基于AToT模型的中国科技部官方微信主题量化分析

张昊东,赵立新

中国科协创新战略研究院,北京 100863

摘要 中国科学技术部已将微信等社交媒体应用于科技管理,“锐科技”是科技部官方微信公众号。与传统科技类报刊、期刊的编辑手段不同,“锐科技”微信公众号在选稿和推送上有自身特点。采用后验概率主题模型对“锐科技”和其子公众号“锐动源”的稿件内容进行量化分析和挖掘,揭示了科技部官方微信选稿内容随时间变化的规律;通过研究科技工作者阅读偏好,展示了文章主题与读者浏览量之间的关系,有助于提高科技类新媒体的编辑和水平,增强科技宣传效果,更好地理解中国科技管理的思路。

关键词 微信公众号编辑;主题模型;锐科技;锐动源

微信是深圳市腾讯计算机系统有限公司开发的新一代即时通信产品,和Facebook、Twitter和Whatsapp一样,是用户量最多的手机应用程序之一。第1个版本于2011年1月发布,是一款免费的跨平台产品,支持Android、iPhone、BlackBerry、Symbian和Windows操作系统。微信提供短信、即时语音、支付、视频会议、游戏、照片和视频分享等服务,截至2015年第一季度,已覆盖超过90%的中国智能手机;2017年,微信每月活跃用户超过9.8亿[1-3]。用户可以注册官方微信公众号并将文章推送给订阅者,与订阅者进行交流。截至2016年1月,政府官方微信公众号超过10万[4]。机构申请的微信官方账号,逐渐成为政府、银行、酒店和医院为人们提供服务的重要平台。

近年来,中国科学技术部(以下简称科技部)逐渐把社交媒体应用到科技管理中。科技部的微信官方公众号“锐科技”于2014年12月15日上线,由科技部办公厅和科技日报社具体管理。“锐科技”主要围绕科技政策、法律法规、科技成果、国际合作、咨询服务等方面进行选题。经过几年的运营,“锐科技”在科技工作者中产生了很大的影响力,在科技宣传中发挥了巨大作用。

随着官方微信公众号的发展,有的机构申请多个微信公众号,将这些微信公众号组成微信公众号

收稿日期:2020-06-16;修回日期:2021-08-20

基金项目:中国科协创新战略研究院院长青年基金项目(2020yzjj-008)

作者简介:张昊东,高级工程师,研究方向为数据挖掘、科技政策,电子信箱:hdzhang06@126.com

引用格式:张昊东, 赵立新. 基于AToT模型的中国科技部官方微信主题量化分析[J]. 科技导报, 2022, 40(6): 110-121; doi: 10.3981/j.issn.1000-7857.2022.06.013

矩阵。微信公众号矩阵中的公众号,可以相互进行流量引导,有利于提高宣传效果。人民日报社建立了完善的微信公众号矩阵,包括人民日报、人民日报评论、人民日报政文等,满足受众不同的阅读兴趣和需求。科技部也逐步建立起微信公众号矩阵。围绕“锐科技”,又建立了“锐动源”“社发科技”等多个子公众号。各公众号根据读者的不同,各有侧重。例如,“锐动源”由科技部资源配置与管理司进行管理,于2016年7月发布文章《我们等待着你开怀大笑的那一天》,正式上线,主题是科技经费管理等内容。截至2018年初,粉丝数量超过35000人。本研究将利用后验概率模型,量化分析特定时期科技部官方微信公众号的推送主题,并通过官方微信公众号矩阵中不同公众号文章的阅读次数了解读者的阅读偏好,利用研究的数据和结论促进科技宣传和决策。

1　对微信和文本分析的有关研究

过去,人们通过广播、电视和报纸等传统媒体了解信息和认识世界,方式和手段比较单一。随着信息技术的进步和互联网的普及应用,新媒体信息更新与传递优势明显,人们获取信息的途径更为多元[5],内容传播从静态到动态,从一维到多维转变[6],以互联网为代表的新媒体进行的媒介技术革命改变了信息的共享生态[7]。根据2018年发布的《中国互联网络发展状况统计报告》,中国网民规模达7.72亿,手机网民规模达7.53亿[8],出现了微信、微博、电子书等多种新的媒体传播和内容分发方式。

有研究人员对微信进行了研究,例如,刘佳静等[9-11]研究了图书馆微信公众号的传播力;董玥[12]研究了智库微信公众平台的传播影响力评价方法;有的文章对高校微信公众号建设进行了探讨[13-14];冀芳[15]对学术期刊微信公众平台进行了评价研究;有的学者研究了报纸等媒体的微信公众号[16-17]。微信可以很好的促进民众和政府的沟通,各政府部门出现了一批有影响力的政务微信账号,例如中华人民共和国外交部的“外交小灵通”、国务院国有资产监督管理委员会的“国资小新”,还有各省市的“平安肇庆”“重庆环保”“平安北京”等。夏保国[18]从技术接受模型的角度探究了微信在政务沟通方面的影响机制;蒋天民[19]认为政务微信有利于政府机关抢占舆论阵地、促进政民互动;李宗富[20]用DEMATEL方法对影响政务微信公众号质量的因素进行分析;刘佐[21]研究了微信公众平台新文章的推荐算法;Qiu等[22]研究分析了微信的生命周期;李松丽[23]对微信热点话题进行了聚类分析;还有张昊东等[24]通过微信来研究科技社团的活跃度。但是对政务微信进行主题文本分析的研究还比较少。

很多研究在文本数据挖掘方面进行了尝试。Zhang等[25]利用中文术语相似度算法开发的QA系统,提高了文本检索的性能。Tseng等[26]把聚类分析、文本分割应用到专利分析中。一些研究把CRF用于分析文本[27-29]。LDA(latent dirichlet allocation)是一种生成统计模型,是一个无监督学习的贝叶斯概率模型,可以识别大规模文档集中潜藏的主题信息。LDA模型在表示文档、减少文本维度和挖掘文档中的隐藏信息方面取得了进展,可被用于专利等内容的主题分析[30]。

