文章编号：1003-0077(2019)07-0001-10

社交媒体话题检测与追踪技术研究综述

张仰森[1,2]，段宇翔[1]，黄改娟[1,2]，蒋玉茹[1,2]

（1. 北京信息科技大学 智能信息处理研究所，北京 100192；
2. 国家经济安全预警工程北京实验室，北京 100044）

摘 要：随着计算机的普及与互联网的高速发展，Facebook、Twitter、新浪微博等社交媒体逐渐成为人们信息交流的主要渠道。然而，由于社交媒体信息具有数量庞大、结构复杂、传播速度快等特点，人们无法从中快速准确地获取想要的信息。于是，话题检测与追踪技术应运而生，它将用户关注的信息从大量无序信息中筛选出来，经过细致的过滤和有效的整合，生成简单、清晰的话题信息，并在此基础上实现对话题的追踪和发展趋势分析。该文对社交媒体上的话题检测与追踪工作进行综述，首先论述了话题检测方面的三类方法，包括基于主题模型的话题检测、基于改进聚类算法的话题检测和基于多特征融合的话题检测；其次，对话题追踪的研究成果进行了介绍，主要分为非自适应话题追踪和自适应话题追踪两大类；最后，列举出社交媒体话题的检测与追踪中存在的问题以及对未来研究的展望。

关键词：话题检测；话题追踪；聚类；主题模型

中图分类号：TP391 **文献标识码**：A

A Survey on Topic Detection and Tracking Methods in Social Media

ZHANG Yangsen[1,2], DUAN Yuxiang[1], HUANG Gaijuan[1,2], JIANG Yuru[1,2]

(1. Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing 100192, China;
2. Beijing Laboratory of National Economic Security Early-warning Engineering, Beijing 100044, China)

Abstract: Social media such as Facebook, Twitter, and Sina Microblog have become the main channels for people to exchange information. To deal with the large quantity, complex structure and the fast transmission speed of social media information, the technology of topic detection and tracking comes into being to generate simple and clear topic information. This paper reviews the work done on social media topic detection and tracking. Firstly, it summarizes three types of topic detection methods based on topic model, clustering algorithm and multi-feature fusion, respectively. Secondly, it introduces the researches on topic tracking in two categories: non adaptive topic tracking and adaptive topic tracking. Finally, it lists the problems in the current topic detection and tracking technology, and discusses the prospects of future researches on social media.

Keywords: topic detection; topic tracking; clustering; topic model

0 引言

随着计算机与互联网技术的蓬勃发展，互联网信息呈现出爆炸性增长，越来越多的人将互联网视为获取信息的最佳平台。如今，我们所处的不再是信息贫乏的时代，而是一个充斥着海量信息的新时代，所面临的问题也从如何获取信息变成了如何在短时间内获取有价值的信息。关键词检索是目前从海量信息中获取有用信息的主要途径，但通过关键词检索得到的信息，其冗余度往往较高，同时有用信息也常常丢失。因此，人们迫切希望有一种方法可

收稿日期：2018-08-30 定稿日期：2018-11-28

基金项目：国家自然科学基金(61772081,61602044)；科技创新服务能力建设—科研基地建设—北京实验室—国家经济安全预警工程北京实验室项目(PXM2018_014224_000010)

以自动处理大量信息并挖掘相关的话题，对话题相关信息进行有效的组织，以便于人们查询。话题检测与追踪(topic detection and tracking，TDT)技术就是在这种需求下应运而生的，它可以帮助普通网民从海量信息中筛选感兴趣的话题信息，也可以帮助相关部门对舆情进行监控。通过话题检测技术发现热点话题，使用话题追踪技术对检测到的热门话题进行后续追踪，这样就可以有效地组织起一个与某话题有关的信息集合，进而可以探索事件中各种信息之间的关系。

本文第1节介绍了话题检测与追踪的发展历程；第2节介绍话题检测技术相关成果与方法；第3节介绍话题追踪技术，从非自适应话题追踪和自适应话题追踪两个方面进行介绍；第4节列举了话题检测与追踪技术中存在的难题，并对该领域的发展前景进行展望。

1 话题检测追踪研究概况

1.1 话题检测与追踪的研究历程

1996年，美国国防高级研究计划署迫切地需要一种可以实现新闻数据流主题判断的全自动化技术，于是就产生了话题检测与追踪技术的概念。话题检测与追踪技术的发展可大致分为三个阶段，如表1所示。

表1 话题检测与追踪技术发展历程

阶段	描述
第一阶段(1996—1997)	作为TDT研究的起步阶段，研究人员确定了一些基本概念，并建立了最早的相关语料库
第二阶段(1998—2002)	相关研究组织开始组织一些测评活动，受到了高校和研究所的关注，TDT研究逐渐走进人们的视野
第三阶段(2002—)	TDT研究走向繁荣，尤其是社交媒体信息量出现井喷式增长后，面向社交媒体的话题检测与追踪研究现实需求旺盛，学术界和企业界在社交媒体TDT领域做出了许多卓有成效的工作

