

基于Altmetrics指标识别的研究热点对比分析 ——以生物学领域为例

迟培娟^{1,2} 陈挺^{1,2,3} 宋秀芳¹ 冷伏海³

(1. 中国科学院文献情报中心, 北京 100190; 2. 中国科学院大学, 北京 100049;
3. 中国科学院科技战略咨询研究院, 北京 100190)

摘要: 研究热点通常采用高被引论文的聚类来识别, 但被引次数需要时间积累。Altmetrics指标具有良好的即时性, 采用相关指标识别研究热点有可能提高时效性。本文重点分析基于Altmetrics指标与高被引论文识别的研究热点之间的异同以及背后的原因, 以探究基于Altmetrics指标识别的研究热点是否适合作为前沿趋势探测和科技政策制定的辅助依据。本文选取生物学领域2013年7月1日—2014年6月30日发表的论文, 分析基于Altmetrics分数、Twitter指标、Mendeley指标、被引次数4个指标识别的研究热点之间的关系。分析发现: 基于Altmetrics分数和Twitter指标识别的研究热点较为接近, 主要反映公众热门话题; 基于高被引论文识别的研究热点更加反映科研人员的关注点, 更大程度上代表了学术界的研究前沿; 基于Mendeley指标识别的研究热点综合了上述两类热点, 既有公众较为关心的科普主题, 也有学术界关注的研究前沿。

关键词: Altmetrics; Twitter; Mendeley; 被引次数; 研究热点

中图分类号: G350; G353.1

DOI: 10.3772/j.issn.1673-2286.2019.05.005

当前科技创新演变加剧, 交叉融合加速, 科技环境愈加复杂。研究热点或研究前沿本质上是指科研共同体高度关注并积极开展研究的领域, 从一定程度上代表了学科的研究重点和未来发展方向^[1]。对于热点和前沿的识别不仅是科研活动中探明知识基础、发掘前沿趋势的必要手段, 也是科技政策制定过程中的重要依据。

从传统的文献计量学角度来看, 研究热点的相关论文表现为高发文量和高被引的特点, 因此目前的研究主要基于主题词和引用关系来探测研究热点。基于主题词的研究热点识别主要通过词频和主题词共现关系来实现^[2-5]。基于引用关系的热点识别主要通过高被引论文的引文耦合^[6]和共被引关系^[7]来实现。基于主题词的热点识别需要大量的人工清洗和判读, 主题词对研究内容的揭示度较低, 而且需要相关主题的发文量积累到一定程度才可以识别。基于引用关系的热点识别可以较为准确地发现研究主题, 目前常采用共被引和引文

耦合方法进行识别。近年来, 科睿唯安公司(原称汤森路透)联合中国科学院每年通过ESI高被引论文的共被引分析遴选研究前沿^[8]。无论是上述哪一种方法, 都需要一定的时间积累发文量或者引用关系, 才可以凸显出研究热点, 而为了更好地进行战略决策, 需要更加快速地识别研究热点。

Altmetrics指标的出现使得快速识别研究热点有了可能。Altmetrics最大的特点是即时性, 短时间内就能获得大量关注和讨论^[9-10]。目前涉及的Altmetrics计量学指标主要包括4类, 即社交媒体平台、文献管理工具、学术服务平台和主流新闻媒体^[11]。Altmetrics当前的研究主要集中在Altmetrics指标与传统文献计量学指标的相关性和对比研究^[12-15]; Altmetrics工具的比较^[16-18]; Altmetrics指标的应用, 如学术评价等^[19-20]。基于Altmetrics指标识别研究热点的相关研究尚未全面展开。Eysenbach^[14]经过研究发现, 高被引论文的Twitter指标在论文发表后30天内就可以积累到较高水

平, 后续变化不大, 这表明该指标有非常好的即时性; 赵雅馨等^[21]尝试使用Altmetrics指标和词频分析法来探测信息与计算科学领域的研究热点; 方志超^[22]使用发文数据、引用数据、使用数据和Altmetrics数据, 基于共词方法对科学计量学领域热点主题进行了识别, 但上述研究没有深入分析采用Altmetrics数据识别的研究热点与采用传统计量学方法识别的热点之间的关系及背后的原因。

