信念网络在话题识别与追踪中的应用研究

吴树芳　朱　杰　著

科学出版社

信念网络在话题识别与追踪中的应用研究

吴树芳　朱　杰　著

科学出版社

北　京

内 容 简 介

向量空间检索模型在话题识别与追踪领域的成功应用，从理论上证明用于信息检索的贝叶斯网络模型亦可用于该领域。信念网络模型属于基于贝叶斯网络的检索模型的一种，它提供了一个灵活的框架，可以有效地归并不同证据。作者尝试将信念网络用于话题识别与追踪模型的构建，为该领域提出新的研究方法。本书在信念网络检索模型的基础上给出四个话题模型，其中第二个动态话题模型归并了新闻话题的初始证据和更新证据，解决了传统静态话题模型、动态话题模型孰优孰劣的问题，有效控制了话题漂移现象。为提高话题识别与追踪的综合性能，对新闻数据预处理阶段的特征选择、权重计算和模型优化进行了相关研究。

本书可以作为高等院校信息管理与信息系统、计算机科学与技术、管理科学与工程、情报学和图书馆学等专业研究生的教材，也可以作为相关领域研究人员的参考书。

图书在版编目 (CIP) 数据

信念网络在话题识别与追踪中的应用研究/吴树芳，朱杰著.
—北京：科学出版社，2017.3
ISBN 978-7-03-051885-9

Ⅰ. ①信… Ⅱ. ①吴… ②朱… Ⅲ. ①网络检索 Ⅳ. ①G254.92

中国版本图书馆 CIP 数据核字(2017)第 039355 号

责任编辑：赵艳春 / 责任校对：桂伟利
责任印制：张 倩 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销
*
2017 年 3 月第 一 版 开本：720×1 000 1/16
2017 年 3 月第一次印刷 印张：9 3/4
字数：200 000

定价：58.00 元

(如有印装质量问题，我社负责调换)

序

话题识别与追踪技术的目的是将杂乱的信息有效地汇总组织起来，当发现新的话题时发出警告，并对其进行及时、实时地控制和引导。在网络技术非常发达的今天，尤其是智能手机的出现，急剧加速了信息传播的速度。危害性言论的爆发式传播会影响社会安全和稳定，故对其进行及时发现、实时控制、有效引导，是当今社会亟须解决的问题之一。从这个角度来说，话题识别与追踪具有研究的理论意义和现实价值，该技术目前已被广泛应用于信息安全、金融证券、行业调研等领域。

话题识别与追踪模型的构建主要借鉴信息检索领域的向量空间模型，缺点是它采用词袋表示话题和新闻报道，不能直观地模拟话题的演化过程，虽然后续研究者对基于向量空间模型的话题建模进行了一系列改进，但其在解决话题漂移问题时仍存在局限性。已有的概率话题模型主要有 LDA 话题模型和语言模型。语言模型的最大缺点是存在数据稀疏问题，虽然目前常采用线性插值技术来解决该问题，但并不能彻底解决。LDA 模型假设不同主题之间是相互独立的，这点与实际情况不符，实际上多个新闻话题之间往往存在关联，并不是完全独立，故将 LDA 用于话题建模存在一定的不足。综上，对于话题建模的研究存在可探究的空间。

《信念网络在话题识别与追踪中的应用研究》是一部关于话题识别与追踪的著作。该书首次采用信念网络进行话题建模，为话题识别与追踪提出了新的研究方法，其意义深远。该书不仅可以使相关领域的读者较为快速、全面地了解话题识别与追踪的研究现状、相关知识、研究手段等，而且更重要的是提出了一些新的研究点，可供相关研究者进行深入研究。

