

基于 LDA 的科研项目主题挖掘与演化分析^{*}

——以 NSF 海洋酸化研究为例

王文娟^{1 2} 马建霞¹

(1.中国科学院兰州文献情报中心 兰州 730000; 2.中国科学院大学管理学院 北京 100049)

摘要 [目的/意义]基于当前对科研项目研究主要以数据统计和计量分析为主,以 NSF 资助的海洋酸化相关研究的项目数据为例,从项目主题的角度研究科研立项的重点和演化,并对未来投入趋势加以预测。[方法/过程]使用 LDA 主题模型对该研究的主题进行分析,计算主题强度分析演化过程,统计 NSF 申请书项目与支撑发表的核心论文主题强度之间的相关性,进而分析 NSF 海洋酸化项目的主题发展趋势。[结果/结论]发现 LDA 模型可以很好的发现海洋酸化的主要研究主题,且 NSF 资助的科研项目与支撑发表论文的主题演化规律具有一致性。使用的方法对科研资助机构资助的项目从主题的角度进行分析,研究国家科研机构资助项目的主题布局 and 变化趋势,对国家科技部门或机构的项目发展部署和规划具有一定的参考和支撑作用。

关键词 LDA NSF 项目 主题演化 海洋酸化

中图分类号 G250

文献标识码 A

文章编号 1002-1965(2017)07-0034-06

引用格式 王文娟,马建霞.基于 LDA 的科研项目主题挖掘与演化分析[J].情报杂志,2017,36(7):34-39.

DOI 10.3969/j.issn.1002-1965.2017.07.007

Topic Detection and Evolution Analysis of Research Project based on LDA

——A Case Study of Projects on Ocean Acidification Supported by NSF

Wang Wenjuan^{1 2} Ma Jianxia¹

(1.Lanzhou Library of Chinese Academy of Sciences, Lanzhou 730000;

2.University of Chinese Academy of Sciences, Beijing 100049)

Abstract [Purpose/Significance] We tried to find a new way based on LDA to detect subjects of projects and their articles to grasp topic distribution and evolution of funding agency and found the topics of projects. [Method/Process] In this contribution we detected the topics of the projects on ocean acidification supported by NSF based on LDA by computing subject-intensity. And we analyzed the evolution of topics of projects on ocean acidification supported by NSF. After comparing the topics of articles supported by those projects and the topics with those projects, we analyzed the topics evolution of the articles. [Result/Conclusion] The results were consistent with topics put forward by Ocean Acidification subcommittee of Ocean carbon and biogeochemical plan. And the evolution of topics of projects is in line with those of articles supported by the projects basically.

Key words LDA projects supported by NSF topics of articles ocean acidification

0 引言

国立科研资助机构项目申请书作为国家资助的科研项目的载体,既体现了不同研究内容上的差异,又从总体上反映了国家的科技发展部署,通过对项目申请

书的主题进行分析,能够快速了解国家科技发展的重点领域和发展趋势。以科研资助机构基金支持发表的期刊论文是基金项目所获成果的直观的体现,结合项目申请数据与产出的论文数据进行分析,既能了解科研基金项目支撑的产出结果,还能从一定程度上掌握

收稿日期:2016-10-21

修回日期:2017-01-08

基金项目:国家自然科学基金项目“基于科学基金项目及知识产出的研究前沿探测”(编号:71373260)研究成果之一。

作者简介:王文娟(ORCID:0000-0001-5233-1387),女,1990年生,硕士研究生,研究方向:情报分析技术;马建霞(ORCID:0000-0002-5401-9992),女,研究馆员,硕士生导师,研究方向:数字图书馆技术、数字资源保存长期、机构仓储、知识技术等。

通信作者:马建霞

该研究领域的发展情况,预测下一时期的科研资助项目的主题变化。

目前,绝大部分学者利用科研项目基金的相关数据,包括申请书和基金支持发表的核心论文数据,主要从科研项目的学科^[1-2]、机构或单位^[3-4]、地区^[5-6]、论文产出绩效^[7-8]等方面展开研究,一般采用数据统计或计量分析的方法对科研项目的立项现状进行概括,或从数量的角度对于基金投入的趋势进行预测^[9],而缺少对未来研究方向趋势变化的分析。

