话题演化研究综述*

钱莉[1] 朱恒民[1,2] 魏静[1]
(1. 南京邮电大学管理学院，南京 210003；2. 江苏高校哲学社会科学重点研究基地—信息产业融合创新与应急管理研究中心，南京 210003)

摘要：话题演化分析对监控话题传播概况、预测其未来发展趋势具有重要意义。本文通过对国内外话题演化相关研究进行系统调研和分析，归纳总结话题演化研究的基础，并从话题强度、话题状态、话题内容与演化路径等多个方面探讨话题演化研究的不同维度，以及话题演化的趋势预测，此外还对话题演化研究的不同方法进行探讨，最后指出现有研究的不足，并对今后的话题演化研究进行展望。

关键词：话题演化；话题强度；话题状态；话题内容；演化路径
中图分类号：G206.3　DOI：10.3772/j.issn.1673-2286.2021.11.008
引文格式：钱莉，朱恒民，魏静. 话题演化研究综述[J]. 数字图书馆论坛，2021 (11)：57-64.

根据中国互联网络信息中心（CNNIC）第48次《中国互联网络发展状况统计报告》[1]显示，截至2021年6月，中国网民数量已达10.11亿人，可见互联网已经成为我国公民获取新闻和发表意见的重要媒介。在网络信息传播中，新闻话题或突发事件的迅速扩散，对政府相关职能部门构成了严峻的挑战。如何快速地跟踪新闻话题或突发事件的后续事态发展，是亟需解决的问题。

“话题”这一概念最早由TDT（Topic Detection and Tracking）评测会议提出，并对其进行了定义：所谓话题（topic），就是一个核心事件或活动以及与之直接相关的事件或活动[2]。而一个事件（event）通常是由某些原因或条件引起的，涉及某些对象（人或物），在特定时间或地点发生，并可能伴随某种必然结果。一般来说，话题就是若干件某事件相关报道的集合，主题则可以看作广泛意义上的话题，即主题可以涵盖多个类似的具体事件或根本不涉及任何具体事件[3]。例如，“社区防控”是一个主题，而“2020年2月10日湖北全省住宅小区实行封闭管理，共同做好疫情防控工作”是一个话题。在英文文献中，话题与主题都有一个共同的表达方式，即“topic”，但是本文将“话题”与“主题”的概念区分开，即新闻事件的话题是由一系列的主题构成。

话题随着时间的推进总是在不断演化的，每个话题都会经历从扩散到衰落的过程，话题之间也会产生漂移或渗透。从大规模网络文本中获取话题及其演化趋势，可以帮助人们掌握话题发展的“来龙去脉”，为监管部门及时应对网络舆情提供科学依据。因此，话题演化研究具有现实的应用背景。近年来，国内外学者对话题演化展开了研究，主要包括话题强度演化和内容演化两条研究路线。话题强度演化是指话题所受关注的程度随时间而变化，通常表现为与主题对应的文档数量；话题内容演化是指文本集中覆盖的主题范围随时间的变化，一般表现为与主题对应的特征词变化[4]。当前，话题演化分析模型常用的方法是将文档划分为不同的时间切片，然后在每个切片中提取主题。然而，这种方法容易导致过多和零碎的主题，且难以判定主题演化方向，对主题演化分析不充分。因此，话题演化仍需进一步探索，尤其是在网络文本领域。此外，还有一些学者尝试拓宽话题演化分析中的维度和深度。例如，Callon等[5]基于共词分析提出用向心度

* 本研究得到国家自然科学基金项目“基于主路径网络的舆情传播态势预测与干预研究——以社会化媒体中舆情为对象”（编号：71874088）资助。

（Centrality）和密度（Density）来分析话题的成熟度和关键性的观点，为话题演化研究提供了新的思路。尽管已有一些工作通过构建话题生命周期来检测话题所处阶段[6]，但将话题状态引入话题演化过程中的研究还非常少。在面对这些问题时，厘清话题演化过程中的复杂性就显得尤为迫切，尤其是话题之间的融合、分裂以及演化路径的分析。