该书共 9 章：第 1 章详尽介绍话题识别与追踪及信念网络的研究现状，并提出将信念网络用于话题建模具备可行性。第 2 章首先介绍话题识别与追踪领域国际上权威的测试集合 TDT 语料的来源、结构及标注方法，然后介绍该领域的评价标准、发展历程、相关概念、实现方法及经典的话题模型。第 4 章、第 5 章提出四个基于信念网络的话题模型，依据模型的拓扑结构，结合贝叶斯概率及条件独立性假设给出条件概率的推导过程。从优化的角度来说，第 3 章、第 6 章、第 7 章可以结为一组。为进一步提高话题识别与追踪系统的综合性能，第 3 章在数据预处理阶段，结合新闻报道的特点提出融合聚类思想的动态互信息特征选择方法。第 6 章提出对话题动态追踪的结果进行二次检测，以降低动态追踪中的误报率，提高系统的综合性能。第 7 章提出时序权重的计算方法，该方法将时序作为直接性因素融入特征权重计算。第 8 章也是作者进行话题识别与追踪研究过程中的研究点，提出基于同一话

题下的事件相似度计算，该方法在事件相似度计算过程中不仅考虑了事件本身的相似性，还考虑它们的话题归属。第 9 章首先对该书的工作进行总结，然后对未来将要进行的研究进行展望。该书全面、透彻地介绍话题识别与追踪的相关知识，并对该领域做出深入研究。

该书内容安排合理，层次性强，符合读者的认知过程，便于理解掌握。基础知识的介绍非常全面，可作为欲了解该领域的初学者的首选。对于每个研究点的介绍，均采用小论文撰写的格式进行内容组织，可使读者在了解相关背景的基础上深入了解该研究点。章节末尾的实验验证体现该书撰写内容的科学性。此外，在章节末尾展望每个研究点可深入研究的内容，为读者提供新的研究思路。

杨秀丹

2016 年 12 月

前　言

话题识别与追踪作为一项旨在帮助人们应对信息过载问题的研究，现阶段主要以网络新媒体中的信息流为处理对象，自动发现话题并把话题相关的内容联系在一起，其实现过程涉及5个子任务，分别是报道切分、首报道检测、关联检测、话题追踪和话题检测。话题识别与追踪技术可用来监控各种语言信息源，在新话题出现时发出警告。在早期研究中，考虑到话题识别与追踪和信息检索的共性，研究者们尝试将信息检索领域的相关技术应用于该领域。但二者之间亦存在不同，例如信息检索一般具备背景知识和先验需求，而话题识别与追踪是在对话题毫无了解的情况下进行识别与追踪，所以完全将信息检索技术移植到话题识别与追踪是不合理的，但是二者的本源性——模型构建基本相通。

针对话题模型的前瞻性研究来自 Allan 等，他们借用信息检索领域广泛采纳的向量空间模型描述话题的特征空间。虽然向量空间模型是目前常用的话题模型之一，但该模型的缺点是不能很好地融合静态话题模型和动态话题模型的优点从而成功解决二者孰优孰劣的问题，且在解决话题漂移问题时能力有限。信息检索领域主要包括三个模型：布尔模型、向量空间模型和概率模型。向量空间模型在话题识别与追踪领域中的成功应用，从理论上验证了概率模型亦可应用于话题识别与追踪。贝叶斯网络模型是重要的概率模型之一，包括推理网络模型、信念网络模型和 BNR 模型。在过去几十年，贝叶斯网络模型已成功应用于信息检索领域，但目前尚未有人将其应用于话题识别与追踪，本书在这方面做了相关研究，试图为该领域提出新的研究方法。

本书结合信念网络模型和新闻报道的特点，给出四个基于信念网络的话题模型 BSTM-Ⅰ、BSTM-Ⅱ、BDTM-Ⅰ和 BDTM-Ⅱ。BSTM-Ⅰ包括三类节点：新报道节点、术语节点和话题节点，弧体现节点间的隶属关系。BSTM-Ⅱ在 BSTM-Ⅰ的基础上加入事件节点，弧的指向和意义不变，为体现核心报道、核心事件的重要性，BSTM-Ⅱ对核心报道、核心事件中的术语权重进行了两次线性提高调整。BDTM-Ⅰ属于动态话题模型，节点类型和弧的意义与上述模型相同，不同的是在话题追踪过程中，其术语层会随着话题的发展而不断更新，重复出现的术语权重采用求和平均的方法更新，新出现的术语作为新的节点插入术语层。以上三个话题模型沿用传统建模思想，具备和以往模型相同的优缺点。BDTM-Ⅱ打破传统建模的思想，运用信念网络模型提供了一个灵活框架的优势，将术语节点分为两类：初始核心术语节点和更新术语节点，并采用析取手段将它们作为两类证据进行归并。依据模型的拓扑结构、贝叶斯