随着时间的推移,文章的主题会发生变化,但LDA无法体现这一点。Rosen-Zvi等[31]将主题变量引入LDA得到了author-topic(AT)模型。Mimno[32]在AT模型的基础上构建了author-persona-topic(APT)模型。Kawamae[33]提出了author-interest-topic模型(AIT)。Author-recipient-topic(ART)模型主要用于社交网络,可以根据人与人之间的关系揭示主题[34]。Group-topic(GT)模型可以把网络中的实体组成群组,并将词汇聚类成不同的主题[35]。LDA、AT、APT模型以静态的方式发现和分析文本潜在的主题。2006年,Blei等[36]开发了应用高斯时间序列的dynamic topic模型(DTM)。在topics over time模型(ToT)中,文本的主题与随着时间戳变化的连续分布有关,而不是马尔可夫假设。

已经有很多基于贝叶斯概率模型的应用。卢盛祺等[37]在在线视频推荐中应用了LDA,根据用户在网站上的活动获取用户偏好,为观众生成个性化的视频集。廖列法等[38]应用LDA对专利文本语料库进行建模,提取文档主题和主题特征词语矩阵,

以达到降维和提取语义链接的目的。有的研究建立了基于LDA模型的信息检索应用[39],在文档生成过程中,添加文档类别标签Y来计算文档中最相关主题的概率。张金瑞[40]提出了一种基于LDA的弱监督算法VB-LDA(latent dirichlet allocation with vector and bigram),并将其应用于文本分类。Tan等[41]用LDA建模了的新闻数据集,通过Bias标准方法选择了最佳主题数。

基于LDA的应用已广泛应用于社交媒体的数据挖掘和信息分类,但缺点是不能表示主题随时间变化。随着时间的推移,微信公众号文章的主题会随着时间进行变化。为解决这一问题,本研究提出利用AToT(author-topic over time)模型建模,分析科技部微信公众号的数据。

2 研究数据

2.1 科技部官方微信公众号数据集

"锐科技"一般每天向订阅用户推送一次或多次。"锐科技"的主题包括科技政策、科技计划、科技监管、科技评估和科技外交等多方面内容。经过几年的发展,科技部的官方微信公众号发展成为"锐科技"矩阵,包括所属部门的数个官方微信公众号。"锐动源"是科技部官方微信公众号矩阵中的一员,发布科研经费、研究项目等方面的信息。

为了量化研究编辑选取文章的思路和倾向,2016年7月29日至2017年3月3日,收集了"锐科技"发布的所有的856篇文章和"锐动源"发布的所有401篇文章,包含科研项目、科研经费、科技政策和科技计划等大量信息。这些微信文章展示了科技部在科学技术管理方面的指导原则,一些热门文章在微信上被转发扩散,大大增加了点击次数,甚至形成了"爆款"。每篇文章的点击量一定程度上显示了中国科技工作者关心的热点方向。

2.2 对收集的微信公众号文章进行数据清洗

微信公众号文章包含文本、图片、视频等多种格式的信息,在本研究中,仅考虑文本信息。在文本包含的词语和句子中,只有一部分词语表示了有意义的信息并揭示文章主题。在对文章的主题进行建模之前,需要从微信文章的语料中删除不包含主题信息的词语。停用词表中的1598个中文词被删除,例如好像、然后、接近、部分、需要、逐渐、重要、快乐、经常、之后等。

因为汉语在词语之间没有空格,所以与处理英文文本有所不同。本研究采用斯坦福大学自然语言处理小组[42]建立的斯坦福单词分词器,将微信文章的中文文本切分成一系列词语。

3 实验和分析

3.1 AToT模型

在本研究中,将AToT模型[43-44]应用到科技部官方微信公众号数据集,构建了一个新的应用来分析科学技术管理方面的文章。表1列出了本应用模型使用的符号[45]。

表1 AToT模型符号

符号	描述
K	主题数量
M	文档数量
V	词项数量
A	作者数量
N_m	文档m中单词的数量
A_m	撰写文档m的作者数量
a_m	撰写文档m的作者形成的向量
θ_a	作者a的主题概率分布
ϕ_k	主题k的词项概率分布
ψ_k	主题k随时间变化的贝塔分布
$z_{m,n}$	文档m中第n个单词的主题分配
$w_{m,n}$	文档m中第n个单词
$x_{m,n}$	文档m中第n个单词的作者分配
$t_{m,n}$	文档m中第n个单词的时间戳
α	多项分布θ的超参数
β	多项分布φ的超参数

图1为AToT模型表示图,具体生产过程为:

1)对于主题$k\in[1,K]$,抽取多项式服从狄利克雷分布,φ_k~Dirichlet(β);对于每个作者$a\in[1, A]$,抽取θ_a~Dirichlet(α)。

2)对于文档$m\in[1,M]$中的每个单词$n\in[1, N_m]$:

(1)取一个作者,$x_{m,n}$~Uniform(a_m);

