DOI:10.13530/j.cnki.jlis.2021006

面向科学社会计算的数据组织与建模方法*

马亚雪　毛　进　李　纲

摘　要　科学系统的动态性与开放性使得针对系统内部海量多源数据的组织与建模难度不断增强,进而提升了科学系统研究的技术壁垒。为延伸科学社会学对科学系统的认知,本文强调科学知识对科学系统运行的影响,将科学视作一个以科学共同体为主体、知识流动为动力的社会系统,探索性地提出科学社会计算的概念,并探究面向科学社会计算的数据组织与建模方法。首先,通过构建三元链路模型,以厘清科学社会内部实体与数据构成,实现对科学社会的抽象化表示;然后,采用多源数据对科学社会中的实体进行数据化表征,并根据三元链路模型进行"实体—属性—关系—数据"的关联,实现科学社会数据建模;最后,以科学社会人际网络和知识扩散研究为例,阐释基于多源数据开展科学社会计算的过程与研究优势。本研究旨在在底层数据与计算技术之间搭建桥梁,为利用多源数据全景化剖析科学系统构成以及开展科学社会计算研究提供方向性的指导。图4。表1。参考文献40。

关键词　科学社会　科学社会计算　数据组织　数据建模　数据融合

分类号　G203

Data Organization and Modeling for the Social Computing of Science

MA Yaxue, MAO Jin & LI Gang

ABSTRACT

Science is a complex, dynamic and open system, which enhances the technical barriers to organize the massive and multi-source data for revealing the operation mechanism of the science system. Extending the definition of the science system in the sociology of science, this study emphasizes the role of scientific knowledge. The science system is regarded as a scientific society where the scientific community works as the main participants, and the diffusion of existing knowledge is the driving force to facilitate scientific creation and production. Based on the idea, the concept of "social computing of science" is proposed to generalize the research on the theories and methods that focus on the complex, changeable phenomena and issues in the scientific society.

To assist the social computing of science, the general framework of data organization and modeling is generated. First, we propose a ternary link model to profile the operating process of the scientific society.

* 本文系国家自然科学基金创新研究群体项目"信息资源管理"(编号:71921002)和国家自然科学基金青年项目"基于学术异质网络表示学习的知识群落发现"(编号:71804135)的研究成果之一。(This article is an outcome of the project "Information Resources Management" (No. 71921002) supported by Science Fund for Creative Research Groups of the National Natural Science Foundation of China and the youth project "Knowledge Community Detection based on Scholarly Heterogeneous Network Embedding" (No. 71804135) supported by National Natural Science Foundation of China.)

通信作者:李纲,Email: imiswhu@aliyun.com,ORCID: 0000-0001-5573-6400 (Correspondence should be addressed to LI Gang, Email: imiswhu@aliyun.com, ORCID: 0000-0001-5573-6400)

The model consists of four types of entities, i. e., the participants, the locations, the activities, and the knowledge that connected the other three components. Through the connection among the entities, the operation mode and potential laws of the scientific society can be more comprehensively reflected and can assist in modeling the scientific society from multiple perspectives. Then, based on the ternary link model, we analyze the path of data organization by summarizing the related data and attributes of the four entities. The extraction of the entities' attributes is regarded as a classification task, based on which the data fusion method of the entities is developed. Finally, from the perspective of the multi-layer network, we explore the relationship among different types of entities, and propose the relationship integration model for the scientific society. Taking the cases on the interpersonal networks analysis and the cross-community knowledge diffusion as examples, the paper illustrates the process and advantages of social computing of science that is based on the fusion of multi-source data.

The originality of this study mainly consists of: 1) A ternary link model is proposed to conceptualize the components and their relationships in the scientific society, which emphasizes the influence of knowledge on other elements. 2) The two research scenarios visualize the concept of social computing of science and provide ideas for the specific application of the proposed data organization and modeling method. 3) This study can build a bridge between the underlying data and computing technology, and provide directional guidance for using multi-source data to comprehensively analyze the composition of scientific society. 4 figs. 1 tab. 40 refs.

KEY WORDS

Scientific society.　Social computing of science.　Data organization.　Data modeling.　Data fusion.

0　引言

科学不是一个“独立变量”,它是嵌套在社会之中、由非常稠密的反馈环与社会相连接的复杂系统[1]。以情报学、科学学为代表的诸多学科从不同角度探究科学系统内部的运行模式以及各类构成要素之间的交互关系[2,3],旨在厘清科学共同体的内部结构,揭示科学知识的增长与演变规律,发现科学知识生产的一般模式,进而为推动科学事业的发展与知识的再生产提供借鉴。

随着科学活动数据可获取性的提升以及数据分析技术的逐步发展,针对科学系统及其构成要素的研究范畴不断拓展。相关工作不再局限于对正式科学交流活动的探索,而是面向更广泛的科学活动场景,从不同维度揭示科学系统结构以及运行规律。其中,有学者从科学评价[4,5]、学术传播[6,7]等角度对学术社交媒体中的科学活动进行研究。另外,相关工作也逐步开始运用机器学习、深度学习等计算方法对科学系统进行分析,以期揭示科学系统中的潜在关系以及内在运行机理[8,9]。然而,不同领域理论与技术方法的交叉借鉴较为困难,同时,科学系统呈现出的动态与开放特征,更是增加了对系统内部海量多源数据组织与建模的难度,提升了科学系统研究的技术壁垒,进而阻碍了相关研究的步伐。为此,本文探索性地提出“科学社会计算”的概念,并探究面向科学社会计算的数据组织与建模方法,旨在在底层数据与计算技术之间搭建桥梁,为利用多源数据全景化剖析科学系统构成以及开展科学社会计算研究提供方向性的指导。