1.2 话题检测与追踪的研究要素

话题检测与跟踪技术中的“话题”与一般的信息技术中涉及的“话题”不同，它表示一个相对具体的“事件”，而不是某一个“领域”。例如，韩美军演、福岛核电站泄露等。下面将介绍话题检测与追踪研究中的四个研究要素，以便更好地理解本文的研究内容。

(1) 话题：通常是指一个由若干个相关子事件或活动组合而成的事件集合。一个话题往往经历事件的产生、发展、演化、消亡四个阶段。例如，寻找森林大火的幸存者、进行灾后重建等，都可以视为与某次自然灾害相关的话题。

(2) 事件：通常是指发生在特定时间、特定地点，具备时间、地点、对象三要素的事情[1]。例如，2001年7月13日，在俄罗斯首都莫斯科，国际奥委会主席萨马兰奇宣布北京成为2008年奥运会主办城市。

(3) 主题：主题的定义相对宽泛，可以简单理解为多个相关话题的抽象描述，但并不涉及任何实际事件。例如，“自然灾害”就是一个主题，“奥运会”也是一个主题。

(4) 报道：报道是指与话题事件相关，包含多个描述语句的新闻片段。例如，据中央气象台消息，10日白天起，持续多日的南部强降雨天气范围继续扩大，强度显著增强，中央气象台1月10日18时发布暴雨红色预警。

在检测追踪技术的文献调研中，本文主要针对话题和事件这两个要素展开。从话题的相关定义可以看出，如果一个目标事件与某个话题内的事件有联系，那么可以认为该事件在该话题的范围内，事件也可以看作话题的一种低粒度的展现。

目前，主要有两种类型的话题，一种是以新闻报道为主体的传统媒体话题，另一种是以微博、Twitter为代表的社交媒体话题。其中，社交媒体话题建立在Web 2.0之上，它与传统媒体话题的区别主要体现在以下三点：①以新闻报道为主体的传统媒体在传播信息时由编辑对信息进行细致的人工处理，话题中心清楚、明确，而社交媒体中的大部分内容是由每一个用户自由创造和编辑的，话题中心远没有传统媒体那么清晰；②社交媒体比传统媒体包含更多的信息，以微博为例，其不仅有转发、评论、点赞等信息，还有标签、影响力、地理定位等诸多非文本信息；③社交媒体较传统媒体而言，口语化倾向更加明显，规范性较差。上面所列举的三个显著区别导致了社交媒体话题的检测与追踪难度比传统媒体更高。

1.3 话题检测与追踪任务

美国国家标准技术研究所为TDT研究设定了五项基本任务，包括：报道切分任务、话题跟踪任

务、话题检测任务、首次报道检测任务、关联检测任务。

1.3.1　报道切分任务

报道切分任务(story segmentation task，SST)要求将原始报道分割成具有完整结构和统一主题的报道。如果有一条包括不同类型信息的报道，报道切分系统需要对报道进行识别并按照要求切分。SST最初针对的是新闻广播报道，其切分方式包括以下两种：一、直接切分音频信号；二、将音频信号转为文本信息后进行切分。报道切分过程如图1所示。

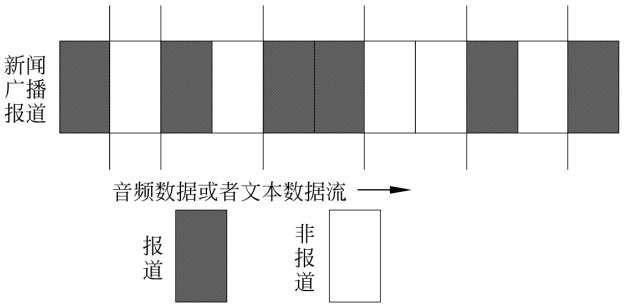


图1　报道切分过程

1.3.2　话题跟踪任务

话题跟踪任务(topic tracking task，TT)是对已知的话题报道进行后续跟踪。由于已知的话题没有明确、详细的描述，描述信息主要是给定的若干篇相关报道。美国国家标准技术研究院为每一个待测话题提供1～4篇相关的报道，同时提供了相应的训练语料来训练跟踪系统和更新话题模型。话题跟踪任务通过计算后续数据流中每一篇报道与话题模型的匹配程度来判断新数据是否属于该话题，从而实现跟踪功能。

1.3.3　话题检测任务

话题检测任务(topic detection task，TD)主要是检测系统中未知的话题。TD任务在构建话题系统时的先验信息非常少，因此，TD系统必须在不清楚话题信息的情况下完成检测模型的构建。同时，构建的检测模型不能仅针对一个特殊的话题，而是应可以检测所有的话题。通过检测模型对后续数据流的检测和识别，找出数据库中没有出现的话题并生成“新话题”[2]。话题检测过程如图2所示。