根据文献计量学的学科生命周期理论,主题相关的论文数量能够反映主题研究的热度和发展水平^[10],文献主题的抽取对研究学科或项目的发展趋势十分重要。文献的主题挖掘有很多方法,如有学者使用共词分析进行研究^[11],或者以共词分析为基础结合社会网络的方法挖掘文献主题^[12];还有学者提出了 TF* PDF 算法^[13],主要通过词汇在不同文献中的频次和结构关系抽取实验数据的主要主题句子,与此类似的还有 TF* IDF 算法^[14-15];另外考虑到文献是由语义表达构成的,有学者提出了 PLSA 算法^[16],将文献看成是在不同的主题表达的向量空间集合,词汇及其概率分布与主题相关。LDA 是以 PLSA 算法为基础由 Blei^[17]提出的主题模型,该算法被提出后得到了广泛的应用并衍生了很多的变体^[18-19],其在实际应用中被证明具有很好的适应性,是目前最常用的文本主题提取方法^[20-21]。

因此,本文利用国立科研机构的项目申请书和项目支持发表的核心论文的摘要文本,采用 LDA 主题模型挖掘国立科研机构支持的研究主题,对以年为时间片资助的项目主题的演化情况进行分析,再对比不同研究主题的论文发表情况并对科研立项的主题趋势进行预测,从而对国家科技下一步的部署方向和趋势提供一定程度的参考和支撑。

1 相关概念

1.1 LDA 模型与主题数的确定 LDA 是文档、主题和词汇组成的三层贝叶斯概率模型,其采用迭代估计的方式计算文档的主题词汇。给定文档集合中有 D 篇文档,共包含 W 个词汇,假设文档集合共有 K 个主题,LDA 模型的主题生成过程如下:

- (1) 对于每篇文档 $d \in D$,根据狄利克雷分布 $\theta_d \sim \text{Dir}(\alpha)$,得到文档 d 的主题分布参数 θ_d ;
- (2) 对于每个主题 $z \in K$,根据狄利克雷分布 $\varphi_z \sim \text{Dir}(\beta)$,得到主题 z 上的词汇的多项式分布 φ_z ;
- (3) 对于文档 D 中的第 i 个词汇 w_{d_j} ,根据多项式分布 $z_{d_j} \sim \text{Mult}(\theta_d)$,得到主题 z_{d_j} ,根据多项式分布 $w_{d_j} \sim \text{Mult}(\varphi_{z_{d_j}})$,得到词汇 w_{d_j} 。

由 LDA 模型的生成过程可知,模型无法直接获得两个狄利克雷分布 θ_d 和 φ_z ,一般而言采用参数估计的方式近似推理参数值,在实际研究中常采用基于马尔科夫链的 Gibbs 抽样算法^[22]。另外在该方法中需要预设 α 和 β 的值,且两者的确定没有明确的范围限制,但考虑在大部分实验研究中采用经验法则,即设置 $\alpha = 50/K$ (K 为文档的主题数量), $\beta = 0.01$ 被认为是在实验过程中具有较好的表现^[23-24]。

1.2 主题强度 主题强度反映了文档主题的热门程度,其演化过程可通过观测主题随时间的变化来衡量。

本文引入文献^[25]熵值加权的方法,根据主题以熵值为文档赋予权值,则具有明确主题倾向的文档得分较高,内容过于宽泛的文档得分较低。熵值是一个很好的可用来计算文档权重的标准,最后再结合文档的权重和文档主题的多项式分布计算主题强度。

$$\text{entropy}(d_m) = - \sum_{k=1}^K \theta_{m,k} \log_2 \theta_{m,k}$$

$$w_m = 1 - \frac{\text{entropy}(d_m)}{\max \{ \text{entropy}(d_1), \dots, \text{entropy}(d_M) \}}$$

$$T_t(z_k) = \frac{\sum_{m=1}^{M_t} w_m \theta_{m,k}}{\sum_{m=1}^{M_t} w_m}$$

其中 $\theta_{m,k}$ 表示文档 m 属于主题 k 的概率大小, M_t 表示在时间段 t 内的文档数目, $T_t(z_k)$ 表示在 t 时间段内 k 主题强度。通过上述公式对文档主题强度按时间片进行划分,时间片段以年为单位,通过对年份主题强度的变化可以观测不同时期申请书文档主题的集中情况,也从侧面反映国家资助项目和论文的主题范围和变化趋势。