由于被引次数是最常使用的传统计量指标, Twitter指标、Mendeley指标等是常见且容易积累的Altmetrics指标, 本文希望进一步探讨依据Altmetrics指标遴选的热门论文与高被引论文之间的关系, 基于Altmetrics指标识别的研究热点与基于传统高被引论文识别的研究热点之间的关系及产生这些异同的原因, 探讨基于Altmetrics指标识别的研究热点是否适合作为前沿趋势探测和科技政策制定的辅助依据。

1 研究方案和分析结果

1.1 研究数据

本文采用Altmetrics.com网站的相关指标, 该网站是目前较为成熟的Altmetrics服务平台之一, 涉及的评价指标较为丰富, 涵盖各类主流指标, 主要包括Twitter指标、Mendeley指标、新闻、文章点评、专利引用等信息, 上述指标被分别赋予不同的权重, 最后合计得出Altmetrics分数^[22]。该平台还免费提供API接口, 便于大样本数据的抓取。

论文数据采用2013年7月1日—2014年6月30日发表的生物学领域论文。采用这一时间段的数据主要原因有: ①Altmetrics.com成立于2011年, 较少涉及2011年之前出版的论文, 运营2年后其论文数据涵盖范围已经较为广泛, 这样可以避免大量目标论文没有分数; ②后续研究涉及被引次数指标的对比分析, 论文的引用次数需要一定的时间积累。论文发表3~4年后, 其被引次数增长趋于稳定。

首先在SCIE数据库中检索指定时间段发表的生物学领域论文并提取被引次数信息, 由于本文要探索研究热点, 虽然“Review”类的文献被引次数高, 但不涉及具体的研究, 因此仅保留“Article”“Letter”“Meeting Abstract”这3种类型的论文, 共获取到10 002篇论文, 并提取其被引次数等信息, 其中有8 841

篇论文有被引记录, 8 529篇有DOI信息; 然后, 利用其DOI信息到Altmetrics.com网站获取相关指标信息, 共有3 940条数据有Altmetrics分数, 提取的具体指标包括Altmetrics分数、Twitter指标、Mendeley指标、新闻指标、F1000指标、Wikipedia指标, 其中大部分论文均有Twitter指标和Mendeley指标, 其他指标空值居多, 因此后续研究主要围绕Altmetrics分数、Twitter指标、Mendeley指标和被引次数展开。

1.2 相关性分析

目前已有的Altmetrics指标与被引次数之间的相关性研究是基于某领域的所有论文集合, 本研究主要目的是识别研究热点, 因此重点分析热门论文的指标相关性。根据二八定律, 本文选取Altmetrics分数排名前20%的788篇论文作为热门论文, 计算Altmetrics分数、Twitter指标和Mendeley指标与被引次数的相关性, 结果如表1所示, 与全部论文相比, 热门论文的3个指标与被引次数的相关系数均有不同程度的降低。这说明在热门论文中, Altmetrics指标与被引次数的相关性要低于总体论文的相关性。笔者将进一步探讨相关性的降低是否会影响采用Altmetrics指标识别研究热点。

表1 热门论文和所有论文的相关性分析结果

指标	相关性	所有论文-被引次数	高影响力论文-被引次数
Altmetrics分数	相关系数r	0.410**	0.150**
	双尾显著性	0.000	0.000
Twitter指标	相关系数r	0.360**	0.168**
	双尾显著性	0.000	0.000
Mendeley指标	相关系数r	0.693**	0.622**
	双尾显著性	0.000	0.000

注: **表示相关性在0.01水平上显著

1.3 研究热点分析

分别按照Altmetrics分数和被引次数降序排列论文, 各取前10%的论文(394篇论文), 合并后去重, 发现有92篇论文重复, 仅占394篇论文的23%, 说明两个指标的前10%论文大部分均不相同, 这也预示着Altmetrics平台的热门论文与SCIE平台的高被引论文存在较大差别。笔者希望进一步探索基于Altmetrics指标与被引次数识别的研究热点之间的关系。

由于基于引文耦合关系识别的热点论文簇对研究热点的揭示度要明显好于高频词^[23],因此本文采用高影响力论文的引文耦合关系来探测研究热点。依次选取Altmetrics分数、Twitter指标、Mendeley指标和

被引次数中排名前10%的高影响力论文各394篇,利用Vosviewer软件的bibliographic coupling功能,对高影响力论文进行引文耦合聚类分析,结果如表2所示。