概率及条件独立性假设，本书给出了上述四个模型计算新闻报道和话题相似度的概率推导过程，用于判断新的新闻报道是否和话题相关。

为进一步提高话题识别与追踪系统的综合性能，本书进行了系统的优化研究。特征选择是话题模型构建的基础，互信息是文本处理领域一种有效的特征选择方法。在基本互信息的基础上，将出现相同高频词的新闻报道进行聚类，计算聚类后术语的互信息，将追踪到的相关报道的发生时间和话题的发生时间量化为时间距离，使其反比影响互信息的动态更新，得到基于聚类的动态互信息计算方法，用于计算新闻报道中术语的权重。为了获得 TDT4 语料中每个话题的初始特征子集规模，给出基于类内距离最小、类间距离最大的目标函数，并采用坐标下降法对其求解，最终完成新闻语料的特征选择。

动态话题模型的典型缺点是误报率较高，优点是其可以体现话题的动态演化过程。如果能在保持动态话题模型优点的同时，降低其误报率，将是该领域的一个新的突破点。本书通过系统分析动态话题追踪的误报原因，提出动态话题追踪中的误报检测。该方法首先分析时间距离、相似度差值、相似话题分布及与核心内容相似度分别如何影响误报检测，然后通过将这四项内容线性调和得到误报检测因子的计算方法，用于判断追踪到的相关报道是否属于误报，若属于误报，则对部分特征权重进行衰减，并确定模型结构是否需要调整。实验采用 TDT4 语料，结合 DET 曲线验证以上研究内容的合理性和有效性。

专著的出版受到如下项目经费支持：河北大学双一流专项资金项目、河北大学中西部提升综合实力专项资金项目、河北省教育厅青年基金项目“话题特征选择方法研究(QN2015099)”、河北省自然科学基金项目“基于贝叶斯网络的话题识别与追踪方法研究(F2015201142)”、河北省社会科学基金项目“京津冀协同发展网络热点话题发现及其应用研究(HB15SH064)”。

感谢我的博士生导师徐建民先生，是他最初帮我选定了这个研究方向，并在研究工作中对我做出了悉心的指导；感谢本书第二作者中央司法警官学院朱杰博士；感谢我的领导及同事杨会良教授、宛玲教授、杨秀丹教授、郭子雪教授，他们给我提供了进行科研的条件及工作支持；感谢我的师弟王丹青、张猛、武晓波、粟武林、李腾飞，师妹刘畅、孙晓磊等，他们帮我搜集了部分材料并协助我完成了大量的实验。

由于本人水平所限，所做研究尚有不足，欢迎相关研究者批评指正。

吴树芳

2016 年 12 月

目　录

第1章 绪 论

1.1 研究背景

网络等新媒体的出现，使新闻信息的涌现性尤为突出，如何有效地组织这些信息成为目前亟须解决的问题，单纯靠人工手动组织是不现实的，话题识别与追踪技术(Topic Detection and Tracking，TDT)的目的就是解决这个问题[1]。话题识别与追踪的前瞻性研究来自于美国国防高级研究项目局，1996 年，在该单位的倡导组织下，与相关大学和公司一起讨论制定了话题识别与追踪的任务划分和具体的评测方法。从 1998 年开始，美国国防高级研究项目局以及美国国家标准技术研究所资助每年召开话题识别与追踪系列评测会议，采用会议推动研究的方式促进该领域的深入发展。

