

ZHISHI WANGLUO YANJIU DE JINZHAN
YU CHUANGXIN

知识网络研究的进展 与创新

 《图书情报工作》杂志社 编

名家视点 第6辑

知识网络研究的进展与创新

图书情报工作杂志社 编

海洋出版社

2015年·北京

图书在版编目 (CIP) 数据

知识网络研究的进展与创新/图书情报工作杂志社编. —北京: 海洋出版社, 2015. 6

(名家视点. 第6辑)

ISBN 978 - 7 - 5027 - 9148 - 3

I. ①知… II. ①图… III. ①互联网络 - 应用 - 知识学 - 文集
IV. ①G302 - 53

中国版本图书馆 CIP 数据核字 (2015) 第 098759 号

责任编辑: 杨海萍

责任印制: 赵麟芬

海洋出版社 出版发行

<http://www.oceanpress.com.cn>

北京市海淀区大慧寺路 8 号 邮编: 100081

北京旺都印务有限公司印刷 新华书店北京发行所经销

2015 年 6 月第 1 版 2015 年 6 月第 1 次印刷

开本: 787 mm × 1092 mm 1/16 印张: 30.5

字数: 526 千字 定价: 48.00 元

发行部: 62132549 邮购部: 68038093 总编室: 62114335

海洋版图书印、装错误可随时退换

序

由《图书情报工作》编辑部编选的《名家视点：图书馆学情报学理论与实践丛书》第6辑即将由海洋出版社出版发行与广大读者见面。这是一件值得高兴的事情。从期刊的角度，这是编者从大量的已经发表的文章中精心挑选出来的专题文章，虽然均在本刊发表过，但以专题的形式集中出版，是期刊内容与论文内容的一种增值，体现期刊价值的再利用；对作者而言，这是另外一种传播途径，增强研究成果再次被阅读、被利用的机会，实现论文再次得到关注和充分利用；对读者而言，通过专辑而阅读到多篇同一专题的文章，可以高效率地了解 and 跟踪该领域的研究进展，深化对该领域的认识，对于开展深度的研究或应用到实践工作奠定良好的基础。

本专辑共有4册。第一册是《机构知识库的建设与服务推广》，共收录32篇文章，涉及到机构知识库从基本概念、政策、技术、应用、服务的各个方面，也基本涵盖了机构知识库建设与服务的各个方面的问题，也是有关机构知识库国内重要作者研究成果的大汇聚。机构知识库作为开放获取的重要内容和学术机构自主知识产权的管理与服务系统，是知识管理的重要体现形式，也是图书馆业务与服务新的增长点，具有良好的发展前景和战略意义。对图书馆而言，开发、管理、维护机构知识库并提供基于机构知识库分析的情报分析与科研布局咨询，对图书馆业务与服务的转型发展具有十分重要的意义。

第二册是《移动图书馆服务的现状与未来》共收录37篇文章，涉及移动互联网用户阅读行为、移动图书馆服务模式、移动图书馆服务质量控制、国外移动图书馆服务实践进展、移动图书馆需求与评估等方面。移动图书馆服务在国内图书馆界研究成果不少，学界和业界也高度认同，但由于收到诸多因素的制约，实践上的发展并不够普及和深入。随着移动互联网技术的发展和相关设施的普及，移动图书馆建设仍然是一个值得重视并加大投入的一个领域，其发展前景将十分广阔。

第三册是《馆藏资源聚合研究与实践进展》共收录32篇文章，涉及馆藏聚合模式、数字资源语义关联、关联数据与本体、协同推荐、知识图谱、面向下一代的知识整合检索等。馆藏资源聚合是一个前沿性命题，也是图书馆从资源建设走向基于资源的挖掘与服务的必然过程。这些方面的研究对于深

度地利用馆藏数字资源，实现馆藏资源价值的最大化，具有十分重要的现实意义和应用前景。

第四册是《知识网络研究的进展与创新》共收录 31 篇文章，涉及科研合作的网络分析、共词分析、主题演化分析、学科知识结构探测、研究热点聚类、科研合作网络等，体现了学界业界对这些领域的最新探索和应用性研究成果。为科研提供深度的前沿热点揭示和发现服务，对图书馆服务能力的提升具有重大的意义。图书馆（特别是大学图书馆和专业图书馆）需要加大这一领域的研究、研发和应用的投入，加快图书馆向知识服务的转变。