基于此，本文对国内外话题演化相关研究进行了系统调研与总结。首先，按照规范流程对近年来国内外话题演化相关研究进行分析、整合与展示；其次，根据已有研究归纳话题演化研究的基础；在此基础上，从话题强度、话题状态、话题内容与演化路径等多个方面探讨话题演化研究维度，同时讨论了话题演化趋势预测，并总结话题演化研究的不同方法；最后，指出现有研究的不足，并对今后的话题演化研究进行展望。

1 数据和方法

本研究的数据来源分为国内和国外两部分。国内数据来源于CNKI，为了保证论文的质量，以“话题演化”“话题传播”“主题发现”“话题检测”为主题，以中文社会科学引文索引（CSSCI）来源期刊和中文核心期刊收录为范围进行高级检索。在搜索国外“话题演化”相关文献时，本研究首先选择覆盖多个学科领域的综合性数据库Web of Science、Science Direct以及Springer Link等连续动态更新的大型数据库，然后分别以“topic evolution”“evolution path”“topic spreading”为关键词开展主题、标题、摘要和关键字段的搜索；接着使用相同的关键词在Google Scholar中进行搜索，补充了未收录进以上数据库的论文。此外，本研究还查看已有综述文章纳入分析的文献，对现有搜索结果进行补充（如Zhou等[7]的研究），初步得到309篇论文。经过去重获得189篇可能与话题演化相关的论文，但还需要进一步确定其是否符合本研究的综述目标，因而以人工方式剔除条件不符（包括会议摘要、学者随笔等）或信息不全的文献。其次，仅提到该关键词但与研究问题不符的文献也被排除在外。经文献筛选，一共得到符合本文所需的相关论文156篇。最后开展质量评估对每篇文献进行逐一判读，以确保综述对象的质量，最终得到74篇高质量研究论文。这些论文大多发表于近5年，主要为期刊论文，并且广泛涉及图书情报学、计算机科学与技术以及社会科学等多个领域，可见话题演化是一个跨学科的问题且在近年来引起学术界广泛讨论。

从中外话题演化研究的发文量看，随着近年来互联网的快速发展，与之相关的话题演化文献也呈逐年上升趋势（见图1）。首先，本文借鉴文献计量学奠基人普赖斯提出的科技文献增长理论[8]，将话题演化研究分为三个阶段：第一阶段为起步探索期（2001—2011年），文献数量较少；第二阶段是平稳增长期（2012—2017年），文献数量呈稳定增长态势，虽然该时期部分年份发文量略有下降，但总体呈上升趋势；第三阶段为快速发展期（2018—2020年），国内外相关文献数量增长迅速，可见在今后的几年内该研究仍将保持较高的研究热度以及较快的发展速度。其次，国内发文量与国外呈现一致增长趋势，这说明话题演化研究受到世界各国学者的广泛关注，且国内与国外关于话题演化研究的发展趋势是一致的。尽管外文文献数量略高于中文文献，但有4篇外文文献是国内学者发文。

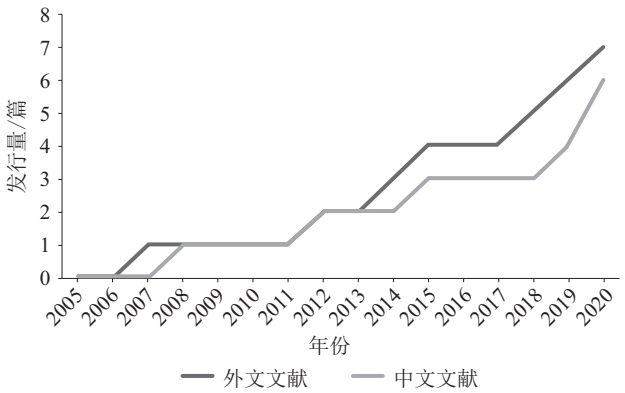

图1 中外话题演化研究年度发文量对比

通过对74篇话题演化研究文献的研读分析，先是介绍话题演化研究的基础，以期对话题演化研究的基本理论与技术手段进行了解。在此基础上，探讨话题演化分析维度，同时讨论话题演化趋势预测，并从中总结话题演化的分析方法，旨在深化话题研究脉络、探讨未来研究思路。

