# 基于文献计量的科学知识图谱发展研究

## The Evolution of Mapping Knowledge Domains Based on the Bibliometrical Method

任红娟[1,2] 张志强[1]

(1. 中国科学院国家科学图书馆兰州分馆 兰州 730000; 2. 中国科学院研究生院 北京 100090)

摘 要 利用文献计量方法综合研究科学知识图谱领域发展状况的研究相对比较少的问题,从构建种子数据集,并利用一级引用在进行数据集扩展的基础上,采用作者共被引和共词分析方法对科学知识图谱研究领域的知识结构进行了划分,并利用逐年演化的高频词共现关系分析了该领域的演化趋势和研究重点,从定量和定性结合的角度对科学知识图谱研究领域进行了全面的描述。

关键词 作者共被引 共词分析 科学知识图谱
中图分类号 G310 文献标识码 A 文章编号 1002-1965(2009)12-0086-05

## 0 引 言

科学知识图谱[1],是将传统的文献计量方法与现代的文本挖掘和复杂网络、数学、统计学、计算机科学方法以及可视化技术等有机地整合在一起的一种综合分析科学发展的知识发现方法。它利用一定的方式把抽象数据映射在 2D 或 3D 的图形中,从宏观、中观、微观各个层面来揭示一个领域或者学科、主题发展的概貌,使得人们能够从各个角度全面地审视一个学科的结构和研究热点、重点等信息。由于图形展示方式,非常符合人的认知习惯,而且比起文本信息,人的大脑能够处理更多的图形信息,因此科学知识图谱的方法越来越受到各个领域研究者的关注。

虽然自文献计量学出现以来就有了科学知识图谱发展的雏形[2-4],但其真正的发展是在 20 世纪 90 年代末。十多年来科学知识图谱方法体系在不断的发展和扩充,如今处在一个什么样的发展阶段?主要研究哪些内容?领域的主要领军人物是谁?领域的研究结构如何?这些信息到目前都还不十分明了。虽然国内外的学者对于科学知识图谱的发展都进行了综合的研究[5-7],但主要还是从描述和内容总结的角度出发的。从定量的角度来研究科学知识。

图谱领域的发展,尤其是领域的演化趋势还很少有人涉及。因此,本文主要从计量学的角度对科学知识图谱研究领域的主要参与主体和研究的内容以及研究的动态进行了深入的研究。

## 1 科学知识图谱主要代表人物和主要研究内容分析

1.1 种子数据集的构建和基础分析 选取汤姆森科技的 SCI 和 SSCI 数据库,利用"science mapping" or "mapping science"or"visualiz * knowledge domain *" or "knowledge domain * visualiz *"or "mapping knowledge domain *"为检索词进行主题检索,选取数据库收录的所有年限的数据,检索结果共得到 70 篇文献。去除与该领域不相关的遥感和地理学领域的文献并经过去重处理,共得到 50 篇文献,是该领域的研究成果的集中代表。从这些文献的来源期刊的学科分布来看,主要集中在信息科学和图书馆科学、计算机科学和信息系统学科领域。

从学科的产出分布来看,德莱克斯大学的陈超美是这个领域最高产的作者,论文数量达到了 8 篇,可以说是科学知识图谱领域的领军人物。他不但比较早就开始关注和研究科学知识图谱方法,而且还自行研发了 CITESPACE 一代和二代可视化软件[8],主要用于分析文献、期刊和作者之间的共被引关系。利用 PFNETs、期望值最大化、时间序列等算法,把基于文献的数据转化为多彩的可视化图谱。目前这个软件已经在知识图谱和文献计量领域被广泛采用。

排在第二位的是 Small,他是科学计量和文献计量领域的重要代表人物之一。从 Small 的几个代表作品来看,在知识图谱领域他主要关注大科学的图谱,而不是学科专业知识的展示和揭示,着重宏观知识图谱方

收稿日期: 2009-05-16 修回日期: 2009-07-23

作者简介: 任红娟(1979-),女,博士研究生,研究方向为情报分析、文献计量和战略情报; 张志强(1964-),男,教授,博士生导师,研究方向为战略情报、地球科学、生态经济学。

