



新世纪高等学校规划教材·信息管理核心课系列

北京师范大学优质课程配套教材

网络计量学

肖 明◎主编



Webometrics



北京师范大学出版集团

BEIJING NORMAL UNIVERSITY PUBLISHING GROUP

北京师范大学出版社



新世纪高等学校规划教材·信息管理核心课系列

北京师范大学优质课程配套教材

网络计量学

肖 明◎主编

Webom



北京师范大学出版集团

BEIJING NORMAL UNIVERSITY PUBLISHING GROUP

北京师范大学出版社

图书在版编目 (CIP) 数据

网络计量学/肖明主编. —北京: 北京师范大学出版社, 2017.3
(新世纪高等学校规划教材·信息管理核心课系列)
ISBN 978-7-303-21553-9

I. ①网… II. ①肖… III. ①计算机网络—文献计量学—高等学校—教材 IV. ①G257 ②TP393.07

中国版本图书馆 CIP 数据核字 (2016) 第 282848 号

营 销 中 心 电 话 010-62978190 62979006
北师大出版社科技与经管分社网 www.jswsbook.com
电 子 信 箱 jswsbook@163.com

出版发行: 北京师范大学出版社 www.bnup.com
北京市海淀区新街口外大街 19 号
邮政编码: 100875

印 刷: 北京京师印务有限公司
经 销: 全国新华书店
开 本: 787mm×980mm 1/16
印 张: 20
字 数: 399 千字
版 次: 2017 年 3 月第 1 版
印 次: 2017 年 3 月第 1 次印刷
定 价: 39.80 元

策划编辑: 张自然 陈仕云	责任编辑: 张自然
美术编辑: 刘 超	装帧设计: 刘 超
责任校对: 赵丽杰	责任印制: 赵非非

版权所有 侵权必究

反盗版、反侵权举报电话: 010-62978190

北京读者服务部电话: 010-62979006-8021

外埠邮购电话: 010-62978190

本书如有印装质量问题, 请与出版部联系调换。

印制管理部电话: 010-62979006-8006

内 容 简 介

网络计量学是在文献计量学、社会网络分析、网络科学等相关学科的基础上发展起来的一门重要的新兴学科，也是当前国际学术界最活跃的研究领域之一。本书从理论、方法、专题三个层面构建了网络计量学的内容体系，全面、系统地论述了网络计量学的理论基础、经典规律、分析方法、工具软件、专题研究等内容。本书可作为高等院校信息管理与信息系统、信息资源管理、图书馆学、情报学、档案学、出版发行学、科学学、科技管理等专业的网络计量学及相关课程的教材，也可作为广大信息工作者、图书情报与档案工作人员、科研人员、评价人员以及管理人员的业务参考书。为方便学生自学，本书增设了学习目标和课后思考题；为方便老师教学，本书免费配备了电子教案。

前 言

网络计量学是在文献计量学、社会网络分析、网络科学等相关学科的基础上发展起来的一门重要的新兴学科，也是当前国际学术界最为活跃的研究领域之一。

本书是课题组成员在从事国家社科基金项目“基于语义识别的引文分析理论、方法与应用研究（项目编号：16BTQ073）”、国家社科基金项目“基于多方法融合的中外图书馆学情报学知识图谱实证研究（项目编号：11BTQ019）”、北京师范大学 MOOC 课程建设项目“网络信息计量与评价（项目编号：02200-3122121J1）”研究成果以及在讲授“信息计量学”“网络信息计量与评价”等课程的教学、科研基础上写成的。全书分为 9 章，从理论、方法、专题三个层面构建了网络计量学的内容体系，全面、系统地论述了网络计量学的理论基础、经典规律、分析方法、工具软件、专题研究等内容。