1.3.4　首次报道检测任务

首次报道检测任务(first-story detection task，FSD)是要在时序报道流中检测出各种话题的第一篇报道。总的来讲，FSD与TD有相似之处，但是

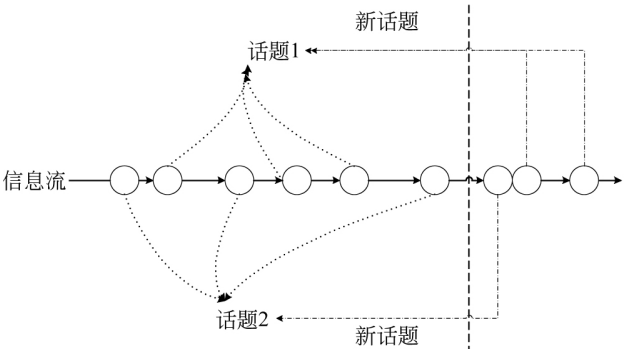


图2　话题检测过程

FSD的结果是某话题的第一篇报道，而TD的结果是关于某一话题的一系列报道，可以说，FSD是话题检测系统的基础和前提。

1.3.5　关联检测任务

关联检测任务(link detection task，LDT)是判断两篇报道是否属于同一个话题。与TD相同，LDT也没有先验信息辅助判断。所以，LDT系统必须能够自己分析报道所描述的话题，并通过对比话题模型来判定两篇报道的话题相关性。

2　话题检测技术

2.1　基于主题模型的话题检测

话题检测首先是在离线的静态文本中提出的，而静态文本的话题检测一般都是基于LDA(latent Dirichlet allocation)主题模型或者改进的LDA主题模型。LDA主题模型将一篇文档理解成由若干隐含主题组合而成，而隐含主题通过文档中一些特定词语来体现[3]。一般情况下，隐含主题被视为词的一种概率分布，单个文档可以由多个隐含主题按照一定比例来构成。本节将介绍三个典型的改进LDA主题模型，分别是有监督潜在狄利克雷模型(supervised LDA，sLDA)、标签潜在狄利克雷模型(labeled LDA，L-LDA)、在线潜在狄利克雷模型(online LDA，OLDA)。有监督学习与无监督学习最主要的区别在于，有监督学习不仅将训练数据传给计算机，还将带标签的数据传给计算机。在话题检测与追踪领域，有监督LDA模型的效果要明显优于无监督LDA模型，故仅介绍有监督LDA模型。sLDA[4]是一个可以添加额外属性的话题检测模型，与普通的LDA模型的区别是，sLDA含有一个甚至多个文本标签，可以通过文本标签对建模过程进行监督。Zhang等[5]使用经整理过滤后的北弗

吉尼亚州和纽约市2016年300万条的Twitter正文和评论数据,进行交通事故的话题发现。在进行sLDA主题建模时,添加了上述两个地区的高速公路事故记录和15 000个环路探测器的交通数据。实验结果证明,66%以上的事故可以通过事故日志找到,80%以上的事故能够从探测器中的交通数据找到,其检测准确率比LDA和SVM更高。既然sLDA需要通过标签的建立来实现话题检测,那么如何寻找最合适的标签便成为了最大的问题。于是,Ramage等[6]在2009年提出了L-LDA模型,这是一个基于多标签文本的主题模型,通过将标签直接映射到主题的方法以实现文档的多标签决策。但是,L-LDA模型没有考虑到人为设置的文档类别标签和通过主题模型提取出来的标签之间的差异性,从而导致模型与文档数据无法充分拟合,泛化能力较差。例如,LDA模型在进行话题提取时,经常会生成我们无法理解的内容,如果简单地把生成的内容与文档进行匹配关联,就会导致检测准确率的下降。周先琳[7]对新浪微博短文本进行预处理后,使用改进后的VSM特征选择方法对文本特征进行选择,并构建动态L-LDA模型。基于4万多条预处理后的新浪微博文本进行实验,可以发现,动态L-LDA模型与LDA模型相比,前者在微博动态文本主题挖掘方面有明显的优势。同样的,为了解决L-LDA模型无法充分拟合和泛化性能较差的问题,江雨燕等[8]提出了一种可用于文档多标签判定的改进L-LDA模型,该模型定义了类别标记在独享主题、共享主题之间的映射关系,这样的映射关系可以更加真实地反映文档的生成过程。基于新浪微博数据的实验表明,该模型可以有效地解决类别标记在共享主题和独享主题中分析困难的问题。

前面介绍的几种模型都是在静态数据下进行实验的,但在真实情况下,数据通常不是静态的,而是以在线文本数据流的形式存在,所以,将时间属性引入LDA模型后就构建了OLDA模型。该模型为了保证主题的延续性,将范围广泛的主题进行一定的缩小,对即将消失的话题在时间粒度上做出延续,减轻了主题演化过程中的偏差问题。余本功等[9]提出了一种改进的双通道OLDA模型,该模型一方面改进了文档中主题分布与词分布之间的遗传度,另一方面改进了词概率的计算方法,有效解决了因为新、旧主题混合和冗余词较多而导致的新兴主题检测困难的问题。