法的理论和应用研究。排名第三、第四位的 Borner 和 Boyack 也是科学知识图谱研究领域中非常关键的人物。尤其是 Borner, 她进行了很多文献可视化方面的研究。2003 年, 美国科学院组织的"mapping knowledge domains"讨论会, 她就是主要的组织者之一, 而这次会议的召开也揭开了我国科学知识图谱研究的序幕, 大连理工大学的刘则渊教授正是由于捕获到了这次会议的内容才引发了他对于科学知识图谱研究的浓厚兴趣。

从领域研究的主要参与机构来看, 德莱克斯大学、布鲁内尔大学、美国的圣蒂亚国家实验室、荷兰的伊拉兹马斯大学和中国的大连理工大学 WISE 实验室都是研究成果比较丰富的机构。

由于本文获取的数据量相对比较少, 而且很多文献计量的研究领域虽然没有采用知识图谱这个术语但是从事的同样也是这方面的研究, 因为这些数据对于全面了解科学知识图谱研究领域并不充分。但是扩大检索词就会检索到很多不相关的信息, 即使选取更多的关键词也会出现很多漏检和误检的情况, 因此本文从目前科学知识图谱研究方法的融合思想出发, 基于内容和引用结合的思想, 通过引用这些核心文献的引文来进行数据集的拓展, 这样不但避免了大量不相关信息的混入, 也融入了很多由于术语的不同而内容相似的内容, 扩充了分析数据源, 有利于更加完整和全面地把握领域的发展状况。而且利用引文作为数据拓展方法能够把更多前沿的研究内容引入, 是一种比较理想的领域数据集构建方法。

1.2　数据集的扩展和分析　本文利用这 50 篇初始文献作为种子, 把直接引用它们的文献也融入到科学知识图谱发展分析的数据源当中, 也就是只进行一级的拓展, 并没有从引用它们的文献的引用文献再进行拓展。因为网络过大的话又会使得很多不相关的污染信息源融入到分析的数据之中, 不仅增加了分析的复杂度, 而且也会使得得出的结论过于离散, 起不到聚合分析的效果。经过去重处理之后, 共得到 416 篇文献。

1.2.1　高产作者分析。我们利用 HistCite[9] 对这些文献的作者进行分析, 发现产出最高的作者仍是陈超美, 与种子数据源得到的结论是相同的, 这也说明了陈超美在科学知识图谱领域的突出地位, 而且这个数据集的构建是基于引用种子源得到的, 这也在一定程度上印证了陈超美不光产出多, 而且影响力也非常大。而排在第二位的是 Thelwall, 他是网络计量学研究的代表学者之一, 在网络计量领域有着极其重要的地位, 他主要关注文献计量学在网络环境中的方法的拓展。排在第三位的仍是 Borner, 虽然在信息科学和图书馆学领域, 学者的合作, 尤其是基于合著来揭示的合作现象在整个领域来说趋势很不明显, 但是从 Borner 的作品来看, 她与陈超美、Boyack、McCain、Klavans 等知名的可视化和计量学方法的专家都进行过合作, 从该数据集来看, Borner 的作品合作率达到了 84.2%, 而且近年来合作出版的论文数量在逐渐的增多, 虽然不能直接地说明合作促进了 Borner 的研究的产出数量和质量, 但是也可以在一定程度上有所印证。

从产出 TOP10 的作者来看, Garfield 和 Small 具有特殊性, 原因在于他们的作品发表在 20 世纪 80 年代和 90 年代居多, 这是由于科学知识图谱是在文献计量和科学计量的基础上发展起来的一种研究方法, 而 Garfield 和 Small 在科学计量学的泰斗地位在全作者分析当中就被凸显了出来。

从机构产出来看, 排名前 10 名的机构分别为德莱克斯大学、印第安纳大学、伍尔夫汉姆普顿大学、圣蒂亚国家实验室、布鲁奈尔大学、阿姆斯特丹大学、俄克拉荷马州大学、皇家图书馆和信息科学学院、亚利桑那大学、科学信息学会, 这些机构在计量学研究和科学知识图谱研究领域都是世界领先的。