表2 通过各种指标识别的研究热点对比

研究热点	Altmetrics分数	Twitter指标	Mendeley指标	被引次数
热点1	动物遗传多样性与进化	物种多样性	物种多样性研究	肿瘤治疗和基因组编辑
热点2	海洋生态保护和物种保护	海洋生态保护和物种保护	海洋生态保护	脱落酸等植物信号转导
热点3	恐龙生物进化	恐龙生物进化	人类活动对动物行为的影响; 进化生物学研究	蛋白质晶体结构解析
热点4	人类健康及疾病	涉及动物的药物实验数据偏差	遗传学和认知研究(肿瘤治疗和基因组编辑)	古地球化学和热力学研究
热点5	动物行为研究	人类微生物组学研究	-	线粒体抗病毒蛋白和肥胖基因
热点6	人类及灵长类合作、分享、社交等行为的认知研究	-	-	长链非编码RNA研究

基于Altmetrics分数识别的研究热点主要有6个,从子学科领域的分布来看,主要分布在环境科学、进化生物学、动物学和行为学等领域。具体研究包括:温室气体对海洋生态的影响,海参等高价值海洋生物的灭绝,非禽类恐龙的进化模拟,恐龙体重演变,肥胖的影响,肠道益生菌与关节炎的关系等,狗对颜色的敏感性,转基因鲑鱼对其他鱼类的影响等内容。从以上分析可以看出,这些研究主要涉及环保、恐龙、健康、动物行为、转基因等公众话题,十分容易引起公众的关注。

通过比较可以发现,基于Altmetrics分数和Twitter指标识别的研究热点更为接近,均涉及海洋生态环境保护、物种多样性、恐龙进化、医药健康等公众较为关注的话题。基于被引次数识别的研究热点主要反映科研人员的视角,与学术界广泛认可的科技前沿识别结果能够吻合,如基因组编辑技术和癌症免疫治疗均入选*Science*年度十大科学突破^[24]和《2015研究前沿及分析解读》,植物信号转入选《2015研究前沿及分析解读》^[8]。可以看出基于被引次数识别的研究热点更具专业性,往往代表了科技界研究的前沿,与基于Altmetrics分数和Twitter指标识别的热点明显不同。基于Mendeley指标识别的研究热点除了公众关心的海洋生态环境保护、物种多样性、恐龙进化等话题,还涉及一些学术界广泛认可的研究前沿,如CRISPR代表的基因组编辑技术、癌症免疫治疗等遗传学领域研究。

2 分析与讨论

本文基于3种Altmetrics指标和被引次数识别了2013—2014年生物学领域的研究热点,对比不同方法识别的研究热点的异同,下面就结果进行分析讨论。

基于Twitter指标识别的研究热点主要反映公众科普话题。Twitter是美国主要的社交网络及微博服务网站,与其他社交媒体相比,学者更愿意在Twitter上介绍自己或他人的最新研究成果,Twitter也因此成为Altmetrics研究的焦点。已有研究表明,推文多少与被引次数的多少呈现弱相关性^[25]。本文的研究结果也进一步佐证了这一结果,广大Twitter用户和研究人员的关注点存在明显差异。Twitter用户以普通民众为主,科研人员只占小部分,因此能引起广大用户兴趣的科普话题容易成为热点。

基于Altmetrics分数识别的研究热点与基于Twitter指标识别的研究热点较为相似。Altmetrics分数是综合Twitter、新闻、Google+、Facebook等多种数据计算出来的,王睿等^[26]对Altmetrics分数较高的论文分析后发现,从各底层数据源的数据量来看,Twitter的数据量最多,可以推测Twitter数据对Altmetrics分数的影响较大,这可能是二者识别的研究热点相似的原因。同时由于Altmetrics分数还涉及其他多种数据,因此二者识别的研究热点还存在一定差异。

基于Mendeley指标识别的研究热点既有公众较为

关心的科普主题,也有科研人员关注的研究前沿。这是因为Mendeley是一款被广泛使用的文献管理软件,同时也是一个在线的学术社交网络平台,因此它的使用主体既包括科研人员,也包括科学爱好者和公众。从本文研究的情况来看,基于Mendeley指标识别的研究热点可以部分反映研究前沿,与基于高被引论文识别的研究热点部分相同。

