情报学报　2022 年 9 月　第 41 卷　第 9 期
Journal of the China Society for Scientific and Technical Information, Sept. 2022, 41(9): 930-944

DOI: 10.3772/j.issn.1000-0135.2022.09.004

# 基于媒介比较的学科新兴主题动态识别——altmetrics与引文数据的融合方法

段庆锋，闫绪娴，陈　红，刘东霞

（山西财经大学管理科学与工程学院，太原　030006）

**摘　要**　社交网络媒介具有时间优势，非常适合于发现处于新兴阶段的学科主题。新兴主题演化在社交媒介和出版媒介呈现差异化，表现为altmetrics指标相对于引文指标的领先优势，由此考虑将两种媒介指标落差的快速形成及拉大作为识别新兴主题的关键依据。基于此，提出媒介比较视角的学科新兴主题识别方法。首先，融合altmetrics和引文数据，构建可用于时序比较的主题媒介活跃度相对落差指标rgap；其次，采用突发性检测算法，探测主题时序指标rgap的突发状态，以揭示新兴主题涌现及其演化过程；最后，通过情报学领域的实证分析说明该识别方法有效可靠，基于媒介比较指标的识别方法表现出良好的识别能力与一定程度的及时性优势，发现以推特及博客为代表的高覆盖高流行性指标更有助于获得满意分析效果。

**关键词**　学科新兴主题；媒介比较；altmetrics；引文

# Dynamic Identification of Emerging Topics in Discipline Based on the Comparison between Different Types of Media: A Method Combining Altmetrics and Citations

Duan Qingfeng, Yan Xuxian, Chen Hong and Liu Dongxia

(*School of Management Science & Engineering, Shanxi University of Finance and Economics, Taiyuan*　030006)

**Abstract**: The timeliness advantage of social networking media helps in identifying emerging scientific topics. However, the evolution of emerging topics in disciplines is relatively different between social media and publishing media in that their altmetrics tend to dominate citations during the period of emergence. We believe that the fast-growing gap occurring between these two indicators is a key basis for distinguishing scientific emerging topics from other topics. As such, we propose a method to recognize the underlying emergent topics in disciplines via comparison among different media. First, by combining altmetrics and citations, we devise the "rgap" indicator (a gap indicating the relative differences of media in terms of the activeness of topics) to conduct sequential comparisons. Second, we employ the Burst Detection Algorithm to detect the burst status of topics, using the sequence of the "rgap" indicator, which can help identify the emerging topics and show their process of evolution. Finally, we conduct an empirical analysis in the field of information science, and the empirical analysis has proven that this identification method is effective and reliable. The identification method based on a media comparison index showed good identification ability and timeliness advantage. We also found that a satisfactory re-

**收稿日期**：2021-05-15；**修回日期**：2021-07-27

**基金项目**：教育部人文社会科学项目"基于学术社交媒体的学科新兴趋势识别研究"（20YJA870005）；国家社会科学基金项目"供给侧改革背景下提升资源型企业经济韧性的关键要素、效应评估与实现路径研究"（20BGL100）。

**作者简介**：段庆锋，男，1977年生，博士，副教授，硕士生导师，主要研究领域为科技情报，E-mail：dqf01@sina.com；闫绪娴，女，1978年生，博士，教授，博士生导师，主要研究领域为智能决策；陈红，女，1972年生，博士，教授，博士生导师，主要研究领域为科技创新；刘东霞，女，1975年生，博士，副教授，硕士生导师，主要研究领域为科技创新。

sult can be achieved to some extent when using altmetrics indicators that are characterized by high level of coverage and prevalence, such as tweets or posts.

**Key words:** emerging topics in disciplines; comparison among different media; altmetrics; citations

## 1　引　言

新兴主题是学科前沿动态涌现的语义表达，是认知学科内在机制、支撑科技战略决策的重要基础。与其他一般性主题相比，新兴主题更强调时间维度的“新兴”属性[1]，很大程度上是学科发展趋势的凸显，是反映其未来应用价值的重要因素，更是形成科学预见能力的基础。尤其，新兴主题对分析识别程序的语义敏感性与时间响应性能力要求更高[2]。虽然互联网传播及在线预印本极大地加速了传统出版媒介对学科前沿动态的响应性，但客观存在的出版周期从根本上制约了针对学科新兴主题的及时响应[3]。特别地，以引文为代表的文献计量指标及相关分析工具在学术评价中得到了广泛应用与普遍认同[1]，但是滞后性弊端也限制了分析结果的前瞻预见能力。

如此背景下，借助社交网络媒体数据及指标开展学科新兴主题分析的必要性愈加凸显。大量研究及实践应用表明，社交网络对于社会热点及潮流趋势具有极强的捕捉能力与敏感响应性[4]。这种信息传播与反馈能力已经在舆情引导、社交营销等多领域得到实践检验。学术活动及成果传播同样也日益嵌入社交网络，新成果及重大科学事件第一时间在社交媒介传播与发酵，为观察科学前沿动态提供别样途径。作为面向学术文献的新型计量方式，altmetrics 指标通过定量途径刻画了学术成果在社会网络及相关数字媒体的传播及反馈程度，其学术价值与应用潜力迅速得到大量关注[5]。近年来，altmetrics 在科学监测及预见方面的应用潜力日益受到重视，不过主要焦点在指数本身的合理有效性方面，强调学术评价层面的功能价值[6]。

基于社交媒体对于新兴趋势的敏感能力，从媒介比较的独特视角，提出学科新兴主题的动态识别方法。本文贡献体现在以下方面：①基于新兴主题在不同媒介形成的差异性活跃分布，提出媒介比较落差的学科新兴主题识别方法，发挥了社交媒介先于传统出版媒介的时间分析优势。②借助主题 LDA（latent Dirichlet allocation）模型，将单篇论文层面指标扩展至主题层面，构建主题 altmetrics 指标，基于此构建了可用于时序比较的媒介相对落差指标，以满足动态分析需求。③采用突发检测算法，探测主题媒介相对落差时序指标的突发性增长，从全局动态视角揭示新兴主题的分布与演化。

## 2　相关研究

理解新兴主题内涵并把握其关键特征，并形成可操作化方法，是从复杂科学现象及事件中识别出高决策价值新兴主题的基础与关键。Rotolo 等[7]基于文献梳理，将新兴主题特征归纳为：新颖性（novelty）、相对高增长性（relative rapid growth）、凝聚性（coherence）、显著影响力（prominent impact）、不确定性和模糊性（uncertainty and ambiguity）。识别主题新颖性的方法相对明确，从生命周期初始阶段的基本共识出发，多数文献以要素（如论文、专业术语）的时间属性为考察依据，将最新出现，或者与最新知识要素存在紧密联系的学科主题纳入考察范畴，也有部分文献从技术创新属性角度进一步判别主题的新颖程度[8]。高增长性几乎是学科主题研究文献中被使用最多的识别特征，新兴主题知识的高增长能力往往外化为各种度量指标与结构的变化，识别方法较为多样，如增长型指数法[9]、S 形增长曲线拟合法[10]、科学文献或者主题词聚类规模的变化[11]等。显著影响力同样是对甄别出高价值新兴主题非常关键的特征，但常被视为隐含条件，因为那些被寄予高影响力期望的新知识与新技术理应被识别程序优先挑选出来。不过针对显著影响力的评估更加适用于回溯性分析，面对当下新兴成果的展望性分析通常更多地需要借助于专家的主观性经验预判。凝聚性、不确定性和模糊性相对而言被讨论较少，学科新兴主题发展经历可以被视为凝聚性不断加强、不确定和模糊性相对减少的过程，虽然存在初步认知，但总体上，可操作性度量指标与挖掘识别算法研究相对不足。

学科新主题分析主要建立在学术成果产出数据之上，如学术文献、专利等。基于成果出版媒体数据的识别方法多样，原理上可以粗略分为两种类型，一是通过特征指标的数量变化识别新兴主题的增长状态，二是通过结构性变化考察新兴主题的演进规律。①反映学科状态的度量指标设计与分析是

关键，论文、作者、引文、词语等都是指标构建的基础要素。通过焦点指标的纵向比较，可以形成主题的新兴增长程度研判。还有，以生命周期理论为指导，曲线拟合法也是识别主题生命周期状态（如新兴成长阶段）的重要途径[12]。特别地，Kleinberg[13]提出的突发检测（burst detection）算法得到了科技情报领域学者广泛关注，可以将新兴主题的动态趋势建模为文档或词语流的状态转化，并在CiteSpaceII、SCI2等代表性科学计量与情报分析软件中得到了成功应用。②科学系统的结构性变化是探测新兴主题的重要途径。文献的引文关系是科技情报的重要分析工具，基于此可以形成多种有效网络分析[14]，如引文网络（citation network）、共被引网络（co-citation network）、引文耦合网络（bibliographic coupling network）。通常将合适的文献聚类概念化为不同学科主题范畴，通过分析这些聚类的演化有助于发现目标主题，如近期涌现的，或者近期被频繁引用，或者频繁引用新知识的文献簇。共词分析（co-word analysis）也是常用的探测新兴知识的有力工具，通过文本挖掘模型，基于词的共现关系，形成的共词网络能够从语义层面揭示学科主题结构，通过词语的凝聚、分离、新生、消亡等演化关系，达到探测主题新知识的分析目的[15]。