本书主编讲授“网络计量学”课程多年，积累了丰富的课程教学经验，曾获得北京师范大学教学教育成果奖一等奖、优秀教学奖等多项奖励。本书初稿内容已在北京师范大学进行过多轮次的教学试用，受到学生们的欢迎和普遍好评。作为一个多学科交叉的研究领域，本书具有以下特色：①理论联系实际。本书针对若干热点问题进行研讨。部分案例直接改编自本书主编主持完成的多个相关科研项目，能够引领学生了解科研前沿，增强专业学习的兴趣。②学以致用。本书各章均可安排相应的实验实践环节。不仅能使学生熟练掌握网络计量学的基本理论与方法，而且能将理论应用于实践，做到学以致用。

因此，本书可作为高等院校信息管理与信息系统、信息资源管理、图书馆学、情报学、档案学、出版发行学、科学学、科技管理等专业的网络计量学及相关课程的教材，参考学时数为 32 学时。由于各院校、各专业的课程在内容设计、课时要求、教学条件等方面不尽相同，所以建议使用本书的老师们从各自实际出发，在现有教学内容的基础上进行适当增减。本书也可作为广大信息工作者、图书情报与档案工作人员、科研人员、评价人员以及管理人员的业务参考书。

本书由肖明教授组织编写，编写人员主要由北京师范大学、中国科学技术信息研究所、加拿大阿尔伯塔大学、北京科技大学、中央财经大学等单位从事网络计量学或信息计量学教学和科研的相关老师组成，具体分工如下：第 1 章（赵党志、肖明）、第 2 章（佟贺丰、肖明）、第 3 章（袁军鹏、肖明）、第 4 章（刘兰、肖明）、第 5 章（肖明）、第 6 章（袁军鹏、肖明）、第 7 章（杨冠灿、肖明）、第 8 章（刘恩涛、肖明）、第 9 章（邱小花、李国俊）。全书最后由肖明进行统稿和修改。另外，特别感谢王珊珊、石泽顺参与了

本书 PPT 课件的制作与修改、参考文献的查找与校对等工作，他们的耐心和细心为本书增色不少。最后要衷心感谢本书写作过程中所参考的全部参考文献的原作者！

此外，本书能够顺利出版，还应该特别感谢北京师范大学政府管理学院学科建设基金的资助以及北京师范大学出版社张自然等编辑的帮助。

为方便阅读，本书首次尝试插注二维码形式。由于作者水平有限、编写时间紧迫，不足和错误之处在所难免，恳请读者不吝批评指正。

肖 明

2016 年 10 月

目 录

第 1 章 网络计量学引论	1
1.1 网络计量学的产生背景	1
1.1.1 网络化的信息组织与检索	2
1.1.2 网络学术传播系统带来的机遇与挑战	6
1.2 网络计量学的体系结构	10
1.2.1 链接结构分析	10
1.2.2 网络服务器日志分析	12
1.2.3 网络信息计量学分布规律	13
1.2.4 其他研究	13
1.3 网络计量学研究面临的问题	14
思考题	16
第 2 章 网络计量学的理论基础	17
2.1 网络计量学的信息学基础	17
2.1.1 信息论	17
2.1.2 熵、互信息与自信息	18
2.2 网络计量学的社会学基础	20
2.2.1 马太效应	20
2.2.2 人类行为与最省力法则	21
2.2.3 社会网络分析	22
2.3 网络计量学的情报学基础	25
2.3.1 知识网络	26
2.3.2 知识地图	29
2.3.3 信息可视化	32
2.4 网络计量学的网络科学基础	34
2.4.1 Web 理论	34
2.4.2 小世界理论	37
2.4.3 复杂网络	39
思考题	40

