中国科学: 信息科学 2020 年 第 50 卷 第 7 期: 957–987
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
领域大数据知识图谱专题 • 评述

科技大数据知识图谱构建方法及应用研究综述

周园春[1,2], 王卫军[1,2], 乔子越[1,2], 肖濛[1,2], 杜一[1,2*]

1. 中国科学院计算机网络信息中心, 北京 100190
2. 中国科学院大学, 北京 100049
* 通信作者. E-mail: duyi@cnic.cn

收稿日期: 2019–12–03; 修回日期: 2020–02–19; 接受日期: 2020–04–28; 网络出版日期: 2020–07–13

国家自然科学基金重点项目 (批准号: 61836013)、国家自然科学基金专项项目 (批准号: L1924075)、中国烟草总公司科技重大专项 (批准号: 110201901027(SJ-06))、科技部创新方法工作专项 (批准号: 2019IM020100) 和北京市科技新星计划 (批准号: Z191100001119090) 资助项目

摘要 以研究科学创新与演化规律为目的的科学学近年来迎来了进一步的发展, 科技大数据领域知识图谱在其中发挥了重大的作用. 本文将从科技大数据知识图谱构建及应用研究角度, 对科学学研究过程中发挥重大推动作用的科技领域知识图谱技术进行系统、深入的综述, 阐述科技大数据知识图谱构建过程中涉及的科技实体抽取、科技实体消歧、科技关系抽取、科技关系推断等问题, 对科技实体推荐、科技社区发现、科技实体评价、学科交叉以及学科演化等科技大数据知识图谱分析挖掘方法进行系统梳理, 并给出科技大数据知识图谱未来的研究及应用方向.

关键词 科技大数据, 科技领域知识图谱, 科学学, 科技数据挖掘, 图神经网络

1 引言

以研究科学创新与演化规律为目的的科学学 [1] 近年来受到研究人员的不断关注, 通过对科学与技术研究过程中的合作网络特征、问题发现与选择、引文特征等的研究, 极大地促进了科技的发展. 科学学的研究涉及论文、发明专利、科研人员、科研团队与机构、项目等各类与科技研究相关的数据, 这些数据通过各种关系进行相互关联, 形成了典型的海量、异质 [2] 的科研大数据 [3] 网络. 如何构建该异质网络, 并挖掘网络中的关系, 成为科学研究的热点问题.

基于知识工程长期的研究与实践工作, 谷歌于 2012 年提出了知识图谱的概念 [4]. 谷歌知识图谱的表现形式是以提高人工智能可解释性为主要目标的智能搜索引擎, 其本质上是一种以实体语义为核心的语义网络, 能够从关系的角度提供分析及解决相关问题的能力. 近年来, 以应用需求为驱动, 知识图谱在智能问答、搜索推荐、常识推理等应用场景下取得了较好的效果. 针对通用知识图谱以及中文通用领域知识图谱的构建及应用, 文献 [5~7] 已经进行了系统、全面的综述. 然而, 领域知识图谱与通

引用格式: 周园春, 王卫军, 乔子越, 等. 科技大数据知识图谱构建方法及应用研究综述. 中国科学: 信息科学, 2020, 50: 957–987, doi: 10.1360/SSI-2019-0271
Zhou Y C, Wang W J, Qiao Z Y, et al. A survey on the construction methods and applications of sci-tech big data knowledge graph (in Chinese). Sci Sin Inform, 2020, 50: 957–987, doi: 10.1360/SSI-2019-0271

用知识图谱相比, 由于领域数据的特点及在精准度、专业性、时效性等方面的不同要求, 存在需要解决的特定问题. 领域知识图谱面向特定领域构建知识网络, 能够将知识网络赋能医疗、教育、科技等以知识密集型领域为代表的特定领域. 科技领域知识图谱面向科学技术领域, 构建以科技项目、学术论文、专利、科技动态等为主要数据源, 以科技成果、科研人员、机构、科技项目、主题词等为主要实体, 以支持面向科技领域的学科分析、影响力评价、关联挖掘为主要目的的领域知识图谱.

科技大数据知识图谱与其他领域知识图谱相比, 在数据源上的可获得性、更新频率、数据质量等各不相同, 这导致了在科技大数据知识图谱构建过程中涉及的知识抽取、实体对齐、知识推断等也需要不同的方法. 同时, 由于科技领域本身的领域特点, 其在应用需求上面临学科分析、影响评价等特定领域需求, 也需要不断的研究. 为进一步系统化科技大数据领域知识图谱的研究, 本文从科技大数据知识图谱构建及应用的角度, 对科学学研究过程中发挥重大推动作用的科技领域知识图谱技术进行系统、深入的综述, 系统阐述科技大数据知识图谱构建过程中涉及的科技实体抽取、科技实体消歧、科技关系抽取、科技关系推断等问题, 对科技实体推荐、科技社区发现、科技实体评价、学科交叉以及学科演化等科技大数据知识图谱应用进行系统梳理, 并给出科技大数据知识图谱未来的研究及应用方向.

2 科技知识图谱构建及应用体系架构

领域知识图谱的构建方法包含了自顶向下、自底向上以及混合的模式. 其中, 自顶向下的模式是指通过人工整理出领域知识图谱的实体及关系模型图进行知识图谱数据的抽取与知识图谱的构建; 自底向上的模式以各类机器学习方法为主, 实现自动化的构建; 混合模式指结合自底向上的模式及自顶向下模式进行领域知识图谱的构建. 在科技领域知识图谱构建过程中, 预先定义的实体与关系模型图能够极大地提高科技领域知识图谱的构建质量及应用效率. 图 1 呈现了常见的科技实体及其之间关系的模型图 (部分实体间关系未在图中呈现), 包括了科研人员、出版物 (会议、期刊)、科技成果 (期刊论文、学位论文、会议论文、专利、图书等)、科技项目、机构、主题词、学科等实体, 以及科研人员之间的合作关系、科技成果之间的引用及被引关系、项目与科技成果之间的资助关系等, 同时在实际设计时还需对科技实体进行属性的扩充, 如科研人员的姓名、年龄、性别、职称等属性信息, 构建科技知识图谱实体及其之间关系的模型图. 在科技知识图谱的实体及关系模型构建中, 各科技实体之间的关系是完成相应的科技数据分析与挖掘任务的重要组成部分, 如一些重要论文发表以后对作者影响力、期刊影响因子、新兴前沿学科发展的影响等是不断动态变化的, 此类科技实体之间的成果引用关系、作者合作关系、主题词同时出现在同一论文中的共现关系或影响等是科技知识图谱分析的重要基础. 科技知识图谱利用三元组或属性图的方式进行知识的存储和表示, 通过形成的网状结构数据完成科技实体推荐、评价, 以及学科发展趋势、学科交叉趋势等研究, 其价值不仅体现在助力科研人员寻求研究方向、合作伙伴等, 也体现在通过发现科学研究的相关规律, 辅助科研资助机构从宏观上评估、制定相应的政策去引导科学技术的发展.

科技知识图谱构建及应用体系架构主要包括科技领域数据源、科技大数据知识图谱构建、基于科技知识图谱的分析与挖掘、基于科技知识图谱的洞察与发现、基于科技知识图谱的应用系统及工具、支撑科技知识图谱的大数据技术以及支撑科技知识图谱的标识技术, 如图 2 所示. 在大数据技术及标识技术的支持下, 基于海量、多源、异构的科技领域数据源实现科技大数据知识图谱的构建, 并进一步实现基于科技知识图谱的分析与挖掘; 在科技大数据知识图谱及各类分析、挖掘方法的基础上, 进一步实现对科技领域的洞察与分析, 并提供基于科技大数据领域知识图谱的应用.

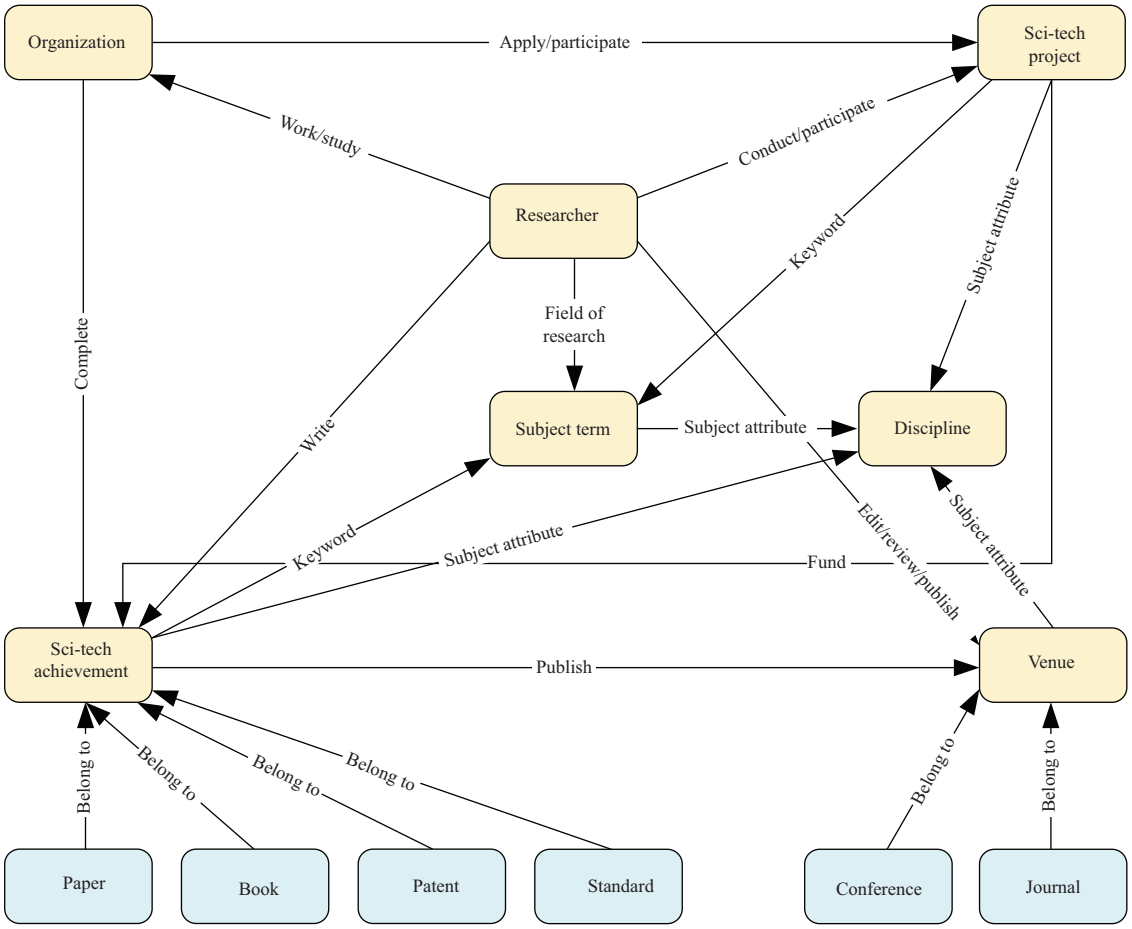

图 1 (网络版彩图) 科技知识图谱实体及其关系模型

Figure 1 (Color online) Entity and relation model of the sci-tech knowledge graph

(1) 科技领域数据源: 与科技领域相关的各类数据资源, 包括科技论文、专利、科技项目等各类结构化、半结构化及非结构化数据源, 为科技大数据知识图谱构建提供数据支持.

(2) 科技大数据知识图谱构建: 利用各类科技领域数据源, 根据面向科技领域数据的特点, 研究科技实体抽取、科技实体消歧、科技关系抽取和科技关系推断等科学问题, 同时解决不同的科技大数据知识图谱的融合和跨语言实体的对齐问题.

(3) 基于科技知识图谱的分析与挖掘: 利用构建的科技大数据知识图谱, 研究科技实体推荐、科技社区发现、科技实体评价、学科交叉、学科演化等科学问题, 提供包括跨领域科技实体与关系问答等应用功能, 同时为科学学的洞察与发现提供数据及分析方法的支持.

(4) 基于科技知识图谱的洞察与发现: 面向科学学的主要问题, 利用科技领域数据源及科技大数据知识图谱及关键技术, 并结合各类社会学、心理学、经济学等不同学科, 研究包括科研网络洞察、学术生涯评价、科学问题发现等主要问题.

(5) 基于科技知识图谱的应用系统及工具: 面向科技领域应用, 服务社会公众、科研人员、科研机构、资助机构、政府部门、企业等不同角色, 研究系统及相应服务工具.

(6) 支撑科技知识图谱的大数据技术: 科技大数据领域知识图谱的存储、计算等需要结合各类大

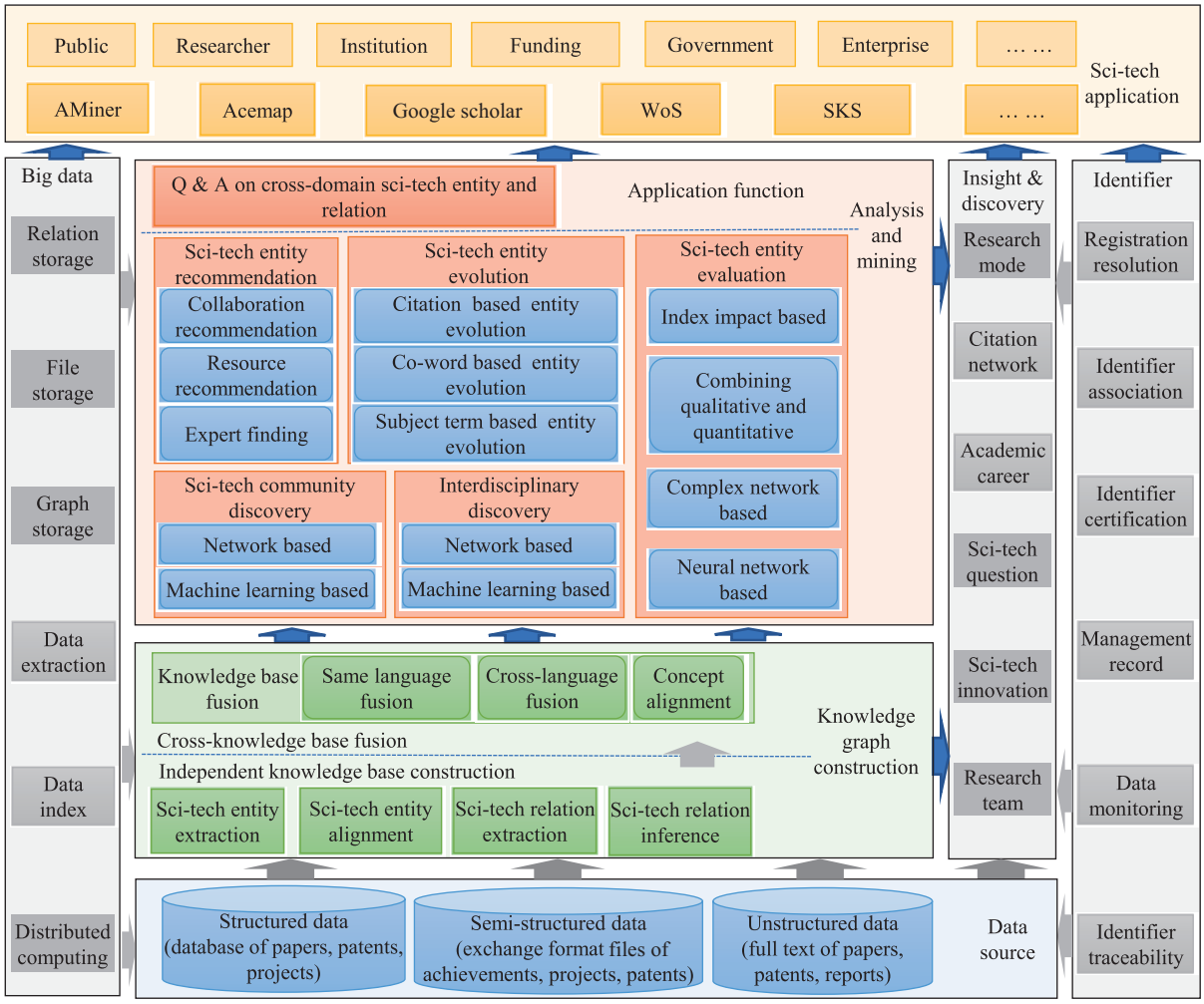

图 2　(网络版彩图) 科技知识图谱构建及应用框架

Figure 2　(Color online) Construction and application framework of sci-tech knowledge graph

数据技术, 包括关系存储、文件存储、图存储等大数据存储技术, 和数据抽取、数据索引以及分布计算等大数据计算技术. 由于领域知识图谱本质上是一种关联网络, 需要对关联数据的存储和计算进行针对性的研究.