近年来，学术活动的网络化与社交化背景下，以社交网络为代表的新型媒介给科技情报带来互补性分析优势。以替代计量学altmetrics以及社交网络挖掘分析为代表的理论与方法成为研究热点。以Thelwall、余厚强、邱均平、赵蓉英等为代表的国内外学者深入分析了altmetrics内在机理及特征基础[16-18]，为其拓展性应用奠定了理论基础。基于altmetrics指标的设计出发点，许多学者从单篇论文层面探讨了altmetrics对于多维学术影响力的揭示能力[19-20]，不论是特定学科领域还是不同平台及指标类型选取，大多实证研究发现altmetrics与传统文献计量指标具有中等程度相关性，但存在表征维度的差异性，刻画了超出学界范畴的社会影响力[21]。在单篇学术文献影响力基础上，可以形成更加宏观的学科领域前沿探测。王菲菲等[22]结合传统文献计量指标和altmetrics指标，基于LDA主题模型，提出5个学科前沿探测指标，通过综合性评估方法挑选出有价值主题。牌艳欣等[23]采用altmetrics指标数据，采用z指数法识别出包括突发性新兴主题在内的4种类型主题，以情报学为例的实证检验了altmetrics数据的学科动态探测能力。可见，altmetrics指标在学科探测方面的潜在优势逐步得到重视。Small等[24]指出，altmetrics比传统文献计量及挖掘分析具有更大的即时分析能力，是重要的潜在研究方向。王贤文等[25]较早基于下载数据提出了替代计量在科学趋势实时探测方面的应用价值。

相对于学术评价方面的丰富成果，altmetrics数据在学科监测及预见方面的研究相对较少，大多只是利用相关指标的普通统计分析应用，缺乏主题层面的深度语义探讨。已有学科前沿分析大多基于传统学术成果出版媒体，将出版媒介数据与社交网络媒介数据深度融合的研究不多。面对学科趋势复杂问题，altmetrics数据还存在有待克服的分析局限，比如，单篇文献层面计量指标难以直接用于主题分析，作为存量型指标并不适用于动态时序分析等，这也正是本文尝试探讨的环节。

## 3 识别原理

### 3.1 学术社交媒介优势

在时间维度，学术社交媒介对于学科动态具有高响应性能力——即时性。相比于文献出版的固定周期，作为非正式学术场合的社交网络能够对成果内容与学术事件产生即时性响应。相比于传统互联网（Web 1.0）环境下学术文献的存储电子化及传播网络化（如电子文献数据库），社交关系能够进一步加强传播效应，焦点信息能够在复杂社交网络（Web 2.0）形成“病毒式”扩散，产生指数式加速传播。这样，学科前沿及热点能够以最快速度通过社交平台在学界甚至更广泛社会层面形成最大化传播。相对而言，引文分析虽然已成为科技情报界成熟分析范式，但时滞性缺陷极大地制约了其对于新兴趋势的捕捉能力。

在内容维度，学术社交媒介对于学科动态具有高响应能力——敏感性。敏感性反映了学术社交网络在内容层面的灵敏性，即对于传播学术内容具有自主化的甄别与挑选能力[26]。相关研究发现，被学术社交平台广泛关注的论文不但具有高流行度，而且往往表现出较高的学术水准[27]，这与altmetrics指标与引文指标存在中等程度相关性的认知相吻合。高水平论文社交平台的涌现是用户群体的集体选择结果，每个用户自主地对科学成果及事件产生偏好性，并表现为特定网络行为，而社交关系（如好友、粉丝、推荐等）的交织互动不断将集体选择结果强化与放大。段庆锋等[28]采用协同过滤模型刻画

学术信息传播的社会化过程，网络用户间形成相互的学术推荐，形成高价值高流行信息的过滤结果，这种信息过滤机制事实上形成了对学术信息的集体性价值判断。从群决策视角看，altmetrics 指标就是所有相关用户偏好的集结体现，通过用户的“投票”（每次针对焦点主题的社交事件）形成群体智慧。另外，社交网络的马太效应能够将有价值弱信号加以放大，适合于发现潜力巨大但还未被学界大众普遍认知的学科新兴主题。

### 3.2　新兴主题的媒介比较与识别

按照生命周期理论，相对高增长已经成为识别新兴主题的共识性特征。但是，这种区别于其他生命周期阶段的相对高增长特征在不同媒介观察下可能存在时间与程度上的差异。比较视角下，通常认为社交媒介具有明显的时间响应优势，流行迅速，但是持续周期较短，而出版媒介则相对呈现相反特点。Ortega[29]研究了不同媒介指标的生命周期，发现基于推特和博客的提及（mentions）指标反映最迅速，读者人数次之；阅读和下载数具有最长活跃期，而引文指标的活跃呈现最迟钝。可见，社交媒介指标相对于引文指标具有即时性优势。围绕具有流行价值的新兴主题，学术社交媒介通过各种社交事件（如评论、转发等）通常能够快速形成突发性热点涌现，而出版媒介需经历滞后性出版周期才能形成后续文献和引证行为。学科新兴主题在上述两种媒介快速增长的不同步导致 altmetrics 指标的短期比较优势，先行性 altmetrics 指标的领先优势迅速拉大，形成突发性指标相对差距。而随着主题新兴性减弱，趋向成熟，后发性引文指标开始快速增长，相对落差将逐步缩小。当主题进入成熟阶段，这种相对落差不再显著甚至自然消弭。

altmetrics 指标先行于引文指标并形成活跃落差的现象为识别新兴主题提供了独特途径，如图 1 所示。在两种媒介指标增长速率差异消失之前的 $t_2$ 时间段是重要的探测窗口期，其中增长率差距 gap 达到峰值的 $t_1$ 时刻为新兴主题提供了关键识别机会，因为差距 gap 的快速拉大是新兴主题动态在不同媒介的差异化体现。上述认知为识别学科新兴主题提供了理论启发。学科新兴主题识别的关键在于落差 gap 的发现，通过数据挖掘算法，检测指标 gap 的突发性增长，并将涌现出这样特征的主题视为新兴主题的重点考察对象。

不同媒介的综合比较能够弥补单一媒介的视角局限，形成互补性方案，更有利于甄别新兴主题。一方面，学术社交媒介丰富多元语义与文献计量信息融合，有助于构建高探测能力的识别系统；另一方面，基于媒介差距 gap 的识别方法有助于提升新兴主题的识别准确性。基于前述生命周期理论分析可知，媒介差距 gap 的快速增长通常与主题新兴阶段相关联。相比于单一媒介的增长性分析，以不同媒介差距 gap 为识别基础的方法更有助于降低新兴主题与传统热门主题混淆的风险。

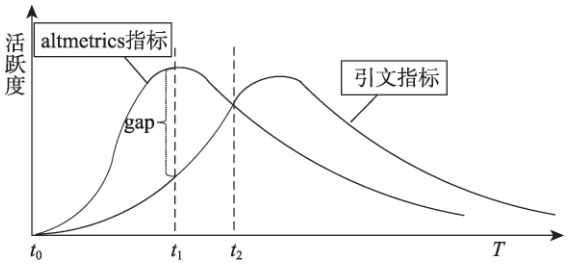


图 1　新兴主题的媒介活跃差距示意图

## 4　研究方法

### 4.1　指标设计

#### 4.1.1　主题层面指标

为了反映论文主题在社交网络受到的关注程度，构建主题层面 altmetrics 指标。原有 altmetrics 指标针对单篇论文，难以直接定量刻画主题特征。由此，以论文层面 altmetrics 指标为基础，设计实现主题层面指标的映射方法是关键。聚焦于 altmetrics 指标形成机制，考虑学术社交事件、论文、主题之间的逻辑关系，形成不同层面指标转化依据，如图 2 所示。假设每篇论文旨在反映若干主题，且与不同主题的关联程度存在差异，即不同论文具有不同的主题分布。altmetrics 指标是针对论文的社交事件计数，同时这些焦点事件也是针对论文所代表的主题，因为用户对于论文的理解以主题认知为基础，驱动社交行为的用户兴趣、情感及偏好等事实上都围绕主题形成。由此，通过论文与主题关联可以建立起社交事件与主题间联系。考虑到用户对于不同

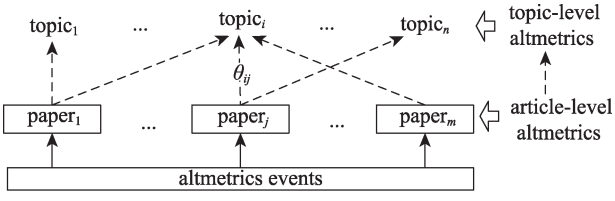


图 2　主题层面 altmetrics 指标构建原理

主题的关注兴趣程度存在差异，可以将面向主题的社交事件加权累计，以度量刻画主题 altmetrics 指标。

主题层面 altmetrics 是以主题为焦点的学术社交事件计数。第 $i$ 个主题在第 $t$ 年的 altmetrics 指标 TLA 定义为

$$\mathrm{TLA}_{it} = \sum_{j,\ \text{for}\ T(j)=t} \mathrm{altmetrics}_j \cdot \theta_{ij} \tag{1}$$