第3章 网络计量学的经典规律	41
3.1 信息增长规律	41
3.1.1 指数增长模型	42
3.1.2 逻辑增长模型	45
3.1.3 国内外信息增长规律研究的进展	46
3.1.4 信息增长规律应用举例	49
3.2 信息老化规律	51
3.2.1 信息老化规律概述	51
3.2.2 信息老化规律应用举例	56
3.3 布拉德福定律	58
3.3.1 布拉德福定律的基本内容	58
3.3.2 布拉德福定律的发展与趋势	63
3.3.3 布拉德福定律的前提条件与局限性	66
3.3.4 布拉德福定律应用举例	67
3.4 齐普夫定律	71
3.4.1 齐普夫定律的形成和确立	71
3.4.2 齐普夫定律的基本内容	73
3.4.3 齐普夫定律的前提与局限性	73
3.4.4 齐普夫定律应用举例	74
3.5 洛特卡定律	77
3.5.1 洛特卡定律的基本内容	78
3.5.2 洛特卡定律应用举例	81
思考题	82
第4章 网络计量学的数据来源	83
4.1 传统数据来源	83
4.2 引文数据来源	88
4.2.1 国外引文数据库	88
4.2.2 国内引文数据库	94
4.2.3 我国港台地区引文数据库	98
4.2.4 网络引文数据来源	99
思考题	103
第5章 网络计量学的工具软件	104
5.1 统计分析软件	104
5.1.1 SPSS 软件	105

5.1.2	SAS 软件	118
5.1.3	MATLAB 软件	128
5.1.4	R 语言	132
5.2	信息可视化软件	138
5.2.1	Pajek 软件	138
5.2.2	Ucinet 软件	141
5.2.3	BibExcel 软件	147
5.2.4	HistCite 软件	155
5.2.5	CiteSpace 软件	159
	思考题	169
第 6 章	网络计量学的分析方法	170
6.1	引文分析方法	170
6.1.1	引文分析概述	171
6.1.2	引文分析主要指标	181
6.1.3	引文分析法的应用	184
6.2	网络日志分析	188
6.2.1	网络日志分析概述	188
6.2.2	网络日志分析应用软件	190
6.3	网络链接分析	194
6.3.1	网络链接分析概述	194
6.3.2	网络链接分析的研究方向	196
6.4	文本挖掘	197
6.4.1	文本挖掘概述	198
6.4.2	自动分词原理与算法	199
6.4.3	文本特征提取原理与算法	201
6.4.4	文本分类的基本思想	203
6.4.5	聚类算法的基本思想	204
6.4.6	关联规则的基本思想	206
	思考题	210
第 7 章	网络引文分析	211
7.1	网络引文分析概述	211
7.1.1	网络引文分析含义	211
7.1.2	网络条件下引文分析方法的挑战	212

7.1.3 网络引文分析与传统引文分析的比较.....	214
7.1.4 网络引文的可获得性	216
7.2 网络引文类型	217
7.2.1 P-P 网络引文的研究.....	218
7.2.2 P-W 网络引文的研究	218
7.2.3 W-P 网络引文的研究	220
7.2.4 W-W 网络引文的研究.....	221
7.3 网络引文分析的应用	222
7.3.1 学科信息门户资源采选	222
7.3.2 网络信息资源生命周期研究	223
思考题	224
第 8 章 网络搜索引擎.....	225
8.1 网络搜索引擎概述	225
8.1.1 网络搜索引擎的概念	225
8.1.2 网络搜索引擎的分类	226
8.1.3 网络搜索引擎的发展	227
8.2 网络搜索引擎技术	231
8.2.1 网络搜索引擎的系统架构	231
8.2.2 网络搜索引擎的关键技术	232
思考题	253
第 9 章 知识网络图谱.....	254
9.1 基于引文网络的引文知识图谱	254
9.1.1 引文知识图谱概述	255
9.1.2 引文知识图谱的发展	255
9.1.3 引文知识图谱构建举例	256
9.2 基于语义网络的语义知识图谱	264
9.2.1 语义知识图谱概述	264
9.2.2 语义知识图谱研究框架	265
9.2.3 语义知识图谱举例	271
思考题	276
参考文献	277

第 1 章

网络计量学引论

【学习目标】

1. 了解在互联网环境下网络计量学面临的诸多机会和挑战。
2. 掌握网络计量学的主要研究分支概况。
3. 熟悉当前网络计量学研究面临的主要问题。

网络计量学 (webometrics) 一词是由丹麦学者埃尔蒙德 (T. C. Almind) 和英格沃森 (P. Ingwersen)^① 在 1997 年发表的一篇文章中首次提出的, 它是以信息计量学为核心, 逐步发展成一门交叉学科。