1　科学社会与科学社会计算

1.1　科学社会的内涵

自然科学诞生以来,科研人员对科学本质

的认知大致经历了三个阶段,科学最早被定义为一种系统化、有条理的知识;随着科学事业的发展,科研人员对科学的认知逐渐深入,将其视为一种包含更多内容的社会活动和驾驭自然的力量,并且开始强调科学与技术、生产、经济等现象的联系;之后以默顿及其学派为代表的科学社会学提出科学是一种社会体制(Social Institution),使得科学的社会属性得以突显[10]。

将科学视为一种社会体制开展研究,可以从社会关系、结构和环境的角度揭示科学的起源与发展,然而,这一研究视角却悬置了科学的关键要素——知识[11]。科学共同体与科学知识之间存在"共生关系",科学共同体作为科学知识的创造者与发现者,其所参与的所有科学活动均围绕科学知识展开;相反,科学知识是促进科学共同体形成与发展的动力,以科学出版物为代表的知识载体更被广泛用作科学共同体成员学术影响力的评价依据。从科学共同体或科学知识的单一视角探究科学的变革与发展,均难以全面呈现科学系统的内部结构与运行规律。为此,本文在科学社会学的基础上延伸性地提出"科学社会"的概念,将科学视作一个以推动知识创新与发展为目标、以科学共同体为主体、以科学知识交流与生产为核心活动的复杂社会系统。

科学社会与其他社会系统一样表现出复杂的结构特征以及动态演化的特性,科学共同体成员通过学术交流等科学活动建立联系,使得知识得以在科学社会中流动并加以利用,进而促进科学知识的传播与再生产。在此过程中,由于科学活动的多样性以及构成要素的多粒度特征,科学社会中的关系通常具备多重性与传递性,并且随着衡量粒度的改变,要素间关系的强弱也会随之变换[12,13];另外,科研人员的加入与退出、科学共同体的融合与分裂以及科学知识的进化与衰退,都将导致科学社会内部结构产生变化,使得针对科学社会的研究需面向具体的研究场景构建多元关系模型,以呈现科学社会内部复杂的关系结构与运行模式。

1.2 科学社会计算的兴起

现代信息网络技术的运用极大地提升了社会的运行效能,却也加剧了社会系统规模和运行过程的复杂性、交互性、实时性与数据海量性,使得社会系统中的管理与控制变得异常复杂,从而削弱了传统研究方法的作用。在此背景下,社会计算理论的出现,为有效应对复杂和动态变化的新兴社会问题提供了现代化的方法与手段。王飞跃在总结社会计算实质(即兼具"集成深度计算""群体广度计算"和"历史经验计算")的基础上,认为社会计算的主要动机和目的是"把传统上受限于语言层次和静态的知识,不管是书本上还是社会上、解析型还是经验型、历史的还是现实的,都使之数字化、网络化和动态化,并用于各种复杂社会问题的建模、分析和决策支持"[14];他提出"社会计算旨在在社会问题和计算技术间架起桥梁,从基础理论、实验手段及领域应用等各个层面突破社会科学与计算科学交叉借鉴的困难"[15]。

科学社会作为一个复杂社会系统,以其为对象的研究本质上隶属于社会计算的范畴,特别是在信息网络技术的助力与社会需求的变更下,科学社会正在经历着由传统闭合式科学向高度开放科学范式的根本转型[16],这使得针对科学社会的研究更符合社会计算研究的特征。延续王飞跃等学者对社会计算的认识,本文提出"科学社会计算"这一概念,以概括当前为应对和解析科学社会复杂多变的现象和问题而开展的理论与方法研究。具体来说,科学社会计算指面向科学社会活动、过程、结构、组织及其作用和效应的计算理论与方法,其目标是实现对科学社会的建模与分析,以揭示其内部运行模式与规律。

当前数据可获取性的提升以及分析技术的发展,为科学社会计算营造了良好的研究环境。一方面,诸如 Altmetric. com 等数据平台的构建,使得越来越多的研究开始采用"文献数据+"的模式,在文献题录数据的基础上,不仅引入社交媒体数据[17]、地理空间数据[18]、科研人员简历

数据[19]、交通数据[20]、访谈数据[21]等多源数据,还不断提升所用数据的量级,为实现面向科学社会问题的集成深度计算和历史经验计算提供了丰富的数据支持;另一方面,随着自然科学、计算机科学、社会科学相关学者的加入,来自不同学科的理论与方法被广泛运用到科学社会研究中,使得学科交叉研究逐步深化,实现对面向科学社会的群体广度计算的方法支持[22]。然而,科学社会计算的研究同样离不开理论模型的支持。为此,本文提出面向科学社会计算的数据组织与建模路径,如图1所示。首先,通过构建科学社会概念模型,厘清科学社会内部实体与数据构成,实现对科学社会的抽象化表示;其次,采用多源数据对科学社会中的实体属性进行数据化表征,实现实体层面的数据组织;然后,根据科学社会抽象化表示模型进行"实体—属性—关系—数据"的关联,实现科学社会数据建模;最后,面向具体的研究场景,融合多源数据,实现对科学社会中的现象进行分析,以揭示科学社会运行模式与规律。