(7) 支撑科技知识图谱的标识技术: 高效的标识技术, 可以对科技知识图谱的构建及分析挖掘提供支持. 这包括对于标识的注册解析、关联、认证、管理备案、流量监控以及标识溯源等方面的研究.

其中, (4) 基于科技知识图谱的洞察与发现, 更多地从经济学、心理学、物理学等角度进行更加深度的洞察; (6) 支撑科技知识图谱的大数据技术, 在知识图谱的存储、计算等方面支持了科技大数据知识图谱, 与传统知识图谱的存储、计算技术上的技术与方法类似; (7) 支撑科技知识图谱的标识技术, 目前已经广泛在科技实体识别、认证等领域进行了成熟、成功的应用. 为更为聚焦, 本文着重对 (1), (2), (3), (5), 即科技领域数据源、科技大数据知识图谱构建、基于科技知识图谱的分析与挖掘、基于科技知识图谱的应用系统及工具等 4 个部分展开介绍.

3 科技知识图谱构建

3.1 科技实体抽取

科技数据具有不同类型的来源和结构, 对不同类型的数据进行知识抽取的方法各有不同, 科技领域的数据类型可分为结构化数据、半结构化数据以及非结构化数据. 结构化数据具有较好的组织特性, 而半结构化和非结构化的科技数据往往是由异构数据混合组成, 具有多来源和多类型的特征, 因此建立一个多来源异构数据的集成实体抽取系统尤为重要 [8,9]. 科技实体抽取的主要目的是根据名称规范元数据设计, 通过系统从集成融合的科技数据中抽取不同的实体信息, 例如出版物信息、作者信息、专利信息、机构信息等. 对于结构化的科技实体抽取, 通常使用基于规则匹配的抽取方法; 非结构化的科技实体抽取通常使用基于机器学习方法; 半结构化的科技实体抽取方法则两者皆有, 通常结合这两种方法.

结构化的科技数据指的是使用关系型数据库 (relational database management system, RDBMS) 进行表示和存储, 表现为二维形式的科技数据, 他们往往各项之间存在明确的关系名称和对应关系. 因此可以将其转化为 RDF (resource description framework) 或其他形式的知识库内容. 例如, 我们只需将这样的数据中的特征中的作者项等与需要构建的知识库中的命名一一对应即可.

半结构化的科技数据往往指的是类似于网页、科研数据主页、PDF (portable document format) 文档等本身存在一定的结构但往往需要进一步提取和整理的数据. 半结构化数据的特点是种类繁多且缺少固定和严格的模式, 一方面, 要对每种类型的半结构化科技数据发掘固定的抽取模式, 另一方面, 要对不同类型的数据抽取的实体进行链接. 目前对于半结构化的科技数据往往是针对大量相似页面中的数据进行抓取和识别来进行实体抽取, 例如对谷歌学术数据的爬取, 需要通过一种称为包装器 (wrapper) 的工具从中提取需要的数据, 包装器是一个能够将数据从 HTML (hypertext markup language) 网页中抽取出来, 并且将它们还原为结构化的数据的软件程序. 对于一般的有规律的页面, 可以在包装器中使用归纳规则的方法, 如使用正则表达式, 或者运用网页的相似元素抽取数据的重复模式来提取网页中的科技实体元素. 同时也可以通过包装器归纳这种基于有监督学习的方法, 自动地从标注好的训练样例集合中学习数据抽取规则, 用于从其他相同标记或相同网页模板抽取目标科技实体 [10]. 在相关工作上, 文献 [11] 提出了一个两步的科技实体抽取策略, 首先使用基于查找的抽取方法将半结构化的文献中的片段与已有的数据库比对识别出所有的候选实体, 然后通过候选实体的上下文关系将候选实体与已有的实体匹配. 文献 [12] 提出了一种基于规则的方法, 从科学出版物的 PDF 文件中根据文本信息抽取布局上的分块实体. 除了基于规则匹配的方法, 基于机器学习的方法也广泛应用于半结构化数据的科研实体抽取, 文献 [13] 基于支持向量机的模型从 PDF 格式的科学论文数据中抽取 8 种类型的科研实体. 文献 [14] 提出了一个基于 web 的平台 SmartPub, 该平台可从不同的科技数据来源 (例如 DBLP (database systems and logic programming) 和 arXiv 数据库) 中提取特定领域的命名实体, 同时对于罕见的实体类型, 可以在使用很少的人工监督情况下使用多种命名实体识别方法训练模型并抽取实体. 文献 [15] 提出了一种新颖的无监督的集体推理方法, 将实体从非结构化的生物医学文献全文中提取, 并将其链接到已有的知识库中的实体. 文献 [16] 使用两种模型: 条件随机场和最大熵方法, 从包含文本文档和外部参考注释的 131 页细菌种类的网页实验数据中抽取生物医学实体.

对于非结构化的科技数据的实体抽取往往是指对科技文本数据的领域命名实体识别, 其目的是识别出文本类型的科技数据中表示命名实体的成分, 这类技术随着自然语言处理和文本挖掘技术研究的

不断深入, 已经有了长足的发展. 早期的领域命名实体识别主要使用基于规则和统计模型的方法, 这类方法对于像科技数据这样的领域知识往往有着很高的准确率, 但需要领域专家构造出基于科技实体的模式集合. 基于机器学习的科技实体识别往往使用分类模型, 将不同类型的科技实体作为给定的类别, 通过有监督的模型在训练数据上进行训练得到分类器, 主要思路是先识别出文本中所有命名实体的边界, 再对这些命名实体进行分类, 主要使用的方法有传统的概率模型如马尔可夫 (Markov) 模型、深度学习模型 (如 long short-term memory, LSTM). 除此之外, 卷积神经网络 (convolutional neural network, CNN)、混合神经网络 (hybrid neural network, HNN) 等深度学习方法也被成功用来解决命名实体识别问题, 并取得了较好的结果 [17]. 命名实体识别方法被应用在各种非结构化的科技数据中, 文献 [18] 使用支持向量机模型来实现生物医学实体的抽取, 通过领域专家对本体的概念、属性和关系的标注文本样本对模型进行训练, 这样就可以对新的文本中的医学实体进行标注. 文献 [19] 在使用科学文章的全文的基础上, 使用多种实体自动提取算法构建多种实体网络, 并将其应用于 4 种应用场景: 多产作者发现和重要生物实体发现、有意义的关键字发现和重要话题发现. 文献 [20] 提出了一种基于条件随机场的对于地理文本的命名实体识别方法. 文献 [21] 使用遗传算法的搜索能力抽取生物医学领域的命名实体, 并使用基于条件随机场和支持向量机的方法对实体进行分类. 文献 [22] 提出了一个基于最大熵标记器对天文杂志文章中的天文命名实体的识别方法. 文献 [23] 提出了一种基于 Bi-LSTM (bi-directional long short-term memory) 和 CRF (conditional random fields) 与基于特征的命名实体知识库相结合的实体抽取方法, 抽取与生态修复技术相关的实体.

3.2 科技实体消歧

在科技领域知识图谱构建过程中, 通过科技实体识别技术识别出有效的科技实体之后, 需要实体对齐技术对不规范的实体名称进行消歧. 由于出版形态中具有不同的表达形式, 或相同的表达形式可能会指向多个科技实体, 给科技实体的准确识别带来极大的困难. 消除名称歧义、精确识别与定位科技实体, 是信息组织与检索、科技评价、知识服务等的重要基础.

科技数据的实体对齐中的实体主要指的是人员、机构、期刊的实体消歧, 这些实体之间的歧义主要分为两种形式, 第 1 种是同实体不同名 (synonyms), 这类歧义的产生原因是对于同一个实体名称的不同写法或者不同的存储模式造成该实体名称的不统一, 例如对于某位作者的姓名, 可能会产生名和姓前后不一, 有时是缩写有时是全称的不同字符串. 另一种歧义是同名的不同实体 (homonyms), 又以人员举例, 在数据库中会存在相同姓名但不是同一个人的作者, 但如果不加标识, 数据库会把属于这个名字的作者归为同一个人 [24]. 以作者姓名为例的两种歧义形式如图 3 所示. 在科技数据库中, 这些实体名称的歧义往往会造成文档检索准确率的不足, 因此命名实体消歧的技术是实体抽取过程中必不可少的一个步骤. 其次由于同一作者、机构可能发表多篇文章或撰写多件专利, 同一项目可能资助多篇文章或专利, 同一期刊会刊载多篇文章, 因此抽取的科技实体会在多篇文章或专利等科技成果中出现, 即会存在重复现象.

给定一个通过实体抽取得到的需要消歧某类科技实体的集合 $E = \{e_1, e_2, \ldots, e_{|E|}\}$, 其中有些科技实体属于同一个实体, 其中假设 k 为这类实体真实的判别个体数. 则科技实体消歧的目的是将 E 中的元素聚类到 k 个不相交的簇 $C = \{C_1, C_2, \ldots, C_k\}$, 使得每个簇里边的实体属于同一个科技实体, 而不同簇中的实体不属于同一科技实体.

通常对于这些实体的消歧思路主要是, 将有歧义嫌疑的所有命名实体集合在一起, 根据每个实体本身除名称之外的其他特征进行实体间的相似性对比, 然后通过聚类或者分类的方式将这些实体划分为不同的簇, 同一个簇的命名实体就代表是同一个实体. 具体来说, 首先可以使用作者、机构、期刊的

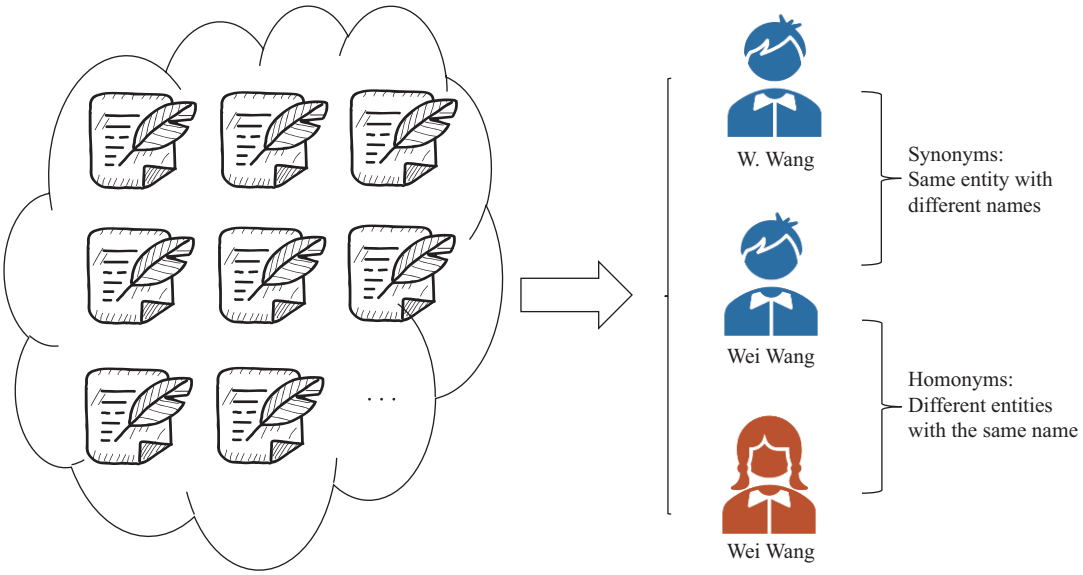

图 3 (网络版彩图) 作者姓名的两种歧义形式

Figure 3 (Color online) Two ambiguous forms of the author's name

唯一 ID 标识符对实体进行预去重的操作, 接着, 可以使用实体的一些属性特征比较不同实体之间的相似性, 例如使用标题、作者、出版年份等信息确定期刊或会议论文是否属于同一篇, 使用机构名称、国家、城市、邮编、地址等特征来匹配是否是同一机构. 同时也可以使用基于表示学习的方法将不同的实体转化成向量的形式, 通过聚类算法根据向量之间的相似度, 生成最终的聚类结果, 进而产生消歧结果 [25].

作者名消歧一直以来都是科研实体消歧中的常见问题, 许多论文对这一问题提出了解决思路. 这些消歧方法的过程通常是通过不同的模型例如概率模型、表示学习模型, 计算出论文的相似性, 再使用聚类模型将这些论文实体划分给不同的作者实体. 文献 [26] 提出了一种基于概率模型的论文相似度计算度量, 计算成对的相似性, 从而衡量论文中的同名作者实体是否匹配. 文献 [27] 通过将查询提交给搜索引擎引入 web 数据, 在作者的出版物网页上提取有助于歧义消歧的有用信息, 再根据层次聚类的方法, 消除作者实体歧义. 文献 [28] 提出了一种基于增量最近邻聚类模型的对于新引入的出版物记录进行增量作者名消歧的方法. 文献 [29] 提出一种结合全局和局部信息的新颖的表示学习方法和一种端对端的簇大小估计方法来对作者消歧, 并将人工注释纳入到消歧过程. 文献 [30] 对于每个需要消歧的作者名构建多种不同的出版物关系网络进行表征学习, 然后使用 HDBSCAN (hierarchical density-based spatial clustering of applications with noise) 和 AP (affinity propagation) 聚类对作者实体进行消歧. 文献 [31] 提出一种视觉分析系统帮助人员消除作者歧义, 这个系统可以量化以及可视化待消岐名字与数字图书馆中已有名字的相似性. 除了作者名消歧之外, 文献 [32] 提出了用于消除机构组织名称歧义的系统. 文献 [33] 提出一种半监督的方法, 利用基于社区的科学术语本体结构, 消除科学术语的歧义. 文献 [34] 提出了一种对科学论文文本中的蛋白质缩写实体进行识别、消歧和存储的框架.