1.2.2　作者共被引分析。这 416 篇文献的参考文献中共有 6419 个作者, 被引参考文献为 12682 篇, 篇均作者人数不足 2, 由此也可以看出科学知识图谱领域的合作程度相对比较低。选取被引频次大于等于 50 的 42 位作者作为分析的对象, 表 1 列出了被引频次排名前 10 名的作者和被引次数。根据这些高被引作者的共被引数据, 构建作者共被引矩阵, 并利用 COSINE 方法对这个共现矩阵进行标准化处理, 利用多维尺度分析方法和凝聚层次聚类方法, 对这个矩阵进行可视化显示, 如图 1 所示。这 42 位作者根据相似程度共分成两个大类, 其中 Shiffrin 自成一类, 这是由于 Shiffrin 本身并不是研究计量学的, 也和科学知识图谱相关的技术研究、应用研究没有交集, 他作为 2003 年美国科学院"mapping knowledge domain"讨论会的组织者, 和 Borner 一起写了《mapping knowledge domains》一文, 这篇综述性的文章提高他的被引频次, 而他本人的研究领域是心理学。对第二大类进一步细分, 分成了 4 个子类。其中 Wise 是一个独立的结点, 他主要从事研究可视化技术。图 1 中最大的一个子类由很多高被引的作者组成, 他们是科学知识图谱、科学计量学研究的主力。其次是由 The Wall、Borgman 等人为代表的一类, 他们主要关注新的科学交流形式的研究。而另外一类是主要是从事复杂网络的特性和相关的技术研究。综合的来看, 从 42 个高被引的研究内容得到科学知识图谱 5 个相关的子领域, 分别研究心理学、科学计量、计量学在新的交流环境下的拓展、可视化技术和复杂网络。由于被引作者有 6000 多个, 而本文选取的数

量仅占 6/1000 左右，虽然总被引次数占到了 18.5%，但这些高影响力作者并不能从整体反映领域的全面的知识结构。而我们将被引频次阈值设为 20，利用社会网络方法进行分析，结果分成了 11 个子类，主要包括文献计量学、网络计量学、可视化技术、社会网络分析、人工智能和人机交互、资源的分类和聚类、统计学、社会结构技术、管理科学，此外还有部分比较分散的研究主题，例如图书馆利用的可视化研究、数字图书馆和物理学等。比起前面的结构划分更加细化，内容也更加丰富。但是，从总体上来看，由于科学知识图谱植根于计量学，所以主要还是围绕计量学的方法和应用拓展方面的研究。

表 1 科学知识图谱领域被引频次前 10 名的作者及被引频次

| 排名 | 作者 | 被引次数 |
|---|---|---|
| 1 | Small H | 349 |
| 2 | Chen CM | 240 |
| 3 | Garfield E | 234 |
| 4 | White HD | 231 |
| 5 | Price DJD | 171 |
| 6 | Borner K | 162 |
| 7 | Leydesdorff LA | 132 |
| 8 | Boyack KW | 126 |
| 9 | McCain KW | 118 |
| 10 | Vanraan AFJ | 103 |