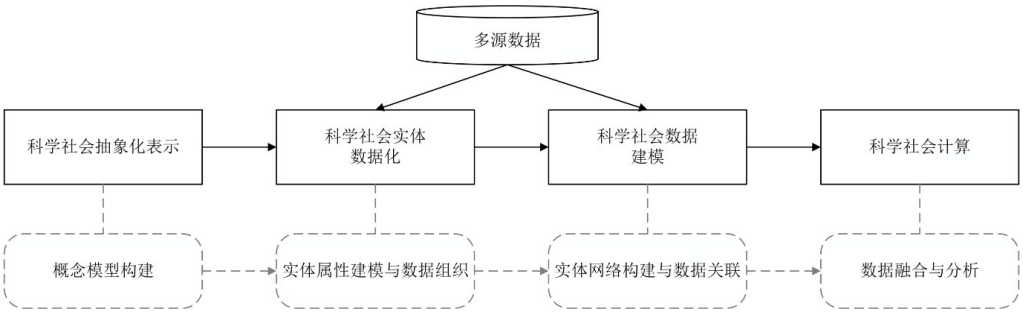

图1　面向科学社会计算的数据组织与建模路径

2　科学社会抽象化表示

科学社会的运行过程可抽象为由知识连接人、地点、活动而形成的三元链路模型(见图2)。人是三元链路的起点,通过在特定地点开展相应的活动,实现知识在科学社会中的循环流动。人、地点、活动与知识是科学社会中具有多种表现形式的四类实体,通过它们之间的关联能够较为全面地反映科学社会的运行模式与潜在规律,实现对科学社会的多角度建模。

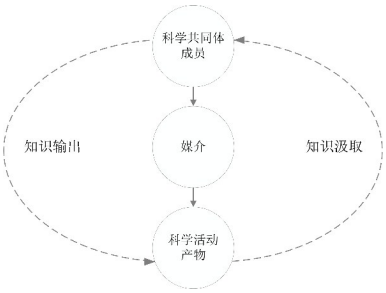

图2　科学社会三元链路模型

人、地点与活动共同构成科学社会运行过程中的主要链路。"人"是科学社会研究的主要对象,既可以泛指各类科学共同体,又可专指特定的科学共同体成员,发挥着推动科学社会运行的作用。其中,科学共同体是由具有共同范式的科研人员所构成的团体,主要包括科学学派、无形学院,科学学会、科学协会、研究会及各类实体研究机构[23,24]。科学社会中的"地点"具有多种表示形式,包含如纸质期刊、电子期刊、社交媒体、会议在内的各类物理与网络空间,科学共同体(成员)将这些地点作为媒介,开展学术出版、科学合作与引用以及学术报告等科学活动(活动),实现知识的输出与汲取。在科学社会的研究中,科学活动通常由论文、社交媒体博文、演讲信息等各类活动产物进行表征,以记录具体的活动类型和内容。

知识是一类特殊的实体,可通过关键词、主题、学科分类、数据集、关键方法、关键理论和领域实体等知识单元表征[25]。知识实体贯穿于科学社会运行的每个环节,构成了其他三类实体的知识属性。通过科学共同体(成员)、媒介以及科学活动产物的关联可以初步得到知识实体间的联系,但知识实体间的层级结构却需结合领域知识从全局的角度获取。

3 科学社会实体数据化

实体数据化旨在通过融合多源数据实现对科学社会中的实体进行数据化表征,进而形成较为完整的结构化数据单元,服务于科学社会数据建模与后续计算工作,其过程包括多源数据汇聚、实体属性结构化以及实体数据融合三个核心任务。

3.1 多源数据汇聚

多源数据汇聚的目标是对科学社会中的数据进行有效组织,以服务科学社会建模与计算。学术文献是获取科学社会中实体属性与关系数据的主要途径,其中,文献题录数据凭借其结构化、易获取的特征更是被广泛用于科学社会的研究。随着开放获取的发展以及数据处理能力的提升,学术文献全文本数据也被逐步运用到相关工作,以期更加准确地测度和评价学术影响力并透视作者的引证动机[26]。另外,学术社交媒体的出现以及学术使用社交媒体人数的增长,使得社交媒体数据同样被用于科学社会研究。Sugimoto 等对八类学术使用社交媒体进行总结,发现它们在科学社会的运行中发挥不同的作用,能够从不同角度为科学社会研究提供数据支持[27]。除此之外,图书、专利文献、新闻报道、政策文件等也能为科学社会研究提供相应的数据。