其中，$\mathrm{altmetrics}_j$ 表示第 $j$ 篇论文的 altmetrics 指标；$\theta_{ij}$ 表示第 $i$ 个主题在第 $j$ 篇论文中相对重要程度；$T(j)$ 是返回值为时间的函数，表示第 $j$ 篇论文的发表年份。指标 TLA 反映了主题内容在学术社交网络的流行度，体现了该主题的社会影响力，其取值越大，反映受到的网络用户关注越多。

同理，可以对引文指标做相同方式变换处理，构建主题层面引文指标 TLC。第 $i$ 个主题在第 $t$ 年的 TLC 定义为

$$\mathrm{TLC}_{it} = \sum_{j,\ \text{for}\ T(j)=t} \mathrm{TC}_j \cdot \theta_{ij} \tag{2}$$

其中，$\mathrm{TC}_j$ 代表第 $j$ 篇论文的被引；其他变量含义同公式(1)。类似地，指标 TLC 反映了该主题内容在学界受到的重视程度，体现了该主题的学术影响力。

公式(1)和公式(2)中，参数 $\theta$ 是主题 $i$ 在论文 $j$ 的相对权重，反映主题和论文之间联系，如何估计该参数是指标构建的关键。考虑到动态视角下主题概念的演化性，这里采用 Blei 等[30]提出的动态主题模型 DTM（dynamic topic models）加以估计确定。DTM 模型是经典主题 LDA 模型[31]在时间维度的拓展，能够克服静态模型无法刻画主题演变的局限，更加适用于本文针对新兴主题生命周期特征的动态分析[32]。DTM 是生成式概率模型，将主题建模为单词分布 $\varphi$，将文档建模为主题分布 $\theta$，通过变分推断或者随机采样的方法估计得到模型参数。不同于 LDA 模型，DTM 模型将影响 $\varphi$ 和 $\theta$ 取值的 Dirichlet 分布超参数不再设为固定值，而是建模为随时间变化的高斯分布，以刻画随时间演化的主题分布。DTM 模型的动态设定更利于准确地揭示随时间变化的主题，而估计的参数 $\theta$ 更适用于动态分析。将拟合参数 $\theta_{ij}$ 代入相应公式，可以得到聚焦于主题的度量指标。

#### 4.1.2 媒介活跃度落差指标 gap

为了刻画揭示新兴主题在不同媒介的动态差异，构建反映主题媒介活跃落差的指标 gap。基于学术社交网络和出版媒介对于新兴主题的不同响应能力，将指标 gap 定义为学科主题在某时段内 altmetrics 指标（TLA）与引文指标（TLC）的相对差，

$$\mathrm{gap}_{it} = \frac{\mathrm{TLA}_{it} - \mathrm{TLC}_{it}}{\mathrm{TLC}_{it}} \tag{3}$$

其中，$\mathrm{gap}_{it}$ 代表第 $i$ 个主题在第 $t$ 年的媒介活跃度落差；$\mathrm{TLA}_{it}$ 和 $\mathrm{TLC}_{it}$ 分别由公式(1)与公式(2)给出。指标 gap 取值越大，反映主题在两种不同媒介的活跃度差异越大，即主题在学术社交网络的关注程度超过学术出版媒介的程度越大。

#### 4.1.3 媒介活跃度相对落差指标 rgap

基于时序的动态分析是揭示新兴性的有效途径。但是，不同年份的指标 gap 难以直接比较。因为 altmetrics 指标和引文指标都是存量型指标，指标数值从学术文献发表开始持续累积，意味着出版久的论文拥有更多指标累积时间，越可能取得时间累积优势。还有，社交网络近年发展迅猛，用户规模呈指数式增长，社交方式不断创新多样化[33]，这些导致不同年份 altmetrics 指标平均水平差异很大，跨时域 altmetrics 指标缺乏可比性。指标 gap 由 altmetrics 指标和引文指标构成，自然同样存在难以比较的问题。

为了克服上述因素引致的比较偏差，借鉴相对技术优势指数 RTA（relative technology advantage）[34]的设计思路，对指标 gap 加权修正，以适用于时序比较场景。具体地，第 $i$ 个主题在第 $t$ 年份的媒介活跃度相对落差 $\mathrm{rgap}_{it}$ 定义为

$$\mathrm{rgap}_{it} = \frac{\mathrm{gap}_{it}}{\sum_i \mathrm{gap}_{it}} \Big/ \frac{\sum_t \mathrm{gap}_{it}}{\sum_i \sum_t \mathrm{gap}_{it}} \tag{4}$$

其中，$\dfrac{\mathrm{gap}_{it}}{\sum_i \mathrm{gap}_{it}}$ 反映了主题 $i$ 在第 $t$ 年的指标 gap 优势；$\dfrac{\sum_t \mathrm{gap}_{it}}{\sum_i \sum_t \mathrm{gap}_{it}}$ 反映了考虑所有年份下主题 $i$ 的指标 gap 期望优势；两者比值即 $\mathrm{rgap}_{it}$ 衡量了主题 $i$ 在第 $t$ 年的媒介活跃度落差相对优势。指标 rgap 值大于 1，说明主题 $i$ 的媒介活跃度落差在第 $t$ 年呈现相对优势，取值越大，优势越显著，否则相反。该指数能够进行横纵向比较，适用于识别新兴主题所需的动态分析。

### 4.2 识别方法

按照新兴主题的生命周期理论认知，altmetrics

指标在时间上先行于引文指标，两者落差能够成为识别学科新兴主题的有效判别依据。若主题的媒介活跃度落差突然呈现，甚至出现快速明显增长态势，则被视为新兴主题涌现信号，应该将其纳入重点考察范畴。这种状态的突然变化可以有多种分析方法，如指标直接比较、拟合法等。新兴主题处于生命周期初始阶段，不同阶段呈现差异状态，可以通过全时域比较的状态转换揭示新兴特征的突发涌现。由此，本文采用 Kleinberg[13]提出的突发性检测算法识别媒介活跃度落差的突然增大。

该算法采用状态机模型来对突发事件进行建模，假设系统在不同状态间转换，形成马尔科夫决策过程，可以从低级状态跃升到突发状态。以基准（baseline）和突发（burst）组成的两状态系统为例，将基准状态的期望概率定义为焦点事件发生的概率，$p_0=R/D$，其中 $R$ 为焦点事件发生数量，$D$ 为全体事件发生数量；突发状态的期望概率定义为 $p_1=s\times p_0$，其中参数 $s$ 为常数，决定了两种状态的差距，若 $s$ 取值增大，则只有焦点事件处于更高概率水平，才能认定系统进入突发状态。系统在某时间点处于哪个状态取决于两个方面因素：状态拟合程度 $\sigma$ 和状态转换困难程度 $\tau$。

（1）具体地，$t$ 时刻状态 $i$ 的拟合度 $\sigma$ 定义为

$$\sigma_t(i)=\ln\left[\binom{d_t}{r_t}\cdot p_i^{r_t}\cdot(1-p_i)^{d_t-r_t}\right] \tag{5}$$

其中，$r_t$ 和 $d_t$ 分别代表 $t$ 时刻焦点事件与全体事件的发生数量；$i$=0，表示基准状态，$i$=1，表示突发状态。拟合度 $\sigma$ 反映了焦点事件的发生比例与潜在状态期望概率水平的吻合程度，吻合程度越高，处于该状态的可能性就越大。与基准状态相比，突发状态的拟合度越高，当前处于突发状态的概率亦越高。另外，针对突发时刻 $t$，还可以计算其突发强度 $\text{weight}_t$，定义为 $\sigma_t(1)-\sigma_t(0)$，反映了突发时刻系统处于突发状态相比于基准状态的拟合度提高程度。$\text{weight}_t$ 值越大，说明突发程度越强烈。

（2）具体地，系统状态转换困难度 $\tau$ 定义为

$$\tau_t=\gamma\cdot I(i_t-i_{t-1})\cdot\ln(T) \tag{6}$$

其中，$T$ 为时间跨度数量；$\gamma$ 为常数；$i_t$ 和 $i_{t-1}$ 分别为 $t$ 时刻与 $t$–1 时刻的状态；$I(\cdot)$ 为指示函数，当输入值大于等于 0 时，直接返回输入值，否则为 0。设定时间跨度共有 $n$ 期，当系统从前期基准状态（$i$=0）进入当期突发状态（$i$=1）时，状态转换困难度为 $\gamma\ln(n)$；当保持状态不变或者从突发状态返回基准状态时，转换难度为 0。常数 $\gamma$ 反映了状态转换的阻力，该参数设定越大，就越难以实现从基准状态向突发状态的跃迁。