虽然本专辑的这 4 本书只是从《图书情报工作》近年来发表的文章精选出来的，但也可基本上代表国内学界对相关问题的最新研究成果和图书馆界实践上的探索与创新，具有学术上的引领和实践上的示范作用。尽管研究者还不够多，研究水平也还有待提升，实践应用也处于探索阶段，但也能显示作者们对这些领域的贡献以及潜在的广泛应用价值。

期待这些研究成果能通过这一专辑的出版，对推动国内的学术和实践产生应有的作用，引起更多的图书情报机构的重视，引发更多的研究人员的后续研究，并不断走向深化。在此也感谢所有作者的智慧和贡献，感谢海洋出版社的倾心出版，感谢编辑部同仁所付出的努力。

初景利

《图书情报工作》杂志社社长、主编

中国科学院文献情报中心教授，博士，博士生导师

2015 年 4 月 23 日 于北京中关村

目 次

方 法 篇

多数据源科研合作网络分析方法及实证研究	王 飒 吕 娜(3)
多词共现分析方法的实现及其在研究热点识别中的应用	高继平等(16)
细粒度语义共词分析方法研究	王忠义(29)
全局视角下的科研领域特色知识点提取	陈 果(47)
基于特色关键词的科研机构研究主题揭示:方法与实证	陈 果(58)
一种基于共词网络社区的科研主题演化分析框架	程齐凯 王晓光(70)
基于形式概念分析的学科知识结构探测	
——以图书情报学为例	刘 萍 吴 琮(80)
科研合作视角下的学科知识流动分析方法研究	
——以药物化学学科为例	徐晓艺 杨立英(114)
优化战略坐标方法在科研选题中的应用研究	
——以那他霉素纳米乳新型纳米乳创制为例	李雅等(133)
典型农业前瞻案例中情景分析法的应用分析	邢颖等(148)
面向情报获取的主题采集工具设计与实现 ...	谷 俊 翁 佳 许 鑫(159)
一种基于时序主题模型的网络热点话题演化分析系统	
.....	廖君华 孙克迎 钟丽霞(177)

应 用 篇

专利与技术创新的关系研究	徐 迎 张 薇(195)
国内外图书情报学认知结构比较研究	张 斌 贾 茜(208)

基于作者文献耦合分析的情报学知识结构研究	宋艳辉 武夷山(226)
学科交叉研究热点聚类分析	
——以国内图书情报学和新闻传播学为例	闵 超 孙建军(241)
运用重叠社群可视化软件 CFinder 分析学科交叉研究主题	
——以情报学和计算机科学为例	李长玲 刘非凡 郭凤娇(258)
基于文献时间特征的学科主题演化分析方法研究	
——以图书情报学领域为例	沈 思 王东波 张 祥 张文博(272)
基于层次概率主题模型的科技文献主题发现及演化	王 平(290)
基于动态 LDA 主题模型的内容主题挖掘与演化	胡吉明 陈 果(307)
维基百科知识分类结构演化规律研究	徐胜国 刘 旭(317)
科学计量学主流研究领域与热点前沿研究 ...	赵蓉英 郭凤娇 赵月华(329)
微群核心用户挖掘的关联规则方法的应用	王和勇 蓝金炯(348)
基于内容分析的用户评论质量的评价与预测	聂 卉(360)

机 构 篇

基于社会网络分析的科研团队发现研究	李 纲 李春雅 李翔(377)
科研机构研究主题的测度	
——以我国情报学领域为例	张发亮 谭宗颖 王燕萍(394)
科研机构对新兴主题的贡献度可视化研究	
——以中美图情科研机构为例	安璐等(407)
跨学科团队协同知识创造中的知识类型和互动过程研究	
——来自重大科技工程创新团队的案例分析	王 馨(421)
元网络视角下科研团队建模及分析	李 纲 毛 进(436)
我国高校科研合作网络的构建与特征分析	
——基于“211”高校的数据	柴 玥 刘 趁 王贤文(452)
基于 h 指数族的科研机构评价及其改进	
——以黑河流域资源环境领域研究为例 ...	韦博洋 王雪梅 张志强(465)