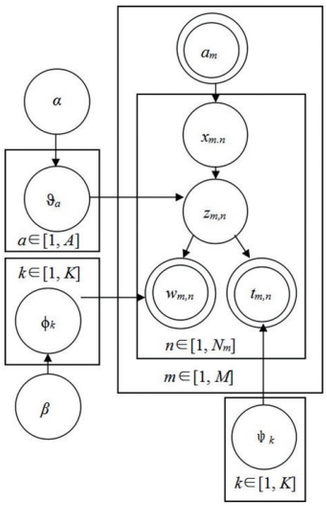


图1　ATOT模型

（2）取一个主题，$z_{m,n}$~Multinomial($\theta x_{m,n}$)；

（3）取一个词语，$w_{m,n}$~Multinomial($\varphi z_{m,n}$)；

（4）取一个时间戳，$t_{m,n}$~Beta($\psi z_{m,n}$,1, $\psi z_{m,n}$,2)。

在本应用中，每篇文章都有一个ID，相当于文章的作者。在本研究中，将collapsed Gibbs采样算法应用于参数估计[46]。ATOT的条件概率是

$$P\left(z_{m,n}=k,x_{m,n}=a\middle|w,z_{\neg(m,n)},x_{\neg(m,n)},t,a,\alpha,\beta,\psi\right)\propto\frac{n_k^{\left(w_{m,n}\right)}+\beta_{w_{m,n}}-1}{\sum_{v=1}^{V}\left(n_k^{(v)}+\beta_v\right)-1}\times\frac{n_a^{(k)}+a_k-1}{\sum_{k=1}^{K}\left(n_a^{(k)}+a_k\right)-1}\times\mathrm{Beta}\left(\psi_{z_{m,n,1}},\psi_{z_{m,n,2}}\right)\quad(1)$$

式中，$n_k^{(v)}$表示将词语v分配给主题k的次数；$n_a^{(k)}$是作者a被分配给主题k的次数；$z_{\neg(m,n)}$，$x_{\neg(m,n)}$是除了$w_{m,n}$之外，分配给所有单词的主题、作者变量。

参数估计中，$n_a^{(k)}$是$\boldsymbol{A}\times\boldsymbol{K}$的矩阵；$n_k^{(v)}$表示$\boldsymbol{K}\times\boldsymbol{V}$的矩阵。$\varphi$和$\theta$的估计如下，

$$\varphi_{k,v}=\frac{n_k^{(v)}+\beta_v}{\sum_{v=1}^{V}\left(n_k^{(v)}+\beta_v\right)}\quad(2)$$

$$\theta_{a,k}=\frac{n_a^{(k)}+\alpha_k}{\sum_{k=1}^{K}\left(n_a^{(k)}+a_k\right)}\quad(3)$$

采用矩估计算法估计Ψ[47]，s_k^2和$\bar{t}_2$是主题k采样的方差和均值，

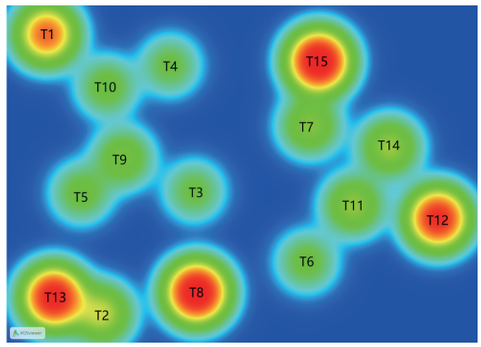
$$\psi_{k,1}=\bar{t}_k\left(\frac{\bar{t}_k\left(1-\bar{t}_k\right)}{s_k^2}-1\right)\quad(4)$$

$$\psi_{k,2}=\left(1-\bar{t}_k\right)\left(\frac{\bar{t}_k\left(1-\bar{t}_k\right)}{s_k^2}-1\right)\quad(5)$$

3.2　科技部微信公众号数据分析结果和讨论

模型不仅能够发现“锐科技”推送微信文章的潜在主题，而且能够挖掘主题随着时间的推移不断变化的规律。

通过计算，从微信文章中获得了有关主题。选取部分区分度较好、意义明确的主题通过图表进行重点分析。图2显示了“锐科技”共15个主题的分布，主题名称见表2。在热力图中，主题8和主题13颜色较红，热度高于周围的主题，这说明“锐科技”大量微信文章的内容与国家科研项目、科技奖励以及科技成果转化有关，聚集度较高。与主题13高度相关的文章包括《科研|促进科技成果转移转化行动有关重点政策摘编(4)》(相似度0.967631)、《创新|2016 众创空间孵化双创梦想》(相似度0.963021)等文章。与主题8高度相关的文章包括



红色代表主题热度较高，绿色代表主题热度较低

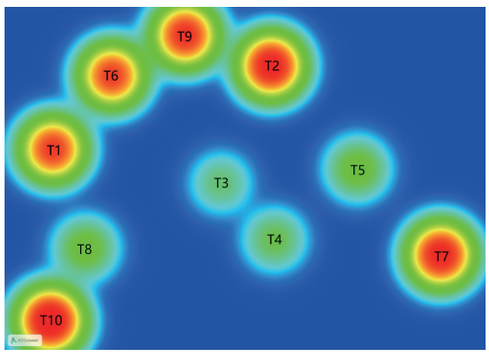
图2　“锐科技”主题分布热力图

表2　“锐科技”部分主题

符号	名称
主题3	科技部党组
主题6	科研资金
主题8	国家科研项目和科技奖励
主题9	航天航空
主题11	国际合作
主题13	科技成果转化
主题14	量子和空间科学

《关于组织国家科技重大专项“大型油气田及煤层气开发”2017年度课题申报的通知》(相似度0.983712)、《重点专项巡礼|聚焦国家重点研发计划之农业领域(林业资源篇)》(相似度0.975418)等文章。主题11和主题14的热度不高,表明关于这2个主题的文章较少。主题11是关于国际合作,包括《中国科技部-美国农业部农业科技合作第十四次联合工作组会议在广东珠海召开》《国际合作|中韩核聚变双边合作联合协调委员会第四次会议在韩国釜山成功召开》等。主题14关于量子和空间科学,包括《前沿|量子通信卫星:让“悄悄话儿”悄悄说》《人类到底可不可以穿越回过去?》等。