综合上述两个方面因素，构建全时域的状态成本函数，定义为

$$\text{cost}=\sum_{t=1}^{n}(\tau_t-\sigma_t(q_t)) \tag{7}$$

其中，$q_t$ 为 $t$ 时刻状态。假设 $q_t$ 按时间排列构成状态序列 $Q$，若存在某个状态序列 $Q^*$ 能够使成本函数 cost 取最小值，则 $Q^*$ 为最优状态序列，因为该状态序列能够最好地解释焦点事件的发生及状态转换情况。考虑函数 cost 最小化为优化目标，每个时刻 $q_t$ 都存在两种可能状态（基准状态和突发状态），最优状态 $Q^*$ 是一种动态规划问题，采用 Viterbi 算法可以快速求得最优状态序列 $Q^*$[13]。

以媒介活跃度相对落差 rgap 为分析数据，采用突发性检测算法，可以探测哪些主题及在什么时刻指标 rgap 发生突然性增大。以主题 $i$ 为例，以取整后的主题指标 $\text{rgap}_{it}$ 代表焦点事件数量 $r_t$，所有主题的指标 $\text{rgap}_{it}$ 和代表全体事件数量 $d_t$，通过算法可以求解出该主题的突发状态时刻及突发强度。对于全部 $n$ 个主题，通过 $n$ 次探测分析，可以分别得到各个主题的突发状态分布。基于主题媒介落差的突发性分布，开展系统性综合分析，可以形成媒介比较视角的学科新兴主题识别方法与决策支持。

## 5　实证分析

### 5.1　数据来源及处理

选取情报学为实证学科领域。数据包括文献元数据与 altmetrics 指标两个部分，采集处理过程亦由对应的两个阶段构成，以文献 DOI（digital object identifier）号为线索实现两个部分数据的一对一匹配。

首先，从引文数据库 Web of Science（WoS）中检索情报学领域近年文献，并从中获取所需引文指标及相关元数据。检索思路为通过代表性期刊，获取情报学领域相关学术文献。参考相关实证文献，经过比对筛选，选取 5 种代表性期刊为分析对象，包括 *Information Processing & Management*、*Journal of the Association for Information Science and Technology*、*Scientometrics*、*Journal of Informetrics*、*Information & Management*，它们在情报学科中具有较高影响力与期刊影响因子，基本能够代表本学科的前沿成果动态。为了适应动态分析需要，检索文献跨度为 7 年（2013—2019 年）。考虑到检索时由于收录

机制 2020 年文献数据存在缺失，故未包含 2020 年相关数据。以 WoS 数据库为检索源，通过高级检索界面，检索上述 5 种期刊在 7 年间发表的论文，选取文献类型为 article，查询时间为 2020 年 8 月，得到文献数据共计 4969 条。从中抽取相关信息，包括标题（TI）、摘要（AB）、期刊（SO）、年份（PY）、DOI 号（DI）、被引（TC）。

其次，获取样本文献的 altmetrics 指标数据。2011 年成立的网站 altmetric.com 是最早的 altmetrics 数据提供商之一，具有应用广泛、免费、开源、覆盖率高、指标丰富诸多优点，尤其面向研究人员开放提供数据接口，能够满足本文数据需求。以文献 DOI 号为线索，通过 altmetric 网站 API 接口，编写 python 程序采集数据，对 JSON 格式数据进行解析得到指标数据。altmetric 网站无法监测覆盖所有文献，由此删去缺失 altmetrics 指标的文献记录，经过数据清洗，最终得到匹配成功数据 2740 条。数据包含多种类型的 altmetrics 指标，包括博客（posts count）、推特（tweets count）、阅读（readers count）等反映特定类型及来源的计数指标，还有在各项指标基础上构建的加权总指标（altmetric mention score），能够反映学术文献的总体影响力。

## 5.2 模型设定及指标特征

基于动态模型 DTM 的主题抽取是分析基础。以论文摘要为文本对象，采用 python 软件包 gensim，调用 LdaSeqModel 函数，构建主题模型。其中，主题数是模型超参数，需要人工设定。确定最优主题数的定量方法主要有 perplexity 和 coherence 两种。研究发现 perplexity 方法对主题模型的识别效果不佳，与专家判断的主题结果相差较大，甚至呈现负相关关系[35]，而 coherence 方法更多地考虑了主题上下文，如单词的共现关系，能够较大程度弥补 perplexity 方法的不足。基于此，采用 coherence 方法确定模型主题数量。图 3 给出了设定不同主题数量的 DTM 模型 coherence 得分，通过比较分析选取得分高的主题数，最终设定模型主题数为 36。

采用设定模型分析摘要文本，最终构建并识别出情报学领域 36 个主题，如表 1 所示。这些主题基

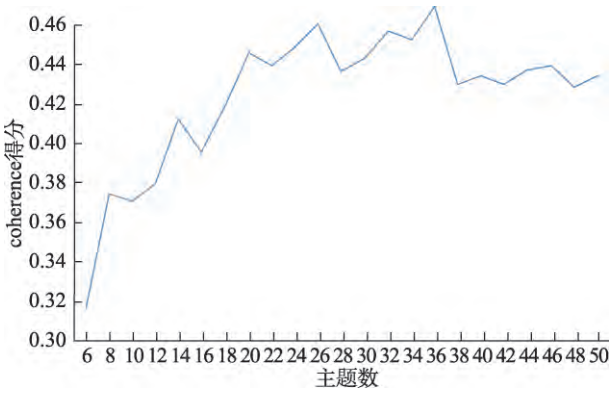


图 3 不同主题数设定的模型 coherence 得分

**表 1 主题及代表性单词或词组**

| 主题 | 代表性单词或词组 | 主题 | 代表性单词或词组 |
|---|---|---|---|
| 0# | information retrieval, collaborative, similarity | 18# | software, books, citation, impact, indicators |
| 1# | bibliometric, sleeping beauties, impact | 19# | co-citation, collaboration, Eugene Garfield |
| 2# | h-index, e-index, academic influence | 20# | machine learning, academic, models |
| 3# | productivity, peer review, research evaluation | 21# | humanities, citation, social sciences |
| 4# | gender, scientific productivity, efficiency | 22# | impact factor, journals, authorship, citations |
| 5# | plagiarism, impact, misconduct, retraction | 23# | field normalization, impact indicators, indicator |
| 6# | collaboration, globalization, interdisciplinary | 24# | natural language processing, semantic analysis |
| 7# | evaluation, performance, co-word analysis | 25# | information behavior, information seeking |
| 8# | peer review, bibliometrics, citation | 26# | webometrics, bibliometrics, informetrics |
| 9# | bibliometrics, countries, China, impact | 27# | bibliometrics, evaluation, impact, assessment |
| 10# | co-authorship, scientific collaboration, network | 28# | innovation, technology, science, patent |
| 11# | productivity, university, rankings, performance | 29# | journal impact factor, citation, bibliometric |
| 12# | altmetrics, twitter, social media, f1000 | 30# | bibliometric analysis, big data, global trends |
| 13# | networks, nanotechnology, collaboration | 31# | network analysis, citation, interdisciplinarity |
| 14# | open access, journals, institutional repositories | 32# | scopus, wos, google scholar, coverage |
| 15# | scientometrics, citations, prediction | 33# | wikipedia, intellectual structure, social network |
| 16# | information, behavior, services, library | 34# | interdisciplinary, matthew effect, correlation |
| 17# | university rank, journal rank, bibliometrics | 35# | privacy, online, perceptions, word-of-mouth |

本覆盖了该领域的主要研究内容及对象，具有较清晰语义与概念边界。上述主题集合是识别分析目标对象，有待于从媒介比较视角揭示其新兴特征。

将总指标（altmetric mentions score）代入公式(1)，由此计算出36个主题的相关指标。表2给出了按年份统计的主题指标gap和rgap的均值及方差。可以看出，主题活跃度的媒介落差随着年份而逐步增加，2017年达到最大均值（10.91），是2013年的2倍，之后两年均值下降与指标累积时间短有关；另外，方差呈现大幅波动，印证了该指标时序差异大且不适用于大跨度时域比较的特点。相对指标rgap克服了上述问题，样本整体分布稳定，能够满足算法要求。

**表2　主题层面指标统计概况**

| | | 2013年 | 2014年 | 2015年 | 2016年 | 2017年 | 2018年 | 2019年 |
|---|---|---|---|---|---|---|---|---|
| gap | 均值 | 4.93 | 5.58 | 9.69 | 6.61 | 10.91 | 9.70 | 7.59 |
| | 方差 | 8.78 | 16.06 | 61.76 | 16.41 | 38.70 | 16.39 | 23.11 |
| rgap | 均值 | 1.02 | 1.01 | 0.99 | 1.02 | 0.97 | 1.01 | 1.00 |
| | 方差 | 0.30 | 0.44 | 0.36 | 0.14 | 0.14 | 0.14 | 0.29 |