3.3 科技关系抽取

关系抽取的目标是从文本中抽取到两个实体之间的二元关系类型, 表示为 ⟨实体 1, 关系, 实体 2⟩

这样的三元组形式, 通常是命名实体识别的后续任务. 科技文献中蕴含着大量的科技实体, 借助科技关系抽取在科技实体之间建立的关联链接, 进而构建科技知识图谱, 有助于揭示科技实体之间的联系、结构、交叉、演化、合作等科技关系.

在结构化和半结构化的科技数据中, 科技关系的抽取往往伴随着科技实体抽取通过规则匹配的方法而自动生成. 例如, 在抽取作者实体和论文实体的过程中, 通常可以自动化地在论文实体与该论文的作者实体之间抽取出一条撰写关系. 基于模式匹配的关系抽取方法需要科研领域的专家运用语言学知识和科技领域专业知识构造出基于词语、词性或语义的模式, 通过对需要处理的科技实体进行模式匹配, 最终抽取到实体之间的关系. 这种方法的适应性往往较差, 并且建立模式的过程往往费时费力.

对于非结构化的科技数据, 则通常需要对抽取到的不同实体之间的关系进行预测与分类以完成关系抽取的步骤, 科技实体之间的关联关系存在着多种结构, 包括树形结构、网络结构, 关系也分为多种类型, 如同一性关联、隶属性关联、相关性关联等. 这为科研领域的关系抽取带来了一定的挑战.

给定一系列抽取到的科技实体 $E=\{e_1,e_2,\ldots,e_{|E|}\}$, 假定这些实体之间的关系集合为 $R=\{r_1,r_2,\ldots,r_{|R|}\}$, 其中, e_i, $e_j\in E$, $r_k\in R$. 则科技关系抽取的任务是为了得到一个函数 f, 满足

$$f(e_i,e_j)=\begin{cases}r_k, \text{ if } e_i,\ e_j \text{ are related by relation } r_k,\ e_i,\ e_j\ \in\ E,\\ \emptyset, \quad \text{otherwise.}\end{cases} \tag{1}$$

基于规则匹配的实体关系抽取方法被应用于许多的科技实体抽取方法中, 其中, 文献 [35,36] 各自提出了一种从科学文献中自动抽取蛋白质相互作用关系的方法. 文献 [37] 提出了一种从科技文献文本中抽取生物相互作用关系网络的方法. 文献 [38] 提出了一个名为 PKDE4J 的实体与关系抽取系统, 该系统在高度灵活和可扩展的框架中继承了基于字典的实体抽取和基于规则的关系抽取. 同时许多工作提出了基于机器学习和深度学习的实体关系抽取方法. 文献 [39] 提出了一种基于卷积神经网络的科技实体关系抽取方法. 文献 [40] 构建了一种化学领域里的 PDF 文件、文档, 以及 HTML 网页中的化学实体与图像之间的对应关系抽取方法. 文献 [41] 使用无监督的基于多项式核的模式聚类, 对生物医学文献进行文本挖掘, 进而抽取医学实体关系. 文献 [42] 使用了集成学习的方法, 提取文献中的药物相互作用关系.

3.4 科技关系推断

目前, 通过实体和关系抽取技术所构建的知识图谱是高度不完整的, 尤其对于数据来源高度非结构化的科技知识图谱 [43]. 在科技知识图谱中, 科技关系推断的主要目的是为了发现除了关系抽取所生成的实体间关系以外的实体间互动、相关的关系信息, 而知识图谱存在的主要意义也是支持非查表 (look-up) 形式的问答功能. 科研关系推断的方法有基于符号学与逻辑推理的方法、基于 RDF 网络模型的方法以及基于神经网络模型的方法. 其中, 基于神经网络模型的方法, 更容易推断出一些隐藏且复杂的关系.

传统方法主要是一些基于规则的方法, 在早期, 对于多关系复杂网络模型的学习方法主要是通过 ILP (inductive logic programming) 与其抽取的 Horn 从句进行推理. SHERLOCK 是第一个从万维网中抽取先序 Horn 从句的系统. 早期的关系推断通过这些方式进行 [44]. 还有些方法利用张量 (tensor) 相乘 [45,46], 这些方法可以看作是在一系列的证据与事实中进行推理和理解. 或者比如 Schoenmackers 等利用富含关系的 Horn 从句对实体进行网络构建 [44], 然后利用这个蕴含关系信息的网络模型进行跨越多个实体的关系推理. 另外一些传统方法通过网络模型, 如 “马云 – 创办 – 阿里巴巴 – 位于 – 杭州” 这一关系, 可以依照传统的符号规则方法推理得到 “马云 – 位于 – 杭州”, 基于符号规则的方法例

如 PRA (path ranking algorithm) [47] 等这一类利用随机游走思想的算法, 由于在推理过程中专注于利用符号规则进行关系推断, 故其在知识图谱的关系推断中能够取得较好的泛化效果.

RNN (recurrent neural network) 模型解决关系推理时的主要思路是通过获取路径对应的语义特征表示, 从而使任意两个实体的关系都可以用其路径的语义特征表示. 文献 [48] 通过 Embedding 这一表示学习方法, 将离散的高维向量映射到连续的低维向量从而获取到低维空间的实体语义表征向量, 并将该低维向量的相似性解释为语义的相似性, 为使用 RNN 进行关系推理奠定了基础. 文献 [49] 利用 RNN 捕捉每条边关系的低维连续向量表示, 并通过一个组合函数, 将路径的最后输出的隐藏层表示以及路径上的各隐藏层表示组合为一个起点到终点的低维连续向量表示, 从而以此表示这条路径的信息. 但这样的方式在解决关系推理问题时准确度较低, 这类方式被称为 Path-RNN. Path-RNN 方式对实体间的全部路径进行评分并取其中的某一条评分最高的路径, 通过模型获取其低维连续向量表示, 而忽略其他的路径, 这样不但浪费了计算效率, 同时也忽略了其他路径包含的信息. 文献 [50] 利用负采样技术加快了模型的收敛速度, 同时对两个实体间 $\{E_s, E_t\}$ 的全部路径 $\{S_1, S_2, \ldots, S_N\}$ 进行了带评分的池化操作, 对路径中的实体的概率密度进行包含取前 K 个 (top-K)、取均值 (mean)、对数指数和 (Log-Sum-Exp, LSE) 等不同的池化 (pooling) 操作. 并对不同的预测关系特征, 利用相同隐藏层进行学习 (类似于基于共享参数的多任务学习). 同时对于每个实体, 利用实体最常出现的 N 种类别, 为实体的类别赋予表示向量从而表示该实体, 并且代入路径进行模型训练, 使得其在效果上领先于 Path-RNN 方法, 同时训练效率也能更高. 文献 [51] 将问题描述为获取一个解释、支撑问题和推论的子图, 并利用一个简单的信息检索系统加上 BERT (bidirectional encoder representations from transformers) 模型, 构造了一个可解释的关系推理系统.

4 基于科技知识图谱的数据分析与挖掘

基于科技知识图谱的数据分析与挖掘, 包括科技实体推荐、科技社区发现、科技实体评价、学科交叉研究及学科演化研究. 本节对 5 类基于科技知识图谱的分析与挖掘的应用场景进行分类总结, 表 1 [52~120] 从应用场景、具体应用场景、使用方法等维度, 对相关应用研究类文献进行了梳理. 对于每一类应用场景, 横向维度表示其研究方法, 纵向维度表示具体应用场景, 进而列出相关研究文献. 从表 1 可以看出, 部分研究方法维度和应用场景维度交叉处缺少相关研究文献, 主要原因为在交叉处尚未发现相关研究或者不适合进行相关研究. 如基于项目的评价主要通过定性和定量相结合的方法进行研究, 关于其他方法进行项目评价的研究尚未发现; 学科间知识转移的研究体现的是知识的流动, 很难通过成果内容独立进行相关研究等. 除此之外, 在各类不同的基于科技知识图谱的数据分析与挖掘应用场景中, 科技实体推荐、科技实体评价等场景, 由于其问题定义清晰、明确, 存在较多的应用研究论文, 而学科演化、学科交叉等场景的研究论文较少. 但我们发现, 在图 2 中基于科技知识图谱的洞察与发现上, 近两年来有较多高水平研究成果集中在学科演化、学科交叉等场景, 这体现了场景的重要性, 也会成为未来几年应用研究的突破点.

4.1 科技实体推荐

科技数据中有许多类型的实体, 包括作者、论文、专利、期刊、机构等, 对于这些实体有许多的推荐场景. 实体推荐指的是系统通过使用相关的算法对已知的数据信息和某些寻求推荐的实体进行分析, 进而得到推荐结果. 可将寻求推荐的实体看作为用户实体, 将推荐的信息看作推荐对象. 实体推荐算法主要分为以下 3 种: 基于关联规则的推荐的核心思想是从大量的数据中寻找满足一定支持度的

表 1　科技知识图谱数据分析及应用

Table 1　Data analysis and application of sci-tech Knowledge Graph

Sci-tech entity recommendation	Application scenario	Association rule based	Content based	Collaborative filtering based	Hybrid approach
	Collaboration recommendation	[52, 53]	[54, 55]	[56, 57]	[58]
	Resource recommendation	[53, 59]	[60, 61]	[62, 63]	[64]
	Expert finding	[65]	[66]	[67, 68]	[69∼71]
Sci-tech community discovery	Application scenario	Collaboration based	Achievement citation based	Keyword based	Hybrid approach
	Similar interest researcher community	[72, 73]	[74]	[75]	[81]
	Similar research topic community	[76]	[77∼79]	[80]	[81, 82]
Sci-tech entity evaluation	Application scenario	Index impact based	Combining qualitative and quantitative	Complex network based	Neural network based
	Scholar impact evaluation	[83∼88]	[89, 90]	[91∼94]	[95]
	Project evaluation	–	[96, 97]	–	–
	Other entity evaluation	[98]	[90, 99]	[91∼93, 100]	[95]
Interdisciplinary research	Application scenario	Collaboration based	Citation based	Achievement content based	Hybrid approach
	Knowledge transfer	[101]	[102, 103]	–	[104, 105]
	Topic discovery	–	[106]	[107, 108]	[109]
	Pattern discovery	[110]	[111]	[112, 113]	[114]
Discipline evolution research	Application scenario	Citation based	Co-word based	Keyword based	
	Sci-tech community evolution	[115]	–	–	
	Discipline and topic evolution	–	[116∼118]	[119, 120]	

频繁实体项集, 然后再依据置信度从中找到强关联规则, 最后便可以根据该规则向用户实体推荐其关联实体. 基于内容的推荐算法的理论依据主要来自于信息检索和信息过滤. 该方法的基本原理就是根据用户实体已有特征, 获得用户实体的目标画像, 然后再将推荐对象的特征与用户实体的目标画像进行比较, 最后将比较结果相似的推荐对象推送给用户实体. 协同过滤推荐的基本原理是基于用户已有的关联对象建立用户实体的特征, 进而通过计算寻找与用户实体具有相似特征的实体, 再根据这些实体的关联实体去向用户实体推荐其未关联的目标实体.

(1) 科技合作推荐. 随着科学技术的快速发展, 科学研究的专业性和复杂性使得科技合作的重要性日益凸显, 科研人员之间与科研机构之间的科技合作成为推动科学发展的重要方式, 科研人员之间往往需要与其他人合作完成某些科研课题或者论文, 而科研机构之间往往需要在项目上密切合作. 科技合作推荐就是解决这样的问题的一种技术, 它利用论文、专利、项目等现有科技数据资源, 通过数据

挖掘、推荐系统等技术发现相关的、潜在的合作者, 为科研管理人员或研究人员推荐他们关注领域的研究人员 [121], 为科研机构推荐他们可能潜在的合作机构. 该方法能够帮助他们快速发现并了解领域相关的研究机构、人员和内容, 促进开展进一步的交流合作, 带来更好的知识和资源共享, 提高科研质量, 加快科研进程, 取得更高水平的科研成果.

文献 [52] 提出了一种使用共同作者网络推荐学术社交网络中合作者的新方法, 使用了一种加权间接规则挖掘方法. 文献 [54] 提出一种基于论文共同作者学术关系的推荐系统. 该推荐系统应用科研人员所著论文的关键词对科研人员进行建模, 通过论文共同作者等学术关系计算科研人员之间的相似度以进行推荐. 文献 [58] 提出一种利用作者出版物内容和作者之间的协助网络的协作者推荐模型, 使用主题聚类确定学术领域, 并采用随机游走模型, 计算研究人员的特征向量. 文献 [55] 利用主题分布模型提出了一种跨领域的合作者推荐模型. 文献 [56] 提出了一个结合了作者协作实体嵌入网络和层次分解模型的上下文感知的作者协作者推荐模型. 文献 [57] 结合基于内容的过滤技术与主题建模算法, 根据作者的职业、作者之间的协作、医学主题词术语和相关研究人员的工作效率提出了一个生物医学推荐系统.

(2) 科技资源推荐. 科技资源包括论文、专利等科技实体资料, 也包括期刊信息等知识性资料. 科技资源推荐就是向科研人员推荐相关的科技资料. 例如引用文献推荐 [122]、投稿期刊推荐 [60]、专利推荐 [123] 等. 随着科技文献数量和期刊会议数量的增长, 研究人员在寻找科技资料时如何准确地、快速地找到自己想要的资料成为一个具有挑战性的任务. 通过自动化的推荐技术可以帮助研究人员从数量庞大的数据库中智能地筛选出合适的文献资料. 以引用文献推荐为例, 这类推荐技术一方面可以帮助科研人员寻找到内容与当前文献相符的参考文献, 另一方面可以根据科研人员自身的用户特征找到适合其引用的与其期望层次相符的文献, 以此提高推荐的精度和质量.

文献 [53] 提出了一种基于 Apriori 算法和 PageRank 算法的改进的混合加权关联规则挖掘算法, 通过用户行为分析及其权重计算进行相关文献推荐. 文献 [59] 提出一种基于神经概率模型的引文推荐方法, 通过共同学习引文上下文和引用论文的语义表示, 得到给定引用上下文时引用相关论文的概率. 文献 [62] 提出一种基于引文上下文的协同过滤引文推荐方法, 通过关联挖掘技术确定引文的共现关系, 通过论文对之间的相似性关系, 预测相应的引文. 文献 [63] 提出一种基于协同过滤和子图局部排序的方法的引文推荐模型. 文献 [60] 提出了一种个性化的学术期刊会议推荐模型, 向作者推荐高质量的学术期刊和会议. 文献 [64] 提出了一种基于信息过滤和信息汇总模型的方法, 向相关研究人员和从业者推荐合适的研发项目. 文献 [61] 提出一种基于异质图表示学习的个性化科技论文推荐方法, 首先根据提取的内容信息构建作者和论文简介, 通过词嵌入技术和基于元路径的相似性度量学习到异质图的节点表征, 并根据作者和论文的相似度来完成论文推荐.