图3显示了“锐动源”共分为10个主题,主题名称见表3。在热力图中,主题2和主题10颜色较红,具有较高热度,表明关于国家自然科学基金和资金监督的文章较多。与主题2相关的文章有2017年1月6日的《2017年国家自然科学基金面上项目(医学科学部)指南》(相似度0.991857)、2017年1月5日的《2017年国家自然科学基金面上项目(生命科学部)指南》(相似度0.986079)等文章。和主题10相关的文章有2016年8月11日的《不看是损失!<关于进一步完善中央财政科研项目资金管理等政策的若干意见>问答》(相似度0.984329)、2016年8月8日的《权威|教育部:科学事业费重大项目分三类实行预算制》(相似度0.971951)。主题4的热度较低,表明“锐动源”在2016年7月29日至2017年3月3日关于此话题的文章较少。与主题4相关的文章有2017年2月13日的《第十三回!2017年重点专项预算编制全集之设备费篇》、2017年2月14日的《第十四回!2017年重点专项预算编制全集之——材料费篇》等。



红色代表主题热度较高,绿色代表主题热度较低

图3 “锐动源”主题分布热力图

表3 “锐动源”部分主题

序号	名称
主题1	科技部重点专项2017年度申报
主题2	2017自然科学基金
主题3	国家重点研发计划等项目资金管理
主题4	2017年重点专项预算
主题8	科研经费管理改革
主题10	经费监督

表2和表3列出了部分区分度较好、意义明确的“锐科技”和“锐动源”主题的名称。“锐科技”4个主题(表4—表7)分别是科技部党组、研究经费、国际合作与空间科学。“锐动源”的4个主题(表8—表11)是科技部重点专项2017年度申报、2017自然科学基金、国家重点研发计划等项目资金管理和2017年重点专项预算。对于表4—表11的每个主题,列表给出10个高概率词语,直方图显示2016年7月29日至2017年3月3日主题随时间的分布,给出了拟合贝塔曲线。

表4 “锐科技”主题3(科技部党组)的主题分布

词语	概率
改革	0.04466210
党	0.01339500
中央	0.01047510
政治	0.00840684
制度	0.00682523
人民	0.00573028
总书记	0.00573028
治党	0.00439199
利益	0.00427033
会议	0.00354036

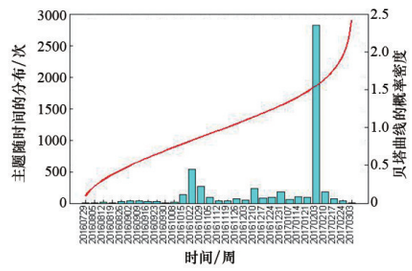


表5　“锐科技”主题6(科研资金)的主题分布

词语	概率
项目	0.03106070
管理	0.02422030
预算	0.01756000
费用	0.01188970
意见	0.01017960
高校	0.00864948
业务费	0.00729940
间接	0.00603933
劳务费	0.00594933
财政	0.00540930

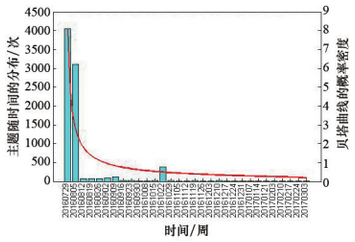
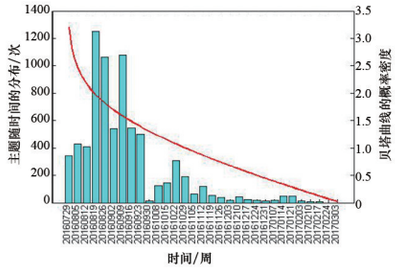


表6　“锐科技”主题11(国际合作)的主题分布

词语	概率
合作	0.03753700
会议	0.01688240
交流	0.01562180
论坛	0.01116120
代表	0.00699152
国际	0.00699152
青年	0.00660364
东盟	0.00640970
聚变	0.00485818
双边	0.00485818



如表6所示,有关国际合作的主题主要出现在从2016年8月19日开始的一周。展开分析,“锐科技”的文章中,有关中国和韩国核聚变双边合作的文章发布于2016年8月26日;中国科技部与美国农业部农业科技合作的文章《中国科技部—美国农业部农业科技合作第十四次联合工作组会议在广东珠海召开》发布于2016年8月25日。由此可见,通过AToT模型对文本进行主题分析,得到的结果和实际文章内容是对应的,显示了很好的准确性,表7—表9的主题也验证了这一点。

表7　“锐科技”主题14(量子和空间科学)的主题分布

词语	概率
量子	0.01666720
卫星	0.01337460
通信	0.00777716
实验	0.00738204
地球	0.00461625
反物质	0.00441869
宇宙	0.00395772
发射	0.00389187
粒子	0.00323335
时间	0.00323335

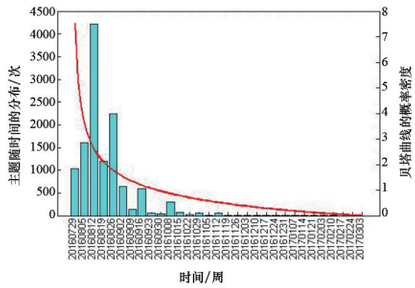


表7显示关于空间科学的主题主要出现在2016年7月29日至2016年9月8日。8月17日,文章《刘延东:锐意进取　勇攀高峰　推动量子科学研究实现跨越式发展》和《前沿|量子通信卫星:让“悄悄话儿”悄悄说》是关于量子通信的。在8月27日,文章《到底什么是反物质?为什么要研究反物质?》解释了为什么要做关于反物质的研究。

表8表明“锐动源”关于科技部重点专项2017年度申报的文章大多出现在2016年9月26日至2016年10月23日。例如,10月13日,《“数字诊疗装备研发”试点专项2017年度项目申报指南+形式审查条件+指南编制专家名单》发布;10月14日,《来了!科技部发布新能源汽车等14个2017年国家重点研发计划项目申报指南》发布。