通过 LDA 技术可以得到主题 z 上的词汇的多项式分布 φ_z ,在 NSF 项目支持发表的论文数据中可以生成文档-词汇矩阵 M_d ,由此可以计算每一篇论文在不同主题上的分布情况 $\theta_{m,k} = M_d * \varphi_z$,再结合上文主题强度计算公式计算出论文在不同的主题上的演化趋势。

2 实验及分析

2.1 实验数据 本研究选用美国 NSF 资助的海洋酸化相关的项目申请书和 Web of Science 收录的期刊以 NSF 基金支持发表的有关海洋酸化的期刊论文为研究数据。目前全球海洋正处于 5 500 万年以来海洋酸化速度最快的时期,2003 年发表在 Nature 杂志上的文献^[26]首次探讨了海洋酸化问题引起了科学界的广泛关注,伴随 2004 年主题为“高浓度二氧化碳环境下的海洋”国际研讨会的举行,海洋酸化成为全球海洋

研究的热点课题^[27]。NSF 是美国重要的科研项目资助机构,关注公众健康、食品安全、水质质量、环境、教育等各方面的问題,海洋酸化研究是其资助的一个重要分支。选取该部分数据,研究其主题的演化情况具有一定的代表性和实际意义。

以检索式“(ocean* OR seawater* OR marin*) AND (acidif* OR (carbon* OR Nitro* OR sulf*) AND PH))”对项目题名和摘要进行检索,另外再以 NSF 海洋酸化的项目编码“1382”进行检索,将两部分的结果进行综合并删除重复项。以同样的检索式在 Web of Science 中进行检索并对基金机构进行筛选获得 NSF 项目支持发表的核心论文数据。由于基金的摘要信息中除研究主题的内容外还包括整个项目的研究意义、作用、承担机构等其他信息,因此本文借鉴文献^[28]中的清洗规则,人工删除摘要中无实质性意义、与研究主题无关的信息,获得独立文档量如表 1 所示。然后对独立文档进行去停用词、提取词干处理用于 LDA 主题模型进行计算。项目文档在 LDA 模型进行主题发现的过程中迭代 1 000 次,输出效果最佳的主题的主要关键词,然后对比主题数不同的结果之间的差异,以文献^[29]中的方式确定最终的最优主题数。

表 1 用于 LDA 模型的两组数据

| 序号 | 数据库 | 年份跨度 | 数据量 |
|----|-----|-----------|-----|
| 1 | NSF | 2004-2016 | 357 |
| 2 | WoS | 2008-2016 | 655 |

2.2 结果分析

2.2.1 NSF 项目与核心论文数量分析

以 NSF 项目的起始年份作为其所在年份对总体数量和资助的资金总额进行统计,按照项目类型可分为持续性资助(continuing grant)、基础性资助(standard grant)以及奖学金(Fellowship)三种,其年份变化如图 1 所示。在 2004 年至 2005 年相关的项目还比较少,2006 年至 2008 年明显提高,每年都有将近 20 项的项目被资助,2009 年则出现爆发式增长,2011 年资助的项目数量略少一些,却仍能达到 2008 年项目数量的两倍的水平,可见从 2009 年开始,NSF 一直对海洋酸化问题保持较高的关注度,且其变化趋势未见减缓。在资助类型上,奖学金类型的项目所占比例越来越多,基础资助类型的项目数量每年都在 30 个左右,持续性资助的项目则变化的波动幅度较大。项目资金总额大部分维持在 2 000 万的水平上下波动,2009 年是海洋酸化研究的转折年,2009 年、2010 年、2012 年的拨款额明显高于其他年份。

由于数据来源的原因,笔者只在 WoS 数据库中检索到了 2008 年以后的 NSF 基金支持发表的论文,每年的论文数目如图 2 所示。从其发表论文的变化数量

上来看,总体呈逐渐增加的趋势,2012 年属于一个特别值,论文的数量较少,原因可能受 2011 年资助金额与项目数量削减的影响。

以上分析粗略地反映了 NSF 机构每年资助项目的变化情况,具体到每年项目的研究内容主要侧重的方面则无法通过图 1 和图 2 体现,需要进一步对研究的具体内容加以分析。

