# 学术大数据技术在科技管理过程中的应用

**梁英[1]，张伟[1,2]，余知栋[1,2]，史红周[1]**
1. 中国科学院计算技术研究所，北京 100190；2. 中国科学院大学，北京 100190

**摘要**
学术大数据逐步成为提升科技管理水平的重要数据基础。通过调研国内外科技管理信息化的发展现状和特点，总结了学术大数据的发展及应用，分析了学术大数据在科技管理过程应用中面临的问题。结合我国科技管理的应用需求，设计了基于学术大数据的科技管理应用框架。基于知识图谱的学者画像构建技术和基于网络表示学习的相似作者推荐技术，利用多源异构的学术大数据，进行科研布局和资源统筹辅助决策以及科技管理过程中的专家精准推荐和成果评估评价，为提高科技管理效率提供了有效的技术支撑。

**关键词**
学术大数据；科技管理；知识图谱；网络表示学习；专家推荐

中图分类号：TP315 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019037

# *Applications of academic big data in the process of science and technology management*

**LIANG Ying[1], ZHANG Wei[1,2], YU Zhidong[1,2], SHI Hongzhou[1]**
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100190, China

***Abstract***
Academic big data is gradually recognized as an important data foundation for improving the level of science and technology management. Status quo and characteristics of science and technology management informationization were investigated at home and abroad, the development and applications of academic big data were summarized, and the problems were analyzed in applications of academic big data in the process of science and technology management. The needs of science and technology management application were combined, a technology management application framework based on academic big data was designed, scholarly image construction based on knowledge graph and similar author recommendation technology based on network representation learning were applied to assist in improving the overall layout of scientific research and resources utilizing multi-heterogeneous academic big data collection. Effective technical support for improving the efficiency of science and technology management was provided.

***Key words***
academic big data, science and technology management, knowledge graph, network representation learning, experts recommendation

## 1 引言

近年来，我国研究与试验发展（R&D）经费投入增速加快，据《2018年国民经济和社会发展统计公报》显示，2018年研究与试验发展经费支出为19 657亿元，比2017年增加11.6%，与国内生产总值之比为0.021 8。国家科技研发经费对各类科技计划的投入逐年增大，但也存在着重复、分散、封闭、低效等现象以及多头申报项目、资源配置“碎片化”等问题。如何有效管理并最大限度地发挥科技对国民经济和社会发展的巨大促进作用，是当前科技管理面临的重大挑战。为了加强顶层设计，国务院印发了《关于深化中央财政科技计划（专项、基金等）管理改革的方案》（国发〔2014〕64号）文件，通过“统一的信息系统，对科技计划（专项、基金等）的需求征集、指南发布、项目申报、立项和预算安排、监督检查、结题验收等全过程进行信息管理”。在大数据时代，学术大数据不仅是学术信息传递、学术观点交流和科研成果产出的结果，同时，学术大数据也正逐步成为提升科技管理水平的重要数据基础，在科技管理过程中发挥着越来越重要的作用。因此，将大数据的新技术、新工具和新方法运用到科技管理过程中，加强科技管理过程中的数据资源集成分析处理能力，为科研布局和科技决策提供有力支撑，发挥大数据在决策支持、发展战略研究、科技成果产业化等方面的积极作用，已成为应对上述挑战的重要方法。

很多国家在积极探索如何将新技术和新方法应用到科技管理信息化建设中，发挥其在决策支持以及科技成果产业化等方面的作用。美国的科技项目被纳入联邦政府进行统一管理，具有代表性的科技项目管理信息系统包括联邦政府的统一项目管理平台 Grants.gov、国立卫生研究院的eRA系统以及国家科学基金会的FastLane系统。韩国的新税务综合系统（NTIS）通过门户实现国家R&D事业的有效管理，在参与评估活动中，最大化利用已有信息，将重复投资的可能性降到最低。我国从国家到地方，建设了从科研立项到科技成果转化全周期的科技管理信息系统。当前，我国科技管理部门正根据国家政务资源集成共享要求，采用云计算、大数据等技术对各级各类科技资源进行整合共享，以提升科技资源的统筹水平，促进科技创新发展。

科技管理和相关科研活动产生的数据存量大、增速快。据国际科学、技术和医学出版商协会（International Association of Scientific, Technical and Medical Publisher，STM）发布的报告[1]显示，当前全球英文期刊发文量每年约为300万篇，SCI和EI数据库共收录期刊论文260万篇，其中，仅我国论文就有58.9万篇，居世界第二，我国2018年专利申请量为154.2万件，授权量为43.2万件，居世界第1位。学术大数据资源不仅数量巨大，且来源广、复杂性高，具有多元异构的特点。虽然公开发表的学术资源文献可以从传统搜索引擎、Google Scholar等学术搜索引擎、开放存取（open access，OA）期刊站点或中国知网、万方等学术数据库中查询，但目前还没有针对学术领域被广泛认可的覆盖论文、专利、项目等多种不同类型科研行为的多源科研数据收集整合方案。当前科学技术研究日益深入，学科分类和研究领域日益细化，特别是由于科学研究的高度专业性和创新性，各类专家在科技管理过程中发挥着重要作用，选择合适的专家进行专业化的科技管理决策，已成为科技管理过程中的关键环节。目前，全球范围内的研究人员总数已达710万，且正以每年3%~4%的

