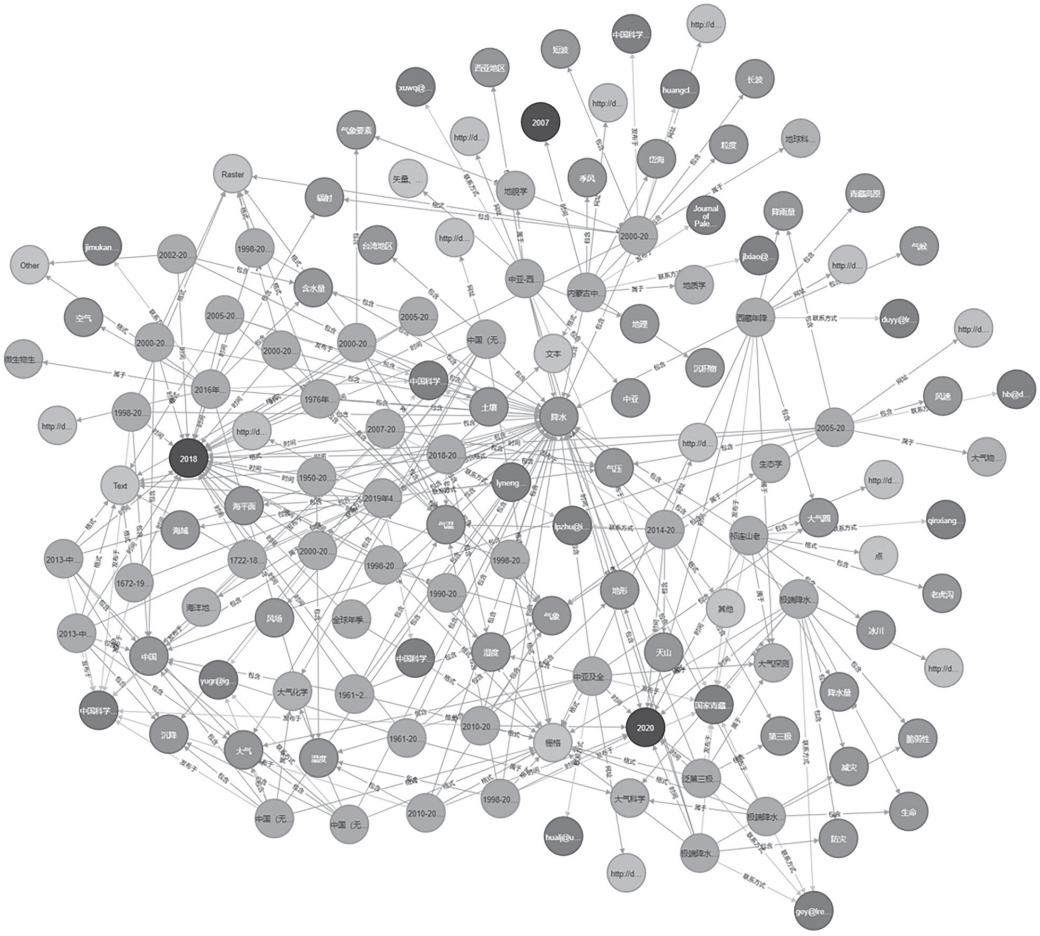

图 8　数据知识图谱

以及顺序结构信息进行的标签提取技术在准确率方面并不高[9]。未来可尝试使用基于长短期记忆网络结构（LSTM）的方法实现标签提取，并对比分析其在地球大数据领域的应用效果。