2 话题演化研究的基础

2.1 理论基础

虽然话题演化最初是在跟踪新闻报道的背景下产生的，但是其理论基础可以追溯到20世纪30年代提出

的“生命周期理论”（Life Cycle Theory）[9]。这一经典理论认为，任何事物都要经历诞生、成长、成熟、衰退和死亡整个过程，也泛指事物的阶段性变化及规律[4]。话题也具有生命周期的基本特征，生命周期理论勾勒了话题的演化轨迹。

Chen等[10]在生命周期的基础上提出了衰老理论（Aging Theory）。该理论认为，话题的生命周期与生物类似，生物拥有丰富的营养，即话题的相关文档增多，生命周期就会延长；反之，当营养耗尽时，一个生命或话题就会消失。换言之，当一个话题刚出现时，人们可能会对它感兴趣，但随着时间的推移，它的关注度逐渐下降。Fang等[11]基于衰老理论，结合话题相关的推文和用户权威构建了一个话题生命周期模型，并将话题划分为婴儿、成长、成熟、衰退和消失五个阶段。同样，谢科范等[12]将网络舆情分为潜伏期、萌动期、加速期、成熟期、衰退期五个阶段来分析网络突发事件，并为相关部门的管理决策提供了理论指导。

2.2 技术基础

近年来，话题演化研究在信息检索和数据挖掘等学术领域引起了广泛的关注。最早的工作可追溯到美国国防部高级研究计划局（Defense Advanced Research Projects Agency）于1996年提出的一种“话题检测与追踪”技术，该项技术是指利用计算机技术从新闻专线或广播新闻等新闻数据来源中自动检测话题，并采用话题相似度计算方法对后续新闻报道中话题的相关内容进行追踪[13]。关于话题演化的研究起始于跟踪具有时间信息的文档的话题趋势[14]，但是早期的TDT研究并未有效地利用语料的时间信息来分析话题随时间的变化[3]。目前，常见的话题演化分析技术路径主要包括以下内容。

（1）基于共词分析的话题演化。共同出现在同一文档或段落中的一对关键词被视为具有共现关系，且共现强度等于关键词的共现频率[15]。共现强度越大，两个词之间的内涵关联性越强，在话题上的一致性越高。计算大规模文档集关键词共现的关系网络能够反映研究话题的结构和演化规律。

（2）基于文本挖掘的话题演化。该方法重在分解文档内容，关注文档内部的特征，实现对文档粒度更小、层次更深、更全面的分析和研究[16]。随着文本挖掘方法的兴起，如何借助话题模型，研究话题随时间的变化以及如何变化，成为话题演化研究热点。LDA（Latent Dirichlet Allocation）话题模型是话题演化研究中最常见的技术[17]。它由Blei等首次提出，是一种混合概率模型，该模型通过最大化词语共现概率来寻找词语聚类，使用狄利克雷分布描述文档生成过程，并对文档的主题数量进行限制。大量研究表明，LDA在不同领域研究热点挖掘[18]、强度演化[19]、趋势预测[20]等方面都取得了良好效果。

3 话题演化研究的脉络

话题演化研究的脉络如图2所示。话题演化研究始于话题检测，即从给定文档集中识别出覆盖的话题，以及不同话题所占的比重，为话题演化分析提供基础。根据收集到的文献资料进行分析归纳，本文认为话题强度、话题状态、话题内容以及演化路径是话题演化分析的主要维度且部分研究只是聚焦于其中的某个或某几个方面。因此，本文将从这4个维度展开深入分析。最后，话题演化研究的主要目的是发现话题演化规律并预测其未来发展趋势，为管理决策提供参考。