2.2 基于改进聚类算法的话题检测

当前,适用于文本领域的聚类算法主要有四种,分别是:基于划分的聚类算法、基于增量式的聚类算法、基于层次的聚类算法和基于图模型的聚类算法。因为基于划分的聚类算法在话题检测与追踪任务中的效率较低,所以本文仅对后面三种聚类算法进行介绍。

2.2.1 基于增量式的聚类

增量式聚类算法是一种高效的处理文本数据流的算法,其中Single-Pass算法较为简单且应用最广。Single-Pass算法是处理流式数据的经典算法,对于输入的流式数据,按照输入顺序依次将每一条数据与已有类别进行匹配,若匹配成功则将该条数据归入该类别,若匹配失败则创建一个新类别来存放该数据,这样就实现了流式数据的聚类。结合微博文本和微博评论信息都是逐步增量产生的特点,下面将对Single-Pass算法在话题检测中的应用进行介绍。

由于Single-Pass聚类算法是随机选取聚类中心的,所以其聚类效率较低,针对这一缺点,李倩[10]提出了一种改进的Single-Pass聚类算法。在聚类中心的选择上,设置邻域半径和最小密度阈值,并根据文档处于邻域半径内的文档数目与最小密度阈值的大小关系来确定初始聚类中心。在相似度的比较上,不是简单地将新文档与类中所有的文档进行比较,而是与主题相似程度最高的文档进行比较,如果其相似度小于设定的相似度阈值,则不需要再与其他文档进行比较,极大地提升了检测效率。叶施仁等[11]提出了一种结合孤立点预处理和Single-Pass聚类的中文微博热点话题检测模型。该模型主要有三部分工作:①优化微博文本的特征选择策略;②提出了微博文本阈值的概念,将主题分散的文本视为噪声并进行过滤;③引入主题词的概念,而主题词是根据中心向量的特征权重确定的。因为该模型加强了对孤立点的处理,同时优化了中心向量的特征选择和相关权重的设置,所以过滤掉了大量的噪声数据,使主题聚类更加准确。不同于叶施仁采用的设置文本阈值来进行噪声过滤的方法,周雪梅等[12]在进行微博话题检测时引入了文本重构的思想,在文本中定义了主题块和细节块两个模块,主题块包括文本的标题和首段信息,细节块包括文本的其余部分和文后的评论信息。因为标题和首段信息往往是文本的总结归纳,最具有区分性,所以用主题

块划分出不同的主话题，而主话题下的小话题则是利用细节块划分。实验语料来自 2015 年 5 月的新浪微博的社会新闻模块，通过实验数据分析得知，当主话题阈值为 0.28，子话题阈值在 0.28 到 0.58 之间时，子话题区分效果基本可以与人工效果媲美。

2.2.2　基于层次的聚类

k-means 算法是一种简单好用的划分聚类算法，但是算法中 k 值的选择和初始聚类中心点的选择是 k-means 算法的重点和难点。不同于 k-means 聚类算法，层次聚类是对样本逐层聚类，直到满足聚类要求，避免了参数设置和聚类中心点选取的难题。

Peixian Chen 等[13]在进行 Twitter 研究的过程中提出了一种称为 HLTA 的分层主题检测方法，这个方法使用分层潜在树模型来模拟单词共现。HLTA 中的每个潜在变量都表示文档的分区，分区中的文档集群即视为主题，而这个主题一定是在属于该主题的文档中以高概率出现，而在不属于该主题的文档中以低概率出现。HLTA 不同于基于 LDA 的分层主题检测方法，虽然两种方法都定义了文档的概率分布，但它们使用不同类型的观察变量和潜在变量。实验结果表明，HLTA 在模型拟合和主题层次结构质量方面优于基于 LDA 的方法。鉴于中文微博具有规模大、话题多、话题无关性强等特点，Xiao Geng 等[14]提出了一种三层混合聚类算法进行话题检测。第一层使用 K-means 算法，对微博文本进行话题聚类。第二层应用凝聚式层次聚类算法，将相同主题的文本结合成小型簇。前两层已经消除了大部分的干扰噪声，第三层再次使用 k-means 算法，对原先分配给错误簇的文本进行重新聚类，实现对聚类结果的修正。

2.2.3　基于图模型的聚类

基于图模型的聚类与其说是聚类算法，还不如说是一种图的向量表示。基于向量进行表示之后，一般可以采用其他的聚类方法得到最后的聚类结果。所以基于图模型的聚类既依赖于向量表示，也与之后采用的聚类算法有关。

