

基于主题模型的科技文献主题演化及优化方法研究综述^{*}

于诗睿 李爱花 林紫洛 陈逸菲 唐小利

(中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100005)

〔摘要〕 目的/意义 梳理主题演化分析方法研究进展, 改善科技文献主题识别和趋势研判效果, 支撑文献信息服务, 为未来深入研究和实践应用指明方向。方法/过程 根据隐含狄利克雷分布主题模型的缺陷归纳其衍生模型, 针对现有科技文献主题演化分析方法存在的不足对相应改善方案进行总结, 最后结合现有研究局限性提出展望。结果/结论 通过全面调研, 为主题演化研究提供启示和借鉴。

〔关键词〕 主题模型; 主题识别; 主题演化; 隐含狄利克雷分布主题模型; 文本挖掘

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2023.08.006

A Review of Research on the Improvement of Topic Model Based Topic Evolution Analysis Methods for Scientific Literature

YU Shirui, LI Aihua, LIN Ziluo, CHEN Yifei, TANG Xiaoli

Library & Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China

〔Abstract〕 **Purpose/Significance** The research progress of topic evolution analysis method is sorted out, so as to improve the effect of topic identification and trend judgment of scientific literature, support literature information service, and point out the direction for future in-depth research and practical application. **Method/Process** The derivative model of latent Dirichlet allocation (LDA) is summed up according to the defects of the topic model, and the corresponding improvement plan is summarized according to the shortcomings of the analysis methods of the topic evolution of the existing scientific literature. Finally, the prospect is put forward according to the limitations of the existing research. **Result/Conclusion** Through comprehensive investigation, the paper can provide inspiration and references for the study of topic evolution.

〔Keywords〕 topic model; topic identification; topic evolution; latent Dirichlet allocation (LDA); text mining

〔修回日期〕 2023-03-25

〔作者简介〕 于诗睿, 硕士研究生, 发表论文 2 篇; 通信作者: 唐小利, 研究馆员。

〔基金项目〕 中国医学科学院医学与健康科技创新工程重大协同创新项目 (项目编号: 2021-I2M-1-033)。

1 引言

在新一轮科技革命和产业变革深入发展的新时代, 学科领域发展局势复杂, 科学技术更新迭代加速^[1]。随着科学成果与创新技术不断涌现, 领域研究发生动态变化, 学科主题演化分析成为科技情报研究中一项重要工作^[2]。探索学科领域的主题识别与发展规律, 有助于有关部门制定学科发展战略规

划，推动科技创新和社会进步；为科研管理机构提供科学决策支持，实现科研资源优化配置和有效利用；帮助科研人员把握学科发展脉络，紧跟热点前沿，准确识别切入点，具有重要研究意义^[3]。

主题模型作为文本挖掘的重要工具，是文献主题分析的主要方法之一，得到广泛应用。本文通过系统梳理主题演化相关研究，重点针对隐含狄利克雷分布主题模型（latent Dirichlet allocation，LDA）的缺陷对其扩展方法进行总结，分析该领域方法方面亟待解决的问题，并提出相应优化办法及未来展望。

表 1 LDA 扩展主题模型总结

模型类别	扩展原因	优点	缺点	代表性研究
整合时间信息的 LDA 模型	考虑时间特征	能够深入了解主题产生和发展过程	主题可解释性低，准确性不高	DTM ^[5] 、TOT ^[6] 、DMM
对主题间关系建模的 LDA 模型	克服不能表达主题相关性缺陷	实现需求分析或主题间相关关系和关联强度可视化，识别主题间层次结构	只能对两个主题间相关性描述，随主题个数增加算法复杂度增大	CTM ^[7] 、HLDA ^[8] 、PAM ^[9]
整合文档元数据的 LDA 模型	利用文档附加信息 弥补无监督学习不足	从多角度分析主题分布情况，提升主题识别准确性	无法分析多标签主题，没有分析语料的语义信息，模型与文本拟合不足	ATM ^[10] 、TAM ^[11] 、LDA 机构－主题模型 ^[12]
融合领域知识的 LDA 模型	融合领域知识弥补无监督学习不足	提升主题的一致性和连贯性	未融合深层次语义信息	融合词特征的 LDA 模型 ^[13] 、RETM ^[14]
考虑文档级别语义结构的 LDA 模型	考虑文档结构信息 弥补“词袋”模型缺陷	提升语义信息表达完整性和准确性	特征表示有限，模型假设过于严格，建模存在偏差，计算复杂度过高	HMM－LDA ^[15] 、BGTm ^[16]