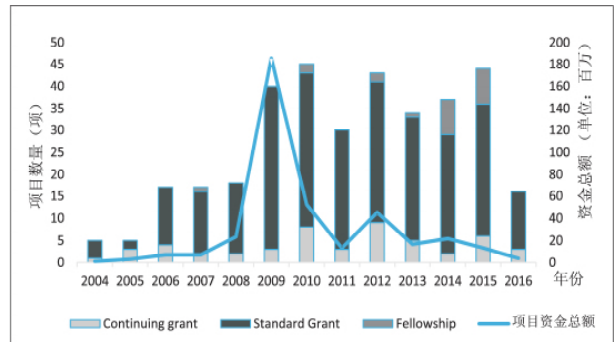


图 1 NSF 资助项目数量与资金总额年度变化图

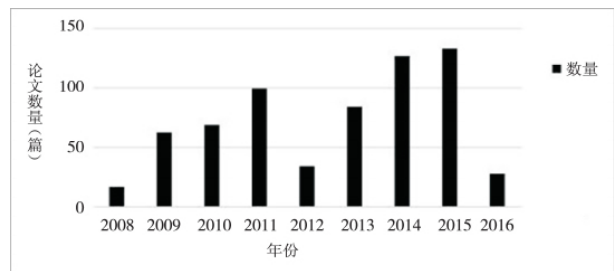


图 2 NSF 基金支撑发表的论文数量变化

2.2.2 NSF 项目主题挖掘与分析

通过 LDA 主题模型计算结果分析并确定最优主题数目为 5,输出每个主题的重要关键词,然后删除其中在 3 个主题中都会出现的对判断主题无参考意义的词汇,如 ocean, carbon 等,最终输出每个主题下前 15 个主要词汇,如表 2 所示。

表 2 LDA 主题模型聚类结果中关键词 Top 15

| 序号 | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|----|----------|----------|----------|------------|-----------|
| 1 | oxid | measur | water | respons | develop |
| 2 | activ | sampl | isotop | coral | model |
| 3 | organ | dissolv | atmosph | effect | ecosystem |
| 4 | microbi | inorgan | surfac | condit | commun |
| 5 | reaction | total | observ | level | studi |
| 6 | abund | seawat | sea | Speci | climat |
| 7 | form | determin | record | temperatur | global |
| 8 | cell | water | season | reef | biolog |
| 9 | sourc | method | sediment | experi | process |
| 10 | repres | situ | cycl | organ | marin |
| 11 | environ | paramet | rate | affect | data |
| 12 | metal | alkalin | estim | predict | result |
| 13 | composit | test | variabl | marin | scale |
| 14 | fluid | develop | dissolut | physiolog | approach |
| 15 | function | ion | global | calcif | identifi |

联系主要关键词的含义可知主题 1 主要关于海洋

酸化与海洋中的有机物与一般生物的关系,包括的关键词有 organ(ic)、microbi(logy)、reaction、cell、function 等。海水中的碳是海洋生物的重要生存环境参数,一部分海洋生物需要海水中的碳酸钙提供构筑外壳和骨骼的原材料,还有很多海洋生物需要保持细胞外液与海水中二氧化碳的浓度在一定的范围内才能保证正常的呼吸作用^[30]。海洋的酸化直接影响到海底生物的生存状况,对不同的生物造成的影响亦是多样化的,这部分的研究集中了大量的对于海洋生物的观测和调查,直观反映海洋酸化的严重性。

海洋二氧化碳系统的测量参数一般包括四个要素:溶解性无机碳含量、碱度、pH 值、二氧化碳和水样本的平衡逸度^[31],主题 2 的主要关键词 dissolv(e)、in-organ(ic)、paramet(er)、sample、alkalin(ic)、ion(ic) 基本包括了二氧化碳系统测量的要素提示词,因此可以判断这一主题主要关于海洋二氧化碳系统的测量。海洋二氧化碳系统的变化与海洋酸化程度息息相关,通过对二氧化碳系统的测量可以评估海底生物的生存状态并通过实验样本的测量研究海洋酸化可供解决的实际方案。

