数据挖掘在专刊组稿策划中的应用

白娅娜[1)] 武英刚[1)] 宫在芹[1)] 代艳玲[2)]

1)煤炭科学研究总院出版传媒集团《洁净煤技术》编辑部;2)煤炭科学研究总院出版传媒集团:100013,北京

摘 要 专题报道是提升科技期刊质量和品牌的重要途径,其中专题策划是关键。以《洁净煤技术》专刊《煤炭燃烧及污染物防治理论与技术》为例,分析数据挖掘在专刊组稿策划中的应用。利用中国知网(CNKI)确定专刊选题方向,利用国家科技报告服务系统确定专刊报道方向,利用百度学术平台、CNKI等挖掘优秀作者资源,对专刊传播效果进行分析。利用数据挖掘准确捕捉了学科热点领域与前沿科学走向,有效把握读者需求,发掘大量优秀作者信息,为专刊的精准约稿提供了保障。据CNKI统计,《洁净煤技术》2015年论文按照被引频次和下载量排序,专刊论文分别占TOP 50论文数的36%和40%。《洁净煤技术》利用大数据选题组稿取得了良好效果,扩大了期刊的学术影响力。

关键词 专题策划;大数据;专刊;数据挖掘

Application of data mining in planning special issues and soliciting contributions // BAI Yana, WU Yinggang, GONG Zaiqin, DAI Yanling

Abstract Organizing featured topics is an important approach to improve the quality and brand of journals, and special topics planning is the key to achieve this goal. Illustrated by the example of the topic "coal combustion and pollution prevention theory and technology" for *Clean Coal Technology*, the application of data mining in planning special issue and soliciting contributions is analyzed. The direction of the topics is determined through CNKI and national science and technology report service system, and excellent author resources are found through Baidu Scholar, CNKI and so on. At last, communication effect of the special issue is analyzed. The results show that hot frontiers of science, reader demands and excellent authors can be obtained through data mining. As for the statistics of citations and downloads of papers in 2015 in CNKI, the ratios of papers in *Clean Coal Technology* special issues are 36% and 40% of the top 50 papers, respectively. In short, planning special issues and soliciting contributions of *Clean Coal Technology* have obtained a good result, and academic influence of journal has been expanded.

Keywords special topics planning; big data; special issue; data mining

First-author's address Editorial Department of Clean Coal Technology, Publication Media Group, China Coal Research Institute, 100013, Beijing, China

DOI:10.16811/j.cnki.1001-4314.2016.06.011

在科技期刊内容越来越同质化的今天,期刊读者资源的争夺将更加激烈,专刊针对特定热点领域的前沿科学技术成果进行专题报道,内容系统,信息量大,实用价值较高。此类专题报道容易得到领域内读者的青睐,是突出期刊特色和优势的重要举措[1-2]。《绿色建筑》从创刊至今,连续推出"绿色地产""超高层建筑的绿色"等16期系列专题,对于推进绿色生态技术发展,完善绿色建筑评价标准等方面起到了重要作用[3]。《放射学实践》近3年成功组织22个专题报道,扩大了该刊在业内的影响,专题论文的篇均被引频次和下载量均明显高于非专题论文[4]。可见,专题报道对提升读者认可度、提高期刊学术影响力以及促进期刊品牌发展具有重要作用,而搞好专题策划是关键。

随着互联网技术的迅猛发展,大数据为精准策划专题、有效把握读者需求提供了可能。通过大数据可以有效捕捉学科热点领域与前沿科学,获得更为准确的读者需求信息,还可以高效挖掘优秀作者的数据信息,为专题的精准、优质、高效组稿提供了强大手段[5-6];因此,随着大数据在各行各业的迅速扩散和渗透,如何利用大数据进行期刊专题策划,深度挖掘选题内容和形式,使其成为期刊亮点[7-8],是每个期刊人需要思考的问题。科技期刊借助中国知网(CNKI)、国家科技报告服务系统、百度学术平台等可获得大数据的互联网平台挖掘有价值的信息,指导期刊选题策划。