Dong 等[15]针对 Twitter 上的突发话题，提出了一种面向突发话题的图模型，该模型可以表示大量 Twitter 用户对突发话题进行传播的拓扑结构。通过该模型可以从宏观上分析突发事件的传播模式，从微观上挖掘突发事件的传播特点。实验结果表明，通过该方法可以有效地从突发事件中发现新兴话题。不同于微博主体的长文本，在线社交媒体用户每天在评论区会产生大量的短文本评论信息，传统的话题检测对有限的包含大量信息的文档有良好的效果，但是对海量的包含信息量低的小文本见效甚微。因此，Kambiz Ghoorchian 等[16]提出了利用降维和聚类技术的话题检测方法，首先将输入的文档集压缩成一个密集的图，并在图中创建多个稠密的拓扑区域，然后将图分成若干个密集的子图，每一个子图代表一个主题。该方法与标准的 LDA 和 BiTerm 方法相比，不仅保持了更好的精度，而且执行速度快一个数量级。

传统的主题检测方法通过挖掘语义关系聚合成主题，但是这样的方法忽略了文档间的共现关系。为了解决这个问题，Zhang 等[17]提出了一种混合关系分析方法来整合语义关系和共现关系。具体而言，该方法将多个关系融合成语义图，并使用图分析方法从语义图中检测主题。通过梳理图中的关系，不仅可以更有效地检测话题，还可以利用潜在的共现关系挖掘潜在的重要信息。

由于社交媒体产生的内容大大超出了人工处理这些数据的能力，而已经提出的各种自动主题检测方法，大部分都基于文档聚类和突发检测，它们无法实现对噪声文件的过滤，而对于噪声文件的过滤又是话题检测中的重点和难点。因此，Pablo Torres-Tramón 等[18]提出了一种基于拓扑数据分析的主题检测方法，它将欧几里德特征空间转换成一个拓扑空间，在这个拓扑空间中，被视为噪声的不相关文档的形状很容易与局部相关的文档区分开来。根据点（即文档）的连通性将该拓扑空间组织在网络中，并且根据连接组件的大小进行二次过滤，以达到去除噪声文件、实现话题检测的目的。

2.3　基于多特征融合的话题检测

基于多特征融合的话题检测可以充分地利用多特征数据，实现对话题的精确检测。根据话题检测的方法途径，把多特征分为两大类：一类是基于文本的多特征，另一类是基于非文本的多特征。

2.3.1　基于文本多特征融合的检测方法

基于社交媒体文本特征的方法是指利用微博、Twitter 等新兴社交媒体上的文本消息，根据事件随时间的变化不断对新出现的话题做出检测。

由于中文微博多数为短文本，甚至是超短文本，文本的稀疏性往往导致文本相似度的度量不准确。黄贤英等[19]提出一种基于多维度的微博短文本相似度算法，该算法根据词形相同和词义相近来寻找微博短文本中的公共块，构建基于公共块序列的语

义相似度。利用微博短文本发布时间、转发与评论等信息来修正该语义相似度，形成新的微博短文本相似度算法。最后，将新的微博短文本相似度算法结合 Single-Pass 聚类算法，最终实现对微博话题的检测。

金镇晟[20]利用特征词的时间属性和增长程度这对属性，在传统的 TF-IDF 基础上提出一种改进的特征提取算法，称之为 TF-IDF-KE(term frequency-inverse document frequency-kinetic energy)，用以解决突发性热点话题在聚类时特征不明显的问题。该算法结合动能原理，将特征项的突发值用动能的概念进行描述，并加入权值计算中，提高了突发性特征项的权重，最后通过文本聚类实现了微博的话题检测。该方法描述了文本和特征项所具有的动态属性，实验结果表明，该方法能够有效地提高话题检测的效果。刘志雄[21]针对微博话题的热度、突发性以及时序特征，提出了一种时间窗口下的融合词重要度的微博话题检测方法。该方法首先根据时间属性对微博文本进行分块处理，然后根据词在时间块里的热度进行排序，并选取热度最大的 n 个词作为主题词候选词。其次，以主题词候选词为基础构建词共现网络，并利用社区划分算法对该词共现网络进行主题划分，最后对每个划分社区内的候选词进行重要度排序。利用新浪微博半个月内的三个主题板块数据进行实验，结果表明，通过该检测算法进行社区话题发现有较高的召回率，但是在准确率和漏检率上并没有突出的成绩。

2.3.2 基于非文本多特征融合的检测方法

基于社交媒体文本特征的检测方法主要围绕关键词特征进行，但是随着非文本媒体的盛行，仅依靠关键词特征已经无法满足当前网络环境下的话题检测，结合社交网络中丰富的用户数据(例如，用户行为、好友关系、地理位置、视频等)来进行话题检测就显得尤为重要[22]。

有些微博话题可能在全网范围内并不突出，但是在某一局部地区却是一个热点话题。针对这一情况，李正[23]提出利用地理位置信息进行中文微博突发话题检测，一方面，根据微博空间环境现状，增加适应环境的文本过滤规则，尽可能地过滤冗余数据；另一方面，将微博文本中出现的地点名词与微博所携带的空间地理位置信息进行匹配，并提出“亲历度”的概念，用此概念提高相应微博分词的基础权重，以提高对应用户在该事件上的话语权，从而达到更加精准地获取突发词集的目的。