## 5.3　分析结果

### 5.3.1　*初步分析*

采用指标rgap对主题新兴特征进行整体初步分析。新兴特征可能在不同时间长短下呈现不同涌现规律，由此选取2019年、2016—2019年、2013—2019年3个时间段，分别代表短期、中期和长期，开展比较分析，以揭示主题在不同时间跨度下的动态呈现。分析结果如图4所示，其中2019年媒介活跃度相对落差值绘于右侧纵坐标轴，其余指标绘于左侧纵坐标轴。从长期看，各主题的rgap指标均值波动不大，以2#（h指数及其变种）、9#（各国学术影响力评估）、23#（学术影响力的学科比较）、8#（同行评议与计量方法）为代表的主题成为情报学科领域的稳定热点。这些长期性主题大多属于情报学领域的基本方法或核心问题，相对成熟且获得持续性创新发展。从中期看，以19#（Garfield学术回顾）、33#（Wikipedia知识挖掘）、5#（学术不端行为）为代表的主题成为近些年情报学领域的阶段性热点。从短期看，以19#（Garfield学术回顾）、5#（学术不端行为）、4#（性别与学术产出）为代表的主题成为2019年该领域的短期涌现焦点。

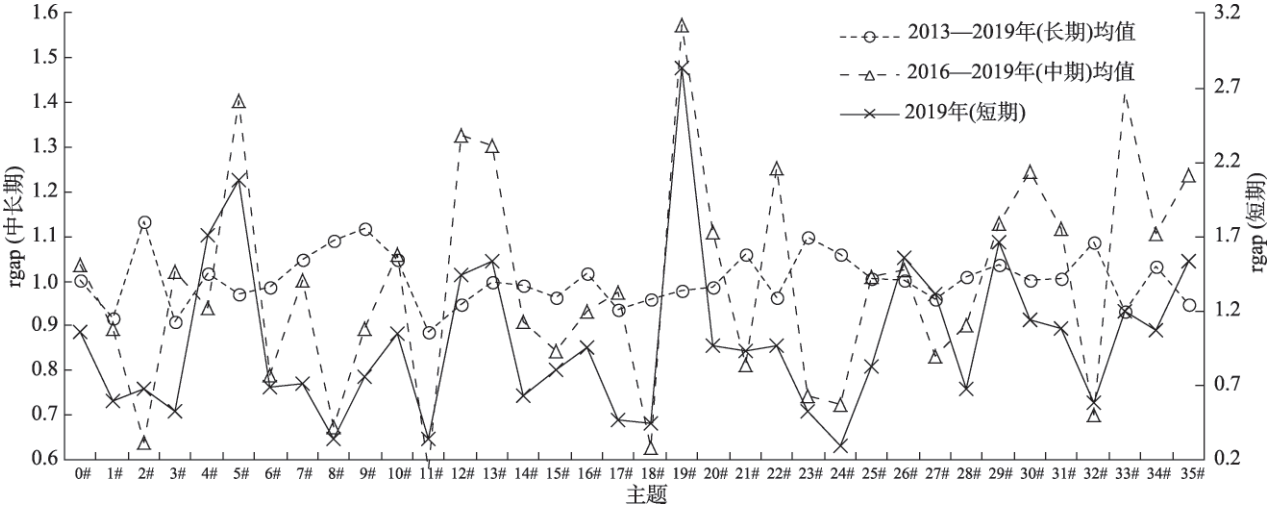


图4　不同时间跨度的主题指标rgap对比分析

可以看出，短期（1年）及中期（4年）分析结果突出了该学科领域的动态性倾向，与长期（7年）分析结果获得的稳定性主题差别较大。这种差异反映了指标rgap在不同时间粒度下的学科动态揭示能力，社交网络（altmetrics）与出版媒介（引文）的活跃度差距能够在中短期被观察和发现，是识别新兴趋势的重要窗口期。通过对比分析，初步展示出指标rgap具有较好的学科动态即时识别能力，相对

于常见的主题模式，对于那些独特主题的涌现更加敏感。

5.3.2 识别分析

为了进一步揭示主题新兴趋势的动态特征与演进过程，采用基于突发性检测算法的新兴主题识别方法，编写python程序，开展实证分析。经过反复尝试，为算法中超参数选取恰当的值，将反映两种状态期望概率比例的参数$s$设定为2，即要求突发状态期望概率是基准状态的2倍以上，还将反映向突发状态跃升困难程度的参数$\gamma$设定为1。图5展示了各主题在不同时点的指标rgap突发特征演化过程，其中横轴代表年份，纵轴代表主题，横条代表指标的突发状态，颜色越深，反映其突发强度越高。后续突发特征演化图（图6~图10）亦采用相同的图形设置。

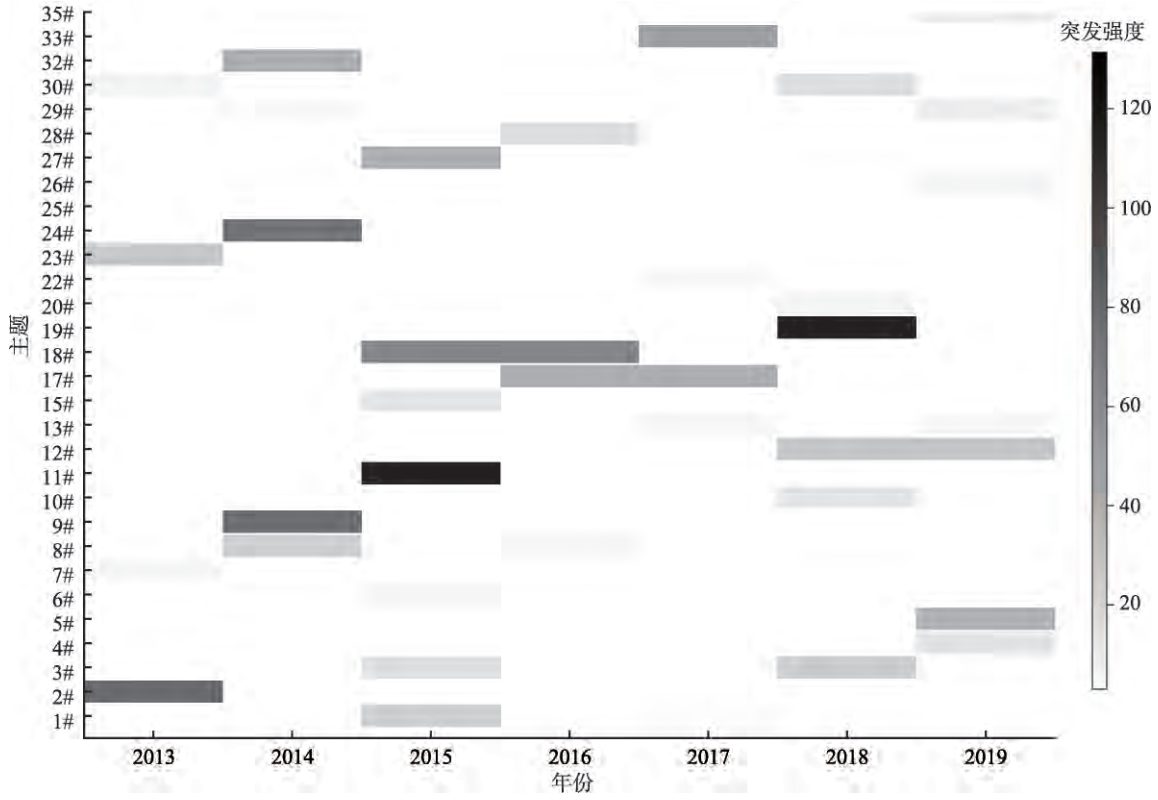


图5 基于指标rgap的主题突发特征演化时序图

从图5可以看出，以11#（2015）、19#（2019）、2#（2013）、9#（2014）、24#（2014）等为代表的主题呈现出强烈的短期涌现特征，在相应年份形成明显的媒介落差。这些典型新兴主题在学术社交媒介形成了远超过出版媒介的关注程度，是近几年应当重点关注的新兴主题。例如，主题11#（大学排名）相关文献共检索出33篇，这些文献对ARWU（Academic Ranking of World Universities）软科排名、CWTS（Centrum voor Wetenschap en Technologische Studies）莱登排名等大学排行榜展开了持续热烈讨论，从2013年开始按年度排列的该主题文献篇数分别为2、6、7、7、4、4、3。探测算法发现该主题2015年呈现突发状态，与文献分布峰值出现时间亦吻合，该分析结果对后续的持续热度具有预示价值。还有，主题19#（Garfield学术回顾）的突发探测也极具典型性，作为情报学与科学计量学奠基人、引文分析创始人，E. Garfield于2017年逝世，期刊*Scientometrics*在2018年出版纪念专刊，Bornmann、Leydesdorff、McCain、Bar-Ilan、Rousseau、Thelwall、White等本领域知名权威学者纷纷发表专门论文。探测算法能够及时发现此类独特新兴主题，也反映出社交网络对于这类高流行潜质新兴主题的较强感知能力。

在最近的2019年，识别出7个具有潜力的新兴主题。主题13#（以纳米领域为代表的合作网络）、26#（不同类型计量学比较）、29#（期刊影响因子探讨）聚集于本领域重要研究方法工具及核心议题。虽然，上述主题基本属于情报学领域的传统主题，但随着学科发展，传统议题也存在新问题、新