《洁净煤技术》2015年第2期专刊《煤炭燃烧及污染物防治理论与技术》就是利用互联网平台大数据进行专刊组稿策划的典型案例。本文论述如何利用大数据确定专刊选题方向与发现优秀作者的技巧。

1 用大数据捕捉选题信息

1.1 利用CNKI确定选题方向 CNKI是面向全行业提供各类知识信息内容的数字出版和知识服务平台,是国内最大的专业数字出版与知识增值服务企业。中文数据量每年更新上亿篇,年均检索20余亿次,年均下载量近10亿次。如此庞大的数据使CNKI基本掌握了我国主要科研机构或研究学者的研究方向和研究热点等信息资源,因此,我们可利用CNKI的大数据筛选专刊的选题方向。

《洁净煤技术》刊载范围主要包括煤炭加工、煤炭高效洁净燃烧、煤炭转化和节能减排等4个领域。利用CNKI统计《洁净煤技术》2005—2014年论文主题

词出现的频次,采用聚类分析方法,识别高频与低频词,以此确定选题主题词。《洁净煤技术》为双月刊,以主题词年均出现 3 次及以上定义为高频词,经过多次数据合并,产生了 3 组群集:第 1 组主题词出现频次 30 次以上,包括煤气化、煤液化、水煤浆、燃煤、燃煤污染物等;第 2 组主题词出现频次为 10 ~ 30 次,包括褐煤干燥、浮选药剂等;低于 10 次划为第 3 组,包括电选、煤岩分析等。因此,初选出现频次最高的第 1 组主题词作为选题方向候选主题词。

然后利用 CNKI 大数据进一步筛选主题词。数据挖掘的属性包括检索条件、发文时间、发文数量、被引频次、下载量等。下载量多少代表读者对该篇文章的青睐程度,最终以各主题词论文平均下载量排序筛选出合适的选题方向。给定检索条件为主题,论文发表年份为 2005—2014 年。结果表明,“燃煤”发文数量最多,为 2 万 17 篇,其次为“煤气化”“水煤浆”“燃煤污染物”“煤液化”,分别为5 982、2 553、1 683 和 955 篇。由于获得数据样本数量太大,定义被引频次在 5 次以上的论文下载量为有效下载量,用于计算主题词论文的平均下载量,经过分析计算,“燃煤”“煤气化”“水煤浆”“燃煤污染物”“煤液化”论文平均下载量分别为 560、395、366、572、413 次,平均下载量排在前 3 位的是“燃煤污染物”“燃煤”“煤液化”。近几年,我国雾霾天气频发,这与我国煤炭燃烧产生的大气污染问题密不可分;因此,通过分析大数据,并结合目前行业形势,专刊选题方向确定为“燃煤”和“燃煤污染物”相关方向。

1.2　利用国家科技报告服务系统确定报道方向　为了确定专刊报道方向,利用国家科技报告服务系统挖掘“燃煤”和“燃煤污染物”方向的研究热点。国家科技报告服务系统是国内目前比较完整记载政府科技基金项目的特种文献,系统将国家支持的科研活动产生的资料等向公众免费开放共享,拥有国家和地方科研计划及科研投入方向的庞大数据,可在线浏览所有公开的科技报告全文。

数据挖掘的属性包括检索条件、立项时间、项目数量等。检索条件选择摘要,立项年份为 2005—2014 年,根据项目数量排名来确定报道方向。“摘要”中分别检索“燃煤”“燃煤污染物”,结果显示共 174 项,剔除结果中的重复项目,最后统计得到实际项目数为 102 项。根据项目研究方向,对挖掘到的数据采用聚类分析方法加工整理,分为 2 个大类,15 小类。2 大类分别为煤炭燃烧和污染物治理:“煤炭燃烧”下分新型循环流化床燃煤锅炉、超临界和超超临界发电技术等 7 个小类,项目总数为 57 项;“污染物治理”下分高效脱硫脱硝技术、多污染物协同控制技术等 8 个小类,项目总数为 45 项。按照项目数量排序(≥5 项)拟定了 10 个报道方向,然后征求编委、专家和读者意见,最终确定专刊题目为《煤炭燃烧及污染物防治理论与技术》,下设煤炭低氮燃烧技术、新型循环流化床燃煤锅炉技术、高效煤粉锅炉技术、超临界和超超临界发电技术、高效烟气脱硫脱硝脱汞、燃煤污染物协同脱除技术、$PM_{2.5}$脱除机理研究和粉煤灰利用等 8 个报道方向。