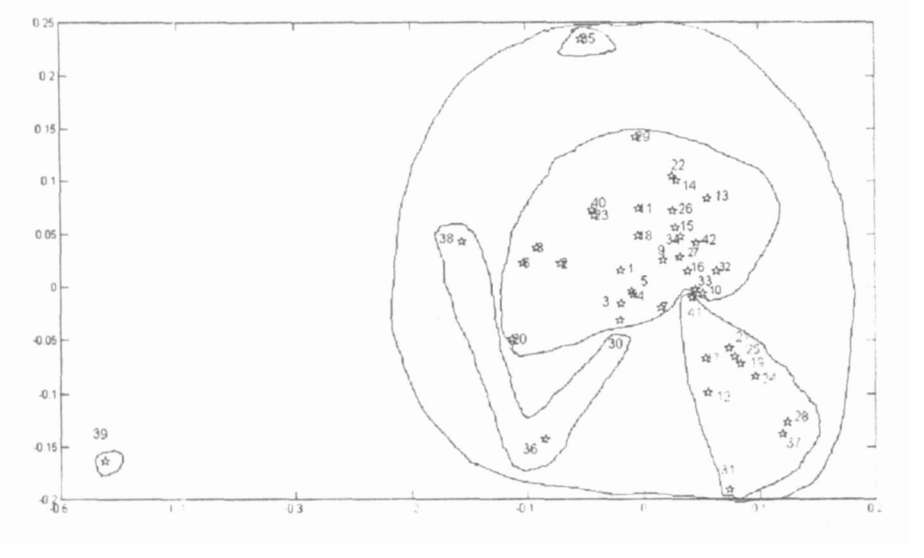

图 1 科学知识图谱领域被引频次大于 50 的 42 位作者关系图

1.2.3 科学知识图谱主要研究内容分析。词是体现文献内容的最小单元，根据高频词就可以在一定程度上了解领域的大体研究内容[10]。本文选择作者关键词作为词频分析对单元，由于作者关键词是作者对研究内容的高度概括，是作者经过慎重考虑所做的选择，而且关键词主要以词组或者短语的形式存在，这些词的逻辑组合，是揭示论文主要内容的很好方式。为了更好地反映科学知识图谱的主要研究内容，本文选择了频次大于 1 的所有关键词，并对关键词进行了同词干词的处理，共得到 158 个关键词，如表 2 所示，是频次大于 5 的 19 个高频词列表。

表 2 科学知识图谱领域频次大于 5 的高频作者关键词列表

| 排名 | 词或短语 | 频次 | 排名 | 词或短语 | 频次 |
|---|---|---|---|---|---|
| 1 | Information Visualization | 22 | 11 | Hyperlinks | 8 |
| 2 | Bibliometrics | 20 | 12 | Knowledge Discovery | 7 |
| 3 | Webometrics | 15 | 13 | Knowledge Visualization | 7 |
| 4 | Citation Analysis | 12 | 14 | Intellectual Structure | 6 |
| 5 | Pfnet | 12 | 15 | Co-citation Analysis | 6 |
| 6 | Scientometrics | 12 | 16 | Co-citation Networks | 6 |
| 7 | Text Mining | 9 | 17 | Citation | 6 |
| 8 | Information Retrieval | 8 | 18 | Network Visualization | 6 |
| 9 | Knowledge Management | 8 | 19 | Factor Analysis | 6 |
| 10 | Mapping Science | 8 | | | |

由表 2 可知，科学知识图谱主要利用文献计量、科学计量以及网络计量的方法，特别是利用引文分析，并利用比较成熟和优秀的可视化算法，例如 PFNETs，对各种抽象的信息和知识利用图形表示的方法展示的一种方法，是一种非常重要的知识发现、知识管理以及信息检索工具，目前主要用于科学文献中的知识发现，用于揭示领域的知识结构和领域的研究重点，利用各种共现方法来形成网络图谱是科学知识图谱最常用的方法。从前 19 个高频词，我们没有看到诸如“mapping knowledge domains” or “visualizing knowledge domain”这类的词，由此可见，虽然没有采用这个术语的很多研究也都属于这个研究的范畴。

根据所选关键词，形成共词矩阵，对这些词利用 TF/IDF 来进行标准化处理，得到共词相似矩阵。由于利用 MDS 方法最多只能处理 100 个对象，因此本文仍采用社会网络分析工具对共词网络进行分析，虽然概念网络不属于社会网络的范畴，但是现在所谓的社会网络分析工具应该被称作复杂网络分析工具，适用于各种对象复杂关系的揭示和分析。得到的图谱如图 2 所示，可以把这些词粗略地分成 9 个连通图，根据每个组件的词，可以看出科学知识图谱的研究分支包括：科学计量学、文献计量学和网络计量学的各种方法在知识图谱中的应用和研究；科学交流理论；图论；认知冲突理论；绩效评价研究；知识和信息的共享研究；社会网络分析、各种领域的应用