主题 3 主要关于海水中二氧化碳以及海底中一些元素的同位素的观测和记录的相关研究,还包括建立全球观测系统的相关研究,从关键词 sea water、iso-top、observ、estim、dissolut(e)、glabal 的可以窥见。海水吸收了大气层中一部分的二氧化碳,日渐上升的大气层二氧化碳浓度对海水造成了不可忽视的影响。对海水中一些具有特殊性质的元素进行标记,特别是同位素的观测,有助于对海洋酸化的变化规律进行探究和预测^[32]。海洋酸化的影响也是全球性的,海水吸收大气中过量的二氧化碳造成海洋酸化是全球范围内需要共同关注的问题,NSF 支持构建全球的检测网络,追踪海洋酸度值的变化情况,评估解决方案的有效性。

主题 4 主要是关于海洋环境中钙盐的变化情况,研究生物钙化作用并预测海洋酸化对其的影响,重要关键词有 coral、reef、effect、affect、calcif(y/ication) 等。海水 pH 值降低会引发海洋生物钙化速率的降低。珊瑚礁生态系统是地球上生物多样性最高且经济效益显著的生态系统,海洋酸化不仅影响其中的钙化生物,还间接影响依赖其生存的植物群、动物群^[33-34]。除珊瑚礁外,海底还有一些其他的重要生物也会有钙化行为,如软体动物、棘皮动物、鱼类等,生物种类的不同对海洋酸化的响应也有所不同,针对不同种类海洋生物的响应的研究,可预测未来海洋将可能发生的变化。

主题 5 主要关于海洋生物模型的构建、模型和数据对比分析,以及海洋酸化与全球生态系统和气候变化之间的关系和影响,重要的提示关键词有 model、

ecosystem、climate、biolo(gy)、data 等。构建模型是研究现实复杂问题的重要手段和方法。海洋系统内的生物种类繁多,不同海域的环境也各不相同,通过构建海洋生物的地球化学模型,对模型和数据进行对比分析可以研究不同生物种类、不同海域之间的异同^[35],更为精准地分析和预测海洋生态对二氧化碳上升引起的气候变化的响应和改变。

2007 年,在海洋碳与生物地球化学计划(OCB) 秘书处成立的海洋酸化小组委员会主持召开了关于美国今后 5-10 年的海洋酸化研究方向,提出了 8 点建议^[36-37]。对比 8 条建议和 LDA 主题模型挖掘的 5 个主要主题,可以发现两者在内容上具有一定的对应关系。建议中提及要关注颗粒无机碳和颗粒有机碳的指标测量、关注海洋二氧化碳系统的检测、开发区域海生物地球化学模型并进行数据对比,分别与本研究发现的主题 3、2、5 一一对应,其他主题内容在建议中也有体现,因此可认为 LDA 模型的主题聚类结果在一定程度上很好地表达了 NSF 资助的海洋酸化项目的主要研究内容。