技术、新视角，持续跟踪与深入探索这些重要内容的新发展亦十分重要。相比而言，主题12#（替代计量学）、5#（学术不端行为）、4#（性别与学术产出）、35#（在线网络隐私）更加体现了短期新兴趋势，除替代计量学为近年普遍关注的新兴热点之外，其他主题相对独特或比较小众。“学术不端行为”与“在线网络隐私”显然是近两年刚开始被学界重视并尝试系统探讨的内容，是情报学对于当前社会现实热点问题的学科回应，虽然并非本学科研究的主流范式与核心议题，但议题新颖且意义重要，反映了本学科未来发展的新兴增长点，具有很大发展潜力与学术探讨空间。另外，“性别与学术产出”是视角独特的研究内容，反映了学界对于学术绩效机制不断深化的认知。不论是刚刚涌现的高学术价值新兴主题，还是常规主题的新兴动态，都预示了学科发展的重要方向，基本说明了本文提出方法对于新兴主题识别的有效性。

### 5.3.3　基于不同指标的识别方法对比

为了进一步探讨本文方法的识别能力及优劣势，表3给出了基于不同指标的探测方法效果比较，包括本文基于媒介相对差距rgap的方法、基于引文指标TLC的方法、基于社交媒介关注强度E的方法。引文指标TLC由公式（2）给定，反映了该主题内容受到学者引用的程度。社交媒介关注强度E采用段庆锋等[28]提出的指标定义，即主题关注热度TLA与相关文献数量的比值，反映了主题受到社交媒体关注的相对程度。类似地，将指标TLC和E分别代入本文构建的识别框架及流程，获得不同指标的探测结果。3种方法都采用相同的指标处理（由公式（4）定义）、突发性探测算法及相关参数设定，以实现不同指标结果的可比性。基于指标TLC和E的方法2和方法3采用了单媒介（分别为文献媒介与社交媒介）的探测思路，而方法1采用跨媒介（文献媒介和社交媒介）活跃性比较的探测思路，它们的结果比较有助于揭示本文所提探测方法的可行性及相对优势。

**表3　基于不同指标的识别方法结果比较**

| 指标 | 方法1：基于媒介相对差距rgap的方法 | 方法2：基于引文指标TLC的方法 | 方法3：基于社交媒介关注强度E的方法 |
|---|---|---|---|
| burst次数 | 36 | 17 | 31 |
| burst主题数 | 30 | 15 | 27 |
| burst强度均值 | 34.45 | 20.97 | 30.38 |
| burst时间均值 | 3.19 | 3.76 | 3.29 |

从召回能力（包括burst（突发）次数和burst主题数）和敏感度（包括burst强度均值和burst时间均值）两个方面，比较不同方法识别结果的差异。①召回能力。burst次数和burst主题数指标反映了尽可能多地对突发性新兴事件加以标记的识别能力，数值越大，意味着遗漏潜在识别目标的可能性越低。相同识别框架下，方法1给出了最多的burst次数和burst主题数，召回识别能力表现最佳。②敏感度。从强度和时间两个维度体现识别敏感性。burst强度均值为所有burst强度的平均值，burst强度越大，说明识别方法对主题波动的反应越强烈，体现了对潜在新兴事件的敏感性。burst时间均值为所有burst发生时间的平均值，这里时间取值采用序数，以2013年为第0年，依次赋值，直到2019年为第6年。指标burst时间均值从时间维度反映了识别敏感度，取值越小，意味着越能够及时发现潜在目标。新兴主题的即时性对于科技战略决策具有重要价值，不但要求对主题内容的敏感性，更强调时间上的及时性。通过表3可以看出，方法1具有最大的burst强度均值（34.45）和最小的burst时间均值（3.19），在上述两个维度表现俱佳。总之，不论是召回能力还是敏感度，方法1都优于其他两种对比方法。

下文从主题个体微观层面，进一步展开方法对比分析。图6和图7分别展示了采用方法2和方法3获得的主题突发特征演化时序图，分别将其与图5所示方法1进行比对，可以更加直观细致地观察不同方法的差异，具体如表4所示。

本文方法1的识别效果优于方法2。①召回能力方面。方法1比方法2多识别出1倍数量主题，包括3#、4#、5#等共16个主题，而仅少识别出主题0#。基于引文指标的方法2虽然探测到了以12#（替代计量学）、19#（Garfield学术回顾）、20#（机器学习）等为代表的新兴议题，但是显然召回率不高，而且漏掉了5#（学术不端行为）、4#（性别与学术产出）、35#（在线网络隐私）等独特新兴议题。通过比较可以发现，本文方法1的目标主题召回能力显然高于方法2，媒介比较指标rgap充分利用了社交媒介的信息活跃优势，对新兴主题具有强敏感性。②及时性方面。表4给出了主题突发事件被探测发现的时间比较。方法1及时地发现了5个主题突发事件，包括2#、8#、18#、28#、30#，比方法2给出的时间分别提前了3、2、3、1、1年。方法1只有主题9#的发现时间晚于方法2。整体上，方法1的探测及时性比方法2表现更好。

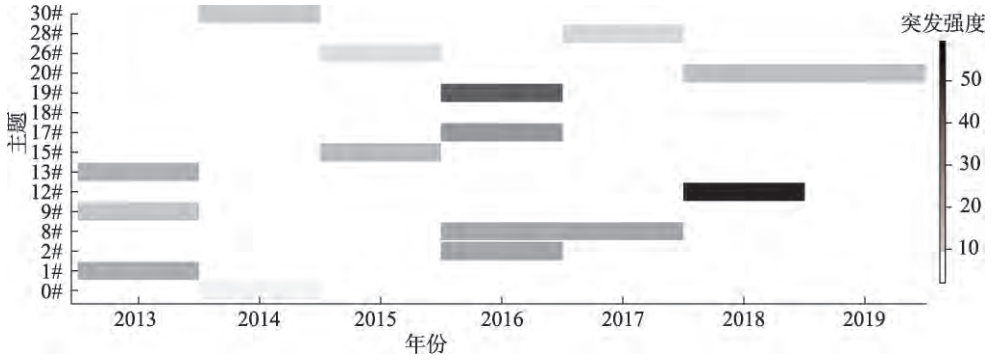


图6 基于引文指标TLC的主题突发特征演化时序图(方法2)

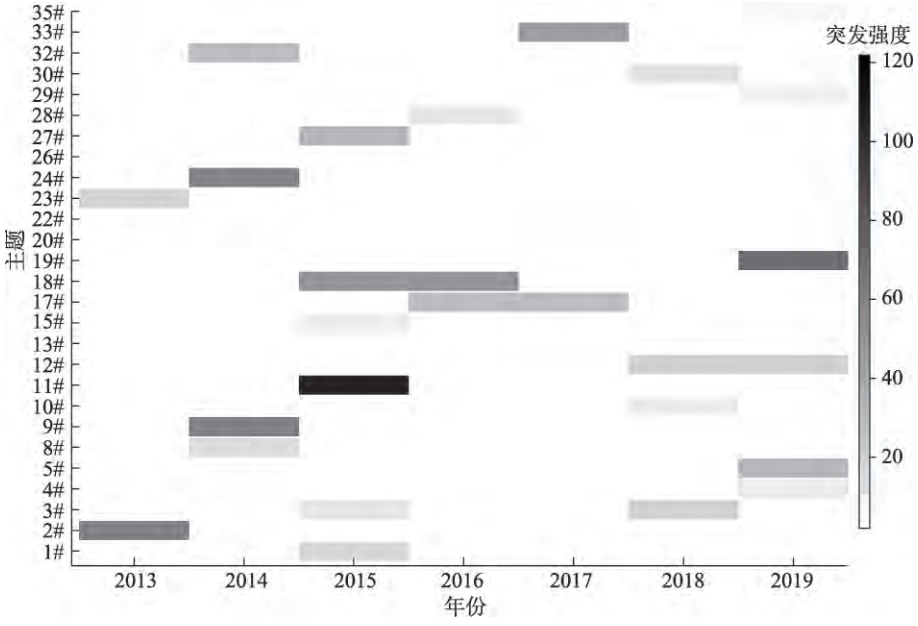


图7 基于社交媒介关注强度E的主题突发特征演化时序图(方法3)

**表4 主题突发事件发现年份对比**

| 方法1与方法2的对比 | | | 方法1与方法3的对比 | | |
|---|---|---|---|---|---|
| 主题 | 基于媒介差距rgap的方法1 | 基于引文指标TLC的方法2 | 主题 | 基于媒介差距rgap的方法1 | 基于社交媒介关注强度E的方法3 |
| 2# | 2013 | 2016 | 19# | 2018 | 2019 |
| 8# | 2014 | 2016 | | | |
| 18# | 2015 | 2018 | | | |
| 28# | 2016 | 2017 | | | |
| 30# | 2013 | 2014 | | | |
| 9# | 2014 | 2013 | | | |

本文方法1与方法3的探测结果几乎相同，召回能力相当，但方法1仍略呈现时间优势。通过图5和图7的对比以及表4内容分析，两种方法探测出几乎相同数量的主题突发事件，而除了主题19#之外，都给出了几乎相同的突发事件发生时间。方法1和方法3探测结果的高度相似在一定程度上与指标定义方式有关，指标rgap与E都为相对指标，反映了信号的相对波动性，而指标TLC是绝对指标，反映了信号的绝对波动性。如上文所述，主题19#（Garfield学术回顾）并非学科常规议题，2018年就已涌现多篇相关文献，方法3虽然探测到了2019年该主题爆发的强烈信号，但并未及时发现更早的2018年波动信号，而方法1显然更加敏感地捕捉到了早期新兴信号，有助于更好地在新兴主题生命周