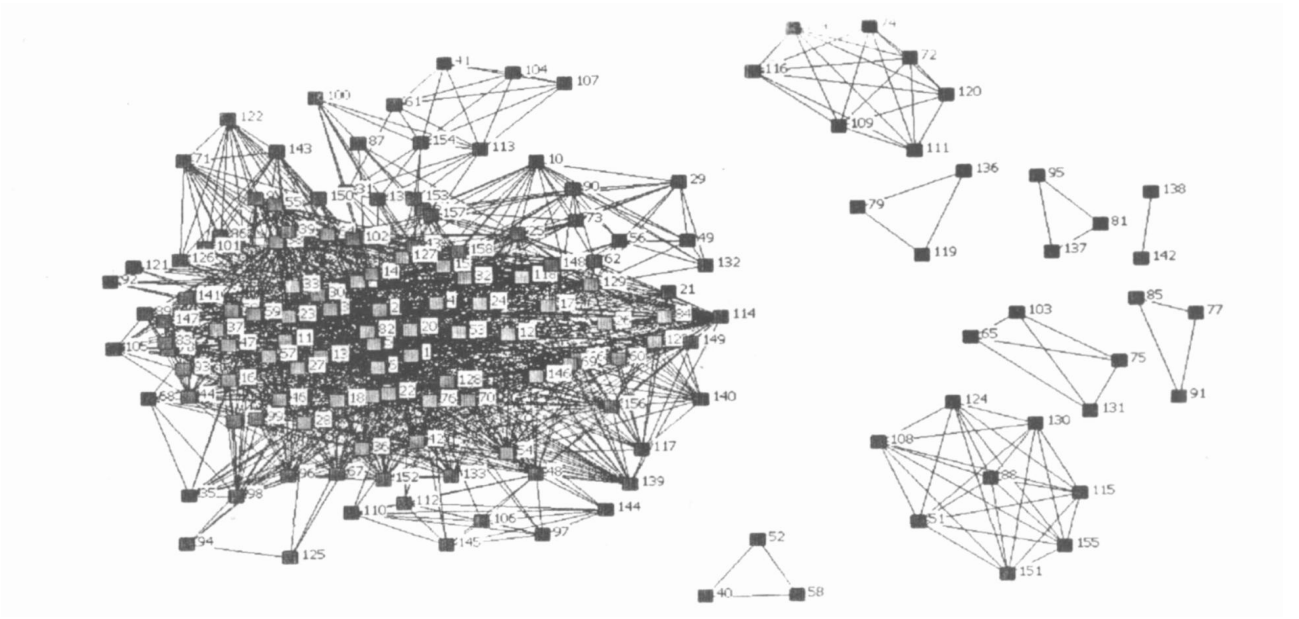

图 2　科学知识图谱研究领域的共词图谱

拓展研究和可视化技术以及各种相似性计算和图演化的方法研究。

从内容分析和引文分析的结果来看, 文献计量学、网络计量学、科学计量学在知识图谱方法中应用, 以及科学交流理论、社会网络分析, 可视化技术研究以及图论是从两个角度都能揭示出来的研究分支, 这在一定程度上说明了科学知识图谱研究并不是文献计量、科学计量方法的简单的可视化方法, 而是有更加宽泛的研究内容, 有许多不同于计量学的特色的内容存在, 是对科学计量学和文献计量学方法的深度拓展。