2　用大数据挖掘优秀作者资源

2.1　查找项目负责人　利用国家科技报告服务系统统计选题的主题词可挖掘出大量相关项目负责人姓名、单位等信息,在前述选题结果的基础上进一步挖掘出在研项目 19 项,并根据项目介绍获得项目负责人相关信息,然后借助百度学术平台对项目负责人进行数据信息的深度挖掘。

关联挖掘规则是一项很重要的数据挖掘技术,百度学术平台的“学者频道”功能正是利用了关联挖掘规则,利用它可以发现与项目负责人关联的专家学者,从而扩大约稿作者来源。利用“学者频道”输入首席或负责人的姓名和单位,可获得发文类型和数量、被引频次等数据,同时可获得关联学者信息及合作次数,编辑部以合作次数在 10 次以上的关联学者也作为约稿对象。如利用国家科技报告服务系统挖掘到中国科学院某位专家承担了一项选题相关课题。通过百度学术平台“学者频道”搜索专家姓名,挖掘出其合作次数超过 10 次的作者 6 人,最终约稿 4 篇,刊出 3 篇。

2.2　定位“四高”作者　“四高”作者指“高影响、高层次、高水平、高产出”的作者[9]。利用百度学术平台搜索“燃煤污染物”,限定发表年份为 2005—2014 年,领域为动力工程及工程热物理,筛选出2 620个结果,根据被引频次排序分为 2 组:第 1 组被引频次 10 次以上,有 106 篇;第 2 组被引频次为 10 次以下,有2 514 篇。从分组结果来看,大量的论文被引次数均小于 10 次,舍弃这些论文作者。

2005 年,美国科学家 Hirsch 提出用 H 指数来测评核心科学家。一般认为,H 指数越高的学者在其学科领域的影响力越大。百度学术“学者频道”已经给出大多数作者的 H 值,可以利用此功能对第 1 组的论文作者进一步筛选,选择 H 指数 10 以上的作者作为“四高”约稿作者,共挖掘出 23 人。

如通过这种方法挖掘到“四高”作者上海理工大学张忠孝,其 2005 年发表的《气体再燃低 NO_x 排放试验研究》被引频次为 54 次,百度学术“学者频道”显示其共发表相关论文 180 篇,被引总频次高达1 108次,H 指数为 17,符合编辑部“四高”作者定位,成功约到了

《复合还原剂脱硝机理与试验研究》一文。

2.3 挖掘优秀年轻作者 优秀年轻作者一般指高校有实力和潜力的副教授、讲师和科研院所的副研究员等。以2005—2011年毕业的博士为挖掘对象,通过CNKI挖掘其博士论文的被引频次,然后根据论文的年均被引频次排序初步筛选出候选的优秀年轻作者,定义年均被引频次为博士论文刊出之日至2014年12月31日论文总被引频次除以年数,该指标可以简单对比不同作者在不同年份所发论文的影响力。

主题词限定"燃煤"或"燃煤污染物",发文年份限定为2005—2011年,跨库选择"博士"论文库,共显示310项结果,通过数据挖掘获得有效博士论文数量198篇。对数据进行统计分析,年均被引频次最大为10次,选择排序在前10%的论文作者为候选优秀年轻作者,共20名。然后利用CNKI"作者发文检索"功能再次筛选,分析博士毕业后发文是否为燃煤相关方向,以此核实作者身份,并根据作者单位再次精准挖掘,分析其发文数量是否满足近3年所发表论文中至少有4~5篇为通信作者或第一作者,且论文学术质量要与本刊论文水平相当或略高的要求,据此挖掘出满足要求的年轻作者8位,并开展约稿。