方 法 篇

多数据源科研合作网络分析方法及实证研究^{*}

王飒 吕娜

(北京理工大学图书馆)

1 引言

随着全球化、信息化、网络化的发展,科技、经济、社会发展问题越来越具有复杂性与综合性,科研合作日益成为科学研究的主流方式。广义的科研领域合作关系是指包括大学、科研院所、政府机构和企业在内的所有科研主体之间在科学研究方面的合作关系,涉及计划、项目、实验室、资源共享、办学等方面的合作。狭义的科研领域合作关系是指从事某学科领域研究的机构、人员两类科研主体在科研资源、项目、计划等方面的合作。合作方式包括师生合作、同一学科或跨学科科研工作者之间的合作、研究人员与顾问之间的合作、机构间合作、国际间合作等。合作成果主要体现为文章的发表与专利的申请,因此从文献计量学的角度对作者的合著关系进行分析是对科研合作进行研究的主要方式^[1]。

自2001年M. E. J. Newman首次用社会网络方法对合著关系网络进行分析研究以来^[2-3],社会网络分析方法被广泛应用到合著关系研究当中。随后,国内外有很多学者利用该方法针对某一研究领域进行作者合著网络/科研合作网络的研究。在科研合作网络实证研究中,基于英文文章的分析多采用Web of Science数据^[4-6]或EI数据,基于中文文章的分析多采用CNKI数据^[7-8]或CSSCI数据^[9-10],而对于中国科研工作者来说,单纯基于英文文章或中文文章的合著网络分析都不能很全面地描述科研工作者的科研合作关系。

多数情况下,科学计量学研究的数据来源比较单一,其主要原因在于各

^{*} 本文系2013年国家社会科学基金项目“科研领域合作关系的识别与关联强度分析”(项目编号:13CTQ024)研究成果之一。

数据来源的差异比较大,难以完全达到预期研究目标的质量要求,随着研究的全面和深入开展,多数据源数据整合研究将成为科学计量学今后重要的研究方向之一^[11]。通过多数据源融合来进行科学计量学分析在引文分析^[11-12]、机构评价^[13]、h 指数计算^[14]、信息可视化^[12]、学科分析^[15]等方面都有相关应用,已有研究表明多数据源融合扩大了数据集合,可以提高分析的准确性和全面性。本文将基于多数据源对科研工作者的科研合作关系进行分析,将不同数据源中提取的合作关系进行加权计算,从而弥补单一数据源合作关系不全面的不足。

2 多数据源合著网络加权模式构建

合著网络中点权的大小能够反映作者科研产出的多少以及在关系网络中重要性及贡献的大小,而边权的大小则反映作者间合作强度的强弱,根据不同的数据源可以建立起不同的合著网络,本文将构建多数据源合著网络加权模式,包括来自不同数据源的点权及边权的融合。

2.1 节点点权的加权融合方法

关于合著论文如何分配作者贡献的问题简单的处理方法主要有两种:①正规计数法,即不论合作者有几个,每位作者都算产出一篇论文;②调节计数法,即根据一篇论文合作者的数量赋予每人相等的份额^[16]。已有合作网络可视化中节点的大小一般表示作者发表的论文数量,节点越大说明发表数量越多,而不考虑合著作者人数带来的影响,本文认为同一数据库中的不同论文的价值是固定的,不因作者数量的不同而不同,采用调节计数法将一篇论文的价值分配给其所有作者。