话题识别与追踪的实现过程包括 5 个子任务[2]，分别是报道切分、首报道检测、关联检测、话题追踪和话题检测，这 5 个子任务中的话题检测、话题追踪子任务受重视程度比较高，相关研究也比较多。从获取信息的角度考虑，话题识别与追踪主要涉及两类问题[3]：信息的识别与集成、信息的采集与追踪。基于话题识别与追踪和信息检索的共性，研究者们陆续尝试将信息检索领域的相关技术应用于话题识别与追踪，并实验验证了该应用的有效性[4]。1998 年，Allan 等首次将信息检索领域的向量空间模型应用于话题识别与追踪，并在后续研究中进行了系列改进[5-7]，目前，大多数研究者均采用向量空间模型及其变形进行话题建模。

贝叶斯网络是概率论和图论相结合的产物，在 20 世纪 80 年代已经成功应用于信息检索领域。信念网络模型[8]是基于贝叶斯网络的信息检索模型的一种，它提供了一个有效的推理机制，在概率推导的过程中，通过对条件概率的不同规定，可以模拟不同的排序方法。向量空间模型在话题识别与追踪领域的成功应用从理论上证明信念网络模型亦可应用于该领域，但此项研究在国内外甚少。本书将信念网络检索模型应用于话题识别与追踪，试图为该领域提出新的突破。话题识别与追踪的应用领域要求系统在达到查全任务的前提下，尽可能提高查准率，为此，本书提出新的话题特征选择方法 DCMI，以提高特征子集对话题的描述能力，由于选择的特征子集是话题模型构建的基础，因此也提高了模型对话题描述的准确性。此外，本书还给出动态话题追踪中的误报检测和时序权重，以提高追踪结果的准确性。

1.2 研 究 现 状

1.2.1 话题识别与追踪研究现状

国外对话题识别与追踪的早期研究主要采用聚类、分类方法，或者二者的结合。相比于国外以统计概率模型为主题的研究趋势，国内的相关研究更侧重于围绕话题识别与追踪的本身特色进行探索研究。话题识别与追踪可划分为 5 个子任务，分别是：报道切分、首报道检测、关联检测、话题追踪和话题检测，其中多数研究围绕后 3 项子任务展开，本节将围绕这三个子任务论述话题识别与追踪在国内外的研究现状[9]。

1. 国外研究现状

1) 关联检测

关联检测的主要任务是检测随机选择的两篇新闻报道是否属于同一个话题。与其他子任务不同的是，关联检测的研究并没有直接对应到实际的应用中，但是它对其他任务的影响是不容忽视的。例如，在首报道检测任务中，首报道检测系统可以通过关联检测判断候选报道与每一个先验报道之间的相关性，从而判断首报道是否属于一个新的话题。传统的基于概率统计的话题识别与追踪的研究认为报道与话题之间、报道与报道之间的相关性，应该通过检验两者之间的共有特征覆盖率来进行计算，即二者之间的共有特征越多，则其相似度越大。因此，大部分针对关联检测的研究都集中于新闻报道的文本描述以及特征选择。Allan 等采用向量空间模型描述新闻报道，根据特征在报道中的分布来计算其权重，并运用余弦夹角公式计算报道与报道之间的相似度[5]。此外，Yamron 等在计算报道之间的相似度时，将参与检测的两个新闻报道分别看做一个话题和一篇报道，采用语言模型计算报道产生于话题的概率，并调换两个报道的角色，从而从两个角度估计它们产生的概率，并依据二者采用 K-L 距离(Kullback-Leibler Divergence)方法将二者融合，得出二者最后的关联度[10]。上述用于关联检测的向量空间模型和语言模型的主要缺陷是特征空间的数据稀疏性，为解决这个问题，研究者们多采用数据平滑技术。另一种解决数据稀疏问题的有效方法是特征扩展技术。在信息检索领域，特征扩展主要应用于查询扩展，其核心思想是将查询中的特征扩展为同义、近义、本体、相关等特征，从而降低数据的稀疏性。Ponte 等采用向量空间模型，并基于特征的上下文语义关系进行扩展，进而执行关联检测任务[11]。特征扩展技术不仅有助于解决数据稀疏问题，同时也可以辅助关联检测系统削弱特征的歧义性。