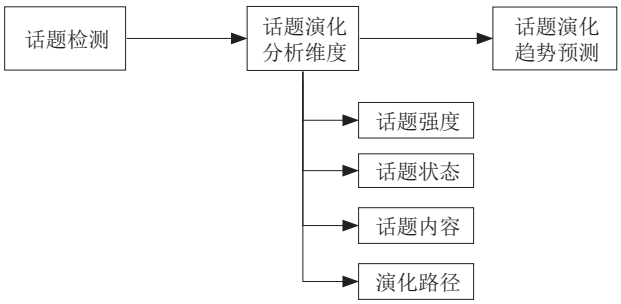

图2 话题演化研究的脉络

3.1 话题检测

话题检测，也称为“话题发现”或“话题识别”，旨在从大规模文档集中找到具有一致语义关系的相同话题。话题检测首先是在静态文本中提出的，大多数静态文本检测方法是基于概率话题模型，如PLSA[21]和LDA[22]。LDA作为PLSA的贝叶斯扩展，是话题演化研究中最流行的一个模型，解决了PLSA的两个问题。首先它的参数不会随着文档集增长而线性增长，具有很好的泛化能力；其次，PLSA是对给定的文档集进行建模，但对于如何将已有的模型应用于新的文档没有直接的办法。

也有一些研究建立了基于机器学习的话题检测方法。例如，Wartena等[23]通过关键词的共现关系来聚类关键词，从而发现话题。为了确定有意义的研究领域，Hurtado等[24]对文档中含有动词的标题进行关联规则分析，并通过删除停用词和动词来检测话题。Chen等[25]提出了一种非参数模型（NPMM）并利用辅助词嵌入来自动确定给定文档是否属于已有主题，进而推断主题编号。此外，Lu等[26]检测了来自共词网络中不同社区的词，在这些社区中，来自某个特定社区的词都属于相同且相互独立的主题。针对数据流连续、动态变化的特征，许多学者提出了一系列有效的解决方案。黄云等[27]针对微博话题检测中需要解决的高维数据、噪声信息以及话题的快速演化等主要问题，提出了一个微博在线话题检测模型（DLM）。贺敏等[28]提出了一种基于时序分析的微博突发话题检测方法。

话题检测的通用技术包括话题概率模型以及机器学习等方法。其中LDA模型最流行，该模型可以从大规模文本中迅速识别主题。由于网络自由文本中包含一些同义词和近义词，如何充分利用文本中词的复杂语义提升话题检测的质量，仍需要进一步探索。此外，互联网充斥着海量信息，且更新速度很快，如何快速识别大规模文本数据中的话题，并跟踪事件发展，成为急需解决的问题。此外，社交媒体上包含大量带有噪声的数据（如广告信息等），对话题检测没有实际意义，甚至给话题检测结果带来偏差，如何有效地从复杂多样的数据中识别出有效的话题，是话题检测的一大任务。

3.2 话题强度演化

话题强度演化表现为话题在不同时间切片中的流行程度，大多数基于LDA模型，将LDA应用在整个文档集合上，然后根据文档的时间信息将文档离散到相应的时间片。对于一个特定的话题，可以在不同时间片中依次考察其话题强度，以显示话题在整个时间轴中的变化情况。例如，Feng等[29]使用LDA来处理不同时间片中的文档集合，计算每个博客上话题分布概率的平均值，从而确定话题的平均热度。这种方法的优点是简单、易于操作，但是由于模型假设文档顺序是可交换的，不能有效地将时间信息与模型结合起来，因此未能充分利用时间信息，从而在同样建模条件下，会出现困惑度值很高的情况。如Wang等[30]提出了一种不同于将时间离散化的主题演化模型（Topic Over Time，TOT），它不再使用马尔科夫假设，而是将每一个主题表示为一个关于时间变量的连续概率分布，每一个主题不仅与词的共现相关，还与文本的时间戳有关，而且主题的内容以及主题之间的关系也是随着时间变化的。

一般而言，讨论一个话题的文档数量越多，这个话题就越受欢迎。由于与人们观察到的文本信息相吻合，这种方法越来越受到重视。Liu等[31]利用过去不同时期话题的频率累积来预测一个话题在未来一段时间内是否会流行。Zhao等[32]提出话题的“成长因子”来预测短期的话题趋势，并认为话题文档数量的增长速度会影响“成长因子”。但是，一篇文档可以包含多个话题，同一特征词或主题词对不同话题的贡献各不相同，因此在话题强度演化过程中，应该考虑特征词或主题词对话题的贡献。例如，李慧等[33]将特征词热度加入微博热点话题演化模型中，可以发现微博热点事件子话题的演化规律。

