·会议综述·

全文本计量分析理论与技术的新进展与新探索

——2019 全文本文献计量分析学术沙龙综述

报告人：章成志[1]　胡志刚[2]　徐　硕[3]　汪雪锋[4]　师庆辉[5]　王　巍[6]

(1.南京理工大学经济管理学院,南京,210094；　2.大连理工大学科学学与科技管理研究所,大连,116024；
3.北京工业大学经济与管理学院,北京,100124；　4.北京理工大学管理与经济学院,北京,100081；
5.同方知网(北京)技术有限公司,北京,100084；　6.励德爱思唯尔信息技术(北京)有限公司,北京,100738)

综述整理：钱佳佳　罗卓然

(1.武汉大学信息管理学院,武汉,430072；　2.武汉大学信息检索与知识挖掘研究所,武汉,430072)

[摘　要]　在 2019 年 11 月 10—12 日于成都召开的“2019 科学计量与科技评价天府国际论坛”上,由章成志等人发起的第二届“全文本文献计量分析”学术沙龙,吸引了百余位专家学者的参与和交流,给参会者留下了深刻的印象。本文通过对沙龙嘉宾的发言和讨论内容进行梳理与总结,将沙龙的主要内容归纳为基于引用行为的学术评价体系、全文本的实体抽取、学术文本关键词自动抽取、新兴研究话题和新兴技术预测、以及全文本数据开放、构建与应用等五个主题,以期揭示国内外全文本文献计量分析在理论与技术方面的最新进展以及发展趋势。

[关键词]　全文本　文献计量　引用行为　情感分析　关键词抽取　实体抽取　新兴话题预测　数据开放

[中图分类号] G350　[文献标识码] A　[文章编号] 2095-2171(2020)01-0111-07
DOI:10.13365/j.jirm.2020.01.111

New Progress and Exploration of Full-text Bibliometric Analysis Theory and Technology
——A Review of the 2019 Academic Salon on Full-text Bibliometric Analysis

Speaker: Zhang Chengzhi[1]　Hu Zhigang[2]　Xu Shuo[3]　Wang Xuefeng[4]　Shi Qinghui[5]　Wang Wei[6]

(1. School of Economics & Management, Nanjing University of Science & Technology, Nanjing 210094;
2. Institute of Science of Science and Science & Technology Management, Dalian University of Technology Dalian 116024;
3. College of Economics and Management, Beijing University of Technology, Beijing 100124;
4. School of Management and Economics, Beijing Institute of Technology, Beijing 100081;
5. Tongfang Knowledge Network Technology Co., Ltd., Beijing 100084;
6. Reed Elsevier Information Technology (Beijing) Co., Ltd., Beijing 100738)

Summary: Qian Jiajia　Luo Zhuoran

(1. School of Information Management, Wuhan University, Wuhan 430072;
2. Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072)

本文引用格式:全文本计量分析理论与技术的新进展与新探索——2019 全文本文献计量分析学术沙龙综述[J].信息资源管理学报,2020,10(1):111-117.

[**Abstract**] The 2019 Tianfu International Forum of Scientometrics and Evaluation was held in Chengdu on November 10-12, as an important part of the forum, the second "Full-text Bibliometric Analysis" academic salon initiated by Zhang Chengzhi and others has attracted and impressed more than 100 experts and participants. This paper summarized the speeches and discussion of the salon and devided the contents into five themes in order to reveal the latest developments and trends in the theory and technology of full-text bibliometrics analysis at home and abroad. Themes of the salon include academic evaluation system based on citation behavior, entity extraction of full-text, automatic extraction of academic text keywords, emerging research topics and technology predictions, and the access, construction and application of the full-text data.