虽然作者署名顺序不同对论文的贡献值也应不同,为了计算方便,这里假设不同作者在同一篇论文中的贡献是相等的,即一篇论文如果由多个作者完成,则每位作者在一篇合作论文中的贡献应为 $1/n$ (n 为该篇论文作者数)。定义 W_v 为节点的点权值,那么节点 i 的点权总和 W_{vi} 可表示为:

$$W_{vi} = \sum_{k=1}^N \frac{1}{n_k} \quad (1)$$

式中 n_k 为第 K 篇论文中合著作者的人数, N 为作者 i 总的文章数,例如某作者总计发表论文 17 篇,其中 1 篇独立完成,11 篇论文为两人合著,5 篇论文为 3 人合著,则该作者的点权值为:

$$W_{vi} = \sum_{k=1}^{17} \frac{1}{n_k} = 1 + 11 * \frac{1}{2} + 5 * \frac{1}{3} = 8.17$$

但如果 N 篇文章来自不同数据库, 公式 (1) 就不能很好地描述作者的贡献值。一般认为, 科研论文的学术价值与其是否被 SCI、EI 检索, 或是否发表于核心期刊等因素有关。本文定义多数据源融合的点权加权表达式为:

$$W_{vi} = \alpha_1 \sum_{k=1}^{N_1} \frac{1}{n_k} + \alpha_2 \sum_{k=1}^{N_2} \frac{1}{n_k} + \alpha_3 \sum_{k=1}^{N_3} \frac{1}{n_k} \quad (2)$$

公式中 α 表示论文被检索及收录情况: α_1 表示被 SCI 或 SSCI 检索, α_2 表示被 EI 检索, α_3 表示发表在中文核心期刊上, N_1 、 N_2 、 N_3 分别表示作者 i 发表的文章被 SCI 或 SSCI、EI 检索及被核心期刊收录的情况。

2.2 边权的加权融合方法

合著网络中两两作者之间的关系强度即两节点之间边权的大小, 取决于合著网络中作者之间的合著次数、合著强度、合著效果等因素。

根据合著次数计算边权, 则两个作者 i 和 j 之间的边权 W_{ej} 计算为:

$$W_{ej} = P \quad (3)$$

其中 P 表示作者 i 和作者 j 合著论文数量。

2001 年, M. E. J. Newman 提出了根据合著次数及合著强度两个因素来计算合著关系网络边权的方法^[3], 假设某篇文献一共有 n 个作者, 那么其中作者 i 就与其他 $n-1$ 个作者进行交流, 作者 i 与其他作者之间的合著强度就分别为 $1/n-1$, 则两个作者 i 和 j 之间的边权 W_{ej} 为:

$$W_{ej} = \sum_{k=1}^P \frac{1}{n_k - 1} \quad (4)$$

式中 n_k 为论文 k 中合著作者人数, P 为作者 i 和作者 j 合著文章数。

2005 年, K. Borner 等人计算合著网络边权时在考虑合著次数、合著强度的基础上, 加入了合著效果因素, 认为不同的合著效果的合著强度也不同, 被引次数多的文献的合著效果大, 则合著强度也大^[17], 作者 i 和作者 j 之间的边权 W_{ej} 计为:

$$W_{ej} = \sum_{k=1}^P \frac{(1 + C_k)}{n_k(n_k - 1)} \quad (5)$$

式中 n_k 为论文 k 中合著作者人数, C_k 为论文 k 的被引频次, P 为作者 i 和作者 j 合著文章数。

同样如果作者 i 和作者 j 的合著论文来自不同数据库, 公式 (3-5) 都不能很好地两个表达两者之间的关系强度, 本文在公式 (3-5) 的基础上, 综合考虑合著次数、合著强度和合著效果 3 个因素, 定义多数据源融合的点权加权表达式为

$$W_{eij} = \alpha_1 \sum_{k=1}^{P_1} \frac{1}{n_k - 1} + \alpha_2 \sum_{k=1}^{P_2} \frac{1}{n_k - 1} + \alpha_3 \sum_{k=1}^{P_3} \frac{1}{n_k - 1} \quad (6)$$

式中 α 表示论文被检索及收录情况： α_1 表示被 SCI 或 SSCI 检索， α_2 表示被 EI 检索， α_3 表示发表在中文核心期刊上， P_1 、 P_2 、 P_3 分别表示作者 i 与作者 j 合作发表的文章被 SCI 和 SSCI、EI 检索及被核心期刊收录的情况。公式 (6) 相对于公式 (5) 更便于操作，也避免了被引频次未归一化带来的影响。