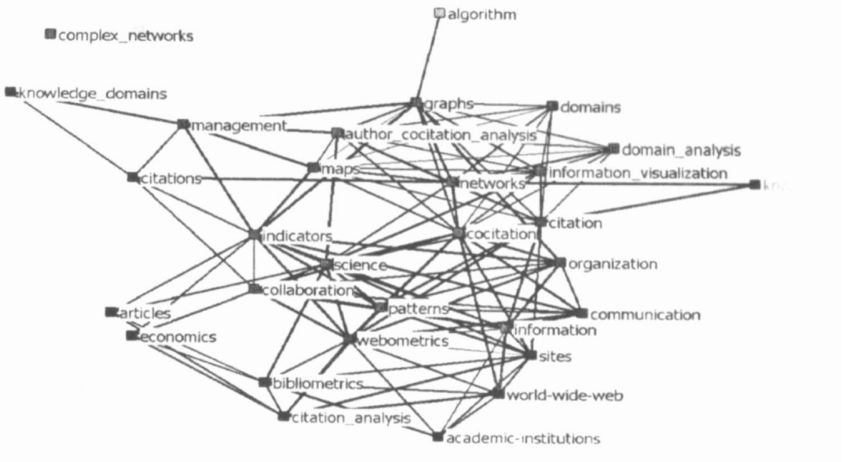


图 3　2006 年 20 个高频词分布图

## 2　科学知识图谱主题演化分析

要了解一个领域的发展现状和发展趋势, 还需要从动态的、演变的观点来考量。本文主要通过词的演化过程来了解科学知识图谱研究领域的发展重点和发展趋势, 研究了 2003 ~ 2008 年近 6 年科学知识图谱的高频关键词以及这些词之间的关系。图 3 ~ 图 5 是 2006 ~ 2008 年前 20 个高频词的关系图, 图中的结点大小表示词的频次, 结点越大表明该词在当年的研究中占有的地位越高, 也就表明了结点代表的研究内容的重要性和突出性。结点之间的线表示词间的共现关系, 与结点连接的线的数量越多表明该词与其它词的联系越紧密。从研究结果来看, 2003 年科学知识图谱研究领域对于网络信息可视化、链接可视化非常关注。2004 年, 在关注网络信息分析的同时, 主要采用利用共被引、作者同被引以及引文分析等方法对许多领域进行的应用拓展分析。2005 年, 领域研究进一步深化, 更加关注可视化的各种算法研究, 共词分析方法运用也在增多, 这也说明很多学者认识到但利用引用揭示学科结构并不完备, 同时复杂网络的分析方法在科学知识图谱研究当中逐步地得以实施。

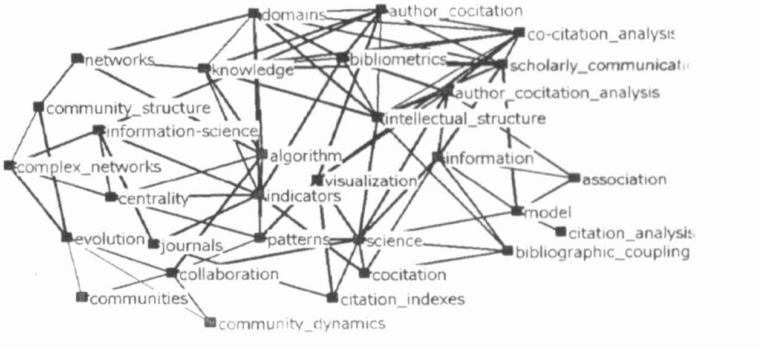


图 4　2007 年 20 个高频词分布图

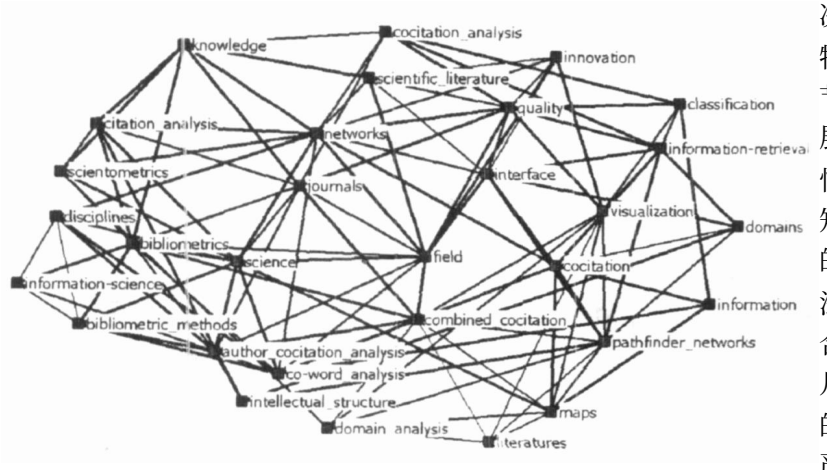


图 5 2008 年高频词分布图

2006 年,相比前几年没有很明显的主题变化趋势,但是对于复杂网络、复杂网络结构的研究不断地加强,这说明科学知识图谱的研究内容在不断的丰富,研究数据也更加复杂。2007 年,比起 2006 年,研究的内容更加多样,关注科学的协作、科学的演化以及科学的知识结构等动态性内容的分析。2008 年,科学知识图谱走向融合的发展方向,把共词、共引以及各种可视化技术、数据挖掘技术融合在一起,注重方法的融合、数据源的融合和数据处理方法的融合,而未来这将是科学知识图谱发展的主流,同时 2008 年的高频词中出现了"创新"一词,这和世界范围内的各种创新思想、创新运动和创新实践是密不可分的,说明科学知识图谱的方法与时代发展的脉搏紧紧相连。