传统的话题检测方法主要集中在单一媒体上，Zhang 等[24]提出将互联网视频和新闻报道中丰富的多媒体信息进行融合，实现跨媒体话题关键词的提取。首先，利用视频相关的文本信息和新闻标题，找出粗加权密集关键词组；然后，利用文本链接和可视化链接细化关键词组并更新权重；最后，将文档与细化的关键词组重新关联以形成与事件相关的文档集。在包含网络视频和新闻图片报道的跨媒体数据集上进行实验，取得了良好的检测效果。在以微博、Twitter 为首的社交媒体中，也存在着大量的视频和图片信息，多媒体信息融合同样可以在社交媒体领域得到应用。

随着社交媒体功能的多样化，越来越多的用户行为信息和时间属性被挖掘出来并用于话题检测。万越等[25]结合微博数据的时序特征以及社交网络用户的行为特征，提出一种动量信号增强模型来进行微博突发话题检测。该文首次提出用影响力因子来修正动量模型的误差。影响力因子是指当前时间点前指定周期内的数据对当前数据的变化的影响，其将作为修正词频序列的依据。通过对比用于检测是否存在突发信号的 MACD 值指标和提前设置好的突发性阈值，判断目标特征词是否是突发特征词。最后，通过 k-means 聚类算法将特征词归类合并，得到突发话题。贺敏等[26]针对微博数据稀疏、微博内容间的关系难以准确度量、微博内容多而杂的特点[27]，提出了基于特征驱动的中文微博话题检测方法。该检测方法通过选取有意义的词或者词组来获取微博特征，将微博的转发数、评论数以及点赞量等文档影响力和关注数、粉丝数等微博博主影响力组成特征影响力属性组并进行建模。最后，根据特征属性划分话题关键特征和噪声特征，并将话题关键特征之间的互信息作为最邻近聚类法的距离度量，通过关键特征的最邻近聚类得到话题结果。根据新浪微博 1 000 个加 V 的活跃博主的 78 万余条微博消息进行实验，相比于传统的 k-means 方法，该方法的检测准确率、召回率以及 F_1 值都有将近 20% 的提升，故该方法有一定的应用价值[28]。Fang 等[28]从话题在时间和空间上局部分布的性质入手，提出了一个基于多视图聚类的新框架 MVTD。该框架通过整合 Twitter 中的语义关系、社会标签关系和时间关系，提出了一种基于后缀树的新文档相似性度量方法和基于后缀树的新关键词提取方法。通过在真实 Twitter 数据上进行实验，发现基于多视图聚类的新框架 MVTD 的聚类性能远远优于单一视

图，并且对于 Twitter 的话题检测有良好的效果。

3　话题追踪技术

话题追踪的主要任务是，在已知目标话题的基础上对后续报道进行持续追踪。由于社交媒体的迅速普及，话题追踪技术应用到了微博、贴吧、论坛、博客等社交媒体平台上。话题追踪可以简单地分为两个步骤：第一步，训练并得到话题模型；第二步，根据得到的话题模型进行判断。该过程如图 3 所示。

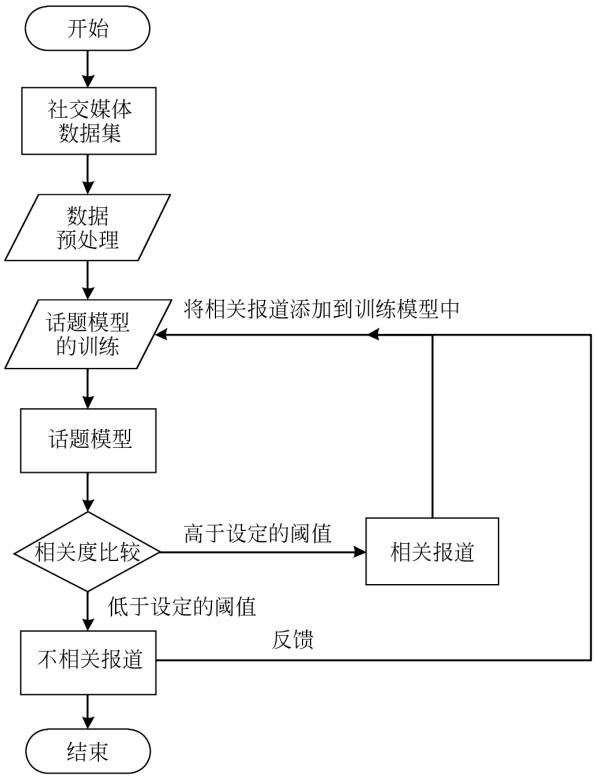


图 3　话题追踪基本流程

本节将话题追踪方法分为非自适应话题追踪和自适应话题追踪两种，自适应话题追踪的优越性在于无指导条件下的自适应能力，而这种自适应能力可以有效地解决“话题漂移”现象。

3.1　非自适应话题追踪

非自适应话题追踪有基于知识和基于统计两种研究思路。基于知识的话题追踪主要是分析报道内容之间的相关关系，并利用与报道内容相关的领域知识对报道进行归类追踪。基于统计的话题追踪主要是利用统计学方法分析报道与话题模型之间的关联程度。