3.2　相似度分析

单纯增加语义标签，还不能实现数据的高效发现和精准推荐。相似度计算是实现这一应用的关键步骤。本文使用Word2Vec方法实现相似度计算，其主要难点是标签向量化和计算余弦距离值。通过解决向量间计算问题，实现了该方法在地学领域的应用。该方法使数据间相关性变得更加紧密，提升地球科学数据间的关联拓展、数据查询功能，增加数据发现率和点击率。此外，通过标签间的关联显示，缩小了数据的查询范围，提高了数据检索的准确性，加快了检索速度。通过相关性分析还可以提升大数据资源分析利用能力，实现数据的可持续增值服务。目前，标签推荐还没有和用户（画像）信息进行结合，无法做到个性化的信息推荐。未来将引入用户偏好信息，提出基于用户偏好的标签推荐模型。这将有利于数据共享平台更高效地进行内容组织和关联，并让每位用户在登录或访问数据平台时，都能看到更贴近本领域的数据标签推荐信息。

3.3　知识图谱构建

结合语义标签建立地球大数据的知识图谱成功验证了预期的设想，总体展示了数据集之间的联系，以及地球大数据多源数据集之间复杂的整体关联性。使用知识图谱的三元组形式整合地球大数据要素，包括数据的学科领域、数据格式、数据标签、发布时间等，实现了知识网络的统一表达构建。如在展示土壤数据时，通过语义标签之间的关联性，很直观地揭示出“地球大数据科学工程”中与土壤数据关联的区域主要集中在青海省、西藏自治区等青藏高原地区，并外延到中亚的哈萨克斯坦、咸海周边，南亚的巴基斯坦，以及东南亚的泰国等区域。这也体现了相关数据关联到泛三极地区研究的实际情况（图7）。在没有知识图谱表达之前，“地球大数据科学工程”难以清晰地把握各项目（课题）实际数据的整体聚合情况。通过知识图谱的表达（图8）可以揭示出这1 169条数据，形成了3 902个节点和10 758条边。进一步可以发现，与标签“中国”连接的边共有175条，每一条边说明有一条数据与其相连，那么本次实验的“地球大数据科学工程”数据中共有175个数据在覆盖范围上是面向“中国”范围的，这反映出国家尺度的数据集占有相当大的比重。在区域性上，与“青藏高原”相关的数据相对也较多。该节点共有73条数据连接关系，区域数据占比较高，这也反映了青藏高原在地学研究领域的重要性。同时还发现，在学科领域上“植被”“地表”“土地”等反映资源、环境、地理等标签关联的数据连接关系最多，而“教育”“旅游”“古生物”“保藏”等人文、教育以及固体地球领域的数据相对关联度较小。在未来的数据规划中，进一步发扬大空间尺度、青藏高原等典型地域的地理、资源、环境数据综合优势，并加强人文领域和固体地球等数据较为分散领域的数据集成。

4　结语

针对地球科学数据语义复杂、数据难以集成管理的问题，本文以地球大数据科学工程汇交数据的元数据信息为对象，进行分词、权重提取等处理，生成语义标签。通过标签之间的相关性进行标签推荐，实现数据语义关联，促进数据服务中的精准推荐。借助知识图谱技术，将这些离散的数据转化为统一的知识形式，以实现“地球大数据科学工程”数据的整体展示。研究实践表明，通过对地学数据赋予更多标签及其相似度计算，实现了规范化语义标签支持下的地球大数据分类管理，方便了用户在海量数据信息中快速定位需要的信息，促进了数据平台中海量数据信息之间的关联，并揭示了多源数据整体的聚集和分散程度，为数据管理的总体规划提供了参考。本文研究结果为进一步实现地学领域用户服务界面的个性化数据展示和结合用户画像的精准化用户数据服务奠定了基础，也为更多学科领域的科学数据管理和精准服务提供借鉴。

参考文献

[1] 郭华东.利用地球大数据促进可持续发展[N].人民日报, 2018−07−30(7).

[2] 石蕾, 高孟绪, 徐波, 等.欧美建设发展科学数据中心的经验及对我国的启示[J].中国科技资源导刊, 2022, 54(3): 31−36, 110.

[3] 王卷乐, 石蕾, 王玉洁, 等.科学数据汇聚的模式分析及对我国的发展建议[J].地球科学进展, 2020, 35(8): 839−847.