期早期阶段给出启示。

总之，上述比较分析说明跨媒介比较指标表现出良好的识别效果。具体地，媒介比较指标比引文指标的召回能力与及时性明显占优，与社交媒介指标相比召回能力相当，及时性方面表现略好。本文构建的媒介比较指标 rgap 在利用社交媒介传播优势基础上，综合考虑不同媒介信号的波动时间差，以期进一步提高探测方法对新兴信号的及时敏感度。对比试验也较好地支撑了通过媒介比较探测新兴主题的分析思路，基本证明了指标构建的合理性与识别方法的可行性。

### 5.3.4　不同类型 altmetrics 指标的适用性

学术社交媒介类型多样，且对于学科动态响应可能存在较大差异性。上述实证分析使用了 altmetrics 总指标（公式(1)），有必要进一步考察不同类型 altmetrics 指标的方法适用性。综合考虑数据可得性与覆盖率，选取适合于学科新兴主题识别场景的 3 个典型指标：推特（tweets）、博客（posts）、阅读（readers）。推特是世界范围重要的短文本社交网络，传播流行能力强，具有反馈迅速的优势，通过推特中学科主题的提及（mention）数量，有助于捕捉学科新兴动态；altmetric 平台实时抓取全球 15000 多博客内容，长文本内容更有利于全面展示学术细节与上下文，通过博客对学科主题的提及情况，可以刻画网络对学科动态的深度关注；聚焦学科主题的读者及阅读数量更大程度上反映了用户的学术兴趣，高阅读数量可能意味着以较大概率形成后续文献引证[36]。

将上述 3 个指标分别代入识别算法，开展主题识别分析，结果如图 8~图 10 所示。通过比较可以发现，基于推特（图 8）和博客（图 9）指标的实例结果更为相近，而且从识别数量与突发演化时间上看都与基于总指标的分析结果（图 5）十分相似。可以看出，上述两种指标（推特和博客）具有很强的探测识别能力，比采用总指标实例得到的突发性主题略多，而且能够更早地探测出部分主题的新兴状态，如主题 10#、17#、26#。阅读指标与其他类型指标相比表现独特，基于阅读指标的识别结果（图 10）与其他实例差异较大，探测出的主题突发状态数量只是其他实例的约 2/3，尤其在最近的 2019 年只探测出 2 个突发性主题，识别出的主题新兴演化过程也差别较大。阅读指标分析的相对低效率很大程度上应与数据来源有关，记录阅读数量或者标记读者数量的网站平台基本以学术社交为专门服务目的，如 Menedely、CiteULike 等，用户相对小众而专业化，其用户数与流行能力都无法与面向大众的通用型社交平台相比，数据覆盖率的不足很大程度上制约了其对新兴动态的分析能力。

总之，通过不同类型 altmetrics 指标的比较分析，验证了该识别方法具有良好的指标扩展能力与适用性。除了阅读指标分析能力稍弱，其他指标（包括总指标、推特、博客）都获得了基本稳定一致的识别结果。本文的研究说明，推特与博客数据更加适合于动态性强的新兴主题识别场景，高覆盖

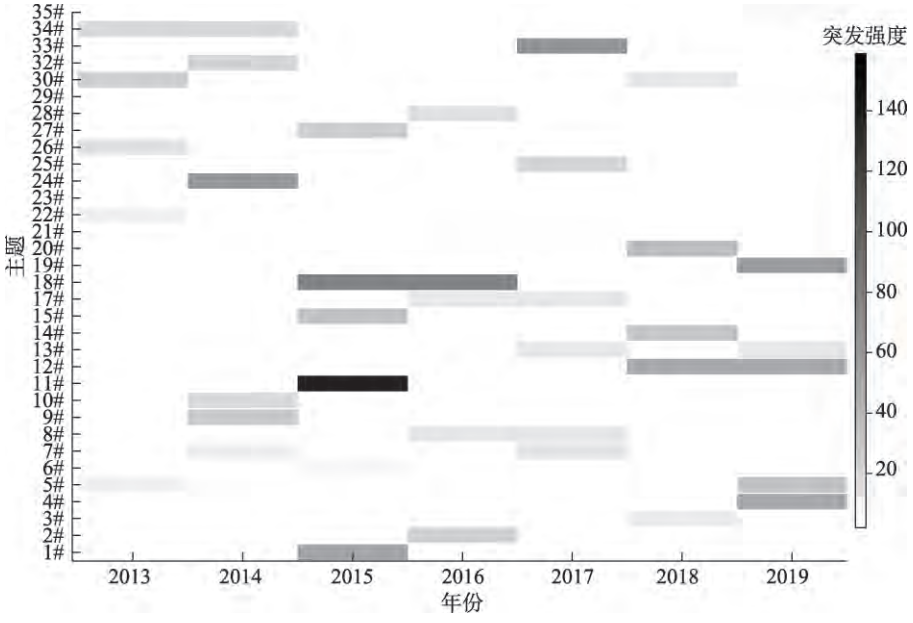


图 8　采用推特指标（tweets）的主题突发特征演化时序图

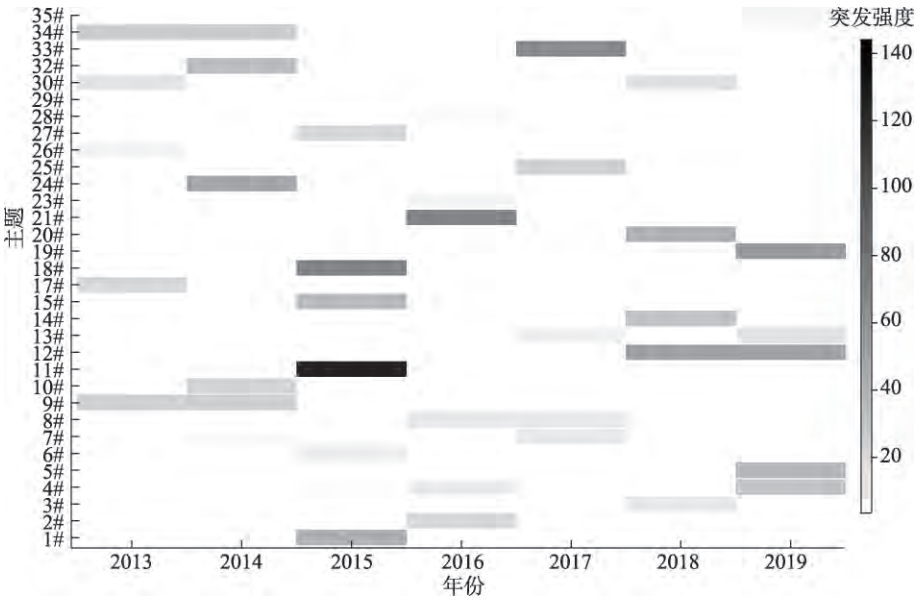


图9 采用博客指标(posts)的主题突发特征演化时序图

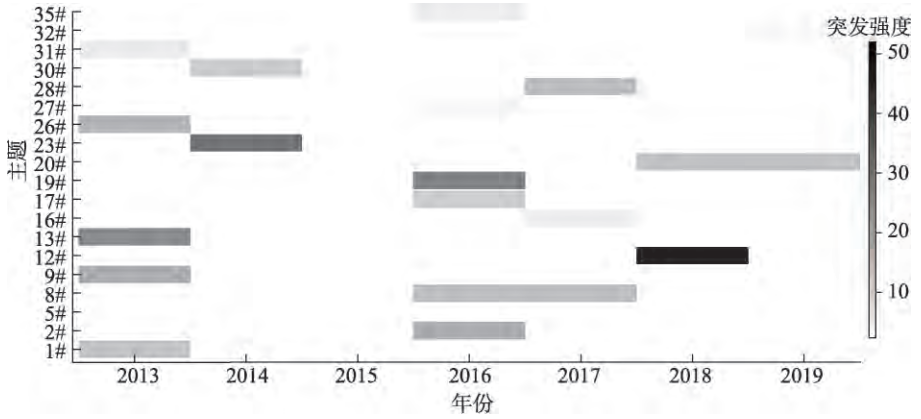


图10 采用阅读指标(readers)的主题突发特征演化时序图

率与高流行度显著的altmetrics指标类型更有利于识别方法获得表现突出的探测结果。

## 6 结 论

及时发现高价值潜力的学科新兴主题是科技情报领域的重要议题。altmetrics虽然已成为情报学领域热点，但在学科前沿趋势探测应用方面相对研究深度不足。本文充分发挥社交网络对流行主题的及时敏感反馈优势，提出媒介比较视角的学科新兴主题识别方法，融合altmetrics和引文数据，设计聚焦于媒介相对落差rgap的学科新兴主题探测算法，并通过实证分析检验方法的有效性与适用性。