表8 “锐动源”主题1(科技部重点专项2017年度申报)的主题分布

词语	概率
申报	0.01386930
专项	0.00833422
国家	0.00766313
指标	0.00700966
重点	0.00691316
研发	0.00661492
内容	0.00629913
考核	0.00592194
计划	0.00529036
示范	0.00495702

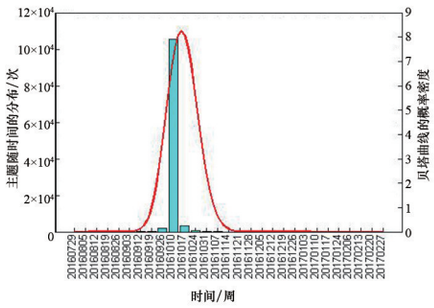


表9 “锐动源”主题2(2017自然科学基金)的主题分布

词语	概率
研究	0.01770740
科学	0.01078010
技术	0.00864786
创新	0.00861838
资金	0.00732135
申请	0.00704622
重点	0.00660405
资助	0.00575901
成果	0.00550354
发展	0.00523824

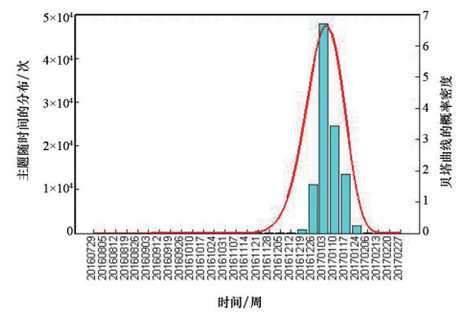


表10 “锐动源”主题3(国家重点研发计划等项目资金管理)的主题分布

词语	概率
资金	0.04319010
项目	0.02069050
预算	0.01825810
财务	0.01308930
验收	0.01202510
费用	0.01172110
重点	0.01141700
支出	0.01020080
概算	0.00837653
科技部	0.00837653

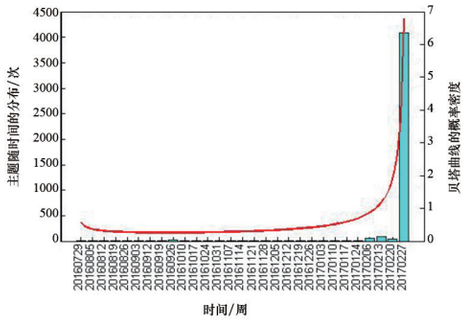


表11 “锐动源”主题4(2017年重点专项预算)的主题分布

词语	概率
预算	0.05807660
课题	0.03756650
设备	0.02444620
支出	0.01660410
劳务费	0.01268300
科目	0.01072250
测算	0.00906363
测试	0.00846039
差旅费	0.00725392
材料费	0.00649987

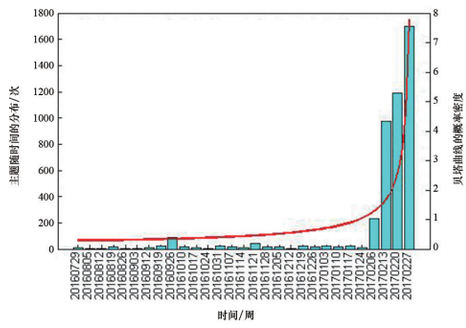


表9显示关于2017自然科学基金的主题主要出现在2016年12月19日至2017年2月5日。2017年1月4日,发布了《2017年度国家自然科学基金项目预算编报须知》《2017年度国家自然科学基金项目限项申请规定》《2017年度国家自然科学基金项目申请须知》。2017年1月5日,发布了《2017年国家自然科学基金面上项目(化学科学部)指南》《2017年国家自然科学基金面上项目(地球科学部)指南》。2017年1月6日,发布了《2017国家自然科学基金面上项目(工程与材料科学部)指南》《2017年国家自然科学基金面上项目(医学科学部)指南》。2017年1月9日,发布了《2017国家自然科学基金重大研究计划项目指南》《2017国家自然科学基金重大项目指南》《2017年国家自然科学基金青年科学基金项目》。

“锐科技”的每篇文章可能包含与15个主题有关的词语,所有856篇文章按单篇文章的阅读量分为2组:一组文章的阅读量大于1000,另一组的阅读量小于1000。每组文章对应的主题信息相加,将每组文章文本的主题信息表示为一个向量,并把向量进行归一化处理。

$$topic_{i\_norm}=\frac{topic_i}{\sum_{j=1}^{15}topic_j} \tag{6}$$

对“锐科技”的2组向量归一化后进行比较,发现2组文章的主题分布不同,这显示中国的科技工作者或者“锐科技”的受众对某些主题更感兴趣,从而致使文章的阅读量不同。把2016年7月29日至2017年3月3日的文章分为2个阶段:第1阶段是2016年7月29日至2016年10月31日(图4),第2阶段是2016年11月1日至2017年3月3日(图5)。研究发现,在不同时期,具有较高阅读量的文章有着不同的主题分布,这表明中国的科技工作者感兴趣的主题随着时间的推移而变化。

2016年7月29日至2017年10月31日,大多数文章是关于主题12科技创新规划、主题14量子和空间科学、主题15重点专项。相比之下,关于主题6科研资金和主题11国际合作的文章阅读量较少。关于主题15重点专项的一些文章具有很高的阅读