速度持续增长。我国是2018年研究人员最多的国家[1]，同时，各类专家涉及的学科领域众多、专家间关系网络复杂，如何从众多研究人员中高效率、高精度、多维度地选择推荐专家，面临着诸多挑战。针对这一问题，本文开展基于学术大数据的科技管理应用框架及关键技术研究，充分发挥学术大数据的价值，为科研布局、资源统筹提供决策支持，同时，为科技管理过程中的科研选题、同行评审评议、科技信用评价、项目过程管理等的科技管理过程提供技术支撑手段。

## 2 研究现状

学术大数据在科技管理过程中的相关应用技术包括学术大数据发展及应用、网络表示学习的发展和应用以及学术评价指标的研究和应用等。

### 2.1 学术大数据发展及应用

互联网的持续发展带来了数据量的爆发式增长，而在学术领域中，可公开收集的数据量也达到了相当可观的程度。随着科技管理信息化的逐步发展，学术数据的来源也变得多种多样，这也对学术数据应用大数据技术进行收集和分析提供了必要的条件。学术数据源包括以下几种。

（1）公开发表的学术文献数据

随着数字图书馆技术的普及，大多数学术论文、专利文档、期刊文章等与学术相关的文献均可从网络上获取文献信息或全文内容。文献的获取途径同样也有多种选择，如利用传统搜索引擎或Google Scholar、Bing学术搜索等学术搜索引擎，在部分OA期刊的网站和OA站点中进行检索，使用中国知网、万方等学术数据库查询所需文献数据。

（2）机构公开的科研管理信息

得益于信息化水平的逐渐成熟，各类项目信息的获取也更加便捷和规范。我国各种项目的公开公示信息可以从国家科技管理信息系统公共服务平台、地方科技管理门户网站、院所企业网站获取。可以获取的信息包括项目指南文件、项目评审专家名单、科技成果报告介绍等。同时，对于专业的科研管理人员，平台也可以提供科技成果的全文展示以及详细分析等功能。

（3）基于众包的数据资源或个人信息展示

由于互联网开放共享的特点，学者个体乃至普通大众均可参与学术信息的构建，因此维基百科、百度百科等基于众包的知识信息库也可以作为一种学术数据。虽然这类数据库具有数据质量参差不齐、存在一定错误等缺点，但经过谨慎的处理，并结合一些有效的数据清洗和冗余处理技术，同样可以发挥巨大的作用。此外，学者个人或实验室课题组的主页介绍、各类学术社交网站等也可以提供一定的学术数据。

在学术科研管理方面，大数据同样带来了一定的变革。传统的学术同行推荐依靠的是学术共同体进行人为的定性评价。虽然评价主体的专业性与权威性均毋庸置疑，但仅凭人力无法做到绝对的客观，推荐的广度也无法做到面面俱到。量化的学术评价自20世纪90年代兴起，但因为其评价指标单一、专业性不强等缺点广受争议[2]。随着大数据时代的到来，大量的数据覆盖支持以及各种大数据分析方法使得学术评价体系更加多元化、全面化、丰富化，这在一定程度上加强了学术评价的公正与客观。同时，如果对科研体制进行正确的改革，大数据也为定性与定量评价的融合提供了一定的契机。

目前，尚未有针对学术领域，覆盖论文、专利、项目等多种不同类型科研行为的

多源科研数据收集整合方案。此外，专家属性繁多、专家间关系网络复杂，实现高效率、高精度、多维度的学术画像仍面临诸多挑战。推荐系统正向着细粒度、个性化的方向发展，这也需要更加精确的方法来保证推荐的准确度与可信度。

## 2.2 网络表示学习的发展和应用

学术数据的快速增长给基于学术数据的挖掘和应用带来了巨大的挑战。数据量的增长增加了数据处理的难度，也意味着数据的维度更加丰富，能够挖掘的潜在信息也越多。学术行为链接了学术数据实体，构成庞大的学术网络。由于学术网络数据呈现出稀疏性特点，传统算法难以进行。网络表示学习提供了一种新的思路，将网络中的节点转化为低维稠密实数向量表示，极大地减少了数据和特征的数量，便于应用算法的开展。

网络数据可以表示成邻接矩阵。传统网络表示学习算法主要针对邻接矩阵做矩阵分解操作，随着数据规模的增大，这种方法不再适用。受到词嵌入经典模型word2vec[3]的启发，Perozzi等人[4]提出了DeepWalk模型，通过随机游走并使用skip-gram训练获得节点向量，从任一节点出发获得多个游走序列，可以描述节点在网络中的特征。由于节点在游走序列中出现的频率符合Zipf幂律分布，与词汇在语料中的分布类似，故skip-gram方法可通过这种方式迁移到网络表示学习中。Grover等人[5]提出了一种概率游走的模型node2vec，将原始DeepWalk的适用范围扩展到了带权图中，并通过概率游走保留了节点的社区特征，在众多数据上取得了更好的效果。Dong等人[6]在metapath2vec模型中提出了基于元路径的游走，将DeepWalk扩展到异质信息网络，更好地捕捉到了潜藏在不同类型节点及连边中的特征。

网络表示学习的结果被应用于大量算法中。推荐系统是常见的应用场景，现阶段多数推荐系统包含大量数据，如何快速找到指定节点的相似节点是常见的问题。通过网络表示学习，节点被转化为向量表示，可直接比较节点的相似度指标。在同质网络中，即节点类型的网络中，阿里巴巴集团的淘宝推荐系统[7]引入了网络表示学习，针对手机淘宝上不同用户的丰富需求，根据用户浏览行为建立网络，并使用skip-gram训练得到向量表示，这种方法在淘宝的推荐系统中取得了极大的成功。在异质网络中，Shi等人[8]提出了基于异质网络的推荐系统，在使用基于元路径的游走获得节点序列后，使用节点过滤规则过滤部分类型的节点，使最终的节点序列更能体现节点之间的关系。