本文提出的识别方法基于两种不同媒介：社交网络媒介与学术出版媒介。按照生命周期理论，相对高增长是识别主题新兴状态的重要特征，但在不同媒介的表现呈现差异化。由此，通过altmetrics指标与引文的差距刻画主题的媒介比较，并采用突发检测算法探测媒介落差的突发性涌现，以揭示学科新兴主题的新兴状态。该方法融合多源数据，将altmetrics指标和文献计量指标相结合，能够利用引文指标的滞后性，并充分发挥社交网络的流行敏感性优势；该方法融合多元语义，altmetrics指标是不同社交事件的加总刻画，其丰富语义源于社交网络“群体智能”机制，有助于提升学科新兴主题的决策支持能力；该方法是动态分析，可横纵比较的主题层面相对指标的构建与修正，能够为时序比较提供基准，而突发性检测算法的选用，能够从时序角度细粒度地揭示主题的新兴演化全局过程。

本文构建了面向主题的altmetrics指标，为刻画主题属性提供了定量化基础。该指标利用主题模型

LDA，依据文档-主题概率分布，将针对单篇论文的 altmetrics 指标映射至相应主题层面，构成聚焦于主题的学术社交事件计量指标，拓展了 altmetrics 指标的应用范畴。

实证研究检验了本文方法的有效性。实证说明本文方法具有及时发现短期涌现新兴学科主题的探测能力，以及对于高价值潜力新颖主题的良好敏感性。值得注意的是，比较分析说明基于媒介比较指标的识别方法表现良好，比如，媒介比较指标比引文指标的召回能力与及时性明显占优，与社交媒介指标相比召回能力相当，及时性方面表现略好。对比试验也较好地支撑了通过媒介比较探测新兴主题的合理性。另外，不同类型 altmetrics 指标的适用性也得到了验证，除阅读指标外，采用总指标、推特指标、博客指标的实例都得到良好且基本稳定的结果，说明选择高覆盖度与高流行度的 altmetrics 指标（如推特、博客）更有利于发挥并体现社交媒介数据对于新兴主题的分析优势。

本文的实证研究建立在 altmetric 网站数据基础之上，后续研究有待于进一步扩展数据来源与类型，检验识别方法的数据扩展性。虽然 altmetric 网站数据具有代表性且应用广泛，但相对于成熟、规范、高覆盖率的文献计量数据库，altmetrics 数据平台还处于快速发展阶段，不同平台的数据来源、分布、采集途径差异较大，而且指标类型多样，甚至有的依赖于特定平台，平台数据内涵与外延的差异化给研究带来局限。另外，本文只利用了指标型数据，而学术社交网络包含了大量语义丰富的文本信息，蕴含了更加细腻的用户学术观点及态度偏好等，这些非结构化数据的深度挖掘与融合分析是有待深入开展的重要研究方向。

## 参 考 文 献


[1] 刘小玲, 谭宗颖. 新兴技术主题识别方法研究进展[J]. 图书情报工作, 2020, 64(11): 145-152.

[2] Xu S, Hao L Y, An X, et al. Emerging research topics detection with multiple machine learning models[J]. Journal of Informetrics, 2019, 13(4): 100983.

[3] 杨金庆, 陆伟, 吴乐艳. 面向学科新兴主题探测的多源科技文献时滞计算及启示——以农业学科领域为例[J]. 情报学报, 2021, 40(1): 21-29.

[4] Yang S L, Xing X, Qi F, et al. Comparison of academic book impact from a disciplinary perspective: an analysis of citations and altmetric indicators[J]. Scientometrics, 2021, 126(2): 1101-1123.

[5] Bornmann L, Haunschild R, Adams J. Do altmetrics assess societal impact in a comparable way to case studies? An empirical test of the convergent validity of altmetrics based on data from the UK research excellence framework (REF)[J]. Journal of Informetrics, 2019, 13(1): 325-340.

[6] Ortega J L. Proposal of composed altmetric indicators based on prevalence and impact dimensions[J]. Journal of Informetrics, 2020, 14(4): 101071.

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

[8] Xu H Y, Winnink J, Yue Z H, et al. Multidimensional Scientometric indicators for the detection of emerging research topics[J]. Technological Forecasting and Social Change, 2021, 163: 120490.

[9] 宋欣娜, 郭颖, 席笑文. 基于专利文献的多指标新兴技术识别研究[J]. 情报杂志, 2020, 39(6): 76-81, 88.

[10] 曹艺文, 许海云, 武华维, 等. 基于引文曲线拟合的新兴技术主题的突破性预测——以干细胞领域为例[J]. 图书情报工作, 2020, 64(5): 100-113.

[11] 刘敏娟, 张学福, 颜蕴. 基于核心词、突变词与新生词的学科主题演化方法研究[J]. 情报杂志, 2016, 35(12): 175-180.

[12] 白敬毅, 颜端武, 陈琼. 基于主题模型和曲线拟合的新兴主题趋势预测研究[J]. 情报理论与实践, 2020, 43(7): 130-136, 193.

[13] Kleinberg J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003, 7: 373-397.

[14] Yang Z L, Zhang W J, Yuan F, et al. Measuring topic network centrality for identifying technology and technological development in online communities[J]. Technological Forecasting and Social Change, 2021, 167: 120673.

[15] Breitzman A, Thomas P. The Emerging Clusters Model: a tool for identifying emerging technologies across multiple patent systems [J]. Research Policy, 2015, 44(1): 195-205.

[16] 余波, 赵蓉英. Altmetrics Top100 论文的演进特征及影响因素分析[J]. 现代情报, 2020, 40(7): 134-143, 151.

[17] 邱均平, 余厚强. 论推动替代计量学发展的若干基本问题[J]. 中国图书馆学报, 2015, 41(1): 4-15.

[18] Khan N, Thelwall M, Kousha K. Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics[J]. Scientometrics, 2021, 126(4): 3621-3639.

[19] 李小涛, 金心怡. 基于 Altmetrics 的《科学计量学》研究热点与前沿分析[J]. 现代情报, 2019, 39(1): 153-160.

[20] 迟培娟, 陈挺, 宋秀芳, 等. 基于 Altmetrics 指标识别的研究热点对比分析——以生物学领域为例[J]. 数字图书馆论坛, 2019 (5): 37-41.

[21] 秦奋, 高健. 基于 Scopus 数据库的 Altmetrics 指标与引文计量对比分析[J]. 情报学报, 2019, 38(4): 377-383.

[22] 王菲菲, 刘明. Altmetrics 视角下的交叉学科研究前沿探测——以医学信息学领域为例[J]. 情报学报, 2020, 39(10): 1011-1020.

[23] 牌艳欣, 李长玲, 刘运梅. 基于 z 指数的 AAS 高关注度学科研究主题识别[J]. 情报资料工作, 2019, 40(6): 30-37.

[24] Small H, Boyack K W, Klavans R. Identifying emerging topics in science and technology[J]. Research Policy, 2014, 43(8): 1450-1467.

[25] 王贤文, 毛文莉, 王治. 基于论文下载数据的科研新趋势实时探测与追踪[J]. 科学学与科学技术管理, 2014, 35(4): 3-9.

[26] Drongstrup D, Malik S, Aljohani N R, et al. Can social media usage of scientific literature predict journal indices of AJG, SNIP and JCR? An altmetric study of economics[J]. Scientometrics, 2020, 125(2): 1541-1558.

[27] Holmberg K, Hedman J, Bowman T D, et al. Do articles in open access journals have more frequent altmetric activity than articles in subscription-based journals? An investigation of the research output of Finnish universities[J]. Scientometrics, 2020, 122(1): 645-659.

[28] 段庆锋, 潘小换. 利用社交媒体识别学科新兴主题研究[J]. 情报学报, 2017, 36(12): 1216-1223.

[29] Ortega J L. The life cycle of altmetric impact: a longitudinal study of six metrics from PlumX[J]. Journal of Informetrics, 2018, 12(3): 579-589.

[30] Blei D M, Lafferty J D. Dynamic topic models[C]// Proceedings of the 23rd International Conference on Machine Learning. New York: ACM Press, 2006: 113-120.

[31] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[32] Xu S, Hao L Y, Yang G C, et al. A topic models based framework for detecting and forecasting emerging technologies[J]. Technological Forecasting and Social Change, 2021, 162: 120366.

[33] Torres-Salinas D, Arroyo-Machado W, Thelwall M. Exploring WorldCat identities as an altmetric information source: a library catalog analysis experiment in the field of Scientometrics[J]. Scientometrics, 2021, 126(2): 1725-1743.

[34] Moed H F, Glänzel W, Schmoch U. Handbook of quantitative science and technology research: the use of publication and patent statistics in studies of S&T systems[M]. Dordrecht: Springer, 2005.

[35] Chang J, Boyd-Graber J, Gerrish S, et al. Reading tea leaves: how humans interpret topic models[C]// Proceedings of the 22nd International Conference on Neural Information Processing Systems. New York: ACM Press, 2009: 288-296.

[36] Akella A P, Alhoori H, Kondamudi P R, et al. Early indicators of scientific impact: predicting citations with altmetrics[J]. Journal of Informetrics, 2021, 15(2): 101128.

（责任编辑 魏瑞斌）