[**Keywords**] Full-text; Bibliometrics; Citation behavior; Sentiment analysis; Keyword extraction; Entity extraction; Prediction of emerging research topics; Data open

1 引言

传统的文献计量学研究大多基于文献的机器可读目录(MARC)或者中文机器可读目录(CNMARC)等元数据进行分析,但用于计量分析的可读目录数据难以获取且数量有限。如今,随着信息科技的飞速发展和开放获取运动(OA)的兴起,以期刊、会议、报告为代表的文献全文本数据的获取途径变得更加通畅,数据量呈现爆发式增长态势。以章节结构、引用信息、图表等为代表的结构化全文信息以及自然语言处理、机器学习和深度学习为代表的文本挖掘技术,为全文文献计量研究的开展提供了良好的数据基础和有效的技术支持,使文献计量学呈现出多元化、全方位、智能化的特点,为文献计量学的研究提供了更加广阔的空间。卢超等对 1970 年以来学术全文本计量分析相关研究话题的论文发表量进行了统计,发现近几年全文本计量分析领域论文数量呈突破式增长[1]。2019 年国际科学计量与信息计量学会议(ISSI)也收录了较多与全文本计量分析相关的研究成果。各种现象表明基于全文本的文献计量研究日渐成为计量学研究的热点,吸引了国内外学者的广泛关注和研究兴趣,并在实体抽取、引文分析、热点预测等领域开展了一系列的探索与研究,取得了一些有代表性的研究成果。

在全文本文献计量研究的热潮下,2018 年 9 月,南京理工大学经济管理学院章成志教授与大连理工大学科学学与科技管理研究所胡志刚副教授在天府论坛曾共同发起了第一届“全文本文献计量分析”沙龙,受到了与会者的热烈欢迎,在今年的天府论坛上,再次举办这一学术沙龙,给与会者带了一场难得的学术盛宴。本次沙龙由章成志、胡志刚以及北京工业大学经济与管理学院徐硕教授共同发起、中国科学院成都文献情报中心科学计量与科技评价研究中心执行主任陈云伟研究员共同组织,特邀北京理工大学管理与经济学院汪雪锋教授、同方知网办公室主任师庆辉经理和爱思唯尔(Elsevier)合作主管王巍博士作为沙龙嘉宾。本次沙龙吸引了来自全国各高校和科研机构的师生、科研人员共 100 余人参加,与会嘉宾现场介绍了各自最新的研究成果,沙龙交流环节会场互动频繁,参会人员纷纷就相关的研究问题进行了热烈的讨论。

通过对本次沙龙内容的梳理和总结,本文将从“基于引用行为的学术评价体系”、“全文本的实体抽取”、“学术文本关键词自动抽取”、“新兴研究话题和新兴技术预测”和“全文本数据开放、构建与应用”五个主题来解读本次沙龙,以剖析和总结全文本文献计量分析的研究现状以及发展趋势。

2 基于引用行为的学术评价体系

学术论文是科学交流的主要方式和科学产出的主要成果,故传统的学术评价方式大都基于学术论文,甚至在某种情况下出现了“唯论文”的评价体系。基于论文的评价方式本身没有问题,但是目前却出现一种“以刊评文”的普遍现象,即以学术论文发表的期刊对论文进行评价,学术文章所在的期刊是否是 SCI、

CSSCI 等核心刊物，其影响因子高低以及期刊分区等，均作为文章的出身决定文章的水平，这种论文评价方式目前也是颇具争议[2]。除以期刊水平来评价论文优劣之外，基于被引次数的评价方式是目前最常用的评价方式之一，但这种评价方式存在的弊端也不容忽视，一是被引次数的滞后性，即文献的被引次数一般需要三年才能达到峰值，这将降低文献评价工作的时效性；二是引用行为的复杂性，引用行为是指科研人员在学术论文中将已有相关研究的观点、成果以引文内容的形式记录下来的举止行动[3]。引用行为是个体一项复杂的社会行为，不同作者的引用目的、情感、位置等千差万别，仅凭引用次数来衡量引用行为，不足以充分解释引用行为的本质意图。