虽然现在网络表示学习被大量应用在各个领域的推荐系统中，但是在学术数据上使用网络表示学习并应用到实践的做法还比较少。针对学术网络的表示学习能够更加充分地挖掘学术数据中蕴藏的丰富语义信息，为解决学术领域的任务提供了全新的视角。

## 2.3 学术评价指标的研究与应用

在科技管理过程中，需要对项目相关负责人与团队以及进行项目评议和审批过程中的评审专家的学术水平进行评估。对于学者学术评价指标的研究，传统上一般基于文献计量学的研究进行，即利用其他文献对学者所著论文的引用情况进行量化考核[9]。最基础的衡量标准为论文的被引频次。考虑到一些学者论文数量很大但实际成果水平不高，被引频次无法真实反映作者水平，一些综合考虑论文数量与被引数量的指标随即被提出，包括H指数[10]、

G指数[11]等评价方法。同时，其他结构的评价方法也有一定的应用价值，如Radicchi等人[12]提出了一种基于PageRank原理进行科学家影响力排名的方法；Ding等人[13]基于全文引用分析，考虑论文中引用出现的次数，进一步分析了研究影响力。

传统的基于引用统计的评价方法也有一些缺陷，如引用计数累积需要一定的时间，作者自引会给评价结果带来偏差，综述类文章引用数偏高会影响论文实际学术价值的判断。一些新的计量方式也被提出，如Altmetrics、Entitymetrics等新型计量学研究。一些研究[14]将论文在Twitter或博客等社交媒体或Google Scholar等文献检索管理工具中被阅读、提及或讨论的次数，作为论文水平评价的标准，很多基于Altmetrics思想的学术评估应用[15]也有一定的发展，例如Altmetric.com可以通过Facebook社交网络统计研究者发表的成果的影响力与贡献度，Plum Analytics则主要针对同行评议方向，全方位地评价学者的影响。

学术评价指标是学术同行评议或评审专家推荐的重要参考因素，而准确客观的评价指标尤为必要。结合传统引用计数方式以及新的文献计量学上的指标，并尝试融入其他基于学术数据的学者评价标准，对构建学者专家的画像有着更为实际的意义。

## 3 框架设计

结合国家重点研发计划的主要管理过程，从备选入库、评审立项、过程评估、验收评审及成果评估等环节，充分利用科研行为产生的海量数据，建立基于学术大数据的应用框架和服务体系，进一步提高科研管理水平和管理效率。

### 3.1 设计目标

为充分利用学术大数据的价值，应用框架的设计目标如下。

- 广泛收集各类科研行为数据资源，并形成共享共用机制，建立统一的数据科研大数据共享资源体系，为科技数据资源的挖掘和综合分析提供数据支持。
- 开展学术论文库、专利库、科技成果库、项目库、专家库、信用库等科技资源“互连互通、共享共用”的建设工作，构建学者画像库和专家评价模型，为项目评审和同行评议专家推荐提供有效的支撑。
- 对国家主要科技计划过程管理对象进行应用研究，探索科技管理过程中基于学术大数据的决策支撑机制。

### 3.2 应用框架

学术大数据科技管理应用主要针对申报用户、各类专家、科研人员、社会公众、企事业单位和管理部门，应用框架分为五部分内容（如**图1**所示），包括应用服务、科技管理过程、画像刻画、数据整合和资源收集。

- 应用服务：为各类用户提供服务，为科研布局、资源统筹等宏观决策提供支持，同时为指南制定、公平公正评审、科研立项等科研管理实施过程提供支撑。
- 科技管理过程：为科技管理过程提供全周期支撑，在科研管理过程备选征集、申报受理、入库凝练、出库立项、实施执行、监督检查、项目验收、成果转化的全周期中，提供入库评审、立项评审、执行监督评审、验收评审及成果鉴定评估等过程管理的支撑。
- 画像刻画：针对学者领域多样、差异较大的特征，对学者进行精准画像构建，提取领域内高水平专家，进行专家全方位评估和多维度排名，建立精准画像库。