鉴于话题追踪方法大多面向新闻、博客和微博等社交媒体，席耀一等[29]针对网络论坛的结构和内容特点，提出基于语义相似度的论坛话题追踪方法。该方法提取百度贴吧、网易论坛、天涯社区等诸多平台的帖子的关键词，分别构建出话题关键词词表和帖子关键词词表，并建立话题与帖子的文本表示模型。利用《知网》的语义框架计算帖子关键词与话题关键词的相似度，当相似度高于设定的阈值时就可以判定该帖子为话题追踪的目标帖。实验证明，该方法的准确率和 F_1 值均高于传统的基于向量空间模型的话题追踪方法，可以广泛应用于论坛领域的话题追踪。不同于前面提到的语义相似度分析，Chen 等[30]提出一种基于语义相关度的微博文本主题跟踪方法来解决微博文本的稀疏性问题。该方法根据微博的结构化信息，以及《知网》的语义关系网络，构建了针对微博的语义关联模型。根据该模型提取文本信息，并以关键词列表的形式表现出来，结合文本相似度的相关理论，综合衡量文本与主题之间的相关性。实验结果表明，该方法比向量空间法和单纯基于文本相似度的方法能更好地降低错误率，大大提升了话题追踪的效果。唐晓波等[31]基于维基语义扩展网络构建出一种微博话题追踪模型，该模型旨在解决微博文本中的语义稀疏性问题和话题漂移性问题。首先，使用维基百科数据进行知识库的构建；其次，利用该知识库对目标微博文本的特征向量进行扩展，经过扩展后的微博文本对事件的描述能力有了很大的提升；最后，通过支持向量机（support vector machine，SVM）进行语义层面的话题追踪。实验结果表明，与传统的 SVM 方法和自适应 SVM 方法相比，基于维基百科进行语义扩展后的 SVM 分类模型有效降低了分类器对初始话题数量的敏感性，同时减轻了话题漂移现象对微博话题追踪产生的影响。

基于统计策略的话题追踪方法主要是根据话题模型与后续报道相关性进行判断，而基于分类策略的话题追踪又是基于统计策略中最常用到的方法。卡内基梅隆大学在话题追踪任务中利用统计策略率先提出了两种方法，分别是 k—最近邻（k-nearest neighbor，KNN）和决策树（decision tree，D-Tree）。

马萨诸塞大学的 Papka[32]采用 KNN 分类算法，将与当前报道最相似的 k 个报道作为最邻近报道，则待测报道所属的话题就由这 k 个报道中出现频率最高的话题来决定。卡耐基梅隆大学的 Carbonell 等[33]采用 D-Tree 算法进行话题追踪，该算法通过训练语料来构建决策树，决策树中的每个中

间节点代表一种决策属性，节点向下的分支则代表一种决策，最终在叶节点得出所属的话题。大量实验和论文表明，基于KNN算法的话题追踪效果要优于D-Tree算法，其原因在于KNN可以通过减少k值来保证追踪的正确率，而D-Tree必须依赖多层树结构得出正确的追踪策略，这样很容易造成漏检和误检。

由于微博信息有变化速度过快、噪声高、文本较短等缺点，所以针对微博的新兴话题追踪的效率一直不高[34]，Huang等[35]提出了一种新兴的微博话题追踪方法，它将新词检测与相关话题挖掘相结合。具体来说就是通过一个基于局部线性的加权回归算法来计算单词的新颖性，同时抑制已有话题的单词新颖性，最后利用单词新颖性和衰落性来追踪新兴的话题。在超过100万条的微博评论数据上进行实验，结果表明该方法在检测新兴话题和追踪现有话题上有着良好的性能。

3.2 自适应话题追踪

非自适应话题追踪是根据少量的话题报道来构建话题模型，进而实现话题追踪。现实生活与之非常类似，用户对突发性话题的了解通常也非常少，而这也是经过训练得到的话题模型不够准确的缘故。因此，研究一种拥有自我学习能力的自适应话题追踪系统(adaptive topic tracking，ATT)就显得尤为重要。自适应话题追踪的核心思想是对话题模型进行自学习，不仅为话题嵌入新的特征，同时可以动态调整特征权重。其优点是可以减小因为先验知识不足而导致的话题模型不完备的问题，同时还可以通过自学习机制实现对话题的持续跟踪。