面向科学社会的多源数据汇聚可通过数据获取、数据精炼和数据特征提取三个核心步骤实现。首先,由于科学社会数据来源与形态的多样,数据获取阶段需秉持渠道多元化、数据互补化、来源可信化的原则,尽可能全面地采集相关数据。其次,根据不同来源的数据类型与质量的差异,在保持数据完整性的前提下,对数据容量进行有效缩减以提升数据质量、实现数据精炼。对于各类数据库中的结构化数据,可根据数据标签之间的语义关联进行数据去重;而对于源自社交媒体等渠道的非结构化数据,则需采用机器处理与人工标注相结合的方法,通过文本匹配实现数据去重,以减少数据冗余。最后,采用多粒度特征提取方法对相关数据进行处理,实现从不同维度解析与透视科学社会数据特征,形成能够直接服务于科学社会计算的基础数据集。

3.2 实体属性结构化

实体属性结构化旨在为科学社会中的实体构建结构化的属性描述框架,实现对各类实体的统一化描述。实体属性描述框架的服务对象是各类实体数据单元,其构建需遵循全面性、针对性与可扩展性的原则,根据各类实体的独特之处从不同角度对实体属性进行描述[28]。科学社会中各类实体常见的属性如表 1 所示。科学共同体(成员)属性描述框架的主体是科学社会中的人或由人组成的机构,可基于科学社会特征利用社交媒体用户画像的方法[29],从基本属性、行为属性和兴趣属性三类维度进行描述;其中,科学共同体(成员)的兴趣属性主要通过知识实体进行描述,用于表征目标实体的研究兴趣。媒介属性描述框架的主体是各类科学活动的场所,无论是物理空间还是网络空间均可采用空间实体属性描述框架中的基本属性和空间属性两个维度进行表征[30];同时,科学社会中的媒介实体还应具备知识属性,以反映在该媒介中可能开展的科学活动的内容。科学活动产物属性描述框架本质上是对科学活动完整事件链的描述,需将科学活动中的人物、地点、时间、内容等因素作为实体属性。知识实体的属性描述框架与前三类实体不同,知识实体由不同层级

的知识单元表征,其属性主要用于描述知识实体隶属的知识领域以及其与领域中其他知识实体的关系。根据四类实体的特征对实体属性描述框架进行逐层细化,可以形成多层次、多维度的描述框架,实现对各类实体的全方位描述。

表 1　实体属性结构化

实体名称	实体属性
科学共同体(成员)	基本属性(自然属性与社会属性)、行为属性(活跃度与影响力)和兴趣属性(长期兴趣与短期兴趣)
媒介	基本属性(自然属性与社会属性)、空间属性、知识属性
科学活动产物	基本属性(自然属性与社会属性)、空间属性、时间属性、知识属性
知识	所属领域、与领域知识实体间的关系(上下位知识实体、同类知识实体等)

3.3　实体数据融合

实体数据融合是将实体数据化的过程,具体做法是基于统一的属性描述框架对实体进行结构化表征后,利用基础数据集中的数据对实体进行建模,形成实体数据单元。在此过程中,如何成功提取并融合能够用于描述实体属性的数据是实现实体数据化的关键。

实体属性数据的提取可视为分类任务,尽管数据汇聚阶段对相关数据进行了初步组织,但仍需根据不同类型的数据特征制定相应的数据提取规则,实现从数据到属性的映射。以文献题录为代表的结构化数据,具有较为清晰的实体划分及“实体—属性”对应关系,可直接采用机器批量处理的方式,抽取原始数据中所涉及的各类实体,实现数据与实体属性的关联;而对于社交媒体博文、新闻报道等各类非结构化数据,由于同一文本中的实体属性归属并不明确,则需结合语义分析技术实现数据到属性的映射。

实体属性数据大多分散于不同来源的数据文本中,通过单一文本提取到的实体属性数据难以形成完整的实体数据单元,需对提取到的实体属性数据进行融合处理,此过程分为两个核心步骤。第一,为保证实体数据单元的唯一性,需对相同的实体进行合并,实现实体层面的融合。具体来说,对于名称相同的实体,可直接通过属性合并实现实体融合;而对于在不同数据文本中名称存在差异的实体,则需先基于实体属性数据间的相似性对实体名称进行消歧,而后根据实体名称进行实体融合。第二,针对融合后的实体数据单元中不同类型的属性数据分别进行数据融合。此过程的核心是将不同来源的属性数据整合到相应的实体属性描述框架中,并对相同类别的属性数据进行数据精炼与特征融合,形成完整的实体数据单元。

4　科学社会数据建模

从网络的视角解析科学社会系统,可将其视为一个由科研人员、项目、论文与思想等实体组成的复杂、自组织且不断发展的网络[31]。科学社会数据建模旨在将独立的实体数据单元连接起来,形成实体关系网络,为科学社会计算提供数据支持。由于三元链路模型能够通过实体的多重组合有效涵盖科学社会中多种类型的实体关系,本文以三元链路模型为基础进行科学社会数据建模,此过程包括实体关系群落的构建和基于多层网络关系的实体关系整合两个核心环节。

4.1　实体关系群落构建

根据三元链路模型构建实体关系群落,有助于初步厘清科学社会中不同类型实体间的关

系,服务于科学社会多层关系网络的构建。实体关系群落的构建是在实体数据单元的基础上提取三元链路关系并实现关系整合的过程。首先,基于实体属性描述框架,采用逆向提取的方式,从科学活动实体出发获取与之相关的人、地点与知识,并通过实体匹配形成完整的三元链路关系。而后,根据实体属性对不同三元链路中的同一实体进行匹配,通过相同实体的合并以及三元链路的关联,实现以特定实体为中心的关系群落的构建。