3.3 话题状态演化

话题状态是指研究话题在其演化生命周期中所处的阶段。当一个话题出现时，人们可能会对它感兴趣。随着时间流逝，话题的演化状态也在不断变化，并展现出一定的特征，如新话题中的关键词数量少，内部关联性弱，与其他话题相关性低；随着话题的成长，内部关键词的数量增加，它们之间的关系增强，与其他话题的相关性增加。因此，一些研究者就话题演化过程中的演化状态进行了研究。为了跟踪一个话题的演化趋势，Du等[34]提出了一种热门话题生命周期模型（HTLCM），并将HTLCM划分为出生、成长、成熟、衰退和消失五个阶段。另外，Callon等[5]基于共词分析提出了向心度和密度的概念，用来表示研究主题的关键性和成熟度。

上述工作多是回溯话题生命周期来识别话题的状态。由于话题演化具有较大的不确定性，对正在传播中的话题来判断其所处的生命周期阶段是非常困难的。少量工作通过设计指标来描述话题当前的状态，但是，如何设计出既能反映话题当前状态以及未来趋势，又能揭示话题潜在发展力的指标，是话题演化状态监测的难点。

3.4 话题内容演化

话题内容演化就是话题内容随着时间的推移而发

生的变化，通常表现为特征词在不同时间切片上的差异，而这种差异主要体现在语义关联方面。Blei等[17]开发了一个动态LDA模型，该模型反映了主题内容的时序变化。胡艳丽等[35]基于话题模型抽象描述文本内容的隐含语义，进而建立话题在时间序列上的内容演化。余本功等[36]利用改进的OLDA模型来应对舆情监控中的话题快速产生和消亡，并且分析得出话题内容演化。陈兴蜀等[37]基于OLDA模型对论坛中的热点话题演化跟踪做了研究。特征词或主题词在话题演化中的重要作用引起了学者的关注。例如，曹丽娜等[38]结合话题热度（强度）变化和内容变化两方面研究天涯论坛，挖掘随时间变化的动态话题链，从词语变化微观角度分析热门事件下公众意见的变迁过程。

话题内容演化是话题演化研究中的一个重要组成部分。随着时间的推进、网民的持续关注和热烈讨论，话题在不断地变化着。若演化后的话题与原有话题在内容上产生了较大的偏移，如何有效地探测和跟踪话题发展过程中的内容变化，是话题内容演化分析的关键问题。

3.5 话题演化路径

演化路径不同于话题内容的演化，它是指研究主题在时间轴上的演化脉络，旨在呈现主题的漂移特征。网络文本中的词汇语义更丰富复杂，这给网络文本话题演化路径研究带来了挑战，一些学者提出了相应的解决方案。Gao等[39]提出了一种新的在线加权条件随机场正则化相关主题模型（OCCTM），该模型利用语义相关性捕捉来自短文本的主要话题和相关子话题的演化路径；张佩瑶等[40]利用K-means算法对主题词向量聚类，得到融合后的主题，进而建立文本集在时间片上的话题演化路径；Li等[41]针对短文本语义稀疏问题，通过引入维基知识库对模型语义进行扩展，结果表明，改进的主题漂移检测方法能够更有效跟踪短文本流中的主题漂移。

对于话题演化的路径分析，上述研究大多是把文档划分为不同的时间片，然后在每个切片中提取主题，再通过计算特征词或主题词之间的语义关联情况来实现话题演化路径分析。但是，时间的分割往往是主观的，一些话题通常存在于多个甚至全部的时间切片中，这种方法将导致话题过多过杂。另外，由于网络本身具有的发散性、渗透性和随意性等特点，使得事件在发展过程中可能朝任何一个方向转换，这导致原有的话题可以衍生出多个与之相关的话题且话题的内容产生较大偏移，而计算不同时间片话题之间相似度的方法难以揭示话题漂移的方向。

3.6 话题演化趋势预测

话题演化趋势预测是话题演化研究的一个拓展问题，是指利用历史数据预测未来的话题演化趋势，可用于挖掘潜在的热点话题等多个方面。目前关于话题演化趋势预测的研究工作主要集中在话题强度预测上。Wang等[42]在建模的时候引入用户的情感，通过马尔可夫随机场和图熵模型计算社区情感能量，然后基于社区情感能量和话题的传播流行度之间的线性相关性来预测话题的流行度。部分研究开始关注话题演化时间序列分析，但对于话题演化的时序分析，主要通过构建话题演化时间序列模型。例如，裴可锋等[43]对话题热度时间序列进行离散化的DTPM模型能够有效提高话题热度预测的精度。