全文本引文分析关注引用行为和引用内容本身。引文内容分析与传统引文分析相比，其更多的是从施引行为所发生的引用语境本身出发进行分析与挖掘，研究引用语境中包含的引用情感、作者动机等[4]。引用情感分析是指对被引文献描述时关于某个实体的观点、情感、情绪以及态度的计算研究，包括正面的、负面的和中性的。

为解决上述问题，胡志刚提出了一个基于引用行为的学术评价新体系，他从论文的被引次数到引用行为，重新定义了引用的授信力[5]。他认为从评价的角度，如果把引用视为一次授信，那么每次引用的授信力是不同的，它既取决于引用者的授信能力，也取决于引用时的授信行为。他将引用行为划分成核心引用得分、正面引用得分、权威引用得分、跨学科引用得分、经典引用得分和方法类引用得分六个维度，分别从引用强度、引用语境、引文作者、引文期刊、引文年份和引用位置进行测度。汪雪锋利用深度学习的方法计算文献的引用情感和引用目的，其研究结果表明融合了 CNN 于 word2vec 的引文情感计算模型的效果在准确度上高于 svm＋word2vec 模型的计算结果，一定程度上表明了将深度学习的算法用于引文情感分析的效果要好于传统的机器学习算法。

3　全文本的实体抽取

学者在进行学术研究时消化一篇已发表的学术成果需要一定的时间，然而学术出版物数量的不断增加提升了追踪学术前沿技术发展的难度。学者们对于知识的需求已经不能满足于以整篇文章为粒度的知识组织方式，文献中所包含的细粒度的实体的抽取成为文献计量、信息检索等领域的研究热点之一。

章成志介绍了丁颖等的一项学术文本实体影响力评价研究[6]。他们将科学论文实体划分成了宏观、中观和微观三个层面，其中宏观层面的实体包括作者、杂志、引用等；中观层面的实体如关键词等；微观层面的实体包括数据集、方法、领域实体等。汪雪锋介绍了斯坦福大学一项关于科研实体分类的研究[7]，他们把实体分为了 15 种类型，包括通用领域和细分领域，通用领域的实体又包括算法、方法、问题、理论、数据集等，而细分领域的实体类型包括基因、病毒、医学等。

研究方法的抽取与评价是全文本实体抽取中一个广受关注的研究问题。国内机器之心研究团队开发了 SOTA 模型，该模型将不同论文中用到的技术方法、模型、数据集进行抽取和整理，并根据准确度等指标进行排序，便于用户快速查找技术任务和数据集。Zha 等人提出了一种挖掘学术论文中算法路线图的方法，以图的形式展示了算法的演化情况[8]。章成志介绍了他们团队的情报学研究方法与技术体系的构建研究，主要是从权威刊物《情报学报》里面抽取出研究方法，在研究中他们将算法和模型都划分到了研究方法的范畴[9]。此外，王玉琢和章成志还进行了基于全文本分析的研究方法实体抽取研究[10]。

汪雪锋介绍了两个关键技术的术语识别模型，一个是纽约大学等提出的基于分布排名和搜索分数的术语识别模型——Termolator[11]，另一个是布兰迪大学研究团队提出的专利中技术术语的识别与抽取[12]。Termolator 模型通过术语组块过程识别潜在技术术语，然后从出现频率、语言结构、被搜索频率计算分布排名和搜索分数，识别出前几位的技术术语。布兰迪大学研究团队提出的专利中“技术术语”的识别与抽取其基本思路是：对专利文档进行分词、词性标注等预处理，挑选候选词，

然后用有监督的机器学习算法对文本预处理后的候选词进行二分类，得出其是技术术语还是非技术术语。

全文本实体抽取可以为研究者提供更细粒度的信息，帮助研究者快速把握文章主要思想和算法等，很好地辅助科研活动。

4 学术文本关键词自动抽取

学术文本关键词抽取是学术文本分析、学术文本语义内涵挖掘等研究方向的重要组成部分。章成志介绍到学术文本关键词的抽取在技术上经历了三个重要的阶段：第一阶段是以词频统计为基础的阶段；第二阶段为传统机器学习阶段；第三阶段是深度学习阶段。另外，随着网络资源的爆炸性增长，针对网络资源的关键词抽取受到越来越多的关注。