由于三元链路的形成使得与之相关的知识实体得以固化,实体关系群落的构建仅考虑科学共同体(成员)、媒介以及科学活动产物三类实体。根据整合中心选择的不同,可形成三种类型的实体关系群落,如图 3 所示。图 3(a)是以科学共同体(成员)(P)为中心建立的实体关系群落示意图,该群落整合了以 P 为主体的所有三元链路,可用于表征科学共同体(成员)(P)在科学社会中参与的所有科学活动及其所建立的各类实体关系;图 3(b)和(c)分别是以媒介(L)和科学活动产物(A)为中心的实体关系群落。三种类型的实体关系群落从不同视角对科学社会中的实体关系进行建模,能够适应不同的研究场景,为科学社会数据建模提供可用的实体关系模型。

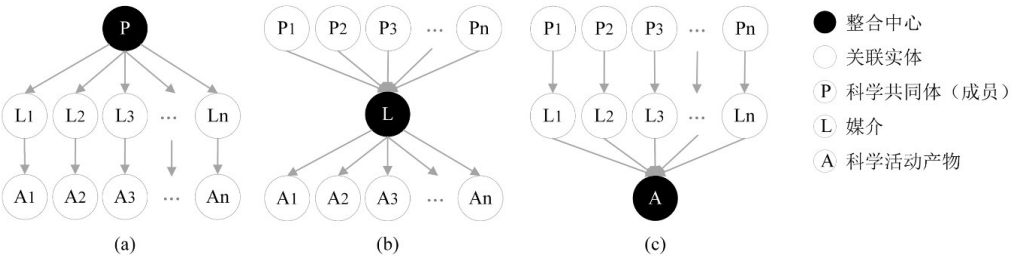

图 3 实体关系群落构建

4.2 实体关系整合

多层关系网络的构建旨在从全局的角度将分散的实体关系群落进行横向连接,以打通科学社会中的“实体—关系—数据”通路,实现科学社会数据建模。实体数据化与实体关系群落构建为科学社会数据建模提供了可用的实体数据单元和以特定实体为核心的纵向关系群落,在此基础上,以科学共同体(成员)、媒介以及科学活动产物三类实体为节点分别构建同类实体关系网络,实现实体数据单元的横向关联,可形成基于三元链路模型的多层关系网络。

多层关系网络本质上是一种元网络模型,它强调网络之间由于目标系统本身运行规律而产生的关联性[32]。网络的构建包括两个核心步骤:第一,以某类实体为中心,根据其属性描述框架获取可用关系类型及数据,构建基础关系整合网络;第二,提取以基础关系整合网络中实体为中心的关系群落,作为其他两类实体关系网络的基本构成单元,根据实体名称对实体关系群落中重复的实体进行去重化处理并构建同类型实体关系网络,以实现实体关系群落横向关联。科学社会中各类实体之间通常具有多重关系,在进行多层关系网络构建时,应尽可能全面地涵盖不同类型的实体关系,以实现科学社会的全景化建模。图 4(a)给出科学社会多层关系网络的示意图,该网络本质上是由不同类型的实体通过横向与纵向连接形成的多模异构网络,能够从不同角度反映科学社会中实体间的关系结构,是科学社会实体关系整合的基础。

科学共同体(成员)、媒介以及科学活动产物均具有知识属性,可采用知识实体集合进行表征。科学社会多层关系网络的构建,使得相

应实体所代表的知识实体集合得以连接,形成以知识实体集合为节点的关系网络并实现知识实体的关联,如图 4(b)所示。知识实体间除了根据科学社会多层关系网络得到应用层面的连接,还可根据知识单元的语义角色建立知识实体层级网络,即科学知识网络,以厘清实体间的功能联系以及知识实体间的上下位关系,实现从专业领域的视角建立知识实体连接,如图 4(c)所示。科学社会多层关系网络以及科学知识网络的构建,能够实现对科学社会中四类实体的全局关系建模以及实体数据的连接,以服务科学社会计算。

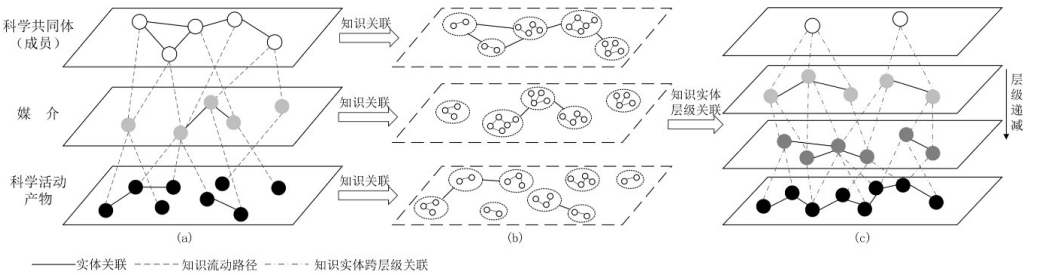

图 4　科学社会关系整合

5　科学社会计算典型实例分析

面向具体的应用场景,在科学社会数据建模的基础上,采用如聚类集成、核融合算法、多关系数据矩阵融合等多种类型的实体关系融合方法,对科学社会中的数据与关系进行融合[31,33],能够有效揭示科学社会中的潜在关系,实现从多角度对科学社会活动、过程、结构、组织及其作用和效应的计算。本文针对科学社会研究领域两个热门研究对象——人际网络和知识扩散,阐述采用多源数据融合的方法进行科学社会计算的优势,以期为科学社会计算相关研究提供参考。