1.1 网络计量学的产生背景

网络计量学是研究测量万维网方方面面的一门学科, 包括网站、网页的组成部分、使用词汇、超链接、搜索引擎的搜索结果等。网络计量学应用信息计量学的理论和方法来研究与万维网有关的现象, 并逐渐成为一门涉及多门学科、快速发展的交叉学科, 对网络计量学的研究已经超越信息计量学的范畴, 对此贡献较大的其他学科包括传播学、统计学和计算机科学。

信息计量学是研究产生于各种社会阶层的各类信息数量特征的一门学科, 其中文献计量学和科学计量学最广为人知。信息计量学具有广泛的应用价值, 其中最重要的应用之一是研究学术传播 (scholarly communication) 的模式, 如学术成果的定量评价、学术创新的历史传播、学术传播的结构分析等, 其他应用包括知识表达、组织与检索。

^①彼得·英格沃森 (Peter Ingwersen) 是丹麦哥本哈根大学 (University of Copenhagen) 皇家图书情报学院教授, 其主要研究兴趣包括: 通信与媒体、定量社会研究、社会科学计算、人文与艺术、网络科学、科学计量学、网络计量学, 曾担任几个欧洲 ESPRIT 项目的专家顾问及多个国际学术会议的组委会主席, 是《信息科学与技术学会会刊》(Journal of the Association for Information Science and Technology, JASIST) 等多家世界著名杂志的编委, 曾获信息科学与技术学会 (Association for Information Science and Technology, ASIST) 杰出教师奖、科学计量学杂志与国际科学计量学与信息计量学学会颁发的普赖斯奖、汤姆森年度杰出人员奖、美国信息科学与技术学会年度研究奖等。

互联网作为一个强大的传播媒介,正在引起信息生产和交换方式的变革。随着各种信息在互联网上的大量涌现,学术传播系统从以学术期刊为中心的模式向网络化新模式发展,信息计量学的理论和方法一方面正在被成功地应用于开发增值信息服务和提高信息组织和检索的效率,另一方面也面临着新的数据和工具给其进一步发展和丰富而带来的机会和挑战。网络计量学可以说是信息计量学在这两个方面发展的集中体现。

1.1.1 网络化的信息组织与检索

由于互联网上的信息一直在飞速而紊乱地增长,探索适当的组织和检索互联网上信息的方式方法一直是而且将继续是非常重要的课题。这方面的努力包括采用引文网络和引文分析等信息计量学的方法和原理进行信息组织与检索。

作者通过引用相关文献以给予所著文献认可已经成为科技创作中的一个传统,引用代表着作者对其撰写的论文与其引用的文献之间关系的一种判断。沿着引文链寻找相关文献已被证明是一种非常有效的文献检索方式,这是因为,这种检索方式是基于引用作者而不是标引人员对文献之间相关性的判断,而作者一般比标引人员对学科及文献内容的理解要深刻得多。另外一个原因是,追寻引用链寻找相关文献可以克服由不同语言及不同的分类和标引系统所造成的障碍,这对于交叉科学领域的研究极有价值。此外,引文分析可以用来评价学者、出版物、学术研究机构等,还可以用于识别科技研究的前沿领域和趋势。因此,引文分析研究的结果可以辅助信息检索和科学与文献管理。引文索引通过对引文链接的有效记录和组织,可以支持大规模的引文分析研究和沿着引文链进行信息检索的方式。

下面简要介绍应用引文网络和引文分析的原理来提供互联网上增值信息服务。

1. 搜索引擎

在网络时代,许多人每天都要使用搜索引擎在互联网上寻找自己所需要的信息。Google 之所以从众多的搜索引擎中脱颖而出,原因之一就在于它的检索机制成功地采用了引文分析原理,把被链接(backlinking)次数作为衡量网页和网站质量的重要指标之一。具体来说,Google 会根据一个网页的“被引次数”(即被链接次数)来决定这个网页的序位,在其他因素相似的情况下,网页的被链接次数越多,其在结果显示时的排序位置就越靠前。使用这样的检索机制,搜索引擎就可以聚焦于高质量以及与用户需求密切相关的信息。