2) 话题追踪

传统的话题追踪研究主要有基于知识和基于统计两种方法。基于知识的方法要解决的核心问题是分析报道内容之间的关联和继承关系，通过特定的领域知识将相关报道串联起来。基于统计的方法主要是依据特征的概率分布，采用统计策略裁决报道与话题模型的相关性。

基于知识的话题追踪研究中，比较有代表性的方法为 Watanabe 等[12]面向日本语新闻广播开发的话题追踪系统。Watanabe 等通过形如“正如我所提到的……”“正如我所报道的……”和“正如近期发生的……”等领域知识，检测属于相同话题的新闻报道。实验证明，该方法能够显著提高特定知识领域的话题追踪性能。

基于统计策略的话题追踪研究主要借鉴基于内容的信息过滤方法。信息过滤面向静态需求从动态的信息流中识别和获取相关知识，而传统的话题追踪则根据先验的话题模型追踪后续的新闻报道。但是，二者的相似性使许多信息过滤的相关技术都可以有效地应用于传统话题追踪技术。其中，具有代表性的方法为基于分类策略的话题追踪研究，CMU[13]在传统话题追踪评测中采用了 K-最近邻和决策树两种方法。K-最近邻方法首先根据内容的相关性选择当前报道最相似的 K 个先验报道作为最近邻，然后根据最近邻所属话题类别综合判断当前报道属于哪个话题。决策树方法则根据训练语料预先构造话题的决策树，该树形结构中的每个中间节点代表一种决策树形，节点产生的分支则代表一种决策并指向下一层节点，决策树的叶节点代表话题类别，待测的新闻报道通过决策树的逐层节点判断，最终确定其属于哪个话题。K-最近邻方法和决策树方法的主要缺点是先验相关报道的稀疏性，传统的话题追踪任务一般只给定少量的相关报道作为训练(1～4 篇)。数据稀疏性造成 K-最近邻算法无法使待测新闻报道的最近邻涵盖大量正确的相关报道，而根据这些最近邻得到的话题指向往往是错误的。决策树算法在训练过程中无法为每个属性节点嵌入准确的决策条件。总体而言，在传统话题识别与追踪领域，K-最近邻算法的性能优于决策树算法，其原因为前者可以通过缩减最近邻的规模保证跟踪的正确率，而后者则受限于多层属性需要同时产生正确的决策，相关报道稀疏的训练语料使得多数属性本身就不够准确，因而在没有降低漏检率的同时增加了误报率。

针对传统话题追踪，UMass 采用二元分类方法跟踪话题的相关报道，该方法将训练语料分为相关和不相关两类报道，并根据两类报道与话题相关性的概率分布训练线性分类器，后续报道的相关性依据线性判别式进行判断[14]。二元分类方法的优点在于准确率比较高，但必须依赖于训练语料和分类器的选择，通常选择相关度比较高的不相关报道构成反例类别，从而保证分类的灵敏度。与 K-最近邻和决策树算法类似，该方法也存在数据稀疏的问题，为了解决这一问题，UMass 采用了查询扩展技术。Papka 采用 Rocchio 算法实施跟踪[15]。Rocchio 算法的核心思想是话题模型

的经验性构造策略，即假设相关报道中的特征有助于话题的正确描述，因此这些特征在话题中的权重应加强，而不相关报道的特征则倾向于错误地引导话题的描述，因此权重将被削弱。该算法的优点是，依据该算法，传统的话题追踪系统可以运用追踪到的后续报道不断改进和更新话题模型，从而跟踪话题的后续发展。缺点是Rocchio 算法对阈值的依赖程度很高。