注意力机制是一种神经网络模块，用来模拟人阅读和观看时的视觉注意力。Bahdanau等人2014年时将其应用到机器翻译的领域[13]。张颖怡与章成志[14]基于该理论，构建了一种结合注意力机制的神经网络模型，并将该模型用于文本的关键词自动提取，他们使用了Cop等人于2017年构建的GECO数据集（该数据集提供了用户阅读时的眼动数据），在TextRank的基础上融入了眼动的数据[15]，对用户的注意力值进行了关键词的自动抽取，实验结果表明，引入眼动数据能提高关键词自动抽取的性能。

从业界应用来看，爱思唯尔王巍博士在介绍他们的“指纹引擎”工具时介绍了他们的关键词提取方法：从不同的研究领域中提取不同的关键词语，通过自然语言处理技术对特定研究领域的学术论文的标题和摘要进行文本挖掘以发现重要的主题概念，并与相关领域的叙词表进行对比，抽取出具有语义内涵的主题概念。对于每个文档来说，可以通过倒序词频（IDF）的方法平衡文档中主题概念的出现频率和重要性之间的关系，从而在每个研究领域中筛选出具有最高词汇权重的前N个关键词语。

5 新兴研究话题和新兴技术预测

新兴研究话题和技术的识别与预测对于科研政策制定、科研管理以及个人研究方向的确立都有非常重要的作用。自1965年科学计量学之父Price首次提出“研究前沿”概念以来[16]，对学科前沿发展态势的分析及对未来发展的预测一直是科学计量学领域的研究热点，尤其随着学术文献全文本获取难度降低，基于全文本分析的新兴研究话题和技术预测获得学界更广泛的关注。

徐硕介绍了Rotolo等人对新兴技术的定义：一种发展相对较快的、具有一定程度的连贯性和重大影响力的根本性创新技术[17]。他们认为新兴技术应该具备五个属性：①相对较快的增长率；②连贯性；③重大影响（社会经济影响）；④根本性创新；⑤不确定性和模糊性。Wang在这个定义的基础上，界定了新兴研究话题的概念，并提出了识别新兴话题的方法[18]。具体来说，新兴话题应该更多体现科学影响，而不是社会经济影响，并且去掉了不确定性和模糊性属性，因为不确定性和模糊性难以计算。

徐硕团队[19]探索了1965—2019年初之间1607篇论文的引文网络，利用关键路径主路径分析方法（key-route main path analysis approach），结合搜索路径链接数的遍历权重以发现新兴研究领域的知识传播轨迹，结果他们发现新兴话题研究可以分为三个阶段（图1）：第一阶段是1965—1974年，这个阶段引文分析方法刚被提出，为新兴阶段；第二阶段是1974—2015年，引文网络分析比较流行，为探索阶段；第三个阶段是2015年至今，深度学习和机器学习开始被大量采用，为发展阶段。同时，还可以观察到一些研究漂移：第一是研究方法上从基于引用的分析方法到基于机器学习的方法；第二是研究方向上从测量到识别；第三是信息资源上从论文到专利。

此外，徐硕团队[20]采用了Wang对于新兴话题的定义，提出了基于全文本和机器学习模型的新兴话题识别与预测框架。框架基于全文内容和引文数据等，利用DIM和CIM[21-22]主题模型抽取研究话题并计算其连贯性、增长率、科学影响力和创新性，然后通过多任务最小二乘支持向量机[23]预测相应指标的发展趋势。为克服缺乏参考数据的问题，徐硕团队[24]参与了佐治亚理工大学发起的“2018—2019年度新兴

技术预测竞赛”，将原有框架模型中的 DIM 主题模型改成了 TNG 主题模型[25]，基于每个技术涉及的研究人员数量重新计算了影响力指标，获得了第二名的成绩(全球 21 支团队参赛)。