对于已经流行的话题是否会再次流行，Wang等[44]考虑了用户朋友圈、话题类型和突发事件等因素，然后基于高斯混合分布计算在未来时间段内话题再次流行的概率。然而，对话题内容演化进行预测的研究工作还非常少，常用方法是度量特征词或主题词之间的相似度进行话题演化趋势预测，即语义相似度分析。该方法是对文本进行向量表示，然后计算文本相似度，相似度越大，话题演化趋势的可能性越大。因此，如何结合时序分析和语义分析进行话题演化趋势预测，有待于进一步深入研究。

4 话题演化分析方法的比较

话题演化分析方法是指在话题演化研究中所运用的方法或者模型。目前话题演化分析方法，在话题强度、话题状态、话题内容以及演化路径上有各自不同的特点。另外，时间因素也是不可忽视的重要元素，共有3种引入时间方式的不同方法：①将时间作为可观测变量结合到模型中；②在整个文本集上运用话题模型抽取主题，然后按文本的时间信息，后离散分析话题随时间的演化；③将文本集合先按一定时间粒度离散到不同的时间片，在每个时间片上运用话题模型来获取话题随时间的演化。

本节主要对第三部分提到的各种模型方法进行总结比较，并根据话题演化分析的维度，我们选择了代表模型、研究方法、引入时间方式、演化类型等方面来比较，见表1。

表1 话题演化分析方法比较

代表模型	研究方法	引入时间方式	演化类型
TOT	词共现	作为可观测连续变量	强度演化
Griffiths等[45]	马尔科夫链蒙特卡洛方法（MCMC）	按时间后离散	强度演化
OLDA	LDA	先按时间离散	强度演化和内容演化
OCCTM	CCTM	先按时间离散	演化路径
DTPM	LDA	先按时间离散	话题热度预测

5 研究展望

本文综述了关于话题检测，以及话题强度、话题状态、话题内容和演化路径等相关研究工作，并对话题演化趋势的预测进行了探讨。话题演化研究取得了一些进展，但仍然存在一些挑战性课题，同时这也是未来可能的研究方向。

首先，话题演化研究中挑战性课题之一就是识别出贯穿时间周期内的话题，在此基础上实现话题强度、状态、内容和路径的演化分析。目前，大多数方法是基于划分时间片，通过计算不同时间片中话题的相似性来获得演化的话题，这种方法会产生过多、不连贯的话题，且不能有效解决话题演化时的漂移现象。

其次，已有话题演化的相关研究常采用的词共现分析并不能有效处理复杂语义的词汇，也没有考虑到不同特征词对主题的贡献度差异。此外，大多话题演化状态研究是通过生命周期理论辅助进行状态识别，几乎没有对正在演化中的话题状态进行识别或预测。因此，充分挖掘自由文本中词汇的丰富语义关系和重要程度，设计有效的话题检测方法和演化状态指标，是话题演化的未来研究方向之一。

最后，已有的话题演化趋势预测相关工作多是预测话题强度，很少对话题内容演化趋势进行预测。内容演化趋势预测是指对下一阶段话题的漂移方向，甚至是新衍生的主题进行预测，这为相关部门有效监控信息传播提供了科学依据，是话题演化研究的又一方向。但是，话题演化过程并没有统一、通用的模式，受到诸多不确定因素的影响，给话题内容演化预测带来了巨大挑战。

参考文献

[1] 中国互联网络信息中心. 第48次中国互联网络发展状况统计报告 [EB/OL]. [2021-09-15]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202109/t20210915_71543.htm.

[2] ALLAN J, CARBONELL J G, DODDINGTON G, et al. Topic detection and tracking pilot study final report [EB/OL]. [2021-09-03]. https://max.book118.com/html/2017/0724/123975092.shtm.

[3] 张仰森，段宇翔，黄改娟，等. 社交媒体话题检测与追踪技术研究综述 [J]. 中文信息学报，2019，33（7）：1-10，30.