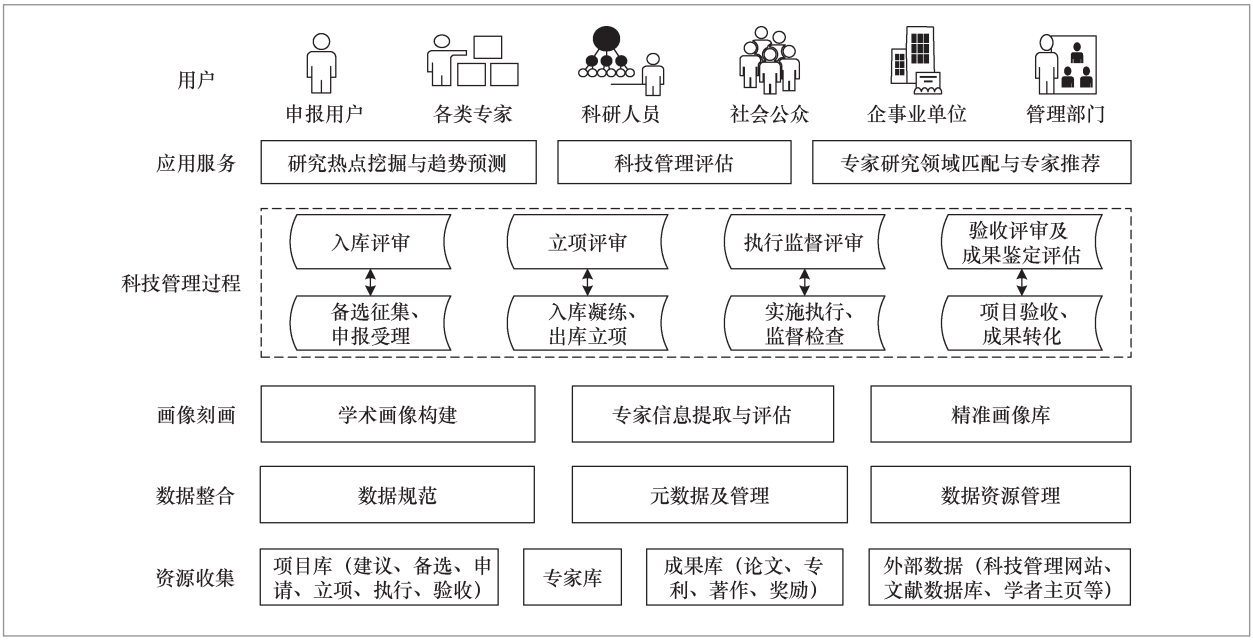


**图1 学术大数据科技管理应用框架**

● 数据整合：整合项目库、专家库、成果库和外部数据等多源异构的数据，基于数据规范、元数据，构建统一的异构资源集成框架。

● 资源收集：收集科研人员和科技专家的基本数据与相关联的文本数据（论文、专利等），并进行融合、实时推送与更新。

学术大数据科技管理应用框架中的“科技管理过程”涉及各种业务环节，这些业务环节主要依据国发〔2014〕64号《关于深化中央财政科技计划（专项、基金等）管理改革的方案》。该框架中“应用服务”的目标是让合适的人做合适的事，利用技术手段辅助支撑科技管理过程的业务流程，比如通过热点挖掘与趋势预测应用服务辅助指南的征集过程，利用专家推荐技术辅助科技管理过程各阶段的评审专家的遴选等。

## 4 关键技术

为了解决学术大数据的收集、学者画像构建和同行评审专家推荐等问题，开展了多源异构学术大数据收集整合技术研究、知识图谱与学者画像刻画技术研究以及基于网络表示学习的专家推荐技术研究。

### 4.1 多源异构学术大数据收集与整合

学术数据的来源丰富，为了完整收集所有可能需要的学术数据，本文采用多种方法收集各类数据源中的学术信息，并应用数据整合方案进行多源异构数据的规范，以供进一步的应用。

通过万方、知网、全国报刊索引等数据库进行文献信息的检索，可以获取期刊论文、会议论文、科技报告以及学位论文等文献数据。每条文献数据包含文献标题、摘要、关键词、分类号、发表日期、作者及单位信息等必要或可选信息，同时期刊论文和会议论文也包含各自期刊与会议的具体信息。部分研究[16-18]尝试对文献PDF数据进行元数据抽取，从而获取规格化的文

献信息。

对于项目和专家信息，可以利用爬虫技术，从各级科技管理部门官网、国家科技管理信息系统公共服务平台、国家自然科学基金委员会官网中获取公开的项目指南、立项信息、项目成果报告简介、专家信息等。由于2015年及之前的指南文件组织形式并不规范，因此对于爬取的文件仍然需要进行一定的数据清洗和整理，例如利用TF-IDF文本特征提取方法获得具备足够信息量的关键词句信息，并通过词嵌入方法进行特征值方面的计算与处理。

学术社交网站中的信息也可供收集和采用。参考文献[19]利用SCHOLAT学者网获取了学者之间的社交互动关系的数据，从而进一步分析了学者之间的信任度与研究兴趣；同时，SCHOLAT学者网中学者用户也会提供课题组的介绍以及成员信息，这可作为进一步分析所用的数据信息。

在科技计划管理系统中设有专家库，包括专家教育经历、工作经历、研究内容、研究成果、职称等信息，同时，科技计划管理系统中还存储了各类科技计划项目的申请文档、过程文档等详细资料，这些数据都为有效评估专家的研究领域与学术水平提供了一定的基础。

通过多种数据收集手段，可以收集海量的学术数据，由于数据来源各不相同，整体数据呈异构状态，因此还需要进一步的数据规范化整合。很多研究提出了不同的异源实体整合方法，如科学数据管理系统MOMIS[20]基于基本的通用模型，针对不同结构的数据配置不同的装饰器，进行统一管理；HCONE-Merge方法[21]则对每个来源的实体增加一个WordNet中间层，并进行合并，整合成为统一的数据实体。

中国科学院计算技术研究所研发的科学计划应用数据集成系统（science plan applying data integration system，SPADIS）[22]提出了一种多源异构数据收集、接入、集成的方法和框架。依据科技管理标准规范，制定数据项的名称与格式，将这些条目组成元数据。根据各个数据项之间的逻辑关系进行聚合，将所有对象与元数据构成一个树模型，针对不同的数据源（如利用各种方式在网络中采集的科研数据以及现有的MySQL、Oracle、SQL Server等数据库引擎保存的不同格式的科研项目库、专家库等遗留数据资源）进行封装整合。该系统通过对元数据进行相应的剪枝操作，生成特定的树模型，并配置元数据与数据条目之间的映射关系，从而做到对异构异源数据的规范与管理。通过数据操作接口，利用XML格式进行树模型的最终整合存储，支持动态配置数据库或外部资源库等数据对象。**图2**为SPADIS中学术大数据的收集、存储与整合的架构。