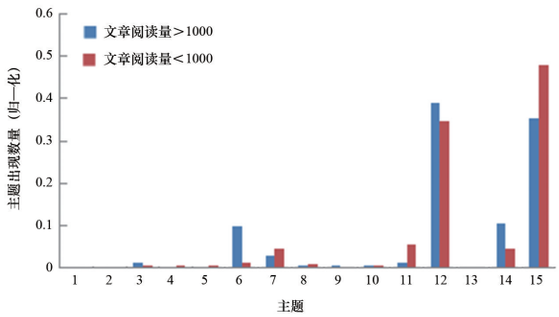


图4　2016年7—10月“锐科技”主题分布

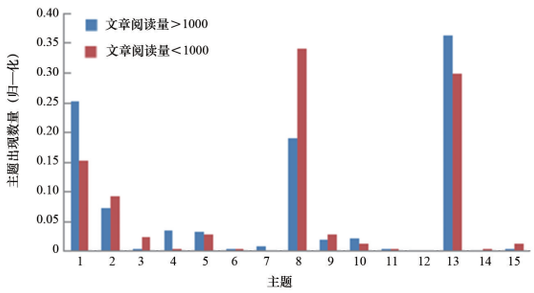


图5　2016年11月至2017年3月“锐科技”主题分布

量,特别是在某些领域,例如:2016年10月14日,文章《通知|科技部关于发布国家重点研发计划新能源汽车等重点专项2017年度项目申报指南的通知》虽然不是当天的头条,但浏览量也达到了2219;2016年10月10日,《科技部发布“干细胞及转化研究”等14个重点专项2017年度项目申报指南》的浏览量达到5718。表明中国科技工作者非常关注2017年重点专项方面的信息,特别是在2016年7—10月。

2016年11月1日至2017年3月3日,主题8国家科研项目和科技奖励很受欢迎。例如,2016年12月2日的文章《关于组织国家科技重大专项“大型油气田及煤层气开发”2017年度课题申报的通知》,浏览量达到5242;2016年11月19日的文章《关于对全国科普工作先进集体和先进工作者拟表彰对象进行公示的公告》,浏览量达到6487。

“锐动源”的每篇文章可能包含与10个主题有关的词语,所有401篇文章按单篇文章的阅读量为2组:一组文章的阅读量大于1000,另一组的阅读量小于1000。每组文章对应的主题信息相加,将每组文章文本的主题信息表示为一个向量,并把向

量进行归一化处理(公式6)。

2016年7—10月,“锐动源”微信公众号主题1科技部重点专项2017年度申报在中国的科技工作者中非常受欢迎(图6)。例如,10月14日,文章《“战略性先进电子材料”重点专项2017年度项目申报指南+形式审查要求+指南编制专家名单》被列为当日的第4条的情况下,阅读量达到2475,远超第3条文章352的阅读量。

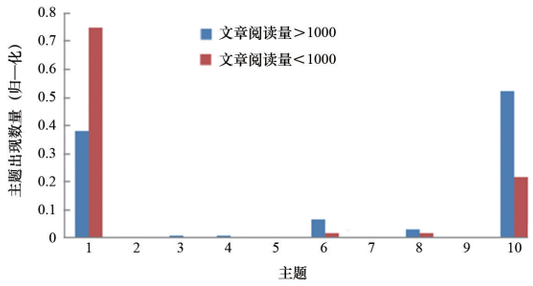


图6 2016年7—10月“锐动源”主题分布

在2016年11月至2017年3月,“锐动源”微信公众号中的主题2自然科学基金、主题7科技发展规划、主题9科技体制改革更受欢迎(图7)。以主题2为例,2017年1月6日,文章《2017年国家自然科学基金面上项目(医学科学部)指南》被列为当日的第2条,仍有5980的阅读量,表明科技工作者非常关注这个话题。

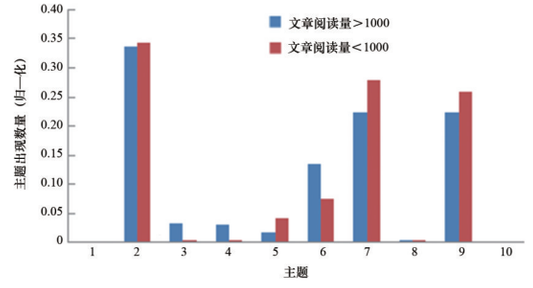


图7 2016年11月至2017年3月“锐动源”主题分布

研究表明,在一段时间内,“锐科技”和“锐动源”的编辑和科技工作者经常关注某些话题,随着时间的推移,他们的兴趣可能随着时间而变化,这也是符合实际情况的。

2016年7月29日至2016年9月2日,这5周的时间分别计算出每一周的“锐科技”和“锐动源”阅读量之和(图8)。除了第1周,“锐动源”出现了阅读量“10万+”文章《图解|一张图看懂中央财政科研项目资金管理新政》,大多数时候,“锐动源”的页面浏览量低于“锐科技”,表明中国的科技工作者不仅仅喜欢阅读有关科技资源配置与管理的文章,还喜欢其他科技主题的文章。

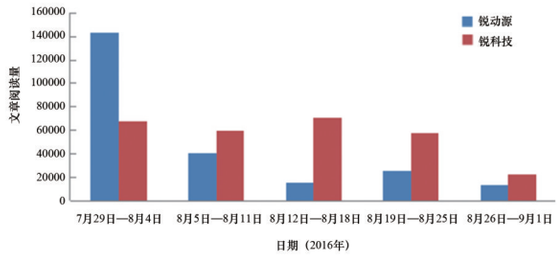


图8 2016年7月29日至9月1日“锐动源”和“锐科技”每周的阅读量对比

4 结论

“锐科技”及其子公众号“锐动源”是科技部官方微信矩阵中的2个公众号,本研究收集了401篇“锐动源”文章和856篇“锐科技”文章进行研究。通过归纳分析发现,和传统科技类报刊、期刊的编辑手段不同,科技部官方微信体现出科技类新媒体的编辑特点。(1) 速度快。“锐科技”编辑会选取重要的科技信息,迅速推出,第一时间让科技工作者获取相关信息。(2) 形式多样。除了文字内容,还包括视频、图片等多种形式。例如,《同属“胖五”家族,“长五B”与“长五”有啥区别?一图了解》一文中,用文字和视频报道了长征五号B火箭发射成功,同时用图片说明了“长五B”和“长五”的区别。(3) 语言活泼。“锐科技”很多文章的语言比较活泼,适合在网络上传播。例如,《选一个字,为大国重器打call》一文,编辑选取的标题非常网络化,可以吸引网上读者点击阅读。(4) 对阅读群体的针对性强。“锐科技”及其子公众号“锐动源”“社发科技”等,主要推送科技类话题,但各公众号的侧重点又有所不同,各公众号有自己的主题和细分读者群,具有一定的针对性。