其他面向传统话题追踪的研究工作还包括话题与报道的相似度匹配算法，比如Dragon 通过基于一元语言模型的文本相似度匹配和基于二项式的相似度匹配衡量话题与报道的相关性[16]。有的研究者则尝试采用聚类方法将话题检测系统转换为话题追踪系统。上述方法对于传统的话题追踪任务能够发挥较好的作用，但由于构造初始话题的数据稀疏，因此无法有效地追踪一段时间之后的话题发展。为解决该问题，出现了自适应话题跟踪。

自适应话题追踪研究主要包括两个方面：基于内容和基于统计。在基于内容的自适应话题追踪研究中，Strzalkowski 等[17]尝试采用文摘技术追踪话题的发展趋势。其核心思想是分别提取话题和报道的文摘代表全文描述，话题与报道之间的相关性通过文摘之间的相似度获得。通常情况下，话题的相关报道在不同历史阶段的侧重点有所不同，因此，话题的发展以初始事件为主线，并以后续直接相关的其他事件和活动为延续。基于上述特点，Strzalkowski 将先验相关报道中的事件主体和相关外延以文摘的形式进行提取和组合，根据这种方法构造的话题模型除了涵盖主题信息外，更注重话题发展的层次结构，从而使得追踪系统可以更好地检测话题的后续进展。该项研究的缺点是其没有嵌入自适应的学习机制，话题模型没有利用检测到的后续相关报道自适应地进行更新。

基于统计的自适应话题追踪研究主要借鉴于自适应信息过滤技术。其核心思想是：自适应话题追踪系统可以根据伪相关反馈对话题模型进行自学习，不仅为话题嵌入新的特征，同时动态调整特征权重。其优点在于削弱先验知识稀疏造成的话题模型不完备性，并通过不断自学习提高自适应话题追踪系统跟踪话题发展的能力。Dragon 和 Umass 是最早尝试无指导自适应话题追踪的研究单位。其追踪系统每次检测到相关报道，都将它嵌入话题模型并改进特征的权重分布，后续报道的相关性则以新生成的话题模型为评估对象，从而实现追踪系统的自学习功能。Dragon 和 Umass 的区别在于，前者把系统看作相关的报道嵌入训练语料，并基于语言模型构造新的话题模型；后者则把所有先验报道的质心作为话题模型，并将先验报道与话题模型相关度的平均值作为阈值，后续跟踪过程中每次检测到相关报道，都将其嵌入到训练语料，并根据上述方法重新估计话题模型和阈值。总体而言，这两种方法并没有在很大程度上提高话题追踪系统的性能。主要原因在于自学习模块对于跟踪反馈不施加任何鉴别地全部用于话题模型的更新，因此学习过程将大量不相关信息也嵌入话题模型，从而导致话题漂移现象的出现。基于这一现象，LIMSI[18]在原有自学习

过程中嵌入二次阈值截取功能，通过设置一个比阈值更高的过滤指标，截取伪相关反馈中相关度较高的报道嵌入话题更新模块，控制话题漂移。

到目前为止，话题追踪子任务的相关研究已经取得很好的效果，但如何更有效地追踪话题的后续发展仍然是该领域有待深入研究的课题。近期更多的研究集中于相关报道的概率分布和话题随时间衰减趋势的估计。未来的研究重心在于如何有效利用新闻语料的时间特征，并分析话题发展在时间轴上的分布。

3)话题检测

针对话题检测的研究比较多，可分为在线话题检测、新事件检测、事件回顾检测和层次话题检测，下面将对其研究现状分别展开论述。在线话题检测的主要任务是检测新话题并收集后续相关报道。通常，在线话题检测系统的检测原理集中于相关报道的聚类算法，即在线监视后续的报道数据流，如果截获与之前聚类得到的话题不相关的报道，则检测到一个新的话题，否则将该报道归为已有的聚类。对于在线话题检测的早期研究主要集中在聚类方法的选择与融合上。在话题识别与追踪评测会议上，参加在线话题检测的所有单位都尝试使用单路径聚类算法对新话题进行检测，此外，CMU 尝试采用凝聚层次聚类算法进行检测，但是获得的效果略差于单路径聚类，而 Paka 则对比了不同聚类算法在在线话题检测中的效果，并尝试融合各自的优点解决在线话题检测的已有问题。