基于被引次数识别的研究热点更加反映科研人员的关注点,这是因为引用行为的主体是科研人员,被科研人员广泛关注和认可的前沿研究更容易成为高被引论文,因此基于高被引论文识别的结果与学术界普遍认可的研究前沿更为相关。

基于上述4种指标识别的研究热点经专业人员判读后,发现基于被引次数识别的研究热点更符合其认知,基于Mendeley指标识别的研究热点部分符合其认知。此前,赵雅馨等^[21]通过对信息与计算科学领域的实例研究认为基于Altmetrics指标识别研究热点符合领域专家的预期,这与本文的研究结果不同,这种情况或许是由于不同学科领域的科研人员在网络上的活跃度和活动轨迹不同造成的,因此还需要其他学科领域的进一步分析验证。

3 结语

本文对生物学领域热门论文的主要Altmetrics指标和被引次数的相关性进行了分析,结果发现,与领域所有论文相比,热门论文各指标的相关性均有不同程度的降低。基于Altmetrics分数和Twitter指标识别的研究热点与领域专家认可的高被引论文研究热点差别较大,基于Mendeley指标识别的研究热点可以部分反映高被引论文研究热点,同时涵盖了一定的公众话题。总体来看,基于上述3种Altmetrics指标识别的研究热点并不适合直接用来作为前沿趋势探测和科技政策制定的辅助依据,基于Mendeley指标识别的研究热点可以结合专家评议等方法作为参考。

本文仅采用生物学领域一年内的论文数据进行研究,可能不同学科、不同时段内的数据分析结果会不同,需要更多的实证分析加以验证。由于分析数据中大部分论文均有Twitter指标和Mendeley指标,其他指标空值居多,因此研究主要围绕Altmetrics分数、Twitter指标、Mendeley指标展开,后续研究可以补充一些新闻数据、政策文件数据和更专业的即时网络数

据,如Faculty1000、Researchgate、科学网等,以便从不同视角加深对Altmetrics指标的理解,更好地利用Altmetrics指标为科研人员和情报人员提供及时、专业的前沿信息。

参考文献

- [1] 王海燕. 未来高影响力科技论文的识别理论与方法研究 [D]. 北京: 中国科学院大学, 2013.
- [2] 龚放, 叶波. 2000—2004年中国教育研究热点与关键词——基于CSSCI的统计分析 [J]. 高等教育研究, 2006 (9): 1-9.
- [3] 国防. 多模态话语分析研究热点及趋势分析——基于文献计量学方法 [J]. 外语与外语教学, 2016 (3): 58-66, 146.
- [4] 王绍卜. 基于复杂网络的港口物流领域研究热点实证分析 [J]. 计算机系统应用, 2014 (7): 246-251.
- [5] 宗乾进, 袁勤俭, 沈洪洲. 国外社交网络研究热点与前沿 [J]. 图书情报知识, 2012 (6): 68-75.
- [6] KESSLER M M. Bibliographic coupling between scientific papers [J]. American Documentation, 1963, 14 (1): 10-25.
- [7] SMALL H. Co-citation in the scientific literature: A new measure of the relationship between two documents [J]. Journal of the American Society for Information Science, 1973, 24 (4): 265-269.
- [8] 中国科学院文献情报中心, 汤森路透. 2015研究前沿及分析解读 [M]. 北京: 科学出版社, 2015.
- [9] ROEMER R C, BORCHARDT R. From bibliometrics to altmetrics: A changing scholarly landscape [J]. College & Research Library News, 2012, 39: 8-9.
- [10] BORNMAN L. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics [J]. Journal of Informetrics, 2014, 8 (4): 895-903.
- [11] LIN J, FENNER M. Altmetrics in evolution: Defining and redefining the ontology of article-level metrics [J]. Information Standards Quarterly, 2013, 25 (2): 20.
- [12] BORNMAN L. Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics [J]. Scientometrics, 2015, 103 (3): 1123-1144.
- [13] HADAS S, JUDIT B I, MIKE T. Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics [J]. Journal of the Association for Information Science and Technology, 2014, 65 (5): 1018-1027.