### 4.2 知识图谱与学者画像刻画

采集的学术科研数据来自多个渠道，要合理地存储，才能够被上层模型和算法高效地利用。本文采用关系型数据与图数据相结合的方式对数据进行存储。关系型数据比较符合人对事物的认知，构建出来的数据也更容易被传统算法利用。图数据是近些年来兴起的数据存储方式，相比于关系型数据，图数据能够表达的信息更加多元化和细粒度化，并在图关系上具备更强的表达能力和更优越的查询速度。对数据采取多种方式的冗余存储，能够为上层算法提供更多的调用方式，该存储方式具备更高的灵活性。

选取MySQL关系型数据库和Neo4j图数据库构建学术网络知识图谱。关系型数据主要包括作者、论文、期刊等字段，以论文为例，关系型数据表中一部分字段的示

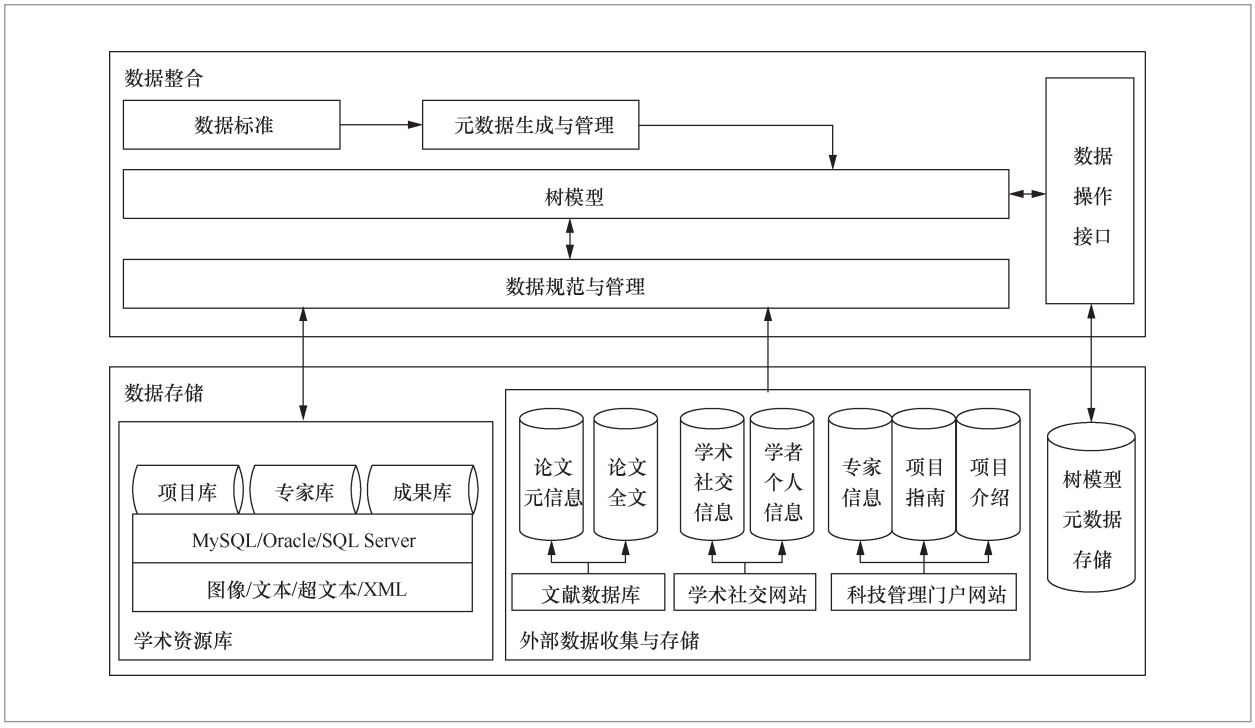


图 2 SPADIS 的学术大数据收集、存储与整合的架构

例见**表1**。

图数据相比关系型数据更能刻画出一些关系，本文使用Neo4j图数据库对学术网络数据构建知识图谱。学术网络中的关系包括作者合作关系、作者发表论文关系、论文出版与期刊关系、作者所在机构关系等。**图3**展示了Neo4j数据库中图数据库实体属性和关系可视化示例，可以看出，同一作者可以属于不同的作者机构，不同的作者之间存在合作关系。不同的关系经过图数据库得到了显式的表达。相比关系型数据库，图数据库的表现能力更强，可以挖掘出数据中更深层次的信息。

表 1 关系型数据表部分字段的示例

| 字段 | 示例 |
|---|---|
| 标题 | 软件测试性设计综述[23] |
| 作者 | 付剑平，陆民燕 |
| 作者机构 | 北京航空航天大学工程系统工程系，100083 |
| 主题 | 软件测试性设计，软件测试性，软件开发周期 |
| 摘要 | 软件测试性设计分为4类，设计时应当遵循的测试性设计原则 |
| 发表年份 | 2008年 |
| 所在期刊 | 计算机应用 |
| 分类号 | TP311.5 |