2.2.3 NSF 项目主题强度演化分析 根据上文的主题强度公式计算并绘制不同主题的年份强度变化折线图,如图 3 所示。

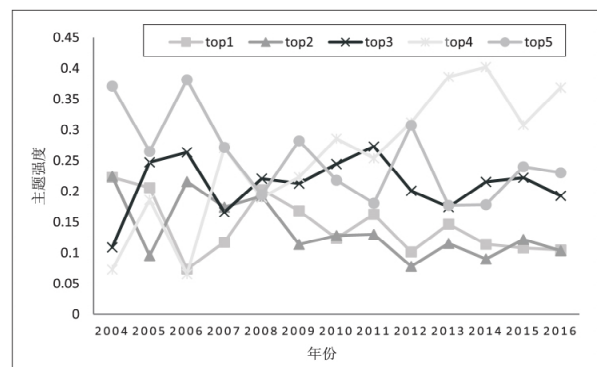


图 3 NSF 资助项目主题的强度年份变化

主题 2 和主题 5 在总体上呈现下滑的趋势。不同的是主题 2 在 2004 年时热度排名第二,2009 年以后其热度已经跌至倒数,说明 NSF 对海洋中二氧化碳系统的测量关注度在降低,原因可能是这一部分的研究已经取得了较为满意的结果。而主题 5 虽然比例下滑但绝对值仍然较高,说明对构建模型模拟海洋生态对海洋酸化的响应的研究是关注热点仅略有减少。主题 4 虽然在 2006 年之前强度普遍处于较低的水平,而在 2006 年之后总体上看强度在逐渐增加,近几年保持在较高水平,代表 NSF 对海洋中生物钙化的研究特别重视,是重点关注的研究方向。主题 1 和主题 3 的强度虽然在某些年份会有异常的表现,但从总体水平看强度变化不大,保持在小范围内变化,其中主题 1 关于海洋酸化与海洋生态中一般生物的研究强度近几年略

有所降低,主题 3 则保持平稳,说明监测海洋中二氧化碳的技术、方法仍在发展。

从强度变化的波动幅度上看,5 个主题的波动幅度则由起始时的大幅度逐渐变为小范围波动,近几年各个主题的热度对比基本差别不大,以 2008 年为对比的关键点,2008 年以前的波动幅度较 2008 年以后的波动高,说明 NSF 对于海洋酸化主题的主要研究方向越来越明确,对各个主题更为规律地把握和控制其资助的比例。

2.2.4 NSF 支撑发表的核心论文主题强度分析

将 NSF 支持发表的核心期刊论文数据的主题进行强度计算,其不同的年份与主题强度的变化情况如图 4 所示。

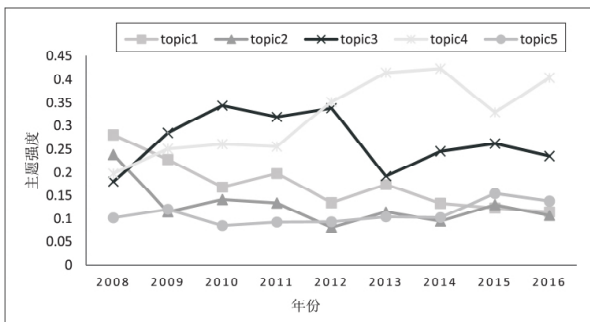


图 4 核心论文主题强度年度变化

对图 4 进行研究发现,主题 2 和主题 5 绝大多数年份里强度都不高,从侧面反映这两个主题在论文产出比率上相对较少,主题 2 在早期是关注的主要研究方向,但 2008 年之后一直处于较低水平。与主题 2 类似,主题 1 的论文强度逐渐呈下滑的趋势。主题 3 的强度变化较多,2008 年至 2012 年呈缓慢上升的趋势,2013 年迅速下降之后有一定回升,但近几年增长率并不高。主题 4 总体的强度保持一定幅度的缓慢增长,可以认为在 NSF 资助的关于海洋酸化的相关研究中,主题 4 的核心论文产出比例最高。

对比 NSF 资助项目的主题强度变化情况,在横向上,核心论文的主题热度变化基本与其保持一致,当项目申请的主题热度发生变化时,则对应以 NSF 基金支持发表的相关主题的论文热度也会减少。在纵向上,主题 5 相关的核心论文热度要低于其在 NSF 项目中的热度,而其他主题的热度排名基本相同。

2.2.5 NSF 项目主题趋势分析

使用 SPSS 工具将 NSF 支撑发表的论文与后一年的科研项目的各个主题的主题强度值进行相关性分析,得到的结果如图 5 所示。通过 Kendall 与 Spearman 的相关性检验均发现论文强度与主题的变化强度存在显著的正相关性,因此可以认为论文的主题强度变化可以用于预测主题强度的变化。