## 3 结 语

本文利用作者共被引和共词分析方法对科学知识图谱领域的结构和研究内容进行了比较详尽的分析。但是对于方法本身的精准性和可靠性没有进行深入的研究,而且共词分析方法只是简单地去除了停用词,没有很好地进行词干的处理以及词权重进一步赋值,因此会对分析的结果准确性有一定程度的影响。在今后的研究中应该更加深入到科学知识图谱方法本身的研究。

科学知识图谱作为文献计量学、科学计量学、网络计量学、可视化技术以及社会网络分析、统计物理学、数据挖掘和人工智能等多个学科方法融合的一个研究领域,无论是研究内容的丰富性,研究方法的广博性还是从应用的广泛性、分析效果的强大性而言,都是一种非常具有发展前景的研究领域。从现阶段的科学知识图谱研究来看,方法的重复应用、数据源选取单一、研究过程缺乏科学性和严谨性都是科学知识图谱研究发展的拦路石,为了能够推动科学知识图谱研究不断地开拓新的局面,就需要有不断探索和创新的精神去解决领域中存在各种各样的不足和难题。特别是对科学知识图谱研究的质量和细节的关注将是推动科学知识图谱更好发展的一个重要研究方向。从对该领域定性和定量的分析,我们可以了解到,科学知识图谱的研究方法的选择具有比较大的随意性,有的作者选取共被引分析方法,有的选择共词方法,有的选择文献耦合方法。对于数据源的选取,从一个或几个数据库获取数据时并没有考证数据的可靠性和代表性,而数据源选取的不严谨,对于分析的结果有很大程度的影响。而在分析的过程对于分析对象选择、阈值的选取、标准化方法以及聚类和可视化方法,甚至结果的分析都带有太多的主观臆断,因此使得整个分析的结果的准确性和可靠性大大折扣。而且一个学科作为一种复杂系统,但从一个角度来进行分析难免会有所偏颇,从对科学知识图谱内容的分析可知,融合的思想在科学知识图谱研究中逐渐地显现,而未来把各种方法有机地结合起来是促进其作为决策和评价依据的主要突破口。总的来说,从科学知识图谱发展的现状和未来的发展趋势来看,提高研究方法的准确性和严谨性,关注方法研究本身的质量控制是未来科学知识图谱一个非常值得关注的研究问题。

## 参 考 文 献

[1] 陈 悦,刘则渊. 悄然兴起的科学知识图谱[J]. 科学学研究,2005,23(2):149—154

[2] D Price. Science since Babylon[M]. Yale University Press,1961

[3] Garfield E, Sher I H, Torpie R J. The Use of Citation Data in Writing the History of Science[M]. Philadelphia: Institute for Scientific Information,1964

[4] Garfield, E. Scientography: Mapping the Tracks of Science[J]. Current Contents: Social & Behavioral Sciences,1994,7(45):5—10

[5] Borner K, Chen C M., Boyack, K W. Visualizing Knowledge Domains[J]. Annual Review of Information Science &Technology,2003(37):179—255

[6] 陈 悦等. 科学知识图谱的发展历程[J]. 科学学研究,2008,26(3):449—460

[7] Morris, S A, Van D V M. B. Mapping Research Specialities[J]. Annual Review of Information Science and Technology,2008(42):213—295

[8] CitespaceII[CP]. [2007—09—30]. http://cluster.cis.drexel.edu/cchen/citespace/

[9] HistCite[CP]. [2009—02—27]. http://www.histcite.com/

[10] 李文兰,杨祖国. 中国情报学期刊论文关键词词频分析[J]. 情报科学,2005,23(1):68—70

(责编:刘影梅)