目前引文分析法仍然是新兴主题探测研究的主要方法，但机器学习和深度学习方法已经开始被应用于新兴主题的识别与探测中。同时，学者们在研究新兴主题时不再局限于论文数据，专利、报告等数据也开始成为分析的对象。

6 全文本数据开放、构建与应用

1950 年引文索引数据库(SCI)的创立，开启了引文分析的新时代[26]。胡志刚在报告如何利用全文本分析构建基于引用行为的学术评价体系以“破四唯”时指出，基于引用行为的评价等于指标体系加上全文本数据，其团队对现有的结构化全文本数据集进行了整理(图 2)。

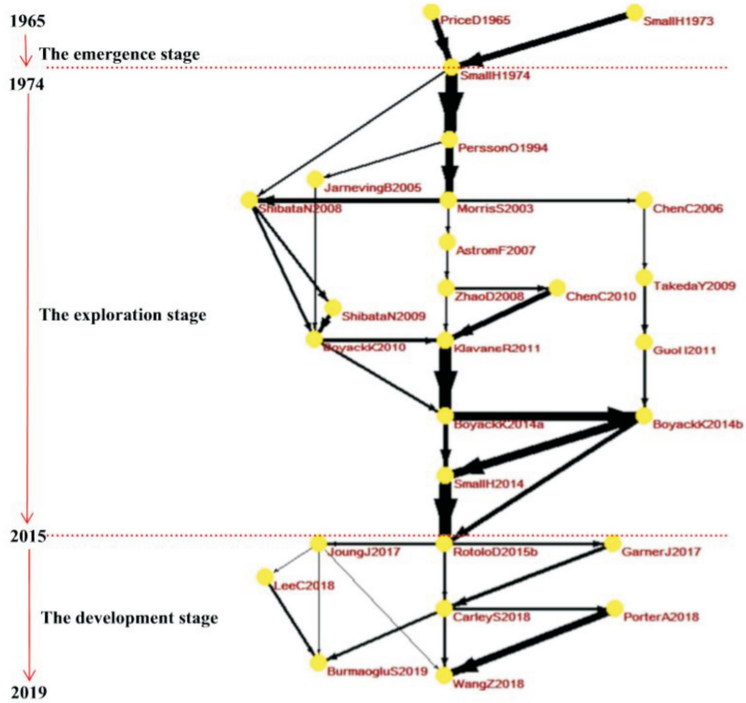

图 1 新兴话题研究领域发展阶段[19]

来源形态	数据来源	期刊量	论文量	结构化格式	学科范围
OA 出版商	BMC	397	38 万+	HTML	全部学科、医学为主
	PLOS Open for Discovery	7	21 万+	HTML / XML	全部学科
	frontiers	64	10 万+	HTML / XML	全学科，含人文社科
	PeerJ	7	0.8 万+	HTML / XML	计算机、生物、化学等
非 OA 出版商	ELSEVIER	2000+	600 万+	HTML / XML	全部学科
	Springer	1200+	46 万+	HTML / XML	全部学科
OA 数据库	PMC	2238	530 万	HTML / XML	生物学、医学为主
非 OA 数据库	CNKI 中国知网 www.cnki.net	9200+	350 万+	HTML	全部学科

图 2 目前国内外可供获取的结构化全文数据集规模

注：此表格节选自胡志刚报告 PPT

Web of Science 是文献计量学研究的主要数据来源之一，但是 Web of Science 数据的引文数据信息量比较少，需要利用文章的数字对象唯一标识符(DOI)想办法获得全文本，但徐

硕研究团队发现 Web of Science 里面很多 DOI 是有错误的，所以还需要进行 DOI 的清洗工作，该团队对 DOI 错误进行了仔细分析，总结了参考文献的几种经典 DOI 错误，并在正则表达式的基础上提出了数据清理的方法[27]。