通过上述分析,由于论文在主题 3 和主题 4 的强

度相对其他主题更大,则预计在未来几年主题 3 和主题 4 将维持稳定高投入水平。主题 1、主题 2 在论文期强度水平在逐年减少,虽然主题 2 的科研立项投入近几年有小幅增长,预计这两个主题会一直保持低强度水平发展甚至衰减。主题 5 的论文主题略有增长,在项目强度中总体呈下滑趋势,预计近几年不会再继续下滑而保持中等水平继续发展。

| | | | |
|-----------------|---|---------|--------|
| | | x | y |
| Kendall 的 tau b | x | 相关系数 | 1.00 |
| | | 显著性(双尾) | . |
| | | N | 40 |
| | y | 相关系数 | .315** |
| | | 显著性(双尾) | .004 |
| | | N | 40 |
| Spearman 的 rho | x | 相关系数 | 1.000 |
| | | 显著性(双尾) | . |
| | | N | 40 |
| | y | 相关系数 | .446** |
| | | 显著性(双尾) | .004 |
| | | N | 40 |

** 相关性在 0.01 层面上显著(双尾)。

图 5 论文及其后一年的项目立项主题强度相关性分析

3 总结

本文利用 LDA 主题聚类算法分析了 NSF 资助的海洋酸化类相关研究主题的演化情况,并且结合 NSF 支持发表的核心期刊论文主题强度情况分析了主要主题的产出效率和演变规律,并对各个主题的未来趋势进行预测。通过对项目主题内容的详细分析,发现挖掘出的主要主题与海洋酸化小组委员发布的海洋酸化研究计划内容上具有一定程度的相似性,并且从项目及论文产出的共同演化能够对未来一段时期的资助趋势进行预估。通过实验分析发现挖掘的主要主题与本文使用的方法具有可行性和有效性,其结果不仅能够预测未来一段时间内的资金资助的比例和发展趋势,还能为政府对制定相关项目计划或政策时提供一定的参考和支撑。

然而,目前在样本中仅仅是对一般的停用词汇进行排除,在数据清理上还有很多有待改进的地方,论文产出的数据局限在核心期刊库里,在未来的研究中如果能够以本文为基础改进数据清洗规则,增加文献分析的产出数据来源,提高语义建模的准确性,加强与相关学科研究人员的交流合作,提升聚类描述主题词汇的精确度和对主题词的理解和解读,相信可以进一步完善分析方法,更准确的揭示科研机构资助项目的主题演化情况。

参考文献

[1] Huang Z, Chen H C, Yan L J, et al. Longitudinal nanotechnology development(1991-2002): National science foundation funding and its impact on patents[J]. Journal of Nanoparticle Research, 2005, 7(4-5): 343-376.