为了防止数据重复造成多次计算加权，当一篇 EI 收录的文章被 SCI 或 SSCI 收录时，按照被 SCI 或 SSCI 收录统计；当一篇核心期刊文章被 EI 收录时，按照被 EI 收录统计。

3 数据分析

3.1 数据采集

本文以“中国科学技术信息研究所”的英文或中文地址为检索词在 Web of Science 平台、EI 数据库、CNKI 平台进行相关检索。选择中国科学技术信息研究所作为案例研究对象的原因是该机构在国内图书情报领域表现突出，并具有一定数量的中英文论文作为研究数据。

以“Inst Sci & Tech Informat China”或“Inst Sci & Technol Informat China”为地址检索词，在 Web of Science 平台 SCI 及 SSCI 数据库检索到该单位 1990 - 2014 年间共被收录论文 112 篇，采集论文全记录。

以“Institute of Scientific and Technical Information of China”或“Inst. of Sci. and Technol. Info. of China”为作者机构检索词，在 EI 数据库检索到该单位 1990 - 2014 年共被收录论文 188 篇，采集论文题目、作者、作者单位、刊名等信息。

以“中国科学技术信息研究所”为单位检索词，在中国知网 (CNKI) 中国学术期刊网络出版总库检索到该单位 1990 - 2014 年间在核心期刊上共发文 1 662 篇，采集数据包括论文题目、作者、作者单位、刊名、关键词、摘要。需要说明的是，CNKI 并没有收录所有的核心期刊，可能造成检索结果比实际发文数量要少。

本文尽可能全面地检索论文，但因地址规范及收录问题，本文数据仅用于科研合作网络的案例分析，其中涉及到单位及个人的数据仅供参考。

3.2 数据处理

编写程序对采集到的数据分别进行结构化及去重处理，对作者发文数量以及作者合作关系信息进行提取。得到 SCI 及 SSCI 收录论文 112 篇，涉及作

者 188 人, 作者合作关系 891 条; EI 收录 (非 SCI 及 SSCI 收录) 论文 165 篇, 涉及作者 227 人, 作者合作关系 611 条; CNKI 核心期刊去除不含作者信息的论文 44 篇, 被 SCI、SSCI 或 EI 收录的论文 22 篇, 共有论文 1 596 篇, 涉及作者 1 027 人, 作者合作关系 3 411 条。

3.3 数据统计与分析

利用公式 (1) 分别基于 SCI、SSCI、EI 和 CNKI 数据源计算得出各个节点的点权值, 按照点权值大小排序, 列出不同数据源点权值排名前 10 位的节点 (见表 1 - 表 3), 并给出相应节点对应作者的发文篇数及节点度值。

可以看出, 基于不同数据源的统计结果存在很大的差异, 说明一个科研单位中成员发表论文的语种、期刊及论文被收录的情况都存在一定的差异。此外, 在同一数据源发文篇数多的作者, 点权值并不一定高, 说明论文数量固然重要, 但合著人数不同对论文的贡献也会不同。需要说明的是, 表 1 - 表 3 给出的是以“中国科学技术信息研究所”的英文或中文地址为检索机构检出的文献集合基础上的统计结果, 不能保证作者所属机构一定为“中国科学技术信息研究所”, 例如表 1 中的 L. Leydesdorff、F. Y. Ye、R. Rousseau, 表 2 中的章成志、穗志方, 表 3 中的化柏林、俞立平所属机构就不为或不单纯为“中国科学技术信息研究所”, 在一定程度上说明了跨机构科研合作对机构论文产出及影响力带来的影响。