5.1　科学社会人际网络重构

科学共同体成员间的人际网络是科学社会研究的主要对象。学者通过研究科学共同体成员间的共作者网络[34]、引文网络[35]、好友网络[36]等各类人际关系网络,揭示科学社会内部结构及其运行规律。随着机器学习等智能化数据分析方法的引入,相关研究所用数据的量级不断提升,一定程度上实现了针对科学社会中人际关系的大规模计算。然而,鲜有研究通过融合多源数据,从不同角度对科学共同体成员间的人际关系进行建模,以实现对科学社会人际关系网络的重构。

基于多源数据融合的科学社会人际网络重构旨在全方位揭示科学社会中“隐藏”的人际关系,为科学社会内在运行机理及相关规律的发现与阐释提供支持。对于特定科学共同体成员而言,以其为中心建立的基础关系整合网络本质上是由多种人际关系构成的自我中心网络,能够用于反映该实体所具有的多重关系与属性特征。通过面向具体的研究场景,对关系数据进行整合,可从不同角度揭示实体间的关系强度,辅助科学社会人际资本的研究。从全局网络的视角来看,多源数据的融合弥补原有单一数据对科学社会结构呈现的局限,能够从多角度揭示科学社会的人际关系,并厘清科学社会中的团体构成,识别相应的学术团体。例如,社交媒体数据的引入使得科学共同体间的社会关系得以呈现,通过综合实体属性特征并结合共作者数据、地理空间数据等多来源数据,能够较大程度上辅助“无形学院”等科学共同体的识别,并发现跨机构、跨学科甚至跨国家的科学交

流规律。

5.2 知识跨媒介扩散路径解析

知识扩散是知识共享、转移、吸收乃至创新的有机统一,主要包括知识溢出与个体知识创新两个方面[37]。当前针对知识扩散的研究大多集中于对以各类出版物为媒介的学科知识扩散路径与模式的探究,并形成了丰硕的研究成果[38]。随着学术资源在社交媒体中传播数量的增加,已有学者开始对科学知识在社交媒体中的扩散特征与途径开展研究。例如,Alperin 等通过探究 Twitter 中学术论文转发者粉丝网络的构成,发现大多数学术论文主要在单连接的社区中扩散[39]。

信息技术的发展使得知识扩散不再受到媒介的限制,让更多的受众有机会获取并利用相关知识。然而,针对科学知识跨媒介扩散的研究却较为少见,尽管 Filippo 和 Serrano-López 对节能领域学术论文在基金项目、出版物以及社交媒体中的流动进行研究,但其关注的核心是项目或出版物的属性特征对社交媒体扩散能力的影响,而非对完整扩散路径的探究[40]。知识跨媒介扩散研究的对象主要是科学活动产物所代表的知识实体集合。在科学社会数据建模的基础上,快速识别和提取与科学活动产物相关的媒介以及科学共同体(成员),按科学活动开展的事件顺序对扩散媒介进行有向关联,将形成以特定科学活动产物为核心的知识跨媒体扩散路径,可用于研究知识跨媒介扩散的一般规律。在此基础上,进一步关联科学共同体成员,将有助于发现知识跨媒介扩散的守门人与核心推动者。

6 结语

科学作为一个复杂系统,对其内部结构以及运行规律的研究,有助于推动科学事业的发展与知识的再生产。随着计算理论与方法的不断发展以及可用数据的增多,针对科学系统及其构成要素的研究范畴不断拓展,但目前鲜有对此类研究方向的系统性概括,也难以形成统一的研究范式。本文在总结学界对科学本质认知的基础上,将科学视作一个以推动知识创新与发展为目标、以科学共同体为主体、以科学知识交流与生产为核心活动的复杂社会系统,并提出“科学社会计算”的概念以概括当前面向科学社会的各类技术理论与方法。

本文提出从科学社会抽象化表示到科学社会实体数据化,再到科学社会数据建模和基于多源数据融合的科学社会计算的基本路径,实现面向科学社会计算的数据组织与建模。首先,通过构建由知识连接人、地点、活动而形成的三元链路模型,实现了对科学社会运行过程的抽象化表示;然后,基于科学社会三元链路模型,探究科学社会实体数据融合与关系网络构建方法,形成可用于科学社会计算的实体数据单元和多层关系网络,实现对科学社会中数据与关系的有效组织与建模;最终,以科学社会研究领域两个热门研究对象——人际网络和知识扩散为例,阐述采用多源数据融合方法进行科学社会计算的优势。在今后的研究中,可基于本文所提出的数据组织与建模路径,面向具体研究场景开展相应的实证研究,以实现基于多源数据融合的科学社会计算。

参考文献

[1] 殷杰. 当代西方的社会科学哲学研究现状、趋势和意义[J]. 中国社会科学,2006(3):26-38. (Yin Jie. The status quo, trend and significance of contemporary western philosophy of social sciences research[J]. Social Sciences in China, 2006(3):26-38.)