[4] 刘国威，成全. 基于网络舆情生命周期的微博热点事件主题演化研究 [J]. 情报探索，2018，246（4）：15-23.

[5] CALLON M, COURTIAL J P, LAVILLE F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemsitry [J]. Scientometrics, 1991, 22（1）: 155-205.

[6] 曹树金，岳文玉. 突发公共卫生事件微博舆情主题挖掘与演化分析 [J]. 信息资源管理学报，2020，10（6）：28-37.

[7] ZHOU H K, YU H M, HU R. Topic evolution based on the probabilistic topic model: a review [J]. Frontiers of Computer Science, 2017, 11（5）: 786-802.

[8] DE SOLLA PRICE D J. Little Sciece, Big Science [M]. New York: Columbia University Press, 1965.

[9] 王林，王可，吴江. 社交媒体中突发公共卫生事件舆情传播与演变——以2018年疫苗事件为例 [J]. 数据分析与知识发现，2019，3（4）：42-52.

[10] CHEN C C, CHEN Y T, CHEN M C. An aging theory for event life-cycle modeling [J]. IEEE Transactions on Systems Man, and Cybernetics-Part A: Systems and Humans, 2007, 37（2）: 237-248.

[11] FANG M, CHEN Y, GAO P, et al. Topic trend prediction based on wavelet transformation [C] //2014 11th Web Information System and Application Conference. IEEE, 2014: 157-162.

[12] 谢科范，赵湜，陈刚，等. 网络舆情突发事件的生命周期原理及集群决策研究 [J]. 武汉理工大学学报（社会科学版），2010（4）：38-42.

[13] ALLAN J. Introduction to Topic Detection and Tracking [M]. Boston：Kluwer Academic Publishers，2002：1-16.

[14] SMALL H. Tracking and predicting growth areas in science [J]. Scientometrics，2006，68（3）：595-610.

[15] 蔡永明，长青. 共词网络LDA模型的中文短文本主题分析 [J]. 情报学报，2018，37（3）：305-317.

[16] 谭章禄，彭胜男，王兆刚. 基于聚类分析的国内文本挖掘热点与趋势研究 [J]. 情报学报，2019，38（6）：578-585.

[17] BLEI D M，NG A，JORDAN M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research，2003，1（1）：17-35.

[18] 戴长松，王永滨，王琦. 基于在线主题模型的新闻热点演化模型分析 [J]. 软件导刊，2020，19（1）：84-88.

[19] 汪鸿沁泠，巴志超，李纲. 微信群会话话题强度计算及演化分析 [J]. 数据分析与知识发现，2019，3（2）：33-42.

[20] 岳丽欣，刘自强，胡正银. 面向趋势预测的热点主题演化分析方法研究 [J]. 数据分析与知识发现，2020（6）：22-34.

[21] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning，2001，42（1）：177-196.

[22] BASTANI K，NAMAVARI H，SHAFFER J. Latent Dirichlet allocation（LDA）for topic modeling of the CFPB consumer complaints [J]. Expert Systems with Applications，2019，127：256-271.

[23] WARTENA C，BRUSSEE R. Topic detection by clustering keywords [C] //2008 19th International Workshop on Database and Expert Systems Applications. IEEE，2008：54-58.

[24] HURTADO J L，AGARWAL A，ZHU X. Topic discovery and future trend forecasting for texts [J]. Journal of Big Data，2016，3（1）：1-21.

[25] CHEN J，GONG Z，LIU W. A nonparametric model for online topic discovery with word embeddings [J]. Information Sciences，2019，504：32-47.

[26] LU W，WANG J M，HU J M. Analyzing the topic distribution and evolution of foreign relations from parliamentary debates：A framework and case study [J]. Information Processing & Management，2020，57（3）：102191.

[27] 黄云，张彬连，颜一鸣. 基于可区分语言模型的微博在线话题检测 [J]. 计算机应用研究，2014，31（12）：3539-3542.

[28] 贺敏，徐杰，杜攀，等. 基于时间序列分析的微博突发话题检测方法 [J]. 通信学报，2016，37（3）：48-54.