新事件检测逐渐成为辅助话题检测的重要组成部分，新事件检测与首报道检测任务很相似，唯一的区别是前者提交的最新事件可能相关于历史上的某一个话题，而后者必须输出话题的最早相关报道。新事件检测的主流研究来自于 Allan[5]和 Yang 等[19]，他们通过建立一个在线识别系统检验报道流中新出现的事件，进入系统的报道需要与每个已知的事件模型计算相关度，并根据预设阈值判断是否为新事件的首次报道，如果条件成立则根据报道建立新的事件模型，否则将其归为已知的事件模型。传统的新事件检测研究采用基于统计原理的文本表示形式，其中最常用的表示方法是向量空间模型，事件模型和报道的相似度采用余弦夹角公式或者 Hellinger 距离公式计算。统计模型的最大缺陷在于无法有效区分同一话题下的不同事件。话题经常被不同事件触发而重复出现，因为话题描述的是所有相似事件具备的共性，而事件之间的区别集中于时间、地点、人物等实体之间的异同。以“恐怖袭击”话题为例，其中包括“2001 年美国 9·11 恐怖袭击事件”“2002 年印度尼西亚的巴厘岛惨案”和“2004 年马德里系列爆炸案”等。从内容上分析，这些事件的相关报道中都会频繁出现“恐怖分子”“自杀式”“袭击”“损毁”和“死亡”等特征，并且这些特征在报道中出现的频率相对较高。因此，根据传统基于统计的策略，这些特征往往构成事件模型的主体，因而很难区分同一话题下的不同事件。与此不同的是，以名实体为主的特征集合，如“9·11”“美国”“巴厘

岛”和“马德里”等，对于不同事件的区分贡献度更高。Kumaran 等[6]对此进行了相关研究。

事件回顾检测的主要任务是回顾过去所有发生过的新闻报道，并从中检测出未被识别到的相关新闻事件。事件是发生在特定时间和地点的事情，而话题则不仅包含作为种子的事件和活动，同时也包含与其直接相关的事件与活动。事件回顾检测的任务是辅助话题检测系统回顾整个新闻语料，从中检测出相关于某一话题而未被识别到的一类新闻事件，其研究的必要性来源于新闻报道的波动性。首次提出事件回顾检测并给出定义的学者是 Yang，他采用凝聚式聚类算法与批平均聚类算法相结合的策略，将近似于同一话题模型的相关事件综合在一起作为话题检测的结果，从而使话题检测系统具备了回顾相关事件的能力[19]。虽然独立于事件回顾检测方向的相关研究较少，但由于回顾事件检测和新事件检测中都涉及未知事件的识别与发现，因此很多学者尝试使用新事件检测中的相关研究同时处理回顾事件检测问题。

层次话题检测是面向话题检测中两种不恰当的假设提出的：一个假设是所有报道与相关话题的近似程度都在一个层次上，另一个假设是每篇报道只可能相关于一个话题。实际上，报道的主题与话题的相关程度往往分布于不同层次，比如“人民币升值”和“建行、交行上市”两篇报道，虽然它们都相关于同一话题“2005 年中国十大金融事件”，但是主题侧重点的差异造成它们与话题的对应程度处于不同层次。此外，这两篇报道都可以分别划分到“汇率制度改革”类和“商业银行改革”类的话题模型当中，因为报道不总是仅仅相关于一个话题，往往不同话题的相关报道存在交集。层次话题检测通常可以采用基于一个根节点的非循环有向图描述话题包含的层次结构。一种解决层次话题检测的方法是凝聚层次聚类算法。其核心思想是计算当前聚类集合中每对聚类的相关度，将满足阈值条件的一对聚类融合成新的聚类，通过反复迭代这一过程，系统最终把话题模型构造成具有层次关系的非循环有向图。层次聚类算法的一个重要缺陷是时间和空间复杂度比较高，对其改进方案的研究来自于 Cutting 等[20]和 Trieschnigg 等[21]。