(3) 专家发现. 在科技知识图谱中, 专家发现主要应用于评阅人推荐 [124], 同时也可以帮助搜索引擎发现某科技领域内的专家, 进而对专家进行排名. 专家发现即发掘在某专业领域内有影响力的、有较高的科研水准的科研人员实体. 以评阅人推荐为例, 评阅专家的学术水平通常由其论文的产出量及影响力决定, 一般用于反映产出量的指标为发文篇数, 反映论文的影响力的指标主要有: 核心期刊、SCI 收录论文、被引频次、期刊影响因子. 此外, 评阅专家的影响力反映在国内外的同行评议和 h-index [125] 等. 同时针对具体的评阅任务, 最好需要专家对该科研方向有相关或相似的评阅经验. 经上述因素的综合评估, 最终系统会推选出排名最高的专家进行评阅.

文献 [65] 提出了一种基于加权关联规则的方法, 利用它来发现共同审稿人之间以及审稿人与作者之间的潜在利益冲突, 以此来分配合适的审稿人. 文献 [66] 利用构建的专长吻合度、学术影响力与社会关联值 3 个变量构建了一个专家遴选、回避与推荐模型. 文献 [69] 开发了一种自动匹配系统, 将研

究计划中基于数据库的具有模糊权重的关键词与进行评估的审阅者进行匹配, 为国家研究管理系统推荐领域专家. 文献 [70] 提出了一种基于词移距离和构造性覆盖算法的分类方法来推荐期刊稿件提交和审阅系统中合适的审稿人. 通过使用提交的稿件关键词和审稿人的专业知识, 评估审稿人相对于提交稿件的专业性. 文献 [71] 利用审稿人候选人与投稿之间的相关性, 候选人的兴趣趋势以及候选人的权威 3 种因素将审阅者分配看作一个整数线性规划问题. 文献 [67] 综合分析科研社交网络中专家所具有的知识信息以及社会关系信息, 并以此为基础, 构建链接预测模型对科研社交网络中的用户进行相关专家推荐. 文献 [68] 将专家发现问题转化为分类问题. 该方法建立了一个知识数据库, 通过协同智能构建的 Web 信息, 表示领域的专业知识特征, 并提出了一种增量学习方法来更新数据库.

4.2 科技社区发现

Girvan 和 Newman 认为[72], 社区是图中节点之间的联系比与图的其余部分的联系更紧密的子图, 或者说子图内的节点连接紧密, 与子图外的节点连接稀疏, 则该子图具有社区结构. 科技社区发现的基础是科技知识网络的构建, 在现有的社区发现的研究中, 科研人员通过作者合作、科技成果引用、关键词共现等关系将科技知识构建为网状结构, 或将上述关系进行混合组成混合网络进行社区发现的研究. 科研网络中, 科技实体及其关系是网络中知识流动和传递的基础, 科研网络中往往将内部连接相对紧密的科技实体对应的子图称为社区, 各社区间科技实体存在交集的部分称为重叠科技社区, 没有交集的则称为非重叠科技社区. 在科研网络中, 通过社区发现方法可以探索科研网络的社区结构, 降低科研行为分析的复杂性. 同一个社区内的科技实体对应着相同研究兴趣或相近研究主题, 社区发现算法是探测并研究科技数据内部的潜在科技规律及研究趋势的常用方法.

目前, 社区发现算法已经有很多相关学者从不同角度提出并逐渐成熟, 文献 [126] 从基于模块度、统计模型、网络节点分类、进化计算 4 个角度对社区发现算法进行相关总结. 在科技数据分析中, 相关研究更多从科研人员合作、成果间引用、关键词共现与日渐成熟的社区发现算法结合进行科技成果的分析. 文献 [72] 通过社区发现算法对 271 位科学家合作网络进行社团划分, 并将其划分为 4 个社团. 文献 [73] 以作者合作网络为数据分析基础, 提出基于矩阵分解学习的科学合作网络社区发现方法. 文献 [74] 通过使用 Blondel 等开发的聚类算法, 将其用于作者共被引网络中作者社区的识别与发现中, 并获得了比以前大多数获取 “聚类簇” 的方法明显的优势. 文献 [75] 提出基于 Word2Vec 模型对作者的文献题名及摘要进行建模, 将作者所使用的关键词表示成语义级别的单词特征向量, 从而将作者研究兴趣表示成矩阵形式进行相似性度量, 并对作者研究兴趣相似矩阵进行层次聚类分析. 文献 [77] 通过社区发现算法分析引文网络社区随时间的变化, 跟踪能源研究中的新兴研究领域. 文献 [78] 通过 Newman 提出的社区结构探测算法对引文网络进行聚类, 分析氮化镓 (GaN) 领域和复杂网络领域知识结构的发展演变. 文献 [76] 将基于拓扑的社区检测方法和基于主题的社区检测方法应用于信息检索领域的作者合作网络, 认为未来的社区检测方法不仅要强调社区与主题的关系, 还要考虑社区与主题的动态变化. 文献 [80] 提出了基于共词网络社区演化分析的研究框架, 基于社区主题表示算法和社区相似度匹配算法, 构建了一个科研主题演化分析模型, 并开发了网络社区演化分析软件 NEViewer. 文献 [79] 通过对引文关系的深入研究与学习, 分析 RWGC (random walk graph clustering algorithm) 算法的思想及其存在的问题, 提出了一种基于相关系数的改进的 RWGC-CC (improved random walk graph clustering algorithm based on correlation coefficient) 算法, 实现科研社区的发现. 文献 [81] 在 LDA (latent dirichlet allocation) 模型的基础上, 提出 “社区 – 作者 – 主题” 模型, 模型根据科研人员之间的合著关系和论文的内容来发现隐性的子社区, 并提取出每个子社区中的研究主题以及每个子社区中有代表性的科研人员. 文献 [82] 利用合作和共被引关系的组合方式, 进行了特定学科领域的科学结

构分析. 可见社区发现相关算法和科技数据分析相结合, 更多是因为通过社区划分可以将原始的、复杂的网络简单化, 从而开展关于科学结构的研究.

4.3 科技实体评价

通过知识图谱去评价图谱内大规模节点的重要性, 并在此基础上发现图谱中的关键节点, 是科技大数据知识图谱需要解决的问题, 也是其他领域知识图谱类数据库需要解决的问题. 关键科技实体的发现不但可以作为各类科技实体推荐系统的基础, 也可以为科技大数据知识图谱提供实体消歧. 同时, 由于科技知识图谱一般来说由上亿级别的实体、关系、属性等组成, 通过评估科技知识图谱内节点的重要程度可以为科技数据库的优化设计提供帮助. 本文将科技实体的评价方法分为基于指数影响力的科技实体评价、定性与定量相结合的科技实体评价、基于复杂网络的科技实体评价以及基于神经网络的科技实体评价.

(1) 基于指数影响力的科技实体评价. 针对人员类别的科技实体重要性评价的传统方法主要基于科研人员论文的发表情况与影响力进行评估. Garfield [127] 于 1955 年划时代地提出了使用论文被引用数进行评价的方式, 之后 Hirsch [83] 提出使用 h-index 作为一种具有较大认可度的高影响力特征测度方法. h-index 方法具有广泛的影响力, 它可以将作出持久而且是重大贡献却未获得与其声望相称奖励的研究者凸显出来. 目前, 包括谷歌学术在内的国际学术机构, 已经基于 h-index 提出了近 30 类扩展指数 [84,85], 从不同的角度增强 h-index 的能力, 如 G 指数 [86]、A 指数 [87,88] 等. 针对科研机构与国家绩效水平的评价, 文献 [98] 通过引入个体、整体以及比较基准的概念, 应用归一化方法, 得出具有可比性的 3 个绩效评价指标: 产出指数、影响指数、效率指数. 并在最后, 分析讨论这 3 个指标在表征个体绩效水平变化方面所具有的优势以及推广应用的可能性.

(2) 定性与定量相结合的科技实体评价. 以科技项目为例, 其生命周期包括了指南征集、指南发布、项目申请、项目评审、项目立项、项目过程管理、项目结题、成果管理、项目后评价等过程. 除使用传统的成果影响力评价外, 一些项目资助机构探索通过建立科学研究评估模型, 确定对科技领域或机构的资金资助. 针对成果类科研实体的评价一般来说是基于一些特定的模型, 这些模型是基于绩效的研究资助系统 PRFSs (performance-based research funding systems) [96] 的重要组成部分, 是一种通过对科研成就进行事后评价而确定资助方案的评价体系. 如, 波兰采用 CESU (the comprehensive evaluation of scientific units) 作为自己的绩效资助评价体系. 另外, 层次分析、模糊评价等管理学方法也被应用于对各种项目进行评价, 但较多融入主观因素. 在国内, 有学者提出基于 AHP (analytic hierarchy process) 和区间模糊 TOPSIS 法 (internal fuzzy technique order preference by similarity to an ideal solution) 的高新技术科研项目评价 [97]. 在对科技项目进行评价时基于指数影响力方法、复杂网络方法与神经网络模型方法相关研究成果比较少, 原因可能是项目之间的关联无法构建起可以支撑复杂模型的网络模型, 同时项目评价的整体主观性比较高, 故在目前的影响力评价模型中出现较少.

在对其他科研实体例如高校、研究机构等的评估中, 文献 [99] 从科研总量、科研质量和前沿研究 3 个角度构建了科研评价指标体系, 利用因子分析法验证了指标设置的合理性, 并针对主观给出的权重为区间数和评价模型为 TOPSIS 的情况, 提出一种基于方差最大化的权重计算方法, 以我国 “985” 高校科研状况为例进行了评价和分析. 在对科研人员进行定性与定量相结合的学者学术影响力评价研究中, 文献 [89] 基于多层级细化的评价指标及赋值标准, 以图书情报领域学者的学术成果为例, 从科研产出、论文影响、载文期刊类别以及两种国家基金项目等方面进行综合分析, 针对文献情报学, 研究了学者的学术影响力和职业发展. 文献 [90] 以 2013 年第 5 批 “青年千人计划” 入选者为研究对象, 在分析了这些入选者统计特征的基础上, 通过计算化学领域入选者影响力分值的大小及其变动情况, 对

表 2　相关模型对比

Table 2　Comparison of related models[a)]

	GNN-based	HAR [93]	PPR [92]	PR [91]
Neighborhood	✓	✓	✓	✓
Predicate	✓	✓	×	×
Centrality	✓	✓	✓	✓
Input Score	✓	✓	✓	×
Flexibility	✓	×	×	×

a) 表格中 "✓" 表示支持, "×" 表示不支持.

相关学者的学术影响力以及学者学术研究进行了评价, 并在此基础上对申报单位与毕业院校对于研究水平的引导进行了讨论.

(3) 基于复杂网络的科技实体评价. 在早期, 知识图谱的概念还不明朗, 研究者倾向于使用图模型去解释现实中的关系模型, 比较有代表性的是 PR (pagerank) [91], PR 已被广泛应用于互联网网站搜索与推荐中. PPR (personalized pagerank) [92] 是另外一类基于图模型的重要性评估算法, 但不同于 PR 算法, PPR 允许用户对图模型中不同的实体进行重要性标注以辅助算法的重要性评估. HAR [93] 将 PR 与 PPR 的思想进行了延伸, 对图模型中不同的关系属性进行了区分, 即将谓语 (predicate) 引入到重要性分析算法中. 这些算法因为其对节点的重要性评估都是基于实体重要性依据关系进行传播这一假设, 无法依据实际情况进行训练, 故在使用中其达到的准确值还有很大的提升空间. 在科学学的相关研究中, 文献的引用网络是评价科技实体影响力的数据分析基础, 文献 [94] 为了量化科学家影响力的演化, 作者构建模型并研究了 2887 位有 20 年以上发表论文经验并且至少发文 10 篇 (每 5 年至少 1 篇) 的物理学家于 1893~2010 年间在 Physical Review 系列期刊发表的论文. 文献 [100] 通过构建模型提出一种揭示论文或期刊产生长期影响力的机制的方法.

(4) 基于神经网络的科技实体评价. 基于神经网络角度的关键科技实体重要性评价, 其主要代表是基于 GNNs (graph neural networks) 的关键实体发现. 知识图谱与一般的图网络不同, 其内部可以由多重的、不同的异构信息组成. GNNs 在实际应用中取得了较好的效果, 并且在图的节点分类任务中, 在多个应用领域达到了目前的先进水平 [128~132]. 通过对图结构数据的学习以及邻居关系的聚合, GNNs 能够利用周围节点的重要程度获得中心节点的重要性, 从而在全局的角度对科技知识图谱中的科技实体进行重要性评价. 通过图神经网络对科技知识图谱的建模能够更加逼近于真实情况, 同时相较于 PR 类的算法能有更好的准确度 [95].

知识图谱可以看作是一类图 $G = (V = \{v_1, v_2, \ldots, v_e\}, E = \{e_1, e_2, \ldots, e_p\})$, 其中的节点 V 与边 E 象征着实体与实体间的关系, 其中 e_p 代表 p 类不同的边 (或者被称为 predicate) 的类型, v_e 象征 e 类不同类型的节点. 重要性度量 $s \in \mathbb{R}_{\geqslant 0}$ 代表一个非负实数, 表达了对应节点的重要度. 给定一个知识图谱 $G = (V = \{v_1, v_2, \ldots, v_e\}, E = \{e_1, e_2, \ldots, e_p\})$ 以及一个重要性度量 $s \in \mathbb{R}_{\geqslant 0}$, 可以利用某一个节点实体在图中子集 S 学习到一个方程 $F(S)$, 并利用这个公式求得所有节点的重要程度. 不同于其他重要性评估指标, 图神经网络角度进行的评估可以利用知识图谱的 "边的类型" 这一重要属性, 同时由于图神经网络的特性, 其也具有良好的泛化性能, 并且一些利用图卷积神经网络的节点表示学习方法可以应对动态的知识图谱, 具有良好的扩展性. 基于神经网络与基于复杂网络的相关模型对比如表 2 [91~93] 所示.

4.4 学科交叉研究

学科交叉是伴随社会和学科自身发展需求而出现的一种综合性科学活动 [133], 在人类发展过程中的很多重要成就都是跨学科领域的研究成果. 科技知识图谱中各种科技数据可以按照网络的形式进行组织, 如共词网络、合作网络、引文网络等, 这为学科交叉的研究提供了基础. 通过对科技知识图谱相关科技实体及关系构成的网络进行分析, 可以发现不同科技领域的交叉现象, 以及预测未来交叉趋势. 目前, 学科交叉分析的研究对象主要为科技成果的引文、作者以及主题 (关键词), 并通过引文分析、作者分析、共词分析 [134] 等进行学科交叉的研究. 通过科技知识图谱, 可以辅助实现学科间知识转移、学科交叉主题发现、学科交叉测度、学科交叉规律发现、交叉学科发现等研究, 进而为学科交叉研究提供助力, 促进相关研究的快速实现.