另外,美国的“开放引文项目”(Open Citation Project, 网址为: <http://opcit.eprints.org>)也创建了一个可以根据被引次数来对搜索结果进行排序的搜索引擎(网址为:

<http://www.citebase.org/>), 这个项目的目标是将美国洛斯阿拉莫斯国家实验室^①的预印本文献库中的每一篇论文与其所引用的该文献库中所有其他文献链接起来, 而这个具有按被引次数进行排序功能的搜索引擎就是该项目的一部分。

2. 引文链接

电子出版的功能远不止提供印刷出版物的电子复印件。新兴的信息技术支持许多传统出版物所不可能具备的新特性, 这些特性可以增加所出版内容和所提供服务的价值。添加价值的有效方法之一就是在参考文献与其所代表的文献的全文信息之间建立可以点击的超链接。这样, 读者就可以在阅读网上文章时自然而有效地追寻这些链接, 在相互关联的文章之间随意浏览选择。引文互联项目(CrossRef, 网址为: <http://www.crossref.org>)、开放URL(Open URL, 网址为: http://www.exlibrisgroup.com/sfx_openurl.htm)以及前面提到的“开放引文项目”都支持引文链接。

引文互联项目是一项出版社之间的合作引文链接服务, 它通过出版社之间的合作, 实现在单枪匹马的情况下不可能实现的一些功能, 进而把用户与其感兴趣的信息内容有效地连接起来。有了这项服务, 用户可以在自己单位已订阅的出版物的范围内, 通过简单地点击引文超链接, 直接从一个出版社的信息内容过渡到被其所引用的另一个出版社的信息内容。迄今为止, 已经有上百万份学术和专业出版物被收入引文互联项目中。

与引文互联项目不同, 开放引文项目不要求参加合作的出版商为其出版物提供源数据并为其出版物中的参考文献建立链接, 而是采用专门的计算机程序自动从各种文献形式的全文文献中为引用及被引用文献抽取源数据, 并为参考文献建立可以点击的链接。这样的服务的确已向实现“把每一个曾经发表的学术著作与其所引用的所有文献或所有那些引用它的文献链接起来”的梦想前进了一步。

在这些先驱者项目的促进下, 网络环境中的学术文献目前已在很大程度上相互链接起来。例如, 当用户通过图书馆的检索系统寻找科研论文时, 就可以通过简单地点击每一条目来获取相应的论文全文, 但前提是该图书馆订购了发表这些论文的期刊; 当用户进入一个学术期刊的网站找到一篇所需要的论文时, 就可以简单地点击文后的参考文献来获取相应的论文全文, 而不用考虑这些论文是否发表在同一期刊上, 但前提是该期刊属于引文互联项目。也就是说, 不仅科研论文中的参考文献与其相应的全文论文已被链接起来, 二次文献(如图书馆目录、图书馆购买的文献数据库)中的条目也已与相应的全文论文链接起来, 这就为科研、教学人员带来了极大的方便。

^①洛斯阿拉莫斯国家实验室(Los Alamos National Laboratory)简称阿拉莫斯实验室(LANL), 以前隶属于美国能源部(DOE), 2007年变更为由洛斯阿拉莫斯国家安全机构(由加州大学等机构共同构成)主管, 位于美国新墨西哥州的洛斯阿拉莫斯。洛斯阿拉莫斯国家实验室是世界上最大的多功能实验室之一, 于1943年开始秘密运转, 当时的第一个任务是“曼哈顿”计划。