2.4 发现优秀老作者 优秀老作者对期刊具有一定的信赖度,向其约稿成功机会较大。利用CNKI影响力统计分析数据库挖掘《洁净煤技术》相关发文作者大数据情况。该功能可以统计2007—2014年本刊论文作者及其发文量。从统计结果来看,共有作者3 116名(所有作者),被引频次在30次以上的作者为67名。针对挖掘到的67名优秀老作者,利用编辑部自有的作者数据库进一步深度挖掘,统计分析与选题方向相关的作者进行重点约稿,共有20名优秀老作者曾经在本刊上发表过相关论文。通过这种方式约稿10篇,最终刊用高水平论文4篇。

3 大数据挖掘组稿传播效果

对于行业期刊而言,好的专题策划可增强期刊对行业发展的指导性和参与性,提升期刊的行业影响力和社会影响力[10]。《洁净煤技术》编辑部利用CNKI、国家科技报告服务系统、百度学术等大数据平台进行数据挖掘,有效指导编辑部选择和确定选题方向,并实现优秀作者的精准发掘,成功组织策划了《煤炭燃烧及污染物防治理论与技术》专刊。专刊系统展示了煤炭洁净燃烧技术及燃煤污染物治理技术等方面的一系列成果,针对具有理论价值和实践意义的热点、难点进行学术探讨,营造了一个公开、透明的科技成果交流共享空间,为解决我国环境污染问题出谋划策、答疑解惑,具有鲜明的针对性和服务性,提高了期刊在业界的参与性、影响力和吸引力。专刊信息量大,针对性与系统性强,为从业者提高理论水平、增强实践能力提供理论支持和决策参考,推动了煤炭燃烧及燃煤污染物治理技术等科技成果的快速转化和开放获取,对企业或地区经济社会发展具有一定的促进作用。

《洁净煤技术》利用大数据挖掘进行专题组稿,将读者调研与市场调研结合起来,弥补了以往根据市场调研数据确定选题模式的不足,大幅度提高了编辑部获取、利用信息数据的效率,准确预测了读者需求,激发选题灵感,优化选题流程,抓住选题重点,使选题过程更科学,选题方向更准确。而且,大数据改变了传统普遍撒网的征稿模式,实现了约稿作者的准确定位,扩大了优秀论文的来源。通过大数据挖掘,专刊刊出高质量论文32篇,基金论文比达到94%。

基于大数据挖掘策划的专刊是对读者偏好的精准化测量和预测,可最大化地满足读者需求;因此,专刊刊出后,通过CNKI统计论文的被引频次、下载量及网站阅读量等指标考察读者需求是否与专刊内容相契合。据CNKI统计,《洁净煤技术》2015年刊出论文183篇:按照被引频次排序,18篇专刊论文进入TOP 50,占TOP 50论文数的36%,占TOP 20论文数的40%,被引频次最高为5次;按照下载量排序,专刊在TOP 50和TOP 20论文中占比均为40%,下载量最高达到808次。期刊主页专刊论文阅读量均大于130次,最高达到212次。从这些指标来看,《洁净煤技术》利用大数据选题组稿取得了良好效果,满足了读者需求,提高了专刊的学术质量和影响力,具有一定的借鉴意义。

4 结束语

专题报道是期刊特色和风格的重要体现,是期刊的亮点,而大数据的迅猛发展和应用为专题的策划、实施、宣传提供了广泛、快捷、便利的渠道。CNKI、百度学术等互联网大数据平台动态演化的知识网络为编辑部利用大数据选题、组稿提供保障,可利用不同的大数据挖掘、分析、处理方法从中获取有用信息,了解期刊学科热点及发展趋势,确定专题策划方向,准确定位高水平作者,实现专刊精准宣传推送等,这将成为科技期刊专题策划的重要发展方向之一。同时,编辑部还应建立、完善自身的专家群库、作者群库及读者群库等大数据,使科技期刊适应互联网时代的发展要求,为期刊的创新发展保驾护航。