学科间知识转移的研究工作中, 相关研究成果从基于合作、引用关系进行学科间知识转移的研究工作. 而对于学科交叉主题识别的研究, 相关研究更多的是从引用、成果内容进行交叉主题的识别, 在文献 [135] 中, 作者认为由于学科专业化程度越来越高, 研究人员更多聚焦于自身专业领域的研究, 故从作者合作角度进行相关研究尚不多见. 文献 [101] 通过调查 3 个分布式的跨学科团队 (一个科学团队和两个社会科学团队) 的成员, 研究结果表明事实知识的交流只是支持团队的许多学习交流的多种知识的一种. 文献 [102] 基于学科间引证关系及作者学科间迁移进行知识转移研究. 文献 [103] 通过学科间引文变化衡量发现新兴跨学科研究领域的出现. 文献 [104] 对中国中医学学科交叉文献的关键词及引文数据, 借助知识图谱可视化软件揭示中国中医学学科交叉研究领域研究重点、知识源流及高影响力作者群. 文献 [105] 采用引文分析、共词分析和社会网络分析方法, 选择计算机科学、科学学两个学科, 对情报学跨学科知识引用情况进行实证研究, 在此基础之上探讨情报学学科的交叉属性. 文献 [106] 通过文献的引用关系来识别学科间的跨学科研究主题. 文献 [109] 使用从数据库获取的社会科学和人文科学的成果数据, 通过分析学科间引用关系、来自不同领域的关键词共现关系去发现新兴的学科交叉主题. 文献 [107] 基于 Rao-Stirling 指数和 LDA 模型进行领域学科交叉主题识别, 并以纳米科技为例验证将 Rao-Stirling 指数和 LDA 模型用于领域学科交叉主题识别的有效性和适用性. 文献 [108] 运用新的测度指标 TI (term interdisciplinary) 来挖掘学科交叉主题, 以情报学为例, 结合社会网络分析和时序分析方法来研究情报学学科交叉主题.

学科交叉测度方面, 文献 [136] 提出从学科多样性、学科聚合性、综合性测度 3 个方面对学科交叉性的测度方法进行总结. 通过科技知识图谱的网状数据关联可以辅助相关测度的计算衡量, 其中, 学科多样性如 Brillouin index [137] 等, 学科平衡性指标如 GC (Gini coefficients) [138] 等, 学科聚合性如中介中心度、网络密度、网络平均路径长度、网络分裂指数等 [135], 这些学科交叉计量指标都可以利用科技知识图谱的数据进行分析获取. 学科交叉的规律性研究和发现方面, 文献 [110] 通过对研究人员之间的协作图分析科学领域的跨学科性, 发现生物技术和自然科学在其出版物和研究项目合作中是跨学科最多的等学科规律. 文献 [111] 对不同国家或范围的出版物的参考文献的年龄分布进行分析, 发现跨学科引用存在时间滞后现象的规律. 文献 [112] 根据澳大利亚研究委员会的 18000 份研究计划, 用计算 IDD (interdisciplinary distance) 学科交叉距离的方法定量分析项目的学科交叉程度, 发现学科交叉距离越大的项目, 获得资助的可能性越低. 文献 [113] 通过收集的材料学论文摘要, 输入到 Word2Vec 相关模型中, 通过分析词之间的共现关系, 去实现新的热电材料的发现. 文献 [114] 利用文献引用、作者合作信息分析了图书情报领域的跨学科研究变化, 发现图书情报领域研究人员最常引用的是本学科的研究成果.

在交叉学科的发现研究中, 文献 [43] 中利用 Word2Vec 词向量技术从学科内具有的关键字中获取

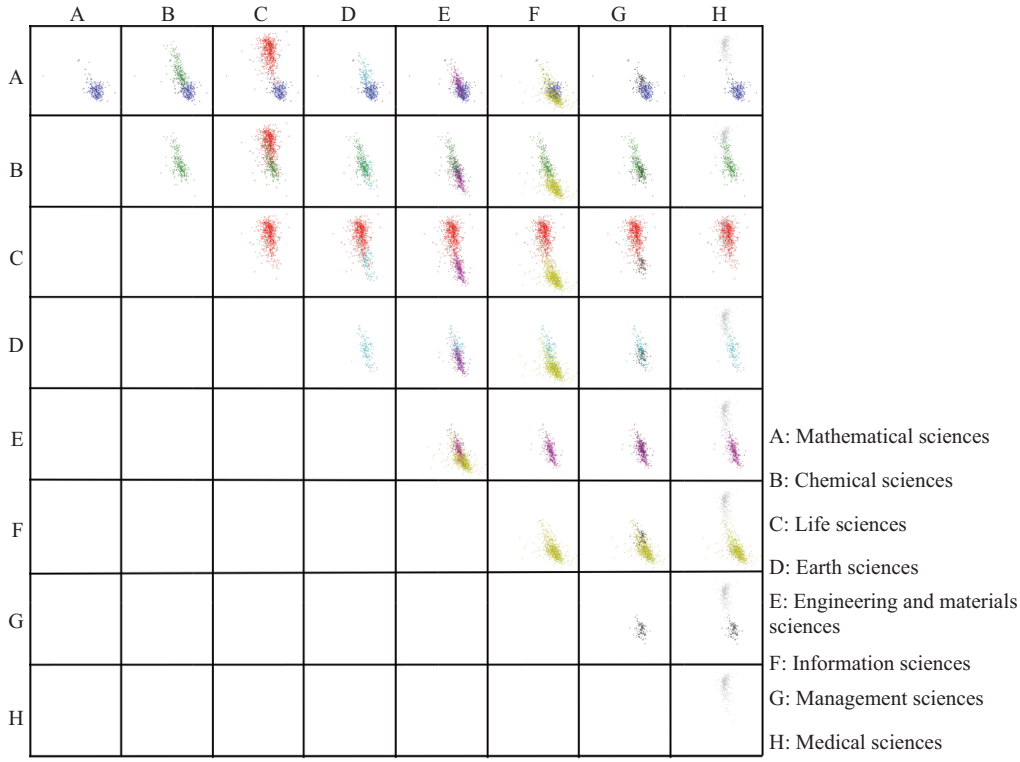

图 4　(网络版彩图) 学科向量化图

Figure 4　(Color online) Discipline vectorized

了对应的低维空间表征, 并基于其二维降维研究了学科的演化过程. 语义上的关联可以解释学科内的话题的变化 [139], 同样的交叉学科的产生可以依据语义上的关联进行发现. 通过将学科表示的语义向量进行二维表示, 可以得到不同学科间的交叉区域, 并能够通过历年的数据与部分已知的交叉学科的状态进行未来的交叉学科发现预测. 基于科技大数据知识图谱的交叉学科发现可以通过使用内建的关键词库解决一般学科语义表征问题中主题词发现的问题. 同样也可以依据知识图谱中海量的语料数据进行训练, 为交叉学科发现提供准确的语义信息. 图 4 所示为依托国家自然科学基金委员会各个学科历年资助项目的项目标题、摘要、关键词、全文等信息, 学习到的各细分学科代码的向量化表示. 从图 4 中, 可直观发现数理学科与工材学科、信息学科、管理学科等在相关细分学科代码下的交叉程度较高, 与生命学科、医学学科、化学学科的交叉程度较低; 生命学科与其他各学科的整体交叉程度比其他学科间交叉程度低. 通过对学科交叉的分析, 可以为科学家、科技政策制定人员、科技项目资助人员等不同角色提供分析与决策的支持.

交叉学科的评价研究中, 交叉学科研究的评价的复杂在于交叉学科定义 [140], 交叉学科在定义上指代着一些具有价值的、覆盖多学科知识面的学科, 它包含了单学科所难以发现的知识. 一般来说, 评价交叉学科首先应该评价的是学科间交换的知识的量 [141], 目前, 关于交叉学科的评价问题主要集中在: 第 1, 由于交叉学科来源于跨越学科间隔的研究, 故难以在目前现有的标准中找到对交叉学科的评价方法; 第 2, 在推选同行进行研究评价时, 由于学科间的偏见的存在, 难以准确找到合适的专家对交叉学科进行客观评价; 第 3, 没有一个基本的度量来对交叉学科的交叉性进行评价. 并且, 由于交叉学科所涉及的概念太多, 这方面决定性的讨论也陷入了困境 [142, 143]. 在对交叉学科所提供的价值的研究

中, 认同度比较大的是交叉学科所带来的主题、视角的宽度 [140]. 将视角放宽于不止一个学科、领域或者研究方向能为学者带来跨领域间的高效知识交换, 同时能够减少并抵制学科内的闭塞与自引用现象, 从长久来看会给学科带来进步并为困难问题提出新的解决思路 [144]. 交叉学科所带来的另外一个价值是其对来自于不同领域的知识的聚合程度, 这意味着学科的交叉性应当将交叉学科的宽度当作一个评价指标而并不应是唯一的评价指标. 聚合程度应当被看作是交叉学科的一个关键价值, 保证了科学进步的方向的正确性 [145]. 最后, 交叉学科所提供的信息交换也是一个重要价值. 交叉学科的信息交换指的是为其相对应的学科的陈旧的知识与观点提供一个新颖的见解与解决方法 [146]. 依据对交叉学科的价值的讨论, 我们可以提出一些对于这些价值的度量方式. 文献 [140] 将学科交叉的评价的度量方式分为多样性和一致性, 这二者用以评价学科主题的宽度与聚合程度. 对于多样性的评价主要是学科的多样性影响学科内部组成成分 (比如说研究者、文章、资助来源), 并使这些成分的分布更加平衡, 或者这些元素的分布将会跨越更多学科领域 [144]. 同样的, 一致性的评价表示于学科内部各类实体间的关联的关系数, 或者说这些实体的联系的紧密程度. 同样也可以根据学科内实体的认知距离在高度的抽象上来审视学科的聚合度与宽度 [147]. 信息交换能力的衡量比较宽度与聚合度来说, 定义更加抽象, 在现有的传统学科领域划分下, 学科间的信息交换能力难以被量化. 传统的学科视角来源于学科体系化的分类系统, 同样在面对新的交叉学科时, 这个系统将会迟缓地采纳新的知识. 针对学科领域, 可以采取热点图的形式将学科研究方向可视化, 以分析其内部有助于交换信息的交叉学科研究领域. 文献 [140] 将管理学与经济学的研究成果、期刊、研究方向等进行可视化, 可以发现聚合在一起的大多是学科本身的核心研究发现, 在边缘的大多是一些与科学、科技以及创新领域相关的研究方向, 这也是一种衡量交叉学科价值的方式.

4.5 学科演化

知识图谱可以看作一种复杂网络形式, 类似于社交媒体网络、电话通信网络, 科技大数据知识图谱也可以利用复杂网络的特性描述图谱内社区的发现、演化、消亡等行为. 通过对科技知识图谱进行网络分析, 可以在较大的视角上获取学科走向、预测学科发展、寻找最新的学科研究点. 同样可以在微观角度上追踪科技合作网络的变化、科技实体间相互交互的行为模式、外部因素对科技实体的影响等. 本文将学科演化分为科研社区演化与学科及主题演化两个主题, 并对文献引用网络、共词网络、主题词方向进行介绍. 在针对科研社区演化的文献中, 由于科研实体的合作主要表示为文献间的互相引用, 故在共词网络与主题词角度上的研究相对比较匮乏, 同样的, 学科及主题演化由于学科的知识主要利用关键词来表示, 故在引用网络上的相关研究相对比较少.

(1) 基于引用的科技实体演变. 通过对科技实体间引用网络的时间变化进行分析, 可以获取科技实体的演变模式. 文献 [115] 中对比了科技实体的引用网络与电话通信交流网络, 将科技实体网络结构演化的基本模式归纳为: 新生、消亡、膨胀、收缩、融合、分解几个过程, 图 5 为该文献中展示的一个社区内的演变事件. 文献 [148] 将网络内的不同实体间的相互作用总结为增续、K- 聚合、K- 分割、建立、消解等不同的事件类型, 并针对时序的异构图的时间的关键行为模式的特性, 围绕统计学的角度进行了讨论, 并基于这些事件对社区间的变化特性进行了总结. 文献 [149] 对按时间变化的动态图谱中的社区的演变进行了讨论.

(2) 基于共词的科技实体演变. 目前在科技实体演变的研究中, 一些科研人员从计量学的角度进行学科演变的研究. 文献 [116] 利用了文献计量和知识图谱的方法, 分析了来源于 CSSCI 的数据, 展示了 1998~2018 年的学科领域建设发展变化情况. 其主要的思路是对领域整体的学者、机构、期刊等进行分布情况的统计和分析, 并利用共词方法对整体学科主题结构与分布进行了可视化. 在文中将学科建

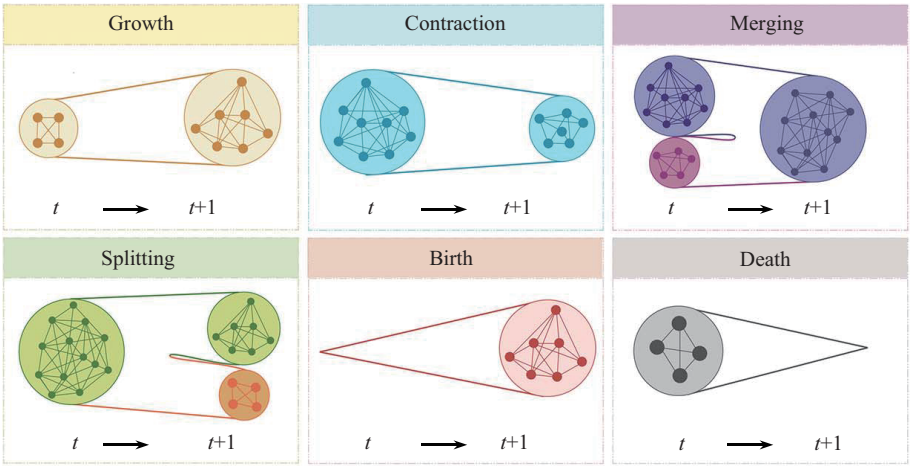

图 5 (网络版彩图) 社区演变事件实例 [115]

Figure 5 (Color online) Examples of community evolution events

设研究的整体发展总结为 “上升、高峰、下降” 3 个不同的阶段. 文献 [117] 利用共词分析方法与社会网络分析方法, 将近 20 年来我国农业工程学科分为 3 个阶段, 并以 9 个重要英文期刊与 2 个重要中文期刊作为样本, 分析并比较了国内外农业工程领域的进展与演化趋势. 文献 [118] 将学科的知识变化抽象为概念的流动, 从文献计量学的角度定义了一种名为概念流的指标, 并以一个共词矩阵 (co-word matrix) 研究计算机与图书馆主题下学科内的知识变化.

(3) 主题词角度的科技实体演变. 在针对科技实体的分析中, 科技主题的挖掘与演化分析是发现科技实体发展趋势的重要思路 [119]. 也就是科技实体的演变可以看作是其主题的演变, 而其主题表现为关键词或者主题词的共现聚集. 对于学科、论文、专利等的主题获取, 可以采集历年的学科关键词作为主题的梗概, 而关键词的词义就可以作为该主题方向; 同样的, 还可以通过统计学习的方式, 从语料中抽取学科主题词. 同时, 一个学科内文献的产出、新关键词的生成等统计信息, 同样可以作为学科性质的科研实体演变的重要指标.