表 1 基于 SCI、SSCI 数据点权前 10 位节点的统计信息

编号	作者	作者中文姓名	发文篇数	度	点权值
1	Zhou Ping	周萍	17	7	8.17
2	Wu Yishan	武夷山	14	26	4.74
3	L. Leydesdorff		10	3	4.67
4	Chen Yingjian	陈颖建	3	0	3.00
5	Pan Yuntao	潘云涛	12	28	2.90
6	F. Y. Ye		5	2	2.83
7	Xu Shuo	徐硕	9	19	2.15
8	R. Rousseau		5	6	2.03
9	Zhang Xiaoyu	张晓宇	2	0	2.00
10	Zhao Zhiyun	赵志耘	9	12	1.78

从表 1 可以看出发文篇数多的作者并一定点权值高，例如潘云涛发文篇数较陈颖建多出很多，但潘云涛的点权值较陈颖建反而要低，分析其原因，是因为潘云涛的大部分文章为合作文章且为多人合作，而陈颖建虽然只有 3 篇文章但都为独立完成，该结果也充分验证了公式（1）中作者数量信息对作者贡献值度量的影响。

表 2 基于 EI 数据点权前 10 位节点的统计信息

编号	作者	作者中文姓名	发文篇数	度	点权值
1	Zhang Chengzhi	章成志	17	18	8.28
2	Liu Yao	刘耀	22	24	6.88
3	Zhang Xiaodan	张晓丹	7	9	4.15
3	Wang Huilin	王惠临	14	22	4.15
5	Su Ying	苏颖	12	8	4.00
6	Qiao Xiaodong	乔晓东	13	30	3.69
7	Sui Zhifang	穗志方	12	12	3.27
8	Zhu Lijun	朱礼军	10	16	3.18
9	Peng Jie	彭洁	9	4	3.00
10	Zhang Zhiping	张志平	8	6	3.00

表 3 基于 CNKI 数据点权前 10 位节点的统计信息

编号	作者	发文篇数	度	点权值
1	武夷山	116	65	52.37
2	陈峰	43	11	31.62
3	化柏林	40	10	29.83
4	潘云涛	90	52	29.02
5	王新新	27	2	26.00
6	郑彦宁	67	47	23.95
7	俞立平	55	7	23.58
8	刘娅	33	25	21.92
9	曾建勋	44	33	21.25
10	庞景安	28	15	20.33

对比表1和表2,表1中周萍与表2中章成志的发文篇数一样,点权值相差不大,但节点度值却相差很多,就论文数据作分析,发现周萍共发文17篇,篇均作者人数为2.2人,其中与L. Leydesdorff合作次数最多(10次),仅合作一次的作者有3个;章成志同样发文17篇,篇均作者人数为2.9人,其中与王惠临合作次数最多(6次),仅合作一次的作者有13人,这说明节点度值与点权值并不是反相关的关系。

相较于表1和表2来说,表3中给出的作者发文篇数和点权值都有大幅提高,说明该机构发文大部分在国内核心期刊上,SCI或SSCI及EI发文数量较少,这代表了国内大部分图书情报研究机构的实际情况。表3中王新新发文数量并不多,但因为大部分为独立完成,所以点权值排在了该表第5位。

根据公式(2)对基于多数据源的统计结果进行点权加权,其中 α_1 、 α_2 、 α_3 分别取1、0.5、0.3,给出基于多数据源统计点权值排名前10位的节点(见表4),并给出相应节点对应作者的发文篇数及节点度值。关于 α_1 、 α_2 、 α_3 的取值问题,综合考虑了科研机构工作量考核系数及职称评定方法来初步确定,该系数可以根据需要在实际操作中进行调整。

表4 基于多数据源点权加权前10位节点的统计信息

编号	作者	发文篇数	度	点权值
1	武夷山	135	79	21.41
2	潘云涛	105	69	12.00
3	陈峰	45	11	10.49
4	郑彦宁	71	54	9.30
5	化柏林	40	10	8.95
6	周萍	17	7	8.17
7	王惠临	60	53	7.98
8	王新新	27	2	7.80
9	曾建勋	49	33	7.54
10	赵志耘	46	22	7.29

表4并不是表1-表3的简单叠加,在进行数据融合时一方面需要对作者姓名进行中英文对照,另一方面根据SCI、SSCI、EI及CNKI数据源的不同,赋予的权值也不尽相同。可以发现分别排进表1-表3中前3位的