- [2] 岳洪江.我国科学基金资助成果学科与机构分布特征[J].情报杂志,2004(12):33-35.
- [3] Zhao D Z.Characteristics and impact of grant-funded research: A case study of the library and information science field[J].Scientometrics,2010,84(2):293-306.
- [4] 张志强,李延梅,柴育成.美国国家自然科学基金(NSF)地学部(GEO)基金资助战略分析[J].地球科学进展,2005(10):1143-1152.
- [5] Gauthier A H.Family policies in industrialized countries: Is there convergence[J].Population-E,2016,57(3):447-474.
- [6] 黄鸣清,韩立炜,吴修红,等.2012年度国家自然科学基金中药学科面上、青年、地区项目申请及资助情况分析[J].中国中药杂志,2013(1):6-9.
- [7] Beaudry C,Allaoui S.Impact of public and private research funding on scientific production: The case of nanotechnology[J].Research Policy,2012,41(9):1589-1606.
- [8] Hillman N W,Tandberg D A,Gross J P K.Performance funding in higher education: Do financial incentives impact college completions? [J].Journal of Higher Education,2014,85(6):826-857.
- [9] Eckhouse S,Lewis G,Sullivan R.Trends in the global funding and activity of cancer research[J].Molecular Oncology,2008,2(1):20-32.
- [10] 邱均平,段宇锋,陈敬全,等.我国文献计量学发展的回顾与展望[J].科学学研究,2003(2):143-148.
- [11] Belter C W.A bibliometric analysis of NOAA's office of ocean exploration and research[J].Scientometrics,2013,95(2):629-644.
- [12] Meichen Z.The research of linked data in domestic and abroad based on co-word analysis and social network analysis[J].Journal of Modern Information,2016(3).
- [13] Bun K K,Ishizuka M.Emerging topic tracking system: Advanced Issues of E-Commerce and Web-Based Information Systems [C].IEEE,2001.
- [14] Ramos J.Using tf-idf to determine word relevance in document queries [C].In proceedings of the first instructional conference on machine learning,2003.
- [15] Aizawa A.Editor an information-theoretic perspective of tf-idf measures [C].Information Processing and Management,2003.
- [16] Hofmann T.Editor probabilistic latent semantic indexing [C].International ACM SIGIR Conference on Research and Development in Information Retrieval,2004.
- [17] Blei D M,Ng A Y,Jordan M I.Latent dirichlet allocation [J].Journal of Machine Learning Research,2003,3(4-5):993-1022.
- [18] Agarwal D,Chen B C.FLDA: Matrix factorization through latent dirichlet allocation [C].International Conference on Web Search and Web Data Mining,WSDM 2010,New York,Ny,Usa,February,2010.
- [19] Wang Y,Bai H,Stanton M,et al.PLDA: Parallel latent dirichlet allocation for large-scale [C].Algorithmic Aspects in Information and Management: 5th International Conference,AAIM 2009,San Francisco,CA,USA,June 15-17,2009.Proceedings.Springer Berlin Heidelberg,301-314.
- [20] Barbieri N,Manco G,Ritacco E,et al.Probabilistic topic models for sequence data [J].Machine Learning,2013,93(1):5-29.
- [21] 胡吉明,陈果.基于动态LDA主题模型的内容主题挖掘与演化[J].图书情报工作,2014(2):138-142.
- [22] Tian J,Guo Z,Huang Y,et al.Automatic image annotation method based on multi-modal topic model [J].Foreign Electronic Measurement Technology,2015.
- [23] 王振振,何明,杜永萍.基于LDA主题模型的文本相似度计算[J].计算机科学,2013(12):229-232.
- [24] Yang Y.Semantic scene modeling for behavior analysis using lda mixture model [J].Journal of Information & Computational Science,2015,12(17):6525-6533.
- [25] 崔凯.基于LDA的主题演化研究与实现[D].长沙:国防科学技术大学,2010.
- [26] Caldeira K,Wickett M E.Anthropogenic carbon and ocean pH [J].Nature,2003,425(6956):365-365.
- [27] Raven J A,Caldeira K,Elderfield H,et al.Ocean acidification due to increasing atmospheric carbon dioxide. policy document 12/05 [J].Science,2005,215(2):1-60.
- [28] 陈挺,韩涛,李泽霞,等.科研项目布局差异对比方法研究——以NSF和EU FP项目为例[J].现代图书情报技术,2015(Z1):89-96.
- [29] 王博,刘盛博,丁堃,等.基于LDA主题模型的专利内容分析方法[J].科研管理,2015(3):111-117.
- [30] 唐启升,陈镇东,余克服,等.海洋酸化及其与海洋生物及生态系统的关系[J].科学通报,2013(14):1307-1314.
- [31] Bockmon E E,Dickson A G.A seawater filtration method suitable for total dissolved inorganic carbon and pH analyses [J].Limnology and Oceanography-Methods,2014(12):191-195.
- [32] Harris D,Horwath W R,van Kessel C.Acid fumigation of soils to remove carbonates prior to total organic carbon or carbon-13 isotopic analysis [J].Soil Science Society of America Journal,2001,65(6):1853-1856.
- [33] 张成龙,黄晖,黄良民,等.海洋酸化对珊瑚礁生态系统的影响研究进展[J].生态学报,2012(5):1606-1615.
- [34] Pandolfi J M,Connolly S R,Marshall D J,et al.Projecting coral reef futures under global warming and ocean acidification [J].Science,2011,333(6041):418-422.
- [35] Waldbusser G G,Brunner E L,Haley B A,et al.A developmental and energetic basis linking larval oyster shell formation to acidification sensitivity [J].Geophysical Research Letters,2013,40(10):2171-2176.
- [36] Ocean Acidification-Recommended Strategy for a U.S.National Research Program 2009[EB/OL].[2016-05-13].http://www.us-ocb.org/publications/OCB_OA_Whitepaper.pdf.
- [37] 石莉,桂静,吴克勤.海洋酸化及国际研究动态[J].海洋科学进展,2011(1):122-128.

(责编:贺小利;校对:王平军)