目前, 除了基于复杂网络的社区演化模型, 一些学者围绕学科的主题语义变化开展相关研究, 文献 [119] 基于 LDA 主题词提取模型, 提出了针对锂电池领域的学科语义的提取, 并在这之上针对学科演化进行了讨论, 将学科内主题的行为分为分裂、融合、继承 3 种主要的主题语义演化模式, 并依据时间窗口进行排列来捕捉学科内的语义演化模式. 文献 [120] 利用了 Word2Vec 词向量方法在摘要中提取语义近似的技术关键词, 同时在跨学科领域文献中提取技术关键词, 最后对于扩展的技术关键词进行聚类与演化分析, 并对农业动物生殖细胞和干细胞调控领域进行了实证分析. 通过利用这些演变模型, 可以依据论文、学科、专利等科研实体的演变趋势, 进行其未来语义的预测, 也就是科研实体的趋势分析.

5 科技知识图谱系统及服务平台

5.1 相关工具及系统

目前, 基于科技文献的各类评估方法及分析方法, 国际上也相应地研发了许多辅助评估分析工具, 如 CiteSpace, Sci2, Gephi, Pajek, Ucinet, VOSviewer 等. 其中, CiteSpace 是美国德雷塞尔大学

(Drexel University) Chen [150] 开发的一款免费的学术文献可视化及分析软件, 软件通过对各种网络结构进行分析, 从而识别新的研究领域、热点及前沿; Sci2[1)]是印第安纳大学 (Indiana University) 的专家研发的一款学术文献分析工具, 提供了包括文献共被引、文献耦合、作者合作等分析以及可视化功能; Gephi [151] 是一款常用的网络分析可视化软件, 主要用于各种复杂网络的分析, 包括链接分析、社交网络等; Pajek[2)]是一个用于绘制网络的软件, 也是分析大型网络的工具, 可同时处理数百万个节点和边, 提供合作网络、引文网络以及数据挖掘等功能; Ucinet[3)]是加州大学 (University of California) 的研究人员开发的社会网络分析软件, 包括中心性分析、子群分析、角色分析、基于置换的统计分析等, 拥有很强的矩阵分析功能; VOSviewer[4)]是荷兰莱顿大学 (Leiden University) 开发的网络分析可视化软件, 可用于生成多种可视化图谱, 如共引关系图、共词关系图等. 此外, igraph [152] 和 NetworkX [153] 是常用的复杂网络分析库, 两者都支持社区发现、复杂网络分析等功能, 是科技数据分析的重要工具. 目前的各种分析工具功能多样, 在帮助科研人员对科技数据进行分析挖掘上节约了大量时间和精力, 推进了相关研究工作的开展.

除了以上的科技成果分析评价工具外, 一些大型出版机构也利用自身的数据优势建立各种评价系统. 其中 InCites[5)]是汤森路透公司 (Thomson Reuters) 在汇集分析 Web of Science 中的 3 大引文数据库的基础上构建的科研评价系统, 提供包括以论文为核心的国际影响力评价、机构关联分析等功能, 同时也提供机构合作推荐等各种更深层次的分析服务. 在国内, 知网、万方等也依托各自在中文学术期刊、学术会议等的数据优势, 研发出相应的辅助分析系统. Google Scholar, Microsoft Academic 基于其自身获取各类论文、专利、报告等数据的索引信息的优势, 构建了以学术成果、科学家为核心的海量学术知识库, 并基于此提供包括引文分析、学术趋势分析、个人及机构影响力分析等功能. 清华大学的 AMiner 系统[6)], 通过融合海量科技论文数据, 构建了以论文和科学家为核心的知识网络, 并提供了包括学者评价、学者迁徙分析在内的特色功能. 中国工程院的 “中国工程科技知识中心” (China Knowledge Centre for Engineering Sciences and Technology, CKCEST)[7)]项目汇聚了包括超 44 亿条论文、行业报告在内的成果数据, 并提供了主题分析、战略咨询、交叉领域分析等功能. 科技资源提供商及资源索引机构利用自身的数据优势去提供更深层次的服务, 是科技资源挖掘分析与应用服务的重要发展方向. 同时, 通过以科技知识图谱的形式对科技数据进行管理, 去构建高质量的科技资源决策分析服务平台也是科技资源管理应用的主要内容.

5.2 典型科技知识图谱

5.2.1 OAG

OAG (open academic graph) 开放学术图谱是将学术图谱 Microsoft Academic Graph 和 AMiner 进行融合后构建的大型学术知识图谱. AMiner 是由清华大学开发的, 是一个为研究人员提供学术文献搜索和挖掘服务的平台. 平台提供学术搜索、专家发现、会议分析、话题趋势分析等多样化服务功能. MAG[8)] (mircosoft academic graph) 是微软开发的学术知识图谱, 目前其学术文献数量达 2.3 亿篇,

1) Sci2 Tool: a tool for science of science research and and practice. https://sci2.cns.iu.edu/.
2) Pajek: analysis and visualization of very large networks. http://mrvar.fdv.uni-lj.si/pajek/.
3) Ucinet. http://www.analytictech.com/ucinet/.
4) VOSviewer. https://www.vosviewer.com/.
5) InCites: an objective analysis of people, programs and peers. https://clarivate.com/webofsciencegroup/solutions/incites/.
6) AMiner. https://www.aminer.cn/.
7) 中国工程科技知识中心. http://www.ckcest.cn/.
8) Microsoft Academic. https://academic.microsoft.com/.

提供搜索、学术排行等功能. OAG 开放知识图谱 [154, 155] 在 2017 年的第 1 版中, 包含来自 MAG 的 166192182 篇论文和 AMiner 的 154771162 篇论文, 并生成了两个图谱之间的相应论文的唯一标识符之间的链接关系 64639608 条. 在第 2 版中, 提供了作者、出版地点、出版物以及相应的数据匹配信息. 该图谱数据可以用于引文网络的研究等, 同时也可用于研究多个学术图谱的整合. 两个知识图谱经过融合处理后, 主要包括论文来源信息、论文描述信息、论文作者信息 3 部分, 可供数据下载[9].

5.2.2 Acemap

Acemap[10]知识图谱是上海交通大学发布的自主研发的新式学术搜索系统, 其学术实体包括论文、作者、机构等, Acemap 将学术实体 (包括论文、作者、机构等) 组织成网络, 通过网络分析和数据挖掘的方法, 最终将学术实体的信息以图像的形式清晰、直观地展现出来, 以帮助研究者归纳整理现存的研究工作. 目前该学术知识图谱涵盖了 2.27 亿篇论文、1.15 亿学者、24000 多期刊会议的学术大数据智库和可视化学术地图系统[11]. 目前为止, Acemap 已经实现了动态引用网络展示、论文聚类、学术族谱、研究者画像、作者和会议主页等功能.

5.2.3 SN SciGraph

SN SciGraph[12] 是学术领域的链接开放数据平台, 汇总了 Springer Nature 和学术领域的主要合作伙伴的数据源. 链接开放数据平台通过不断从期刊/文章、书籍章节、组织、机构、资助者、专利、临床实验、会议、时间、引文、Altmetrics、研究数据等中汇总研究领域的相关信息, 如资助者、研究项目、会议、隶属关系和出版物等信息, 预计数据将达到 20 亿个三元组. 其目标是创建学术领域最先进的关联数据聚合平台, 从内部和外部数据仓储中摄取数据, 将其转换为整个企业和研究领域可重用的知识. 同时该平台提供最新相关数据集的下载[13].

5.2.4 Wizdom.ai

Wizdom.ai[14] (以前称 colwiz) 由 Tahir Mansoori 在牛津大学攻读博士学位期间与他的同事 Rifaqat Ali Shah 共同创立. Wizdom.ai 知识图谱涵盖 1 亿出版物、1.12 亿专利、7100 万作者、7.9 万种期刊、9.9 万机构、9.86 亿概念映射、60 亿事实、7.97 亿引用. Wizdom.ai 通过对相关数据进行分析, 向科研人员提供全球新兴的热点和研究趋势、最前沿的研究机构和人员、引文推荐、个人研究图等.

5.2.5 SKS

SKS [156] (scientific knowledge store) 是中国科学院计算机网络信息中心基于知识图谱相关技术开发的一种面向科技领域的大数据知识图谱平台. 平台通过构建的科技领域知识图谱, 向用户提供科技人员、项目、成果、专家等查询与统计分析, 平台结合传统的文献计量学方法、网络表示学习及神经网络等最新机器学习方法提供包括新技术发现、规划制定、项目立项、成果评价的分析与服务等功能. 实现包括影响力评价、关联挖掘、学科分析、立项评价在内的科技评估. 目前该平台已在烟草科技资源管理、空间科学领域知识图谱与决策平台得到部署和应用.

9) Open Academic Graph 2019. https://www.aminer.cn/oag2019.
10) Acemap. https://www.acemap.info/.
11) Acemap 微博. https://weibo.com/u/2051207867?is_hot=1.
12) SN SciGraph. https://www.springernature.com/gp/researchers/scigraph.
13) Discover research from SN SciGraph. https://sn-scigraph.figshare.com/.
14) Wizdom.ai. https://www.wizdom.ai/.

5.3 典型科技领域数据源

科研人员、科技成果 (期刊论文、学位论文、会议论文、图书、发明专利等)、科技项目、机构、主题词 (关键词)、学科等科技实体, 隶属关系、参与关系、合作关系、引用关系等特定实体间关系是知识图谱构建的基础, 科技数据主要从科技文献数据库、学术搜索引擎、科研人员主页等获取, 并通过对获取的科技数据进行清洗、索引、组织, 形成适合知识图谱的数据格式, 如 MAG 知识图谱 [8] 在 Bing 搜索引擎爬取的数据中提取出学术文献信息, 并将其清洗、组织为适合微软学术知识图谱所需的数据格式.

按照科技数据获取方式的不同, 本文将数据源划分为两类. 第 1 类是通过制定符合法律规定的爬虫爬取机制 (或商业合作) 获取的科技数据, 主要来源于商业性数据库、机构知识库、学术搜索系统、开放获取数据库、科研人员主页等. 商业性数据库主要有国内的中国知网、万方数据、维普等, 国外的 Elsevier Sciencedirect 数据库、ProQuest 学位论文数据库等. 对于机构知识库, 主要包括中国科学院机构知识库网格[15]、台湾学术机构典藏[16]等. 学术搜索引擎也是科技数据的重要来源, 如百度学术搜索、谷歌学术搜索等, 此类平台借助自身强大的搜索引擎技术优势, 通过对各种方式获取的科技资源数据进行组织索引, 是科技知识图谱重要的数据来源之一. 另外 arXiv 作为开放数据获取的重要来源, 提供科技文献全文的下载. 第 2 类是直接提供数据集下载的开放性数据源. 如 DBLP, AMiner 等, 此类平台将相应科技资源的描述信息以结构化的数据方式提供文件下载. 同时一些开放性数据源也为相应的学术研究提供了标准的实验数据集, 用于相应的科学研究工作, 如 AMiner 整理标注的数据集[17]等. 表 3 对目前常用的科技数据源进行了总结.

科技知识图谱是面向领域的知识图谱, 科技数据的质量直接关系到科技数据挖掘及应用分析的效果, 也关系到科技数据辅助决策能力的准确性. 科技知识图谱数据来自不同的系统, 数据资源描述方式各异, 如何将不同数据源的异构数据, 融合为高质量的知识图谱数据是科技数据源获取过程中需要考虑的重要因素. 除了顶层设计时要紧密结合特定科技领域的数据特点构建相应的 Schema 层模型, 来自不同系统的科技数据的融合处理也是科技知识图谱构建过程中数据质量的保证. 其中, 数据歧义问题是科技资源获取后的常见问题, 如同一个人在不同阶段在不同的机构学习工作发表论文、英文文献作者姓名标注方式的差异等都会造成科研人员身份的歧义问题, 除了构建科技实体的标识体系外, 对于来自不同科技资源的数据, 如何尽可能多地为特定科技实体关联更多的上下文信息, 如科技实体属性特征信息、与其他科技实体的关系等, 是辅助解决数据消歧融合的重要途径.

6 展望与挑战

6.1 大数据技术与科技知识图谱应用分析的融合

数据获取、清洗、存储、管理、分析等技术不断发展, 然而海量、多源、异构、异质科技资源的采集、处理流程, 以及基于此的科技知识图谱构建依然面临挑战. 科技知识图谱通过对异构科技资源的数据进行处理及管理, 为科技人员利用科技资源进行分析提供了合适的数据结构. 面对大规模的科技数据, 研究如何构建适合大数据环境下的科技数据分析理论及应用方法体系是一个重要的研究方向. 通过对科技成果、科技人员等科技实体构建统一、可回溯的唯一性标识体系, 研究从多个维度获取科研人员精准科研画像的描述属性从而构建模型, 准确反映科研人员的研究领域及相关影响力, 是科技

15) 中国科学院机构知识库网格. http://www.irgrid.ac.cn/.
16) 台湾学术机构典藏. http://tair.org.tw/.
17) AMiner Dataset. https://www.aminer.cn/data.

表 3　常用科技资源数据来源列表

Table 3　List of common sci-tech resource data sources

Name	Description	Acquisition method	Full text access[a)]	Data provider
CNKI	Literature of multiple disciplines, include journals, dissertations, patents, etc.	Crawling	Subscribe	China Academic Journals Electronic Publishing House Co., Ltd.
WANFANG Data	Literature of multiple disciplines, include journals, dissertations, patents, etc.	Crawling	Subscribe	WANFANG DATA CO., LTD.
Chongqing VIP	Literature of multiple disciplines, include journals.	Crawling	Subscribe	Chongqing VIP Information Co., Ltd.
IEEE Xplore	Mainly covering computer, engineering, electronics and other disciplines.	Crawling	Subscribe	IEEE
ScienceDirect	Mainly covering physical sciences and engineering, life sciences, etc.	Crawling	Subscribe	Elsevier
Scopus	Mainly covering chemical sciences, biological sciences, medical & health sciences, etc.	Crawling	Subscribe	Elsevier
PQDT	Mainly include excellent dissertations from well-known universities in Europe and United States, covering multiple disciplines.	Crawling	Subscribe	ProQuest
Web of Knowledge	Citation indexing of academic literature, covering multiple disciplines.	Crawling	Subscribe	Thomson Reuters
AMiner	Mainly covering computer science, etc., providing downloads of dataset.	Free dataset	Full text link	Tsinghua University
MAG	Mainly covering multiple disciplines, providing downloads of dataset.	Free dataset	Full text link	Microsoft
arXiv	Preprint platform, covering physics, mathematics, computer, economics and other disciplines, providing downloads of full text.	Crawling	Free full text	Cornell University
DBLP	An integrated database system for computer sciences journals and conferences, providing downloads of dataset.	Free dataset	Full text link	University of Trier; Schloss Dagstuhl
Baidu Scholar	An indexing and retrieval platform for academic literature, covering multiple disciplines.	Crawling	Full text link	Baidu
Google Scholar	An indexing and retrieval platform for academic literature, covering multiple disciplines.	Crawling	Full text link	Google
CAS IR Grid	Academic research achievements indexing about research institutions of the Chinese Academy of Sciences.	Crawling	Mostly restricted	Chinese Academy of Sciences
TAIR	Academic research achievements indexing of Taiwan universities.	Crawling	Mostly restricted	Taiwan University

a) 全文获取方式中, “Full text link” 表示资源提供了全文获取的链接, 全文是否免费要看链接的形式, 如谷歌学术资源链接到 PDF 文件时, 往往能免费获取到全文.