3. 引文标引

引文链接使得在互联网上沿着引文链从目前的文章追寻到其引用的早期文章的信息检索成为可能,若要沿着引文链既能从目前的文章追寻到其引用的早期文章,又能从早期文章追寻到后期引用这些文章的文献,引文索引就是必不可少的工具。美国科学信息研究所(Institute for Scientific Information, 简称为 ISI, 现名为 Thomson Scientific) 出版发行的引文索引曾一度是唯一的引文索引。随着互联网在学术传播系统的应用及学术信息电子化的发展,新兴的引文索引不断涌现(如 CiteSeer、CiteBase、Scopus、Google Scholar 等就是其中的典型代表)。

CiteSeer (<http://citeseer.ist.psu.edu>) 由美国 NEC 集团公司的一个研究所创建,现由该所和美国 Pennsylvania 州立大学的信息科学与技术学院(Pennsylvania State University's School of Information Sciences and Technology) 共同维护。该索引基于互联网,结构上类似于 ISI 的引文索引,它自动在网上收集各种类型、对公众开放的计算机科学领域的学术文献,对收集到的文献进行引文标引。尽管该索引与 ISI 的引文索引相类似,却又有很大的不同之处,主要表现为:①ISI 的引文索引主要收录期刊文献,而 CiteSeer 所收录的文献使用互联网作为传播媒体,这个媒体要比期刊的传播速度快,信息流通范围广。②CiteSeer 覆盖各种各样的文献类型(如学位论文、科技报告、会议论文、预印本、期刊文献等),这些文献类型有可能代表着学术传播过程的不同阶段。③CiteSeer 比 ISI 引文索引提供更多的有关被引用文献的信息。例如,ISI 的索引只标引第一著者,只提供发表期刊的缩写名称及发表日期和卷页;CiteSeer 则提供全部著者的名称、被引文章的全文标题、发表期刊的全称及发表日期和卷页。此外,有些文章的全文以及作者的网页也可以获得。④CiteSeer 的制作和维护高度自动化,而 ISI 的引文索引则建立在人工处理的基础之上。⑤CiteSeer 提供引用文献的全文信息,而 ISI 的引文索引只提供引用文献的文摘。⑥CiteSeer 对论文后的参考文献按其自身被引用次数进行排序,但不总是提供一篇论文后的全部参考文献。相比之下,ISI 的引文索引则提供全部参考文献,但不进行任何排序。⑦除了提供 ISI 引文索引所包括的种种信息之外,CiteSeer 还提供许多 ISI 引文索引所没有或根本不可能提供的很有价值的信息。例如,CiteSeer 提供引文环境(citation context)、作者的网页、文章的被引用次数按年代的分布图、基于各种标准(如被共引、被共同浏览、来自相同网页等)的相关论文等信息。

由此可见,CiteSeer 不仅是一个有价值而独特的互联网上的学术文献检索工具,还是一个引文分析的有用数据源。与 ISI 的引文索引相比,CiteSeer 支持更丰富的研究课题。

CiteBase(网址为:<http://www.citebase.org/>)是前面提到的美国开放引文项目所开发和建立的引文索引,这个项目自动从美国洛斯阿拉莫斯国家实验室的预印本文献库的全文文献中抽取引用和被引用数据,通过引文将文献链接起来,并在此基础上建立了引文

索引,支持按文献引用和被引用关系检索文献和排列结果的信息组织与检索方式。

Scopus 是由闻名世界的出版商 Elsevier 开发和生产的一种引文索引。它像 ISI 的引文索引一样只收录期刊文献,主要侧重于自然科学和生物医药科学领域中 1996 年以后出版的论文,但对这些领域中近年来出版的论文的覆盖率接近甚至超过 ISI 的引文索引。它的特别之处在于:提供全文引用文献以及比较完全的有关被引文献的信息,包括论文题名及前七位和最后一位作者的姓名。这些信息对引文分析具有非常重要的价值。例如,使用这个索引来收集引文数据,引文率和共引率的计算就可以不像使用 ISI 的引文索引那样局限于只考虑第一著者,这样的引文分析结果因此也有可能会更公正、更明朗一些。遗憾的是,由于使用 Scopus 收集引文数据的困难性(例如,Scopus 不支持大批量下载检索结果),至今还没有出现基于这个引文索引的大规模引文分析研究。