[4] 林海, 王卷乐.国家重点基础研究发展计划(973)资源环境领域项目数据汇交工作正式启动[J].地球科学进展, 2008(8): 895−896.

[5] 诸云强, 潘鹏, 石蕾, 等.科学大数据集成共享进展及面临的挑战[J].中国科技资源导刊, 2017, 49(5): 2−11.

[6] 郭华东. 地球大数据科学工程[J]. 中国科学院院刊, 2018, 33(8): 818−824.

[7] 王瑞丹, 高孟绪, 石蕾, 等.对大数据背景下科学数据开放共享的研究与思考[J].中国科技资源导刊, 2020, 52(1): 1−5, 26.

[8] SMITH G. Tagging: people−powered metadata for the social web[M]. California: New Riders, 2007: 53.

[9] LEI K, FU Q, Yang M, et al. Tag recommendation by text classification with attention−based capsule network[J]. Neurocomputing, 2020, 391: 65−73.

[10] 李雄. 基于文本的语义标签提取方法研究[D]. 北京: 北京工业大学, 2018.

[11] 陈伟鹤, 刘云.基于词或词组长度和频数的短中文文本关键词提取算法[J].计算机科学, 2016, 12(43): 50−57.

[12] 罗燕, 赵书良, 李晓超, 等.基于词频统计的文本关键词提取方法[J].计算机应用, 2016, 36(3): 718−725.

[13] 李鹏, 王斌, 石志伟, 等.Tag−TextRank: 一种基于Tag的网页关键词抽取方法[J].计算机研究与发展, 2012, 49(11): 2344−2351.

[14] 夏天.词语位置加权TextRank的关键词抽取研究[J].现代图书情报技术, 2013(9): 30−34.

[15] SIGURBJÖRNSSON B, ZWOL R. Flickr tag recommendation based on collective knowledge[C]//Proc. of the Int'l Conf. on World Wide Web. Beijing: ACM, 2008: 327−336.

[16] TANG S, YAO Y, ZHANG S, et al. An integral tag recommendation model for textual content[C]//Proceedings of the AAAI Conference on Artificial Intelligence. California: AAAI, 2019: 5109−5116.

[17] GAO J, HE Y, WANG Y, et al. STAR: spatio−temporal taxonomy−aware tag recommendation for citizen complaints[C]//Proc. of the ACM Int'l Conf. on Information and Knowledge Management. Beijing: ACM, 2019: 1903−1912.

[18] SONG Y, ZHUANG Z, LI H, et al. Real−time automatic tag recommendation[C]//Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. Singapore: ACM SIGIR, 2008: 515−522.

[19] 孙传明, 周炎, 涂燕. 基于混合协同过滤的个性化推荐方法研究[J]. Journal of Central China Normal University, 2020, 54(6): 7.

[20] 魏玲, 郭新悦. 融合用户画像与协同过滤的知识付费平台个性化推荐模型[J]. 情报理论与实践, 2021, 44(3): 188−193.

[21] 董立岩, 修冠宇, 马佳奇. 基于权重调节和用户偏好的协同过滤算法[J]. 吉林大学学报 (理学版), 2020, 58(3): 599−604.

[22] 吴佳婧, 贺嘉楠, 王越群, 等. 基于项目属性分类的协同过滤算法研究[J]. 吉林大学学报 (信息科学版), 2019, 36(4): 470−474.

[23] BELÉM F M, ALMEIDA J M, GONÇALVES M A. A survey on tag recommendation methods[J]. Journal of the association for information science and technology, 2017, 68(4): 830−844.

[24] 郭华东. 地球大数据科学工程数据共享蓝皮书[M]. 北京: 科学出版社, 2019: 12−13.

[25] 徐鹏宇, 刘华锋, 刘冰, 等.标签推荐方法研究综述[J].软件学报, 2022, 33(4): 23.