随着开放获取运动的推动以及自然语言处理等技术的发展，全文本数据的获取变得容易，数据集的构建难度降低，并且业界已经落地了一些相关应用。师庆辉介绍了中国知网支撑全文本计量研究的资源库与知识库建设实践。他指出学术资源深度开发与应用实际上面临三个问题，数字化、数据化和知识化。对此，知网研发了一套数字化处理的机器用于获取全文本数据，可以实现自动完成翻书、扫描、识别、版面分析、pdf 制作、数据库的录入整个一体化的流程，其中应用到的技术细节包括版面的分析和图表公式抽取、基于自然语言处理技术的关键词标引、自动分类等。就平台开放全文数据的难题，师庆辉介绍了他们预期构建的一个从文献库到知识库的操作系统，学者可以在操作系统前端选择全文的相关处理，知网操作知识库各种各样的方案技术或者是算法模型，然后把操作结果返回给学者，学者可以将其作为研究结果，比如研究者可以利用知网平台进行某个领域的热点前沿预测等。王巍博士介绍了爱思唯尔的指纹引擎工具（FingerPrint），它利用自然语言处理技术等进行全文本的分析，可以为用户提供确定关键词、靶向本领域的顶尖学者并寻求合作机会、助力筛选最合适的期刊发表文章和展示个性化学术成果等功能。

7 交流与讨论

本次沙龙六位嘉宾的分享引起了现场老师和同学的强烈兴趣，在交流讨论环节，现场互动频繁，有学者就引用内容范围识别、引用动机、创新性评价、深度学习在全文本分析中的应用等问题与现场嘉宾展开了讨论。此外，本次“全文本文献计量分析”沙龙还吸引了如电子商务、经济管理等其他领域学者的兴趣，他们也就如何开展全文本文献计量的研究请教了现场的六位嘉宾。可见，全文本文献计量分析已经引起了领域内外的广泛关注，这也将为该方向的研究带来更多跨学科、跨领域的新问题和新思考，不断推进该研究领域向前发展。

8 总结

随着全文本数据的开放和文本挖掘技术的发展，尤其是自然语言处理、深度学习等技术的应用，文献计量学的研究对象和理论方法都在不断扩展。在全文本文献计量的热潮下，“全文本文献计量分析”沙龙围绕引用分析、实体抽取、新兴研究话题和新兴技术预测等主题，对全文本文献计量分析的研究和应用现状展开了讨论，为学者们提供了一个思想碰撞、观点交流的平台，也吸引更多的学者了解并参与到全文本文献计量的研究中。

综上所述，目前全文本文献计量的研究热点集中在引用行为、引用情感分析、细粒度的实体抽取、新兴研究话题和新兴技术预测等方向，深度学习等新技术在该研究领域也得到了研究者们的广泛关注和应用。学术评价上，基于全文本分析的评价指标层出不穷，未来文献阅读器中的读者行为（高亮、标注）等也有可能极大地增加基于全文本的评价维度。

致谢：特别感谢中国科学院成都文献情报中心科技处处长、科学计量与科技评价研究中心（SERC）执行主任陈云伟研究员为本次沙龙提供场地支持并进行组织和主持工作。

参考文献

[1] 卢超，章成志，王玉琢，等. 语义特征分析的深化——学术文献的全文计量分析研究综述[J]. 中国图书馆学报，审稿中.

[2] 赵蓉英，王旭. 多维视角下学术期刊影响力评价研究[J]. 情报科学，2019，37(11)：3-10.

[3] 王佳敏，李信，刘齐进. 全文本文献计量分析学术沙龙综述[J]. 信息资源管理学报，2018，8(4)：119-125.

[4] 刘浏，王东波. 引用内容分析研究综述[J]. 情报学报，2017，36(6)：637-643.

[5] 周春雷. 学术授信评价及其应用[M]. 北京：科学出版社，2016.

[6] Ding Y，Song M，Han J，et al. Entitymetrics：Measuring the Impact of Entities[J/OL]. PLoS ONE，2013，8(8)：

e71416.[2019-11-17].https://journals.plos.org/plosone/article? id=10.1371/journal.pone.0071416.