## 4.3 基于网络表示学习的专家推荐

在知识图谱构建的基础上，可实现多种科研管理的任务，其中包括科研热点挖掘、专家推荐、科研成果评估等。目前，在国家科学技术部的专家推荐过程中，专家库拥有约10万名专家的各类基本信息。而传统的专家推荐方式是通过制定一些基础的筛选与回避原则，采用人工方式遴选推荐专家，这种方式效率较低，且难以做到绝对的客观公正。因此，针对科研管理中科研专家推荐面临的这

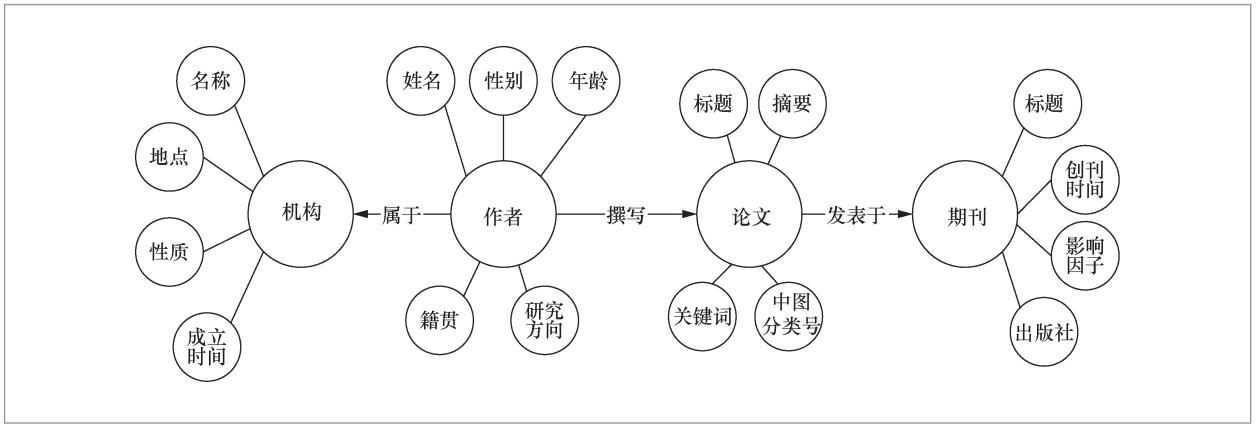


图 3 图数据库实体属性和关系可视化示例

种问题，提出了结合领域推荐和相似专家搜索的专家推荐方法，从而实现专家的精准匹配。

学者领域匹配能够根据推荐的领域匹配申请书对应的项目指南的领域，可以通过相似作者搜索完成最终的专家推荐。学术网络中不同作者和学者的学术行为往往比较复杂，除了研究领域的差别之外，在发表文献数量、权威程度、所在机构等方面存在较大差异。受制于这些复杂特征的多元性、稀疏性等特点，很难通过传统特征工程的方式统一处理，使得传统的相似作者搜索难以体现作者之间深层的关联。**图4**展示了学术网络中作者的多种特征示例，可以观察到学术网络中作者存在大量异构特征，传统方法很难使用一种通用的方式进行处理。

近年来，网络表示学习越来越受到关注。网络表示学习可以将网络中的节点通过机器学习的技术转化为低维稠密向量表示，相比传统邻接矩阵节省了大量存储空间，并包含了更多的信息。通过使用不同的信息构建不同的网络，并进行网络表示学习，能够使节点向量包含不同种类的信息，从而可延用于节点分类[24]、节点聚类[25]、相似搜索[26]等后续任务中。

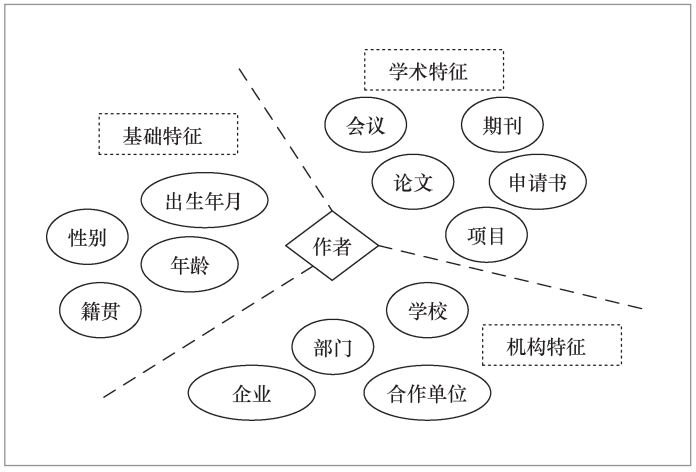


图 4 学术网络中作者多种特征示例

本文使用网络表示学习解决相似作者搜索的问题。如**图5**所示，首先使用网络表示学习将庞大的学术网络中的每个节点转化为实数向量表示。这些向量里蕴含了学者在网络中的结构特征，同时包含了网络中节点的数字和文字形式的特征。获得这些特征向量之后，结合学者研究领域匹配结果获取到的部分科研专家，可以从所有种子学者中筛选候选学者，并直接通过比较学者向量的余弦相似度获得相似作者集合。通过