知识图谱的重要研究方向; 通过对科技知识图谱相关的各类实体、关系的不断优化, 研究适用于科技知识图谱分析、应用任务的分布式存储、索引机制, 提高科技知识图谱的管理效率, 是科技知识图谱

与大数据管理融合的重要研究方向; 通过对科技领域的学科交叉、发展演化趋势等微观角度进行研究, 为相关人员提供决策支持, 需要对海量科技领域数据进行针对性、细粒度的处理, 将智能科技数据融合方法与大数据处理平台进行融合以提供更为高效的科技知识图谱构建, 也是科技知识图谱与大数据处理技术融合的重要研究方向; 通过将大数据相关技术融入到现有的科技知识分析工具中, 为更多的科研人员提供科技资源分析挖掘的通用工具, 提供更加友好的人机交互界面, 是科技知识图谱应用落地的一个重要应用方向; 大数据技术成为支持科技知识图谱的主要方法, 两者的互相融合也将成为未来研究的趋势.

6.2 科技知识图谱支持科技规律洞察与发现

科技资源数量的不断增加、数字化管理及利用技术的不断完善为科研人员开展研究工作提供了极大的便利. 科技知识图谱通过将大量的科技知识关联为网状结构, 为科技资源的规律性知识的洞察和发现提供了支撑. 科学学通过对科学数据的分析, 可以发现科学发展过程背后隐藏的各种知识和规律, 从而为相关科研人员和科研政策制定者提供科学研究的方向和政策制定的依据. 利用科技知识图谱的科技知识, 通过进行科研人员学术研究模式的洞察, 发现科学家学术生涯中的各种影响因素, 引导科学家和相关科研机构制定相应政策为科学家的职业发展提供助力; 通过探究科学研究的潜在未来发展方向, 为相关科研机构或基金资助管理机构制定相应资金支持政策提供决策参考; 通过研究学科间的交叉研究主题等前沿研究领域, 引导科研人员开展相关前沿研究工作, 解决各种复杂科学问题; 通过研究科研人员、科技成果的学术影响力形成模式, 制定合理的学术评价机制, 为管理机构资金资助等工作提供依据; 通过可视化分析技术的研究, 从繁杂的结果中洞察并提取出相关人员能快速理解分析的可视化结果, 也是科技规律洞察和发现不可或缺的研究方向等等. 科技知识图谱为科技规律的洞察和发现提供分析基础, 也为科学学的研究提供更多助力, 是科技知识图谱深层次应用的重要研究方向与动力.

6.3 机器学习方法在科技知识图谱分析的深度应用

从把知识图谱看作一种复杂网络, 使用传统机器学习的方法对知识图谱进行图挖掘分析, 到深度学习方法和图神经网络方法在知识图谱上的进一步应用, 知识图谱的价值和作用得到了进一步的体现. 科技领域知识图谱作为一种特殊的领域知识图谱, 使用通用知识图谱中的分析方法对该图谱进行挖掘分析, 并不能充分地利用科技领域知识图谱中的结构和特点. 对于科技领域知识图谱中特定的结构特征和实体属性, 需要设计特定的机器学习方法或深度神经网络结构, 同时, 对于不同的应用场景, 通常要设计不同的目标函数来学习算法中的参数. 因此, 研究针对科技领域知识图谱的机器学习和深度学习挖掘方法, 有助于科技领域知识图谱的进一步的分析和应用. 另一方面, 随着每年科技文献数据的不断增长, 科技领域知识图谱的容量也在不断地扩大, 这给以复杂网络理论和多层深度神经网络为基础的机器学习方法带来了巨大的挑战, 尤其是在较为复杂的推荐、推理等场景中, 模型需要在准确性和时效性之间进行取舍和折中, 同时如何将上层模型和底层框架进行联合设计和优化是实际应用场景中另外一个亟待解决的问题. 另外, 现有的大多数机器学习模型都是静态的, 缺乏高效处理增量的科技文献数据以及时序性科技文献数据的特点, 近来图神经网络的出现为这个问题提供了一个新的突破口, 但如何提高这类模型的效率以及更深层的可解释性仍然是个挑战.

6.4 科技知识图谱的服务平台建设

基于科技领域知识图谱, 可以为科研人员、科研机构、基金资助机构、出版社、政府决策部门等不同类型用户提供具有针对性的科技知识服务, 例如知识分析洞察、科技领域知识问答、期刊或审稿推荐、研究领域画像等功能. 然而, 要充分地利用科技知识图谱的能力, 一方面需要构建领域完备、时效性强的科技领域知识图谱, 另一方面需要面向实际需求提供具有针对性的知识服务, 这导致了基于科技知识图谱的进一步应用落地还存在一些障碍. 科技知识图谱服务平台, 在数据源层面融合各类开放、可获取的科技领域数据, 同时融合各个需求方特有数据, 通过提供的科技数据获取、实体关系抽取、知识融合、图谱构建及更新等功能构建出来及时更新的科技领域知识图谱; 在应用方面, 通过提供的知识推理、关联挖掘、社区发现等通用知识服务算法与模型, 为不同的角色提供具有针对性的知识服务. 例如, 可以服务于国家自然科学基金委员会及各类基金资助机构的专家智能指派、可以服务于各类期刊、会议的智能评审与评价等, 均在不同层面上体现了科技知识图谱服务平台的思想. 进一步利用各类开放科技数据、科技知识图谱构建及分析方法, 设计兼具通用性及灵活性的科技知识图谱服务平台, 并将科技知识图谱服务平台持续赋能各科技领域知识服务系统, 能够极大地提高科技知识图谱的应用及推广效率, 将是重要且持续的应用方向.

参考文献

1 Fortunato S, Bergstrom C T, Börner K, et al. Science of science. Science, 2018, 359: eaao0185
2 Han J W. Mining heterogeneous information networks: the next frontier. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012. 2–3
3 Xia F, Wang W, Bekele T M, et al. Big scholarly data: a survey. IEEE Trans Big Data, 2017, 3: 18–35
4 Singhal A. Introducing the knowledge graph: things, not strings. Official Google Blog, 2012. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/
5 Yan J H, Wang C Y, Cheng W L, et al. A retrospective of knowledge graphs. Front Comput Sci, 2018, 12: 55–74
6 Wu T X, Qi G L, Li C, et al. A survey of techniques for constructing chinese knowledge graphs and their applications. Sustainability, 2018, 10: 3245
7 Liu Q, Li Y, Duan H, et al. Knowledge graph construction techniques. J Comput Res Develop, 2016, 53: 582–600
8 Liu F, Zhang X L, Kong L H. Research review on the research data repositories. New Tech Library Inform Serv, 2014, 2: 25–31
9 Fayyad U, Haussler D, Stolorz P. Mining scientific data. Commun ACM, 1996, 39: 51–57
10 Huang Y Q, Qi G Z, Zhang F Y. Extracting semi-structured information from the WEB. J Softw, 2000, 11: 73–78
11 Zhao J, Dong K J, Yang L, et al. E-Scholar: improving academic search through combining metasearch with entity extraction. In: Proceedings of 2009 IEEE Youth Conference on Information, Computing and Telecommunication. Piscataway: IEEE, 2009. 247–250
12 Ramakrishnan C, Patnia A, Hovy E, et al. Layout-aware text extraction from full-text PDF of scientific articles. Source Code Biol Med, 2012, 7: 7
13 Kovačević A, Ivanović D, Milosavljević B, et al. Automatic extraction of metadata from scientific publications for CRIS systems. Program, 2011, 45: 376–396
14 Mesbah S, Bozzon A, Lofi C, et al. SmartPub: a platform for long-tail entity extraction from scientific publications. In: Proceedings of Companion Proceedings of the Web Conference 2018. New York: ACM, 2018. 191–194
15 Zheng J G, Howsmon D, Zhang B L, et al. Entity linking for biomedical literature. In: Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics. New York: ACM, 2014. 3–4
16 Grouin C. Biomedical entity extraction using machine-learning based approaches. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), 2014. 2518–2523
17 Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. In: Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg: ACL, 2018. 2145–

2158
18 Takeuchi K, Collier N. Bio-medical entity extraction using support vector machines. Artif Intell Med, 2005, 33: 125–137
19 Amplayo R K, Song M. Building content-driven entity networks for scarce scientific literature using content information. In: Proceedings of the 5th Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), 2016. 20–29
20 Sobhana N V, Mitra P, Ghosh S K. Conditional random field based named entity recognition in geological text. Int J Comput Appl, 2010, 1: 143–147
21 Ekbal A, Saha S, Sikdar U K. Biomedical named entity extraction: some issues of corpus compatibilities. SpringerPlus, 2013, 2: 601
22 Murphy T, McIntosh T, Curran J R. Named entity recognition for astronomy literature. In: Proceedings of the Australasian Language Technology Workshop 2006, 2006. 59–66
23 Ma J X, Yuan H. Bi-LSTM+CRF-based named entity recognition in scientific papers in the field of ecological restoration technology. Proc Assoc Inf Sci Tech, 2019, 56: 186–195
24 Hussain I, Asghar S. A survey of author name disambiguation techniques: 2010–2016. Knowl Eng Rev, 2017, 32: e22
25 Zhuang Y, Li G L, Feng J H. A survey on entity alignment of knowledge base. J Comput Res Dev, 2016, 53: 165–192
26 Torvik V I, Weeber M, Swanson D R, et al. A probabilistic similarity metric for Medline records: a model for author name disambiguation. J Am Soc Inf Sci, 2005, 56: 140–158
27 Pereira D A, Ribeiro-Neto B, Ziviani N, et al. Using web information for author name disambiguation. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries. New York: ACM, 2009. 49–58
28 Santana A F, Gonçalves M A, Laender A H F, et al. Incremental author name disambiguation by exploiting domain-specific heuristics. J Assoc Inf Sci Tech, 2017, 68: 931–945
29 Zhang Y T, Zhang F J, Yao P R, et al. Name disambiguation in AMiner: clustering, maintenance, and human in the loop. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018. 1002–1011
30 Xu J, Shen S Q, Li D S, et al. A network-embedding based method for author disambiguation. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM, 2018. 1735–1738
31 Shen Q M, Wu T S, Yang H Y, et al. NameClarifier: a visual analytics system for author name disambiguation. IEEE Trans Visual Comput Graph, 2017, 23: 141–150
32 Hansen A R, Varon J L, Sinnott-Armstrong N A, et al. Disambiguating organization names. U S Patent 9 779 388. 2017-10-3
33 Prokofyev R, Demartini G, Boyarsky A, et al. Ontology-based word sense disambiguation for scientific literature. In: Proceedings of European Conference on Information Retrieval. Berlin: Springer, 2013. 594–605
34 Atzeni P, Polticelli F, Toti D. A framework for semi-automatic identification, disambiguation and storage of protein-related abbreviations in scientific literature. In: Proceedings of 2011 IEEE 27th International Conference on Data Engineering Workshops. Piscataway: IEEE, 2011. 59–61
35 Thomas J, Milward D, Ouzounis C, et al. Automatic extraction of protein interactions from scientific abstracts. In: Pacific Symposium on Biocomputing 2000, 2000. 541–552
36 Blaschke C, Andrade M A, Ouzounis C A, et al. Automatic extraction of biological information from scientific text: protein-protein interactions. In: Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, 1999. 60–67
37 Skusa A, Rüegg A, Köhler J. Extraction of biological interaction networks from scientific literature. Briefings Bioinf, 2005, 6: 263–276
38 Song M, Kim W C, Lee D, et al. PKDE4J: entity and relation extraction for public knowledge discovery. J Biomed Inf, 2015, 57: 320–332
39 Lee J Y, Dernoncourt F, Szolovits P. Mit at semeval-2017 task 10: relation extraction with convolutional neural networks. 2017. ArXiv: 1704.01523
40 Yan S, Spangler S, Chen Y. Cross media entity extraction and linkage for chemical documents. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2011

41 Quan C Q, Wang M, Ren F J. An unsupervised text mining method for relation extraction from biomedical literature. Plos One, 2014, 9: e102039
42 Thomas P, Neves M, Solt I, et al. Relation extraction for drug-drug interactions using ensemble learning. CEUR Workshop Proc, 2011, 761: 11–18
43 Hu K, Luo Q, Qi K, et al. Understanding the topic evolution of scientific literatures like an evolving city: using Google Word2Vec model and spatial autocorrelation analysis. Inf Process Manage, 2019, 56: 1185–1203
44 Schoenmackers S, Etzioni O, Weld D S, et al. Learning first-order horn clauses from web text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010. 1088–1098
45 Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data. In: Proceedings of International Conference on Machine Learning, 2011. 809–816
46 Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion. In: Proceedings of Advances in Neural Information Processing Systems, 2013. 926–934
47 Lao N, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011. 529–539
48 Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data. In: Proceedings of Advances in Neural Information Processing Systems, 2013. 2787–2795
49 Neelakantan A, Roth B, McCallum A. Compositional vector space models for knowledge base completion. 2015. ArXiv: 1504.06662
50 Das R, Neelakantan A, Belanger D, et al. Chains of reasoning over entities, relations, and text using recurrent neural networks. 2016. ArXiv: 1607.01426
51 Das R, Godbole A, Zaheer M, et al. Chains-of-Reasoning at TextGraphs 2019 shared task: reasoning over chains of facts for explainable multi-hop inference. In: Proceedings of the 13th Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13), 2019. 101–117
52 Koh Y S, Dobbie G. Indirect weighted association rules mining for academic network collaboration recommendations. In: Proceedings of the 10th Australasian Data Mining Conference, 2012. 167–173
53 Chen Z Q, Zhang H L, Ge J, et al. Related document recommending based on weighted association rule mining. New Tech Library Inform Serv, 2007, 2: 57–61
54 Deng S W, Luo Z, Li S R, et al. Scholar recommendation system based on academic relationship of thesis co-authors. Comput Eng, 2013, 39: 12–17
55 Tang J, Wu S, Sun J M, et al. Cross-domain collaboration recommendation. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012. 1285–1293
56 Liu Z, Xie X, Chen L. Context-aware academic collaborator recommendation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018. 1870–1879
57 Guerra J, Quan W, Li K, et al. SCOSY: a biomedical collaboration recommendation system. In: Proceedings of 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Piscataway: IEEE, 2018. 3987–3990
58 Kong X J, Jiang H Z, Yang Z, et al. Exploiting publication contents and collaboration networks for collaborator recommendation. Plos One, 2016, 11: e0148492
59 Huang W Y, Wu Z H, Liang C, et al. A neural probabilistic model for context based citation recommendation. In: Proceedings of 29th AAAI Conference on Artificial Intelligence, 2015
60 Yu S, Liu J, Yang Z, et al. PAVE: personalized academic venue recommendation exploiting co-publication networks. J Netw Comput Appl, 2018, 104: 38–47
61 Ma X, Wang R. Personalized scientific paper recommendation based on heterogeneous graph representation. IEEE Access, 2019, 7: 79887–79894
62 Liu H, Kong X, Bai X, et al. Context-based collaborative filtering for citation recommendation. IEEE Access, 2015, 3: 1695–1703
63 Jia H F, Saule E. Local is good: a fast citation recommendation approach. In: Proceedings of European Conference on Information Retrieval. Berlin: Springer, 2018. 758–764
64 Xu Z Z, Jiang H Z, Kong X J, et al. Cross-domain item recommendation based on user similarity. Comput Sci Inf Syst, 2016, 13: 359–373