3 基于 LDA 模型的科技文献主题演化分析方法缺陷及解决方案

3.1 存在问题

当前，基于主题模型文本挖掘方法的主题识别与演化分析研究已形成比较成熟和完善的体系，但仍存在如下问题。一是主题可解释性差。聚类在同一主题下的词或词组缺乏语义关联，难以对主题内容归纳总结。二是主题语义信息不够具体。在没有上下文的情况下，单纯的名词性词或词组无法反映具体语义信息。三是主题辨识度低。主题下表达语义的层次和深度不易区分、代表性词语难以识别、

2 LDA 扩展主题模型分类

2003 年 Blei D M 等^[4]引入 Dirichlet 先验分布，首次提出 LDA 模型。此后 LDA 模型因其操作简便、扩展性强、过程易于理解等优势成为国内主流的主题识别与演化分析方法。LDA 模型衍生出许多扩展算法，广泛应用于主题聚类、趋势分析、关联分析等方面，主要可分为 5 个方向，见表 1。

词语存在歧义等，加大了主题辨识难度。四是难以识别领域新兴热点主题。主题中判别性词语过于分散、新兴词语缺失，导致新兴热点主题探测效果不好。针对上述问题，本文对相应优化方法进行梳理和总结。

3.2 增强主题可解释性

3.2.1 引入文档上下文信息 引入文档上下文信息可实现语义强化。如 Du L 等^[17]提出一个变体顺序 LDA 模型（sequential latent Dirichlet allocation，SeqLDA），不但能捕捉词序和主题的依赖关系，还能捕捉段落间的依赖关系，从而增强词语间相关性并提升主题连贯性。

3.2.2 融合更多外部知识帮助学习 在建模时充分利用文档元数据等监督信息带来的影响,能够增强主题下词语间的语义关联性,有助于提升主题识别效果。如 Zhou H 等^[18]提出引用-内容-LDA 主题模型,采用双层主题模型,包括父主题的引文信息和子主题的文本信息进行主题识别,能有效发现重要主题及其演化关系。结合领域知识的主题模型获取文档本身以外的信息,如文档集中词汇间的关联关系、层次结构、先验知识等,对主题语义的增强至关重要。如 Allahyari M 等^[19]提出一种基于知识的主题模型(entityLDA, EntLDA),将本体携带的先验知识与主题模型相结合,增加主题连贯性。

3.2.3 基于深度学习的主题模型 深度学习技术的发展,为主题演化分析提供新的解决思路。基于深度学习思想方法的主题模型主要分为结合词向量和神经网络方法两种思路,包括词向量辅助的概率主题模型、神经主题模型、联合训练主题模型和基于文本嵌入的聚类模型4类,在文本特征提取和语义表达方面具有巨大优势^[20]。其中词向量主题模型通过训练低维稠密的词向量刻画词汇间的语义相似度,有助于对文本中的语义概念进行更好的语义表示,进而挖掘出词汇间的关联关系,可有效提高词语的语义一致性。如沈思等^[21]引入主题词嵌入表示模型(topical word embeddings, TWE),联合学习单词和主题的向量表示,从而学习不同主题下词向量的不同表示,提升医学科技报告主题的挖掘精度。

3.2.4 神经主题模型 多以词袋作为模型输入,通过增加网络层捕捉词汇间的语义关系,同时结合稀疏约束解决“主题-词”分布的稀疏性^[22]。如 Card D 等^[23]在主题建模过程中基于神经网络框架结合文档标签等元数据,提升主题识别结果质量。

3.2.5 联合训练主题模型 融合概率主题模型和神经语言模型优势,在原有“文档-主题-词”全局语义关系发现基础上,通过语言模型发现句子级词序间的依赖关系,克服“词袋”假设的局限性。如唐焕玲等^[24]提出 Sem2vec (semantic to vector) 语义增强模型,使用 LDA 模型获得单词与主题间的先验知识,然后通过 Word2vec (word to vector) 模型

将先验知识融入语义词向量训练中,得到增强的语义词向量表示,从而提高文本隐含主题语义的识别效果。

3.2.6 基于预训练语言模型嵌入的潜在空间学习和聚类联合框架 在主题发现中应用文本嵌入并对其聚类有助于挖掘出词间的语义关系。随着 BERT (bidirectional encoder representation from transformers)、RoBERTa (a robustly optimized BERT pretraining approach) 和 XLNet (generalized autoregressive pretraining for language understanding) 等预训练语言模型的发展,利用预训练词嵌入的新框架被提出。如 Meng Y 等^[25]提出 TopClus 方法,一个基于预训练语言模型嵌入的潜在空间学习和聚类联合框架,在潜在空间对“主题-词”和“文档-主题”分布联合建模,得到更加一致、多样化和更具代表性的主题。