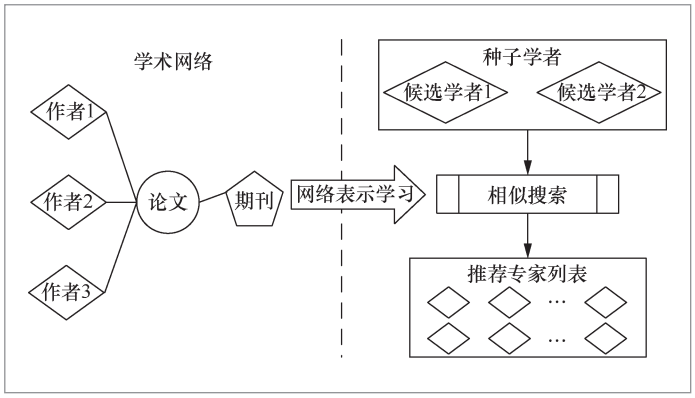


图 5 使用网络表示学习进行专家推荐

这种途径，可以快速寻找到相似专家，并获得最终需要推荐的专家列表。

论文中的关键词描述了该篇论文的研究领域。从论文摘要、标题中提取关键词，并将关键词信息看作网络节点，使用网络表示学习获得其相应的向量表示[27]。关键词的向量表示可用于科研热点挖掘和成果评估，可通过建立时间维度上关键词向量的演变情况来实现。关键词的向量表示可以追踪不同关键词在不同时间内的热度，从而为未来科研资源统筹提供参考。在科研成果评估中，可以通过关键词向量表示得知科研成果的主要技术点，从而达到辅助合理评价的目的。

# 5 预期应用

本节重点介绍学术大数据在科技管理过程中科研布局和资源统筹辅助决策、科技管理过程专家精准推荐、科技管理过程成果评估评价方面的预期应用。

## 5.1 科研布局和资源统筹辅助决策

科研热点挖掘和趋势预测一直是学术界研究的热点，也是进行战略决策和科研投入的基础[28-29]。现阶段相关决策的进行主要依赖于人工审核，依靠业界经验丰富的人士做出下一步主要研究方向和趋势的判断，这种方式受制于人的知识体系，同时产生了巨大的人力消耗。而随着网络表示学习的兴起，可以运用网络表示学习的方式动态地追踪学术界的研究热点，并对下一阶段的热点进行预测，从而减少人为干预，为科研布局和资源统筹提供参考。

## 5.2 科技管理过程专家精准推荐

在科技管理过程中，科研选题（指南制定）、评审立项、执行检查、验收评估、成果鉴定等环节都需要选择适合的专家进行决策。基于学术大数据的各类专家精准画像将有助于在科技管理过程中选出适合的专家。结合科技管理信息系统中的专家推荐功能，系统可以智能化地分析管理需求，并结合管理需求，从专家学术水平、资历经验、专家与评审项目领域匹配、学术道德信用、智能回避原则等多维度选出适合的专家，最大可能地减少人为因素，提升科技管理过程的科学性和公平性。

## 5.3 科技管理过程成果评估评价

在科技管理过程中，各类科研成果的评估评价是一项重要工作，一方面需要对参与评估的候选专家的学术能力、权威度、影响力进行综合考察，通过精准推荐各类专家，选择适合的专家对成果进行评估评价，提升评估评价的权威性；另一方面，也可以基于学术大数据，通过热点挖掘、趋势分析、最新进展分析等，对具体科研成果进行大数据分析比较，全面评判最终成果的创新程度、技术水平与实用价值，为专家最终的评估评价提供参考。

## 6 结束语

信息化已经成为促进经济社会不断发展的关键，学术大数据作为信息化建设的重要组成部分，正在不断创造丰厚的社会效益与经济价值。本文结合我国科技管理过程的应用需求，设计了基于学术大数据的科技管理应用框架，提出多源异构学术大数据收集与整合技术、知识图谱与学者画像刻画技术和基于网络表示学习的专家推荐技术，并应用于科研布局与资源统筹决策、科技管理过程专家精准推荐以及科技管理过程成果评估评价等环节，以全面提高评审专家遴选效率，提升科技管理过程的公平公正性。在未来的研究工作中，笔者将深度挖掘学术大数据的价值，更好地为科技管理过程服务，推动创新型国家转型战略实施。

## 参考文献:


[1] International Association of Scientific, Technical and Medical Publishers. The STM report: an overview of scientific and scholarly publishing[R]. 2018.

[2] 张耀铭. 学术评价存在的问题、成因及其治理[J]. 清华大学学报（哲学社会科学版），2015，30(6)：73-88.
ZHANG Y M. Causes and treatment of academic evaluation[J]. Journal of Tsinghua University (Philosophy and Social Sciences), 2015, 30(6): 73-88.

[3] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// The 26th International Conference on Neural Information Processing System, December 5-10, 2013, Lake Tahoe, USA. New York: Curran Associates Inc., 2013: 3111-3119.

[4] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: online learning of social representations[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2014, New York, USA. New York: ACM Press, 2014: 701-710.

[5] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[C]// The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016, San Francisco, USA. New York: ACM Press, 2016: 855-864.

[6] DONG Y, CHAWLA N V, SWAMI A. Metapath2vec: scalable representation learning for heterogeneous networks[C]// The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2017, Halifax, Canada. New York: ACM Press, 2017: 135-144.

[7] WANG J, HUANG P, ZHAO H, et al. Billion-scale commodity embedding for e-commerce recommendation in Alibaba[C]//The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 19-23, 2018, London, UK. New York: ACM Press, 2018: 839-848.

[8] SHI C, HU B, ZHAO W X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(2): 357-370.

[9] SANDSTRÖM U, SANDSTRÖM E. Meeting the micro-level challenges: bibliometrics at the individual level[C]// The 12th International Conference on Scientometrics and Informetrics, July 14-17, 2009, Rio de Janeiro, Brazil. [S.l.:s.n.], 2009: 845-856.