[7] Babko-Malaya O, Meyers A, Pustejovsky J, et al. Modeling debate within a scientific community[C]// International Conference on Social Intelligence & Technology. IEEE Computer Society,2013:57-63.

[8] Zha H, Chen W, Li K ,et al. Mining algorithm roadmap in scientific publications[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM,2019:1083-1092.

[9] 章成志等. 情报学研究方法与技术体系构建研究[M]. 北京:科学技术文献出版社,2020.

[10] Wang Y, Zhang C. Finding more methodological entities from academic articles via iterative strategy: A preliminary study[C]//Proceedings of the 17th International Conference on Scientometrics and Informetrics (ISSI 2019), Rome, Italy, 2019. International Society for Scientometrics and Informetrics, 2019: 2702-2703.

[11] Meyers A L , He Y, Glass Z, et al. The Termolator: Terminology recognition based on chunking, statistical and search-based scores[J]. Frontiers in Research Metrics and Analytics, 2018, 1384:34-43.

[12] Anick P, Verhagen M, Pustejovsky J. Identification of technology terms in patents[J]. Proceedings of the 9th International Conference on Language Resources and Evaluation,May 26-31, 2014, Iceland. European Language Resources Association (ELRA), 2014: 2008-2014.

[13] Bahdanau D , Cho K , Bengio Y . Neural machine translation by jointly learning to align and translate[EB/OL]. Arxiv, 2014:1409.0473. [2019-11-18]. https://arxiv.org/pdf/1409.0473.pdf.

[14] Zhang Y, Zhang C. Unsupervised keyphrase extraction in academic publications using human attention[C]// Proceedings of the 17th International Conference on Scientometrics and Informetrics(ISSI 2019), Rome, Italy,2019. International Society for Scientometrics and Informetrics, 2019: 2483-2484.

[15] Cop U , Dirix N , Drieghe D , et al. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading[J]. Behavior Research Methods, 2016, 49(2):1-14.

[16] de Solla Price D. Networks of scientific papers[J]. Science, 1965,149(3683):510-515.

[17] Rotolo D, Hicks D, Martin B. What is an emerging technology? [J]. Research Policy, 2015, 44(10):1827-1843.

[18] Wang Q. A bibliometric model for identifying emerging research topics[J]. Journal of the Association for Information Science and Technology, 2018,69(2), 290-304.

[19] Xu S, Hao L, An X, et al. Review on emerging research topics with key-route main path analysis[J/OL]. Scientometrics, 2019. [2019-11-17]. https://doi.org/10.1007/s11192-019-03288-5, 2019.

[20] Xu S, Hao L, An X, et al. Emerging research topic detection with multiple machine learning models[J/OL]. Journal of Informetrics,2019,13(4):e100983.

[21] Gerrish S M,Blei D. A language-based approach to measuring scholarly impact[C]// Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel 2010. Elsevier,2010: 375-382.

[22] DietzL, Bickel S, Scheffer T. Unsupervised prediction of citation influences[C]// Proceedings of the 24th International Conference on Machine Learning. ACM, 2007: 233-240.

[23] Xu S, An X,Qiao X, et al. Multi-task least-squares support vector machines[J]. Multimedia Tools and Applications, 2014, 71(2): 699-715.

[24] Xu S, Hao L, Yang G, et al. A topic models based framework for detecting and forecasting emerging technologies [J]. Technological Forecasting & Social Change. (under review).

[25] Wang X, McCallum A, Wei X. Topical N-Grams: Phrase and topic discovery, with an application to information retrieval[C]// Proceedings of the 7th IEEE International Conference on Data Mining. IEEE, 2007: 697-702.

[26] 邱均平,嵇丽.美国《科学引文索引》与科学评价研究[J].科研管理,2003(4):22-28.

[27] Xu S, Hao L, An X, et al. Types of DOI errors of cited references in Web of Science with a cleaning method[J]. Scientometrics, 2019, 120(3):1427-1437.

(收稿日期:2019-11-17)