3.3 使主题语义信息表达具体化

针对此类问题可考虑将名词性词或词组替换为主谓宾(subject-predicate-object, SPO)三元组结构。SPO三元组能够通过概念及其之间关系形成的语义模式对内容的描述来反映领域主题,提供广泛的语义信息和更精细的语义结构,可有效改善词语歧义、词语表达含义欠缺等语义问题。如 Hu Z 等^[26]以主谓宾三元组作为基础语义单元,通过降维生成相应技术主题,然后以这些技术主题作为分析对象进行主题演化分析。

该类方法虽然提升了主题语义信息的揭示能力,但同时也引入了一个主要缺陷。抽取 SPO三元组时,自然语言处理技术本身的缺陷会导致三元组结构抽取不全,造成重要主题信息遗漏,后续可考虑引入关键词作为补充。

3.4 提升主题辨识度

针对此类问题可考虑使用基于神经网络的主题模型和预训练语言模型^[27]。如 Dieng A B 等^[28]提出 TopicRNN 模型,该模型使用循环神经网络(recurrent neural network, RNN)捕获局部(语法)依赖关系,使用潜在主题捕获全局(语义)依赖关系,

通过两者联合建模提升主题辨识度。Bianchi F 等^[29]使用预训练语言模型的输出增强或替换神经主题模型中基于词袋的文档表示,从而识别出更有意义和连贯的主题。

3.5 探测领域新兴热点主题,实现全面分析

针对此类问题可考虑融合不同类型的数据进行综合分析,如期刊论文、基金项目、专利文献等,从而实现对领域主题更全面的预测分析。如徐路路等^[30]提出一种基于并行 LDA 主题模型(paralleled LDA, PLDA)和多源数据融合相关性分析的新兴主题探测方法,使用项目文本、论文和专利数据,能够快速准确地识别出领域新兴主题。

也可考虑根据新兴热点主题的特点,应用新兴热点主题指标评价方法,从多个方面考量分析,探测出新兴热点主题的潜在阶段及主题内容,提升新兴热点主题识别结果的前瞻性和科学性。如马温如等^[31]在 LDA 主题识别模型基础上引入新颖性、相对增长性、影响性和发展潜力 4 个指标,构建识别框架,能够对新兴技术主题进行有效识别。

4 评价指标

不同主题模型在计算复杂度、计算效率和基于假设方面存在差异,在不同语料库和不同应用程序上表现不同,因此需要一套统一的标准评价不同模型性能。目前对主题模型各方面的评价还没有达成共识。本文总结一套综合的评价标准,包括模型质量、可解释性、稳定性、主题多样性、有效性和灵活性 6 个方面。

模型质量:质量与困惑度指标用于衡量模型学习主题在语料库中生成文档的能力,反映模型预测能力。困惑度越低,模型对观察到的文档解释性越好。模型可解释性:主题表示形式一般是单词的离散分布,因此这些单词应该与单一的语义相关联,评估主题可解释性的方法之一是分析主题的连贯性。模型主题多样性:主题多样性指标描述了识别出的主题在语义上的多样性。模型稳定性:主题建模的一个重要挑战是主题稳定性。高级主题的排名

可能在多次运行后发生显著变化,不同主题的文档分配在整个过程中也可能有所不同。模型有效性:不同算法的计算时间复杂度不同,是用于评估模型开发和运行效率的指标。因为主题模型通常处理大型语料库可能无法装入内存。模型灵活性:灵活性指标无法实现定量分析,不同模型对不同特征语料库的处理能力不同,如预处理步骤和文档中的单词分布都会影响模型性能。

5 总结与展望

5.1 增强主题的语义关联性

主题可解释性差、缺乏语义关联是科技文献主题演化分析中面临的主要问题,在主题模型的基础上考虑融合领域知识和文档元数据是一种有效的解决方法。对考虑融合领域知识的主题模型,现有模型仅融合了词级先验知识等简单信息,没有将领域知识的深层次语义信息融合进来。后续可考虑利用文档中的实体及其关系创建语义图,捕捉更高级的语义信息,也可利用已有的以本体形式构建的外部知识库、叙词表等,使主题易于理解。对考虑结合文档元数据的方法,科技文献除了文本信息外还携带丰富的附加信息,充分利用能够显著提升主题发现效果。其中标题是作者对文档内容的概括,对文档主题具有重要导向作用;文献自组织评论系统中构成的评论网络,包含文档间的关联关系。