65 Cagliero L, Garza P, Pasini A, et al. Additional reviewer assignment by means of weighted association rules. IEEE Trans Emerg Top Comput, 2018. doi: 10.1109/TETC.2018.2861214

66 Li J, Li D, Feng P H, et al. An expert recommendation model based on the speciality, scientific impact of experts, and social relationship between experts and applicants. J China Soc Sci Tech Inf, 2017, 36: 338–345

67 Wang J, Yue F, Wang G, et al. Expert recommendation in scientific social network based on link prediction. J Intell, 2015, 34: 151–157

68 Yang K H, Kuo T L, Lee H M, et al. A reviewer recommendation system based on collaborative intelligence. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. Washington: IEEE Computer Society, 2009. 564–567

69 Shon H S, Han S H, Kim K A, et al. Proposal reviewer recommendation system based on big data for a national research management institute. J Inf Sci, 2017, 43: 147–158

70 Zhao S, Zhang D, Duan Z, et al. A novel classification method for paper-reviewer recommendation. Scientometrics, 2018, 115: 1293–1313

71 Jin J, Geng Q, Zhao Q, et al. Integrating the trend of research interest for reviewer assignment. In: Proceedings of the 26th International Conference on World Wide Web Companion, 2017. 1233–1241

72 Girvan M, Newman M E J. Community structure in social and biological networks. Proc Natl Acad Sci USA, 2002, 99: 7821–7826

73 Shi X H, Lu H T. Detecting community in scientific collaboration network with bayesian symmetric NMF. Data Anal Knowl Discov, 2017, 1: 49–56

74 Wallace M L, Gingras Y, Duhon R. A new approach for detecting scientific specialties from raw cocitation networks. J Am Soc Inf Sci, 2009, 60: 240–246

75 Ba Z C, Li G, Zhu S W. Similarity measurement of research interests in semantic network. New Tech Library Inf Serv, 2016, 32: 81–90

76 Ding Y. Community detection: topological vs. topical. J Informetrics, 2011, 5: 498–514

77 Kajikawa Y, Yoshikawa J, Takeda Y, et al. Tracking emerging technologies in energy research: toward a roadmap for sustainable energy. Tech Forecast Soc Change, 2008, 75: 771–782

78 Shibata N, Kajikawa Y, Takeda Y, et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation, 2008, 28: 758–775

79 Yang X L. An algorithm based on correlation coefficient to find scientific communities. Dissertation for Master's Degree. Dalian: Dalian University of Technology, 2008

80 Wang X G, Cheng Q K. Analysis on evolution of research topics in a discipline based on NEViewer. J China Soc Sci Tech Inf, 2013, 32: 900–911

81 Miao R, Liu L. Community detection in scientific collaboration network. J China Soc Sci Tech Inf, 2011, 30: 1312–1318

82 Larsen K. Knowledge network hubs and measures of research impact, science structure, and publication output in nanostructured solar cell research. Scientometrics, 2008, 74: 123–142

83 Hirsch J E. An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA, 2005, 102: 16569–16572

84 Braun T, Glänzel W, Schubert A. A Hirsch-type index for journals. Scientometrics, 2006, 69: 169–173

85 Ball P. Index aims for fair ranking of scientists. Nature, 2005, 436: 900

86 Egghe L. Theory and practise of the g-index. Scientometrics, 2006, 69: 131–152

87 Jin B, Liang L, Rousseau R, et al. The R- and AR-indices: complementing the h-index. Chin Sci Bull, 2007, 52: 855–863

88 Burrell Q L. On the h-index, the size of the Hirsch core and Jin's A-index. J Informetrics, 2007, 1: 170–177

89 Zhang F. Evaluation and analysis on academic influences of scholars in library and information field based on papers and funds. Inf Res, 2019, 121–128

90 Cheng H J, Xu W T. Academic influence evaluation of the young talents program. Bull National Natural Sci Found China, 2019, 33: 168–175

91 Haveliwala T H. Topic-sensitive pagerank. In: Proceedings of the 11th International Conference on World Wide Web. New York: ACM, 2002. 517–526

92 Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab, 1999. http://ilpubs.stanford.edu:8090/422/
93 Li X T, Ng M K, Ye Y M. HAR: Hub, authority and relevance scores in multi-relational data for query search. In: Proceedings of the 2012 SIAM International Conference on Data Mining. 2012. 141–152
94 Sinatra R, Wang D, Deville P, et al. Quantifying the evolution of individual scientific impact. Science, 2016, 354: aaf5239
95 Park N, Kan A, Dong X L, et al. Estimating node importance in knowledge graphs using graph neural networks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19). New York: ACM, 2019. 596–606
96 Hicks D. Overview of models of performance-based research funding systems. In: Proceedings of Performance-based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings. Paris: OECD, 2010
97 Zhang S H, Sun S D. Evaluation of high-tech research project based on internal fuzzy TOPSIS and AHP. J Shanghai Jiaotong Univ, 2011, 45: 134–138
98 Hua F. Study of indexes for quantitative evaluation of research performance based on discipline benchmarking. Library Inf Serv, 2014, 58: 78–84
99 Shu Y. Design for the indicator system of scientific research evaluation based on factor analysis and variance maximization model. J Intell, 2015, 34: 33–37
100 Wang D, Song C, Barabási A L. Quantifying long-term scientific impact. Science, 2013, 342: 127–132
101 Haythornthwaite C. Learning and knowledge networks in interdisciplinary collaborations. J Am Soc Inf Sci, 2006, 57: 1079–1092
102 Bordons M, Morillo F, Gómez I. Analysis of cross-disciplinary research through bibliometric tools. In: Handbook of Quantitative Science and Technology Research. Dordrecht: Springer, 2004. 437–456
103 Buter R K, Noyons E C M, van Raan A F J. Searching for converging research using field to field citations. Scientometrics, 2011, 86: 325–338
104 Xu H, Pu W Y, Qian A B, et al. Knowledge mapping of Chinese medicine interdisciplinary research field. Acta Acad Medicinae Sinicae, 2015, 37: 93–100
105 Sun H S. Empirical study on knowledge citations of other disciplines to information science. J Intell, 2013, 32: 113–118+100
106 Wang Q. Measuring interdisciplinarity of a given body of research. In: Proceedings of the 15th International Society of Scientometrics and Informetrics Conference, Istanbul, 2015. 372–383
107 Han Z Q, Liu X P, Kou J J. Interdisciplinary literature discovery based on Rao-Stirling diversity indices: case studies in nanoscience and nanotechnology. Inf Sci, 2020, 38: 116–124
108 Xu H Y, Guo T, Yue Z H, et al. Study on the interdisciplinary topics of information science based in TI index series. J China Soc Sci Tech Inf, 2015, 34: 1067–1078
109 Buter R K, Noyons E C M, van Raan A F J. Identification of converging research areas using publication and citation data. Res Eval, 2010, 19: 19–27
110 Karlovčec M, Mladenić D. Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. Scientometrics, 2015, 102: 433–454
111 Rinia E J, van Leeuwen T N, Bruins E E W, et al. Citation delay in interdisciplinary knowledge exchange. Scientometrics, 2001, 51: 293–309
112 Bromham L, Dinnage R, Hua X. Interdisciplinary research has consistently lower funding success. Nature, 2016, 534: 684–687
113 Tshitoyan V, Dagdelen J, Weston L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 2019, 571: 95–98
114 Chang Y W, Huang M H. A study of the evolution of interdisciplinarity in library and information science: using three bibliometric methods. J Am Soc Inf Sci, 2012, 63: 22–33
115 Palla G, Barabási A L, Vicsek T. Quantifying social group evolution. Nature, 2007, 446: 664–667
116 Zhang F L, Liu J J. Visual analysis on research status and themes evolution of discipline construction research in China. Agr Library Inf, 2019, 31: 31–39
117 Guo X Y, Zhang J J, Liu J. Chinese and foreign research progress and trend evolution of agricultural engineering in

the past 20 years: based on the perspective of bibliometrics and social network analysis. Jiangsu Agr Sci, 2019, 47: 1–9
118 Han M Z, Li Y. Research on the evolution of discipline knowledge based on concept flows. In: Proceedings of the 2nd International Conference on Humanities Education and Social Sciences (ICHESS 2019). Paris: Atlantis Press, 2019
119 Guan P, Wang Y F, Fu Z. Analyzing topic semantic evolution with LDA: case study of lithium ion batteries. Data Anal Knowl Discov, 2019, 3: 61–72
120 Wu L, Liang X H, Song H Y. Empirical study of coevolution analysis based on technological keyword. J Modern Inf, 2019, 39: 137–142
121 Yang C. Collaborator recommendation on research social network platforms. Dissertation for Ph.D. Degree. Hefei: University of Science and Technology of China, 2015
122 Beel J, Gipp B, Langer S, et al. Research-paper recommender systems: a literature survey. Int J Digit Libr, 2016, 17: 305–338
123 Zhang P L. Design and implementation of patent recommendation system based on knowledge graph. Dissertation for Master's Degree. Jinan: Shandong University, 2019
124 Ishag M I M, Park K H, Lee J Y, et al. A pattern-based academic reviewer recommendation combining author-paper and diversity metrics. IEEE Access, 2019, 7: 16460–16475
125 Zhao J M, Qiu J P, Huang K, et al. A new scientometric indicator: review on h index and its applications. Bull National Nat Sci Found China, 2008, 23–32
126 Ping S Q. Research on some issues of community detection in complex networks. Dissertation for Ph.D. Degree. Changchun: Jilin University, 2019
127 Garfield E. Citation indexes for science: a new dimension in documentation through association of ideas. Science, 1955, 122: 108–111
128 Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 3844–3852
129 Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graph. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 1024–1034
130 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. 2016. ArXiv: 1609.02907
131 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. 2017. ArXiv: 1710.10903
132 Ying R, He R, Chen K, et al. Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018. 974–983
133 Klein J T. A conceptual vocabulary of interdisciplinary science. In: Practising Interdisciplinarity. Toronto: University of Toronto Press, 2000. 3–24
134 Yang J L, Sun M J. Mining the information about discipline intercrossing from citation index data. J China Soc Sci Tech Inf, 2004, 23: 672–676
135 Xu H Y, Dong K, Kui L. Research on Interdisciplinary Subject Identification and Prediction Methods. Beijing: Scientific and Technical Documentation Press, 2019. 18
136 Xu H Y, Yin C X, Guo T, et al. Interdisciplinary research review. Library Inf Ser, 2015, 59: 119–127
137 Brillouin L. Science and Information Theory. Massachusetts: Courier Corporation, 2013
138 Lerman R I, Yitzhaki S. A note on the calculation and interpretation of the Gini index. Economics Lett, 1984, 15: 363–368
139 Chen B T, Tsutsui S, Ding Y, et al. Understanding the topic evolution in a scientific domain: an exploratory study for the field of information retrieval. J Informetrics, 2017, 11: 1175–1189
140 Huutoniemi K, Rafols I. Interdisciplinarity in Research Evaluation. Oxford: Oxford University Press, 2016
141 Jang W, Kwon H, Park Y, et al. Predicting the degree of interdisciplinarity in academic fields: the case of nanotechnology. Scientometrics, 2018, 116: 231–254
142 Klein J T. Afterword: the emergent literature on interdisciplinary and transdisciplinary research evaluation. Res Eval, 2006, 15: 75–80
143 Klein J T. Evaluation of interdisciplinary and transdisciplinary research. Am J Preventive Med, 2008, 35: S116–S123
144 Stirling A. A general framework for analysing diversity in science, technology and society. J R Soc Interface, 2007,

4: 707–719
145 Repko A F, Szostak R. Interdisciplinary Research: Process and Theory. Los Angeles: Sage, 2008
146 Andrew B, Born G. Interdisciplinarity: Reconfigurations of the Social and Natural Sciences. New York: Routledge, 2013
147 Rafols I, Leydesdorff L, O'Hare A, et al. How journal rankings can suppress interdisciplinary research: a comparison between Innovation Studies and Business & Management. Res Policy, 2012, 41: 1262–1282
148 Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Trans Knowl Discov Data, 2009, 3: 16
149 Zhao Z Y, Li C, Zhang X J, et al. An incremental method to detect communities in dynamic evolving social networks. Knowl-Based Syst, 2019, 163: 404–415
150 Chen C. CiteSpace: visualizing patterns and trends in scientific literature. Retrieved Jan, 2010, 27: 2010
151 Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, 2009
152 Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst, 2006, 1695: 1–9
153 Hagberg A, Schult D, Swart P. Exploring network structure, dynamics, and function using NetworkX. In: Proceeding of the 7th Python in Science Conference (SciPy 2008), 2008. 11–15
154 Tang J, Zhang J, Yao L M, et al. AMiner: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). New York: ACM, 2008. 990–998
155 Sinha A, Shen Z, Song Y, et al. An overview of microsoft academic service (MAS) and applications. In: Proceedings of the 24th International Conference on World Wide Web (WWW'15 Companion). New York: ACM, 2015. 243–246
156 Zhou Y C, Chang Q L, Du Y. SKS: a platform for big data based scientific knowledge graph. Front Data Comput, 2019, 1: 82–93

A survey on the construction methods and applications of sci-tech big data knowledge graph

Yuanchun ZHOU[1,2], Weijun WANG[1,2], Ziyue QIAO[1,2], Meng XIAO[1,2] & Yi DU[1,2*]

1. *Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190*, China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049*, China*
* Corresponding author. E-mail: duyi@cnic.cn

Abstract Recently, the knowledge graph (KG) of sci-tech and big data technology has played a paramount role in the development of the science of science. We carry out a systematic and in-depth review of KG construction and big data technology application in the sci-tech field. Specifically, we explain the issues of sci-tech entity extraction, sci-tech entity disambiguation, sci-tech relationship extraction, and sci-tech relationship inference involved in the construction of sci-tech big data KG, and give a systematized summary of the analysis and mining methods of the sci-tech big data KG, such as sci-tech entity recommendation, sci-tech community detection, sci-tech entity evaluation, interdisciplinary research, and disciplinary evolution analysis. Lastly, we give the future research and application directions of the sci-tech KG.

Keywords big data of sci-tech, knowledge graph of sci-tech domain, science of science, sci-tech data mining, graph neural network

Yuanchun ZHOU was born in 1975. He received his Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, in 2006. He is a professor, Ph.D. supervisor, and the assistant director of Computer Network Information Center, Chinese Academy of Sciences, as well as the director of the Department of Big Data Technology and Application Development. His research interests include data mining, big data processing, and knowledge graph.

Yi DU was born in 1988. He received his Ph.D. degree from Institute of Software, Chinese Academy of Sciences in 2013. He is an associate professor at the Department of Big Data Technology and Application Development at Computer Network Information Center, Chinese Academy of Sciences. His research interests include big data and visual analytics.

Weijun WANG was born in 1981. He received his M.S. degree from Zhengzhou University. He is currently working toward obtaining a Ph.D. degree at University of Chinese Academy of Sciences. His research interests include knowledge graph and data mining.