[2] 盛世豪,徐梦周. 科学学学科发展态势及重点研究领域[J]. 科学学研究,2018,36(12):2154-2159.

(Sheng Shihao,Xu Mengzhou. The development trend and key research areas of science of science studies [J]. Studies in Science of Science,2018,36(12):2154-2159.)

[3] 舒文琛,周恩国,李岱峰,等. 基于合著网络社区发现的情报学研究主题演化分析[J]. 情报科学,2020,38(1):75-81. (Shu Wenchen,Zhou Enguo,Li Daifeng,et al. An analysis of the evolution of informatics research themes based on co-authored network community discovery[J]. Information Science,2020,38(1):75-81.)

[4] Thelwall M,Nevill T. Could scientists use Altmetric. com scores to predict longer term citation counts? [J]. Journal of Informetrics,2018,12(1):237-248.

[5] Nuzzolese A G,Ciancarini P,Gangemi A,et al. Do altmetrics work for assessing research quality?[J]. Scientometrics,2019,118(2):539-562.

[6] Schmitt M,Jäschke R. What do computer scientists tweet? Analyzing the link-sharing practice on Twitter[J]. PLoS One,2017,12(6):e0179630.

[7] Hassan S U,Bowman T D,Shabbir M,et al. Influential tweeters in relation to highly cited articles in altmetric big data[J]. Scientometrics,2019,119(1):481-493.

[8] Hassan S U,Imran M,Iqbal S,et al. Deep context of citations using machine-learning models in scholarly full-text articles[J]. Scientometrics,2018,117(3):1645-1662.

[9] Dey R,Roy A,Chakraborty T,et al. Sleeping beauties in computer science:characterization and early identification[J]. Scientometrics,2017,113(3):1645-1663.

[10] 刘珺珺. 科学社会学[M]. 上海:上海科技教育出版社,2009:1-3. (Liu Junjun. The sociology of science [M]. Shanghai:Shanghai Scientific and Technological Education Publishing House,2009:1-3.)

[11] 卢艳君. 默顿科学社会学:当前困境与未来趋向[J]. 科学学研究,2011(2):9-16. (Lu Yanjun. On the current dilemma and future trend of Merton's sociology of science[J]. Studies in Science of Science,2011(2):9-16.)

[12] 王国胤,张清华. 不同知识粒度下粗糙集的不确定性研究[J]. 计算机学报,2008(9):1588-1598. (Wang Guoyin,Zhang Qinghua. Uncertainty of rough sets in different knowledge granularities[J]. Chinese Journal of Computers,2008(9):1588-1598.)

[13] Pedrycz W,Russo B,Succid G. Knowledge transfer in system modeling and its realization through an optimal allocation of information granularity[J]. Applied Soft Computing,2012,12(8):1985-1995.

[14] 王飞跃. 社会计算的意义及其展望[J]. 中国计算机学会通讯,2006,2(2):28-35. (Wang Feiyue. Social computing:background,significance,and methodology[J]. Communications of the CCF,2006,2(2):28-35.)

[15] 王飞跃,曾大军,毛文吉. 社会计算的意义、发展与研究状况[J]. 科研信息化技术与应用,2010,1(2):3-14. (Wang Feiyue,Zeng Dajun,Mao Wenji. Social computing:its significance,development and research status [J]. E-science Technology & Application,2010,1(2):3-14.)

[16] 陈传夫,李秋实. 数据开放获取使科学惠及更广——中国开放科学与科学数据开放获取的进展与前瞻[J]. 信息资源管理学报,2020,10(1):4-13. (Chen Chuanfu,Li Qiushi. Open access to data makes science benefit more individuals:progress and prospect of open science and open access to scientific data in China[J].

Journal of Information Resources Management,2020,10(1):4-13.)

[17] Didegah F,Bowman T D,Holmberg K. On the differences between citations and altmetrics:an investigation of factors driving altmetrics versus citations for finnish articles[J]. Journal of the Association for Information Science and Technology,2018,69(6):832-843.

[18] Hou J H,Yang X C. The spatial-temporal transfer of scientometrics research topics based on citation analysis[J]. Malaysian Journal of Library & Information Science,2018,23(3):49-68.

[19] Aman V. Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates[J]. Scientometrics,2018,117(2):705-720.

[20] 曾轩芩,韩天尧,康乐乐,等.高铁促进了沿线城市之间的学术合作吗?[J].图书情报知识,2019(1):88-98.(Zeng Xuanqin,Han Tianyao,Kang Lele,et al. Does the high-speed rail promote academic cooperation between cities in China?[J]. Documentation,Information & Knowledge,2019(1):88-98.)

[21] Ynalvez M A,Shrum W. International graduate training,digital inequality and professional network structure:an ego-centric social network analysis of knowledge producers at the "Global South"[J]. Scientometrics,76(2):343-368.