Google Scholar (网址为: <http://scholar.google.com>) 是 Google 的最新发展之一,专门用于搜索学术论文。除了具有 Google 作为搜索引擎的优势以外,Google Scholar 还提供命中文献的被引次数以及引用每一篇命中文献的所有文献等信息。因此,Google Scholar 成为一个利用引文链辅助学术文献检索的有效工具,并且为引文分析研究提供了有价值的工具。研究显示,根据 Google Scholar 提供的被引次数对科研论文进行排序的结果与根据 ISI 引文索引提供的相应数据的结果非常接近,如果只考虑近期的出版物,连被引次数的绝对值都非常接近。

ISI 引文索引虽然奉行其创始人尤金·加菲尔德(Eugene Garfield)的“文献集中法则”(Garfield's Law of Concentration),不可能轻易改变其覆盖面及收录期刊的原则,但它们在新的网络环境中却也发生了一些变化。例如,这些索引目前的万维网版本即是“科学网”(Web of Science),若被引文献恰巧是这些索引的源文献,这些被引条目就可以通过超链接与源文献记录相连。这样,就可以获得这些被引文献的全面信息(例如,所有作者而不只是第一著者的信息)。

4. 信息检索系统

引文索引在学术传播中具有两个重要作用:①独特的信息检索系统;②应用引文分析的方法研究科学技术传播规律的数据源。如前所述,引文分析的结果不仅有助于理解学术传播的结构和过程,而且可以辅助信息检索。例如,评价性引文分析(evaluative citation analysis)的结果有助于检索出高质量的文献及核心著者、机构或国家所发表的文献,而关联性引文分析(relational citation analysis)所产生的论文、作者或研究领域的聚类网络图示结果可以帮助扩展检索提问。ISI 的引文索引及 CiteSeer 等其他引文索引已经通过提供诸如每篇文章的被引次数、共引文献及期刊质量的各种指标等数据,展示出将引文分析结果应用于信息检索系统所具有的独特价值。

所谓“基于信息计量学的信息检索系统”,即寻求充分利用文献计量学的技术和研究

结果,以帮助解决目前信息检索系统所面临的一些问题,这些系统常常使用可视化(visualization)技术动态地显示由某种词汇分析(如共词分析)所产生的概念网络或由共引分析所产生的论文或作者网络。这些网络可以帮助用户获得某一学科领域的概貌和理解概念、论文或作者之间的相互关系,因此有助于检索提问的扩展和检索的精细化。然而,结合评价性和关联性信息计量学分析结果辅助检索的、“充分基于信息计量学”的信息检索系统至今还不存在。

尽管信息计量学特别是引文分析结果在信息检索中的应用价值并不难理解,但对基于文献计量学的信息检索系统的研究大多集中在词汇分析和概念网络上。信息计量学中最广为人知的引文分析技术在信息检索系统中的应用还没有被充分地探讨,这在很大程度上是因为绝大多数信息检索系统收录的信息很久以来根本不包含有关引用和被引用之间关系的信息。随着科研论文全文信息(包括参考文献)越来越多地涌现出来,以及全自动引文标引工具的日益增长、完善,这种由数据源所带来的障碍即将消失,对基于文献计量学的信息检索系统的研究也会大量增长,甚至有可能成为信息检索系统研究的一个新的焦点。林夏、怀特(White)^①及布济德洛斯基(Buzydlowski)等人已经在这方面做出了有益尝试。他们使用 ISI 艺术与人文科学引文索引(Arts & Humanities Citation Index)1988—1997 年的数据,探讨了著者共引分析结果对文献检索的价值。使用他们的实验系统,输入著者名称进行检索就可以获得一幅与该著者共引次数最高的 25 位著者的共引关系图,用户可以用鼠标将其中任何著者的名称拖入检索词输入区,就可以获得该著者的所有文献。