为了定量研究科技部官方微信的主题,本研究

利用AToT后验概率主题模型,对收集到的文本进行分析和数据挖掘,研究科技部官方微信话题随时间演变的规律。所有文本没有经过提前标记,以无监督的方式提取、分析和识别微信公众号的主题。模型显示了“锐科技”和“锐动源”选取文章的主题是如何随着时间的推移而变化的。2016年7月29日之后的一周,“锐科技”的文章更多的是关于科技创新的规划,“锐动源”的文章与资金的监管有关。随着时间的推移,“锐科技”的文章开始介绍重点专项,“锐动源”开始关注科技部重点专项2017年度申报工作。这揭示了公众号的选稿思路和国家科技部的宣传脉络。每篇文章的阅读次数是读者非常重要的反馈。本研究表明,“锐科技”及其子公众号“锐动源”的编辑,在数据采集的2016年7月29日至2017年3月3日这段时间,推送文章的主题主要围绕了国家重点研发计划、科技成果转化、自然科学基金监管等话题,此外,编辑对国际合作等主题的时事科技新闻也进行了及时追踪,例如对中国科技部和美国农业部联合工作组会议内容的推送。同时,中国科技工作者的兴趣随着时间的推移而变化,他们更愿意在特定时间段阅读有关主题的文章,除了关注领导讲话、国际合作等时事科技新闻外,随着医学、生物学、电子、能源等领域重点专项及国家自然基金项目的发布,科技工作者对相关主题非常关注,即使不列在头条,也能有非常高的阅读量,表明了广大科技工作者的兴趣偏好。

本模型量化分析了科技部官方微信公众号的选稿思路和科技部在特定时间的宣传重点及该信息是否被目标受众充分接收,它可以帮助科技部官方微信公众号了解科技工作者在一定时期内的兴趣点和需求,并评估宣传的效果,从而促进中国的科技宣传工作,更好的为科技管理服务。

参考文献(References)

[1] WeChat Wikipedia[EB/OL]. [2018-08-01]. https://en.wikipedia.org/wiki/WeChat.

[2] Number of monthly active WeChat users from 2nd quarter 2010 to 3rd quarter 2016[EB/OL]. [2018-08-01]. https://www. statista. com/statistics/255778/number-of-active-wechat-messenger-accounts.

[3] 孔云, 廖寅, 资芸, 等. 基于微信公众账号的图书馆移动信息服务研究[J]. 情报杂志, 2013, 32(9): 167-170.

[4] WeChat[EB/OL]. [2018-08-01]. http://tech.qq.com/a/20160119/005085.htm.

[5] 侯钰莹. 探析新媒体背景下电视新闻记者的转变与创新[J]. 新闻研究导刊, 2018, 9(3): 204.

[6] 朱杰, 张丽娟. 从“内容为王”的角度看微信公众号的兴衰[J]. 新闻论坛, 2019(1): 7-10.

[7] 张卓, 王飞. 新媒体“社会共享”的崛起与逻辑嬗变[J]. 新媒体与社会, 2017(3): 62-75.

[8] 满静. 全民阅读下新媒体产业的发展浅谈[J]. 出版广角, 2019(1): 43-45.

[9] 刘佳静, 金洁琴, 赵乃瑄. 高校图书馆微信公众号传播力评价研究——以“双一流”大学为例[J]. 图书馆工作与研究, 2019(2): 40-46.

[10] 王海燕. 图书馆微信公众平台传播影响力研究[J]. 图书馆工作与研究, 2015(9): 28-31.

[11] 员立亭. 我国图书馆界微信研究述评[J]. 图书馆工作与研究, 2018(4): 25-30.

[12] 董玥, 王雷, 刘健. 新型智库微信公众平台信息传播影响力评价体系研究[J]. 情报科学, 2018, 36(12): 41-45.

[13] 王鲁峰, 王碧莹, 张玉珊. 高校微信公众号的传播与互动研究——基于15个高校微信公众号的个案分析[J]. 新媒体研究, 2019(1): 17-22.

[14] 岳云鹏. 高校微信公众号建设思考——以河北医科大学官方微信公众号为例[J]. 新媒体研究, 2019(1): 63-64.

[15] 冀芳, 张夏恒. 学术期刊微信公众平台影响力研究——基于5种CSSCI来源期刊的实证分析[J]. 情报杂志, 2016, 35(4): 147-151.

[16] 林琳. 人民日报社的“侠客岛”何以名动微信公众号“江湖”[J]. 中国记者, 2015(2): 44-45.

[17] 李明德, 高如. 媒体微信公众号传播力评价研究——基于20个陕西媒体微信公众号的考察[J]. 情报杂志, 2015, 34(7): 141-147.

[18] 夏保国, 常亚平. 政务微信的沟通机制研究——基于技术接受模型的视角[J]. 国家行政学院学报, 2014(3): 102-106.

[19] 蒋天民, 胡新平. 政务微信的发展现状、问题分析及展望[J]. 现代情报, 2014, 34(10): 88-91.

[20] 李宗富, 张向先. 政务微信公众号服务质量的关键影响因素识别与分析[J]. 图书情报工作, 2016, 60(14): 84-93.