5.2 深度学习技术为主题演化分析提供新视角

随着深度学习技术的发展,基于不同框架神经网络的主题模型已被开发出来,因其具有较好的灵活性、可扩展性、有效性和效率,在主题挖掘中具有广阔应用前景。基于神经网络的主题模型有望与神经结构集成,发挥独特优势;与预训练语言模型相结合,提高性能。未来应结合信息资源管理领域研究,将基于神经网络主题模型的改进与应用工作渗透得更深更广,解决当前研究难点问题。

5.3 拓展分析数据源并实现深度融合

目前主题演化分析研究往往以单一数据源为

主,且融合力度不够深入,一定程度上限制了主题挖掘的准确性和科学性。后续可考虑使用多样化数据,并以多视角剖析领域主题演变影响模式。未来也应考虑引入评价指标考量数据,提出针对不同数据源的一整套指标评价体系,如资助强度、研发机构研究水平等,实现对多源数据主题信息的统一深度融合。

5.4 不同演化分析方法需结合使用,互为补充

在主题演化分析研究方法的选取上,应根据不同方法的侧重点,考虑将多种方法相结合,实现互融互补。如可考虑将主题模型、文本聚类以及深度学习等文本挖掘方法组合进行分析,也可将文本挖掘方法与共现分析、引文分析等文献计量学方法以及社会网络分析方法相结合。

5.5 应考虑领域知识的不确定性

学科领域知识的不确定性是科学发展和演进的内在特点,在主题演化分析中不应被忽略,因此从科学文献中挖掘知识不应仅关注知识本身,还应考虑知识所处语境,这种不确定性是理解学科领域渐进式或颠覆式发展的关键,为分析领域演化趋势提供重要信息。

6 结语

在科学技术飞速发展的大环境下,掌握本学科主题的演化情况对学科发展非常有必要,有助于对学科内部知识结构变化和发展规律的深入分析。近年来有关科技文献学科主题演化分析方法已取得众多研究成果并得到广泛应用,但研究中仍使用以 LDA 及其扩展模型为主的方法。因此本研究针对 LDA 模型存在的不足对其扩展方法进行梳理和总结,针对主题可解释性差、主题语义信息不够具体、主题辨识度低和难以识别领域新兴热点主题 4 大问题梳理出相应的改进策略。尽管学科主题演化分析方法已经形成较成熟的分析体系,但仍存一些不足,未来主题演化分析方法的研究应更加关注提升主题语义关联性、充分借助深度学习技术的优

势、利用多源数据融合分析、利用多种方法优势互补以及考虑领域知识的不确定性等方向,以实现对学科主题演化内容和趋势更准确的分析。

参考文献

- 1 石磊,李君,吴婷.基于深度学习和 LDA 的学科研究前沿主题识别探究[J].江苏科技信息,2022,39(33):33-36,48.
- 2 王金丽.科技文献热点主题提取及演化研究[D].昆明:昆明理工大学,2021.
- 3 梁爽,刘小平.基于文本挖掘的科技文献主题演化研究进展[J].图书情报工作,2022,66(13):138-149.
- 4 BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003, 3(4/5):993-1022.
- 5 BLEI D M, LAFFERTY J D. Dynamic topic models[C]. Pittsburgh: The 23rd International Conference on Machine Learning, 2006.
- 6 WANG X, MCCALLUM A. Topics over time: a non - Markov continuous - time model of topical trends[C]. Philadelphia: The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.
- 7 BLEI D M, LAFFERTY J D. Correlated topic models[C]. Massachusetts: Advances in Neural Information Processing Systems, 2006.
- 8 TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical Dirichlet processes[J]. Journal of the American statistical association, 2006, 101(476):1566-1581.
- 9 LI W, MCCALLUM A. Pachinko allocation: DAG - structured mixture models of topic correlations[C]. Pittsburgh: The 23rd International Conference on Machine Learning, 2006.
- 10 ROSENZVI M, GRIFFITHS T, STEYVERS M, et al. The author topic model for authors and documents[C]. Virginia: The Conference on Uncertainty in Artificial Intelligence, 2004.
- 11 王萍.基于概率主题模型的文献知识挖掘[J].情报学报,2011,30(6):583-590.
- 12 WANG B, LIU S, DING K, et al. Identifying technological topics and institution - topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology[J]. Scientometrics, 2014, 101(1):685-704.
- 13 PETTERSON J, SMOLA A, CAETANO T S, et al. Word features for latent Dirichlet allocation[C]. Vancouver:

- The 23rd International Conference on Neural Information Processing Systems, 2010.
- 14 ALLAHYARI M, POURIYEH S, KOCHUT K J. Combining semantic graph and probabilistic topic models for discovering coherent topics [J]. Web intelligence, 2019, 17 (4): 365–379.
- 15 GRIFFITHS T L, STEYVERS M, BLEI D M, et al. Integrating topics and syntax [C]. Vancouver: The 17th International Conference on Neural Information Processing Systems, 2004.
- 16 WALLACH H M. Topic modeling: beyond bag – of – words [C]. New York: The 23rd International Conference on Machine Learning, 2006.
- 17 DU L, BUNTINE W L, JIN H, et al. Sequential latent Dirichlet allocation [J]. Knowledge and information systems, 2011, 31 (3): 475–503.
- 18 ZHOU H, YU H, HU R. Topic discovery and evolution in scientific literature based on content and citations [J]. Frontiers of information technology & electronic engineering, 2017, 18 (10): 1511–1524.
- 19 ALLAHYARI M, KOCHUT K J. Discovering coherent topics with entity topic models [C]. Omaha: 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2016.
- 20 朱佳晖. 基于深度学习的主题建模方法研究 [D]. 武汉: 武汉大学, 2017.
- 21 沈思, 李沁宇, 叶媛, 等. 基于 TWE 模型的医学科技报告主题挖掘及演化分析研究 [J]. 数据分析与知识发现, 2021, 5 (3): 35–44.
- 22 ABDELRAZEK A, EID Y, GAWISH E K, et al. Topic modeling algorithms and applications: a survey [J]. Information systems, 2022, 112 (1): 102131.
- 23 CARD D, TAN C, SMITH N A. Neural models for documents with metadata [C]. Melbourne: The Annual Meeting of the Association for Computational Linguistics, 2017.
- 24 唐焕玲, 卫红敏, 王育林, 等. 结合 LDA 与 Word2vec 的文本语义增强方法 [J]. 计算机工程与应用, 2022, 58 (13): 135–145.
- 25 MENG Y, ZHANG Y, HUANG J, et al. Topic discovery via latent space clustering of pretrained language model representations [C]. Online: The ACM Web Conference 2022, 2022.
- 26 HU Z, FANG S. An SAO – based approach to technology evolution analysis using patent information: case study—graphene sensors [J]. Data and information science, 2015, 8 (3): 62–75.
- 27 胡雪猛. 基于生成式神经网络的主题模型研究 [D]. 南京: 东南大学, 2021.
- 28 DIENG A B, WANG C, GAO J, et al. TopicRNN: a recurrent neural network with long – range semantic dependency [J]. Computer science, 2016 (11): 1–8.
- 29 BIANCHI F, TERRAGNI S, HOVY D. Pre – training is a hot topic: contextualized document embeddings improve topic coherence [C]. Online: Annual Meeting of the Association for Computational Linguistics, 2020.
- 30 徐路路, 王效岳, 白如江. 基于 PLDA 模型与多数据源融合相关性分析的新兴主题探测研究——以石墨烯领域为例 [J]. 情报理论与实践, 2018, 41 (4): 63–69, 43.
- 31 马温如, 赵钦, 滕珺. 科学技术新兴技术主题识别研究 [J]. 中华医学图书情报杂志, 2021, 30 (12): 1–10.

《医学信息学杂志》版权声明

(1) 作者所投稿件无“抄袭”、“剽窃”、“一稿两投或多投”等学术不端行为, 对于署名无异议, 不涉及保密与知识产权的侵权等问题, 文责自负。对于因上述问题引起的一切法律纠纷, 完全由全体署名作者负责, 无需编辑部承担连带责任。(2) 来稿刊用后, 该稿包括印刷出版和电子出版在内的出版权、复制权、发行权、汇编权、翻译权及信息网络传播权已经转让给《医学信息学杂志》编辑部。除以纸载体形式出版外, 本刊有权以光盘、网络期刊等其他方式刊登文稿, 本刊已加入万方数据“数字化期刊群”、重庆维普“中文科技期刊数据库”、清华同方“中国期刊全文数据库”、中邮阅读网。(3) 作者著作权使用费与本刊稿酬一次性给付, 不再另行发放。作者如不同意文章入编, 投稿时敬请说明。

《医学信息学杂志》编辑部