1.1.2 网络学术传播系统带来的机遇与挑战

传播(communication)在科学研究中的重要作用曾经吸引了从科学社会学家、传播研究人员、科学史学家到信息学家等各类研究人员的广泛注意。他们从不同角度对学术传播进行了研究,其中很多研究使用了引文分析这个信息计量学中众所周知的技术。

信息技术的加速发展,尤其是互联网的快速成长,正在改变学术传播的结构和过程,人们对学术传播的研究兴趣也再次萌发和增长,希望能够了解正在形成的传播形式以及为人们原先所熟悉的以学术期刊为中心的学术传播模式的相似和不同之处。引文分析和其他信息计量学的理论和方法再次被成功地应用于研究这些学术传播中的新现象。

^①霍华德·怀特(Howard D. White, 1936—)是美国著名图书馆学情报学家,现为德雷赛尔大学(Drexel University)信息科学与技术学院(College of Information Science and Technology)名誉教授,其主要研究领域聚焦在信息计量学和科学计量学上,研究兴趣包括文献计量学、共引分析、参考咨询服务评价、参考专家系统、在线搜索等,1993 年获得 ASIST 研究奖,1998 年他和 Katherine McCain 共同获得 JASIST 最佳论文奖,2005 年他因在科学定量研究方面的贡献而获得著名的普赖斯奖。

1. 网络环境中的学术传播

以学术期刊为中心的学术传播系统已有几百年的历史, 学术期刊不仅是学术传播的有效媒介, 还是评价学术贡献、确立科学发现的有用资料。随着互联网的广泛应用, 大部分现有的学术期刊已经开始出版电子版, 新兴的纯电子版学术期刊如雨后春笋般不断涌现。

与此同时, 一些全新的学术传播模式也在不断出现, 开放存取 (open access) 学术期刊和机构数字典藏 (institutional repository) 就是其中的两个例子。开放存取学术期刊与普通学术期刊本质相类似, 唯一的不同之处在于这些期刊免费供所有用户使用。机构数字典藏用于存储由某个或某些机构所拥有的信息资源, 包括教职员工的科研论文、教学资料以及学生的学位论文等。这些资源无须经过漫长的同行评价过程或昂贵的出版发行渠道, 有助于最新科研成果快速与广大用户见面, 从而极大地提高学术传播的效率。

此外, 越来越多的科研和教学人员往往将最新的资料和成果公开发表在自己或工作机构的网站上, 并把参与网上学术交流活动 (如学术邮件列表 listserv) 作为生活和工作不可或缺的一部分。越来越多的资源数字化 (digitization) 项目正在系统化地将许多科技和历史资料全文加以组织整理, 有些项目还供用户免费使用。

可以这么说, 互联网上的学术信息资源已经而且还在继续迅猛地增长, 在网上寻找科研信息已经相当普遍。有些领域 (如高能物理和计算机科学) 的科研人员甚至将互联网作为寻找最新研究成果信息的首选资源, 以便保持与其他科研单位的同行或竞争对手齐头并进的势头。

由此可见, 学术传播系统正在从以学术期刊为中心的模式向更适用于网络环境的新模式发展, 这给信息计量学带来了极大的机会和挑战。

2. 研究学术传播系统发展的新模式

如前文所述, 以学术期刊为中心的学术传播系统已有百余年的历史, 从这一模式向更适用于网络环境的新模式的发展并非易事。例如, 机构数字典藏的优势众所周知, 但由于科研人员一般不会主动地将自己的论文和数据存放到这些资源库中, 现有的数字典藏库存量相当有限, 以至于无法展现其潜在的巨大优势。因此, 研究怎样成功实现机构数字典藏已经成为一个非常重要的研究课题。

应用信息计量学研究学术传播系统从以学术期刊为中心的模式向网络化新模式的发展可以包括这样一些课题: 比较新兴传播模式与传统模式的联系与区别、研究基于互联网的学术传播与以期刊为中心的学术传播之间的信息流通、探索如何将信息计量学为基础的科研评估从期刊扩展到机构数字典藏等。