[10] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the National Academy of Sciences, 2005, 102(46): 16569-16572.

[11] EGGHE L. Theory and practise of the g-index[J]. Scientometrics, 2006, 69(1): 131-152.

[12] RADICCHI F, FORTUNATO S, MARKINES B, et al. Diffusion of scientific credits and the ranking of scientists[J]. Physical Review E, 2009, 80(5): 056103.

[13] DING Y, LIU X, GUO C, et al. The distribution of references across texts: some implications for citation analysis[J]. Journal of Informetrics, 2013, 7(3): 583-592.

[14] PRIEM J, PIWOWAR H A, HEMMINGER B M. Altmetrics in the wild: using social media to explore scholarly impact[J]. Computer Science, 2012, arXiv: 1203. 4745.

[15] ROEMER R C, BORCHARDT R. From bibliometrics to altmetrics: a changing scholarly landscape[J]. College & Research Libraries News, 2012, 73(10): 596-600.

[16] 刘宇, 钱跃. 基于字典匹配和支持向量机的中文科技论文元数据抽取[J]. 工程数学学报, 2012, 29(4): 586-592.
LIU Y, QIAN Y. Metadata extraction from chinese papers based on dictionary matching and support vector machine[J]. Chinese Journal of Engineering Mathematics, 2012, 29(4): 586-592.

[17] 欧阳辉, 禄乐滨. 基于SVM的论文元数据抽取方法研究[J]. 电子设计工程, 2010, 18(5): 4-7.
OUYANG H, LU L B. Research of paper metadata extraction method based on SVM[J]. Electronic Design Engineering, 2010, 18(5): 4-7.

[18] 欧阳辉, 禄乐滨, 钱建立. 基于C4. 5的论文元数据抽取算法研究[J]. 计算机工程与设计, 2010, 31(16): 3708-3711.
OUYANG H, LU L B, QIAN J L. Research of paper metadata extraction algorithm based on C4. 5[J]. Computer Engineering and Design, 2010, 31(16): 3708-3711.

[19] 白如江, 杨京, 王效岳. 单篇学术论文评价研究现状与发展趋势[J]. 情报理论与实践, 2015, 38(11): 11-17.
BAI R J, YANG J, WANG X Y. Research status and development trend of single academic paper evaluation [J]. Information Studies: Theory & Application, 2015, 38(11): 11-17.

[20] TANG H, LIANG Y, CHEN H, et al. Online application of science and technology program oriented distributed heterogeneous data integration[C]//2011 3rd International Conference on Computer Research and Development, March 11-13, 2011, Shanghai, China. Piscataway: IEEE Press, 2011: 363-367.

[21] BENEVENTANO D, BERGAMASCHI S, GUERRA F, et al. The momis approach to information integration[C]// The 3rd International Conference on Enterprise Information Systems (ICEI 01), July 7-10, 2001, Setubal, Portugal. Amsterdam: ICEIS Press, 2001: 194-198.

[22] KOTIS K, VOUROS G A, STERGIOU K. Towards automatic merging of domain ontologies: the HCONE-merge approach[J]. Web Semantics: Science, Services And Agents On The World Wide Web, 2006, 4(1): 60-79.

[23] 付剑平, 陆民燕. 软件测试性设计综述[J]. 计算机应用, 2008, 28(11): 2915-2918.
FU J P, LU M Y. Survey of software design for testability[J]. Journal of Computer Applications, 2008, 28(11): 2915-2918.

[24] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[C]//The 31st Annual Conference on Neural Information Processing Systems December 4-9, 2017, Long Beach, USA. [S.l.:s.n.], 2017: 1024-1034.

[25] WANG X, CUI P, WANG J, et al. Community preserving network embedding[C]// The 31st AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, USA. [S.l.:s.n.], 2017: 203-209.

[26] SHANG J, QU M, LIU J, et al. Meta-path guided embedding for similarity search

in large-scale heterogeneous information networks[J]. Computer Science, 2016, arXiv: 1610. 09769.
[27] ZHANG W, LIANG Y, DONG X. Representation learning in academic network based on research interest and meta-path [C]// International Conference on Knowledge Science, Engineering and Management, August 28-30, 2019, Athens, Greece. Cham: Springer International Publishing, 2019: 389-399.
[28] BOZKURT A, KESKIN N O, WAARD I. Research trends in massive open online course (MOOC) theses and dissertations: surfing the tsunami wave[J]. Open Praxis, 2016, 8(3): 203-221.
[29] JORDAN M I, MITCHELL T M. Machine learning: Trends, perspectives and prospects[J]. Science, 2015, 349(6245): 255-260.

## 作者简介

**梁英**(1962- ),女,中国科学院计算技术研究所高级工程师,主要研究方向为大数据分析挖掘、网络内容安全和隐私保护。

**张伟**(1993- ),男,中国科学院计算技术研究所硕士生,主要研究方向为网络表示学习、学术大数据。

**余知栋**(1996- ),男,中国科学院计算技术研究所硕士生,主要研究方向为物端协同计算、大数据技术。

**史红周**(1971- ),男,中国科学院计算技术研究所高级工程师,主要研究方向为物端协同计算、物联网安全、大数据技术。

**收稿日期:** 2019-06-10
**基金项目:** 国家重点研发计划基金资助项目(No.2018YFB1004700)
**Foundation Item:** The National Key Research and Development Program of China (No.2018YFB1004700)