[22] Usai A,Pironti M,Mital M,et al. Knowledge discovery out of text data:a systematic review via text mining[J]. Journal of Knowledge Management,2018,22(7):1471-1488.

[23] Mao J,Cao Y,Lu K,et al. Topic scientific community in science:a combined perspective of scientific collaboratio-nand topics[J]. Scientometrics,2017,112(4):851-875.

[24] 夏基松.现代西方哲学[M].上海:上海人民出版社,2009:217.(Xia Jisong. Modern western philosophy[M]. Shanghai:Shanghai People's Publishing House,2009:217.)

[25] 张斌,马费成.科学知识网络中的链路预测研究述评[J].中国图书馆学报,2015,41(3):99-113.(Zhang Bin,Ma Feicheng. A review on link prediction of scientific knowledge network[J]. Journal of Library Science in China,2015,41(3):99-113.)

[26] 赵蓉英,曾宪琴,陈必坤.全文本引文分析——引文分析的新发展[J].图书情报工作,2014,58(9):129-135.(Zhao Rongying,Zeng Xianqin,Chen Bikun. Citation in full-text:the development of citation analysis[J]. Library and Information Service,2014,58(9):129-135.)

[27] Sugimoto C R,Work S,Larivière V,et al. Scholarly use of social media and altmetrics:a review of the literature[J]. Journal of the Association for Information Science and Technology,2017,68(9):2037-2062.

[28] 牛力,蒋菲,曾静怡.面向数字记忆的数字文档资源描述框架构建研究[J].档案学研究,2019(4):40-49.(Niu Li,Jiang Fei,Zeng Jingyi. Research on the construction of digital document resource description framework for digital memory[J]. Archives Science Study,2019(4):40-49.)

[29] 刘海鸥,孙晶晶,苏妍嫄,等.基于用户画像的旅游情境化推荐服务研究[J].情报理论与实践,2018,41(10):87-92.(Liu Haiou,Sun Jingjing,Su Yanyuan,et al. Research on the tourism situational recommendation service based on persona[J]. Information Studies:Theory & Application,2018,41(10):87-92.)

[30] 刘岳峰,杨忠智,孙希龄,等.空间实体的时态属性时间语义特征及代数表达框架[J].武汉大学学报(信

息科学版),2013,38(9):1097-1102. (Liu Yuefeng,Yang Zhongzhi,Sun Xiling,et al. Temporal semantic characteristics of spatial entities' attributes and an algebraic framework[J]. Geomatics and Information Science of Wuhan University,2013,38(9):1097-1102.)

[31] Xu H Y,Yue Z H,Wang C,et al. Multi-source data fusion study in scientometrics[J]. Scientometrics,2017,111(2):773-792.

[32] 李纲,毛进. 元网络视角下科研团队建模及分析[J]. 图书情报工作,2014,58(8):65-72. (Li Gang,Mao Jin. Modeling and analyzing research team in the perspective of meta-network[J]. Library and Information Service,2014,58(8):65-72.)

[33] 苏娜,张志强. 科学计量学中多重关系融合方法研究进展及分析[J]. 情报科学,2010,28(9):1309-1313,1318. (Su Na,Zhang Zhiqiang. On the multiple relation fusion research in scientometrics[J]. Information Science,2010,28(9):1309-1313,1318.)

[34] Abbasi A,Chung K S K,Hossain L. Egocentric analysis of co-authorship network structure,position and performance[J]. Information Processing & Management,2012,48(4):671-679.

[35] White H D. Authors as citers over time[J]. Journal of the American Society for Information Science and Technology,2001,52(2):87-108.

[36] Hong W,Zhao Y. How social networks affect scientific performance:evidence from a national survey of Chinese scientists[J]. Science,Technology,& Human Values,2016,41(2):243-273.

[37] 关鹏,王曰芬,傅柱. 基于多 Agent 系统的科研合作网络知识扩散建模与仿真[J]. 情报学报,2019,38(5):512-524. (Guan Peng,Wang Yuefen,Fu Zhu. Modeling and simulation of knowledge diffusion in scientific collaboration network based on a multi-agent system[J]. Journal of the China Society for Scientific and Technical Information,2019,38(5):512-524.)

[38] 董克,张斌. 学科知识扩散网络路径识别研究进展[J]. 情报理论与实践,2017,40(8):139-144. (Dong Ke,Zhang Bin. Research progress on path recognition of subject knowledge diffusion network[J]. Information Studies:Theory& Application,2017,40(8):139-144.)

[39] Alperin J P,Gomez C J,Haustein S. Identifying diffusion patterns of research articles on Twitter:a case study of online engagement with open access articles[J]. Public Understanding of Science,2019,28(1):2-18.

[40] De Filippo D,Serrano-López A E. From academia to citizenry. Study of the flow of scientific information from projects to scientific journals and social media in the field of "Energy saving"[J]. Journal of Cleaner Production,2018,199:248-256.

马亚雪　武汉大学信息管理学院博士研究生。湖北 武汉 430072。
毛　进　武汉大学信息管理学院副教授。湖北 武汉 430072。
李　纲　武汉大学信息管理学院教授,博士生导师,教育部人文社科重点研究基地武汉大学信息资源研究中心主任。湖北 武汉 430072。

(收稿日期:2020-07-28)