上述研究现状围绕话题识别与追踪的研究任务展开论述，实际上，为进一步提高话题识别与追踪系统的性能，国外研究者在研究中考虑了新闻报道的时序性。Yang 等通过考虑文档输入的时间顺序，实现对事件的首报道检测[19]。Brants 等将时序用于术语权重计算，提出 ITF-IDF 模型[22]。Mirza 认为时序信息是新闻语料的重要特点，从不同事件的时间信息中可以挖掘、发现其因果关系，进而提高系统的新事件发现能力，并实验验证了其合理性[23]。Kimura 等在分析舆情形成的过程中，对时序信息进行了量化，给出时间衰减函数，以更好地反映舆情的演化[24]。

2. 国内研究现状

话题识别与追踪作为信息处理领域的研究分支逐步成为国内重要的研究热点。相比于国外以统计概率模型为主体的研究趋势，国内的相关研究更侧重基于话题识别与追踪的本身特色进行探索，包括基于新闻报道的名实体、时序性和层次性展开研究。

名实体是描述话题或报道语义的一类特殊语言单位，它对精确刻画核心内涵和区别不同主题具有重要意义。话题识别与追踪系统应用名实体改进性能的方法主要包括如下两个方面：名实体特征权重的再分配和名实体相关性与其他特征相关性的线性组合。国内较早将名实体融入话题识别与追踪系统的研究来自贾自艳等[25]。赵华等则通过分析英文写作的习惯，自动识别新闻报道中首字母大写和全部大写的特征，其认为该类特征不仅包含名实体，也包含报道需要重点强调的特征，并在此基础上采用相关度加权和的方式评估报道和话题的相关性[26]。上述方法采用经验分配权重或调整相关性线性比例，因此无法保证系统性能的稳定性。张阔等基于χ^2分布统计测试集合中各名实体类别与话题类别的相关性，并将相关度指标融入特征权重计算，提高了新事件检测系统的性能[27]。针对名实体中义同形不同的实体无法匹配的问题，宋丹等面向地点类名实体建立地理树，匹配过程基于两名实体在地理树中路径的覆盖率进行计算[28]，该方法因无法处理诸如人名类等其他实体而存在局限性。在此基础上，骆卫华等基于概念一致性匹配同义的名实体[29]，该方法的缺点在于依赖词典的规模和训练语料的新旧，对于报道流中最新出现的名实体依然无法匹配。

新闻话题起始于种子事件并包含后续相关事件，事件的重要描述特征为时间，话题各个相关报道之间往往具有时序关系。贾自艳建立了统一时间表述方式的机制，在此基础上将当前报道与话题框架下的新事件的时间取差值，并利用该值削弱基于内容获得的相关度[25]。赵华等[26]和金珠等[30]的相关研究也考虑了时间信息，他们希望通过时间信息的使用可以发现话题演化的边界，增强系统对话题的追踪能力。以上研究的缺点是他们提出的方法主观性比较强，而且对语料规模的依赖性也较大，因此系统的稳定性不是很好，为解决这个问题，宋丹提出时间“覆盖矩阵”[28]。洪宇等在相关研究中考虑了时间粒度对系统性能的影响[31]。徐建民等在其相关研究中也考虑了新闻报道的时序性：一是在事件相似度计算中建立了时间模板，将时间信息融入最终的相似度计算中，二是将时间信息视为影响特性权重的直接性因素，提出时序权重的概念及计算方法。[32]。

新闻话题的另一个特点是层次化和结构化。层次化将同一话题下的相关报道组织为宏观到具体的层次体系；结构化则侧重于挖掘和表征同一话题的不同侧面。国内尝试层次话题模型的研究来自于骆卫华等[29]，在研究中，首先基于时序关系对报道进行分组，然后进行组内层次聚类，最后按照时间顺序采用单路径聚类策略合并