[21] 刘佐. 基于微信公众平台的数据挖掘与可视化研究[D]. 北京: 华北电力大学, 2017.

[22] Qiu J, Li Y, Tang J, et al. The lifecycle and cascade of WeChat social messaging groups[C]//Proceedings of the 25th International Conference on World Wide Web. Montreal: ACM, 2016: 311-320.

[23] 李松丽. 基于微信的社会舆论热点挖掘及分析模型研究[D]. 武汉: 华中师范大学, 2016.

[24] 张昊东, 陈锐, 郑凯, 等. 科协系统改革监测平台指标设计和软件平台建设[J]. 学会, 2018, 353(4): 33-38.

[25] Zhang H, Zhu L, Xu S, et al. XML-based document retrieval in chinese diseases question answering system [C]//FTRA international conference on mobile, ubiquitous, and intelligent computing. South Kore: Springer, 2014.

[26] Tseng Y H, Lin C J, Lin Y I. Text mining techniques for patent analysis[J]. Information Processing & Management, 2007, 43(5): 1216-1247.

[27] Wang Z, Zhu L, Xu S, et al. Simple interrogative sentence analysis based on CRF[C]//2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW). Omaha: ACM, 2016: 21-24.

[28] Xu S, An X, Zhu L, et al. A CRF-based system for recognizing chemical entities in biomedical literature[C]//In Proceedings of the 4th BioCreative Challenge Evaluation Workshop. Bethesda, Maryland: BioCreative, 2013: 152-157.

[29] Xu S, An X, Zhu L, et al. A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature[J]. Journal of Cheminformatics, 2015, 7(Supp 11): 9.

[30] Chen H, Zhang Y, Zhang G, et al. Modeling technological topic changes in patent claims[C]//2015 Portland International Conference on Management of Engineering & Technology. Portland: IEEE, 2015: 2049-2059.

[31] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-Topic model for authors and documents[C]//UAI'04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence. Banff: AUAI Press, 2004.

[32] Mimno D, Mccallum A. Expertise modeling for matching papers with reviewers[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California: ACM, 2007: 500-509.

[33] Kawamae N. Author interest topic model[C]//Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva: ACM, 2010: 887-888.

[34] Mccallum A, Corradaemmanuel A, Wang X. Topic and role discovery in social networks[J]. Ijcai, 2005, 30(2): 786-791.

[35] Wang X, Mohanty N, Mccallum A. Group and topic discovery from relations and text[J]. Nips, 2007, 30(2): 1449-1456.

[36] Blei D, Lafferty J. Dynamic topic models[C]//Machine Learning, Proceedings of the Twenty-Third International Conference. Pittsburgh: ACM, 2006.

[37] 卢盛祺, 管连, 金敏等. LDA模型在网络视频推荐中的应用[J]. 微型机与应用, 2016, 35(11): 74-79.

[38] 廖列法, 勒孚刚, 朱亚兰. LDA模型在专利文本分类中的应用[J]. 现代情报, 2017, 37(3): 35-39.

[39] 何锦群. LDA在信息检索中的应用研究[D]. 天津: 天津理工大学, 2014.

[40] 张金瑞. 基于LDA的文本分类研究及其应用[D]. 郑州: 郑州大学, 2016.

[41] Tan C, Wang C. Study on classification of news topic based on LDA model[J]. Computer Knowledge and Technology, 2014, 16: 3795-3797.

[42] Stanford word segmenter[EB/OL]. [2016-11-01]. http://nlp.stanford.edu/software/segmenter.shtml.

[43] Xu S, Shi Q, Qiao X, et al. Author-topic over time (AToT): A dynamic users' interest model[J]. Lecture Notes in Electrical Engineering, 2014, 274: 239-245.

[44] Zhang H, Xu S, Wang Z, et al. Text and data mining of social media in science and technology publicity[C]//Portland International Conference on Management of Engineering & Technology(PICMET). Portland: IEEE, 2017.

[45] 史庆伟, 乔晓东, 徐硕, 等. 作者主题演化模型及其在研究兴趣演化分析中的应用[J]. 情报学报, 2013, 32(9): 912-919.

[46] Griffiths T, Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Suppl 1): 5228-5235.

[47] Bowman K O, Shenton L R. Parameter estimation for the beta distribution[J]. Journal of Statistical Computation & Simulation, 1992, 43(3): 217-228.

Quantitative analysis of topics editing in Ministry of Science and Technology's WeChat official account based on AToT model

ZHANG Haodong, ZHAO Lixin

National Academy of Innovation Strategy, Beijing 100863, China

Abstract Like the Facebook, the Twitter and the TikTok, the WeChat is a well-known mobile instant messaging product with over 1 billion active accounts, and it supports text messages, voices, videos, pictures and payments. The Ministry of Science and Technology of the People's Republic of China makes the science and technology management through various social media, including the WeChat, and the "Rui Ke Ji(Sharp S&T)" is the official WeChat account of the Ministry of Science and Technology, providing information for scientists and technicians. Unlike the editing of the traditional science and technology newspapers and periodicals, the WeChat official account "Rui Ke Ji" has its own characteristics in the manuscript selection and publication (push). In order to see the coverage of the topics of the WeChat articles changing over the time and the relationship between the topics of the articles and the view accounts, a Bias probability topic model, AToT, is used to quantitatively analyze and mine the contents of the articles in the "Rui Ke Ji" and its sub account the "Rui dong yuan(Sharp Dynamic Source)" in this paper. The results reveal where the ministry's publicity lies and how the topics of the articles change over the time. The reading preferences of Chinese scientists and technicians are analyzed and it is shown that there is a certain relationship between the contents of the WeChat articles and the page views. This study helps to improve the editing level of the new media and the science and technology publicity.

Keywords editor of WeChat official account; topic model; Rui Ke Ji; Rui Dong Yuan

（责任编辑　王丽娜）