[29] FENG J，WANGY J，DINGYY. Microblog topic evolution computing based on LDA algorithm [J]. OpenPhysics，2018，16（1）：509-516.

[30] WANG X，MCCALLUM A. Topics over time：a non-Markov continuous-time model of topical trends [C] //Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM，2006.

[31] LIU W，DENG Z H，GONG X，et al. Effectively Predicting Whether and When a Topic Will Become Prevalent in a Social Network [C] //Proceedings of the AAAI Conference on Artificial Intelligence. 2015.

[32] ZHAO J J，WU W L，ZHANG X L，et al. A short-term trend prediction model of topic over Sina Weibo dataset [J]. Journal of Combinatorial Optimization，2014，28（3）：613-625.

[33] 李慧，王丽婷. 基于词项热度的微博热点话题发现研究 [J]. 情报科学，2018，36（4）：45-50.

[34] DU Y J，YI Y T，LI X Y，et al. Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet Allocation-ScienceDirect [J]. Engineering Applications of Artificial Intelligence，2020，87：103270.

[35] 胡艳丽，白亮，张维明. 网络舆情中一种基于OLDA的在线话题演化方法 [J]. 国防科技大学学报，2012，34（1）：150-154.

[36] 余本功，张卫春，王龙飞. 基于改进的OLDA模型话题检测及演化分析 [J]. 情报杂志，2017，36（2）：102-107.

[37] 陈兴蜀，高悦，江浩，等. 基于OLDA的热点话题演化跟踪模型 [J]. 华南理工大学学报（自然科学版），2016，44（5）：130-136.

[38] 曹丽娜，唐锡晋. 基于主题模型的BBS话题演化趋势分析 [J]. 管理科学学报，2014，17（11）：109-121.

[39] GAO W，PENG M，WANG H，et al. Generation of topic evolution graphs from short text streams [J]. Neurocomputing，2019，383：282-294.

[40] 张佩瑶，刘东苏. 基于词向量和BTM的短文本话题演化分析 [J]. 数据分析与知识发现，2019，3（3）：95-101.

[41] LI P P，HE L，WANG H Y，et al. Learning from short text streams with topic drifts [J]. IEEE Transactions on Cybernetics，2017，48（9）：1-15.

[42] WANG X，WANG C，DING Z，et al. Predicting the popularity of topics based on user sentiment in microblogging websites [J]. Journal of Intelligent Information Systems，2018，51（1）：97-114.

[43] 裴可锋，陈永洲，马静. 基于DTPM模型的话题热度预测方法 [J]. 情报杂志，2016，35（12）：52-57.

[44] WANG C，XIN X，SHANG J. When to make a topic popular

again? A temporal model for topic rehotting prediction in online social networks [J]. IEEE Transactions on Signal and Information Processing over Networks，2018，4（1）：202-216.

[45] GRIFFITHS T L，STEYVERS M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America，2004，101（1）：5228-5235.

作者简介

钱莉，女，1997年生，硕士研究生，研究方向：话题演化、舆情传播，E-mail：15852939335@163.com。
朱恒民，男，1974年生，博士，教授，研究方向：数据挖掘、舆情管理。
魏静，女，1982年生，博士，副教授，研究方向：复杂网络、舆情传播。

A Review of Topic Evolution Research

QIAN Li[1] ZHU HengMin[1,2] WEI Jing[1]
(1. School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210003, P.R.China;
2. Jiangsu University Philosophy and Social Science Key Research Base—Information Industry Integration Innovation and Emergency Management Research Center, Nanjing 210003, P.R.China)

Abstract: Topic evolution analysis is of great significance to monitor the general situation of topic spreading and predict its future trend. Based on the systematic investigation and analysis of the related research on topic evolution at home and abroad, this paper summarizes the basis of topic evolution research, and discusses the different dimensions of topic evolution research from the aspects of topic intensity, topic status, content and evolution path, as well as the trend prediction of topic evolution, in addition, it also discusses the different methods of topic evolutionary research. Finally, it points out the shortcomings of the existing research and prospects the future research on topic evolution.

Keywords: Topic Evolution; Topic Intensity; Topic Status; Topic Content; Evolution Path

（收稿日期：2021-09-24）