- [14] EYSENBACH G. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact [J]. *Journal of Medical Internet Research-International Scientific Journal for Medical Research*, 2011, 13: 123.
- [15] PEOPLES B K, MIDWAY S R, SACKETT D, et al. Twitter predicts citation rates of ecological research [J]. *PLoS ONE*, 2016, 11 (11): e0166570.
- [16] ALHOORI H, FURUTA R. Do altmetrics follow the crowd or does the crowd follow altmetrics? [C] // *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. London: IEEE Press, 2014: 375-378.
- [17] LI X, THELWALL M, GIUSTINI D. Validating online reference managers for scholarly impact measurement [J]. *Scientometrics*, 2012, 91 (2): 461-471.
- [18] 王睿, 胡文静, 郭玮. 常用Altmetrics工具比较 [J]. *现代图书情报技术*, 2014 (12): 18-26.
- [19] HAUSTEIN S, BOWMAN T D, COSTAS R. Interpreting “altmetrics”: viewing acts on social media through the lens of citation and social theories [M] // SUGIMOTO C R. *Theories of Informetrics and Scholarly Communication*. Bloomington: De Gruyter, 2016: 372-405.
- [20] 王贤文, 方志超, 王虹茵. 连续、动态和复合的单篇论文评价体系构建研究 [J]. *科学学与科学技术管理*, 2015 (8): 37-48.
- [21] 赵雅馨, 杨志萍. 研究热点探测的替代计量学方法和应用——以信息与计算科学为例 [J]. *情报杂志*, 2016, 35 (11): 39-44.
- [22] 方志超. 基于多元数据对象的科学计量学热点主题识别研究 [D]. 大连: 大连理工大学, 2017.
- [23] Altmetrics about our data [EB/OL]. [2019-03-01]. <https://www.altmetric.com/about-our-data/>.
- [24] BOYACK K, KLAVANS R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? [J]. *Journal of the American Society for Information Science and Technology*, 2010, 61 (12): 2389-2404.
- [25] Breakthrough of The Year 2013 [EB/OL]. [2019-03-18]. <https://www.sciencemag.org/breakthrough-of-the-year-2013>.
- [26] 舒非, HAUSTEIN S, 全薇. 推特 (Twitter) 对中国论文的国际关注度影响研究 [J]. *图书馆论坛*, 2017, 37 (6): 55-60.
- [27] 王睿, 胡文静, 郭玮. 高Altmetrics指标科技论文学术影响力研究 [J]. *图书情报工作*, 2014, 58 (21): 92-98.

作者简介

迟培娟, 女, 1981年生, 博士研究生, 馆员, 研究方向: 情报学。

陈挺, 男, 1984年生, 博士研究生, 助理研究员, 研究方向: 知识提取、科学可视化。

宋秀芳, 女, 1970年生, 博士, 副研究馆员, 研究方向: 学科情报。

冷伏海, 男, 1963年生, 博士, 研究员, 通信作者, 研究方向: 竞争情报、高科技信息分析, E-mail: lengfuhai@casipm.ac.cn。

Identification of Research Hotspots Based on Altmetrics in the Field of Biology

CHI PeiJuan^{1,2} CHEN Ting^{1,2,3} SONG XiuFang¹ LENG FuHai³

(1. National Science Library, Chinese Academy of Sciences, Beijing 100190, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China;

3. Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Research hotspots are usually identified by clustering of highly cited papers, but the number of citations takes time to accumulate. Altmetrics indicators have good immediacy, and it is possible to identify research hotspots with relevant indicators quickly. This paper aims to analyze the research hotspots based on Altmetrics indicators and citation indicators. The papers published in the field of biology from July 1, 2013 to June 30, 2014 were selected to analyze the relationship between the research hotspots identified by the Altmetrics score, the Twitter index, the Mendeley index, and the number of cited times. The research hotspots based on Altmetrics scores and Twitter indicators are relatively close, mainly reflecting the interest of public. The research hotspots based on high-cited papers reflect the focus of research experts and represent research fronts. The research hotspots based on the Mendeley indicator identification include the above two types of hotspots, which are not only the popular science topics that the general public care about, but also the research fronts of research academic concern. The research hotspots identified based on the above three Altmetrics indicators are not suitable for direct use as a supplementary basis for frontier trend detection and technology policy formulation.

Keywords: Altmetrics; Twitter; Mendeley; Cited Times; Research Hotspots

(收稿日期: 2019-04-22)