Khandelwal等[36]是最早进行ATT研究的成员之一，他们根据话题报道构造话题模型，将话题报道与构造出来的话题模型之间的相关度的平均值作为阈值，当有后续相关报道输入时，将其放入训练语料进行训练并重新构建话题模型和阈值。该自适应话题追踪方法有一个很大的缺陷：对于系统反馈不进行任何验证，即反馈信息中包含的相关和不相关报道都会放入训练语料重新训练，这会导致模型更新出现偏差，产生话题漂移现象。针对上述方法可能会造成话题漂移的问题，美国BBN公司的Lo等[37]在其研发的LIMSI话题追踪系统中，采用设置二次阈值的方式来解决反馈信息没有验证的问题。只有在满足反馈阈值的前提下才会把信息提交给系统进行模型更新，反馈阈值的设定有效降低了话题漂移现象的产生[38]。LIMSI系统有静态和动态两种权重更新策略，经实验证明，面对社交媒体的话题追踪时选用动态权重更新策略效果更佳。

有些研究者在微博话题追踪中引入语义信息[39]，刘彦伟[40]将话题中心向量引入话题模型的同时，使用语义相似度对判断结果进行修正，将微博文档划分到对应话题后进行话题中心向量的自适应调整。不同于利用语义信息的自适应话题追踪，柏文言等[41]提出了一种融合用户关系的自适应微博话题追踪方法。首先将追踪时间窗内的推文映射到特征空间，形成候选推文集合，然后根据推文的分布特点和话题追踪的目的对推文特征空间做出变换，最后利用改进的k-means聚类算法对候选推文集合进行二元聚类，划分出相关推文集合。使用Twitter平台的实时数据进行实验，结果表明，该方法能够及时追踪话题的热度变化和话题焦点的演变，同时也可以提高微博话题追踪的稳定性。

因为话题的演化过程与时间紧密相关，Fuling Hu等[42]提出了一个事件—时间关系模型来研究话题跟踪任务，该方法主要通过识别和挖掘后续报道中的事件—时间流，将事件的时间属性引入向量空间模型，并将该模型应用于话题跟踪的相关决策，最后根据时间属性重新调整特征向量的权重分配，实现自适应话题追踪。实验结果显示，在DET曲线性能评估系统平台上，该模型能够比非自适应话题追踪模型更加准确地跟踪话题事件的演化过程。

4 社交媒体话题检测追踪研究展望

中文语义信息复杂多变，想要通过机器对文本信息进行深层挖掘就显得格外困难。另外，针对目前热门的社交媒体，又出现了海量短文本，甚至超短文本的挑战。因此，有许多方面的问题需要解决。

(1) 海量信息问题。由于社交媒体数据量庞大，且更新速度快，如新浪微博在2017年有3.76亿月活跃用户，1.65亿日活跃用户，每天发送微博数目超过1亿条，所以建立针对社交媒体的流数据处理系统是一个亟需解决的问题。为了解决上述问题，可以在原有算法的基础上，结合Hadoop、Spark等大数据分析工具对微博数据进行处理和

分析。在话题检测追踪的任务中，需要研究出有效的针对大规模文本数据的快速聚类算法，以应对这一挑战。

（2）噪声干扰问题。社交媒体中充斥着大量的广告信息，这些广告不仅包括公司的推广信息，还有很多个人用户的商品买卖信息，这些广告噪声对话题检测与追踪基本没有实际的意义，甚至会使检测结果出现一定程度的偏差。针对垃圾邮件问题，研究者提出了许多垃圾邮件检测算法。在未来的研究中，可以将这些算法改进，并应用到微博数据中。同时，鉴于微博的广告中有很大一部分是商品信息，可以将各大电商网站的商品信息作为微博广告库的扩展信息源，这样可以省去大量的人工广告信息标注任务。

（3）多源信息传播问题。在当前的社交网络中，大部分话题检测追踪的研究都是针对单一数据源的，如国外的 Twitter 或者中国的新浪微博。虽然它们的数据量巨大且更新迅速，但是如果忽略了社交媒体平台间转发、分享等功能，就会遗失许多其他来源的话题信息，导致无法全面地获取新兴话题以及话题的演变过程。因此，可以考虑在检测过程中加入关联网站信息，全面反映网络中目标话题的分布态势。

（4）非文本信息问题。针对热门话题的检测，应该同时考虑文本信息和非文本信息对检测过程的贡献。近期的研究中，有人将社交媒体的时序特征和用户细节特征等非文本信息用于话题检测模型的构建，取得了一定的效果，但是，目前的研究只简单利用了用户的权威度和评论转发数等用户行为特征[43]。在以后的研究中，可以将用户的影响力信息、用户参与社交媒体互动的行为信息等特征纳入话题检测与追踪的研究中。

（5）结果评估问题。话题检测与追踪结果的评价方法除了传统的准确率、召回率、F_1值外，第三方使用效果、人工评估也是经常用于评价的指标。第三方使用效果指的是将话题检测追踪模型应用于文本分类、信息检索等方面，利用文本分类的效果、信息检索的准确性对话题检测追踪的效果进行间接评估。在实际工作中，人工评价其实是最可靠、适用范围最广的评估方式，这也是话题检测追踪领域一个亟需解决的问题，即找到一种自动的适用全领域的评估方法。

（6）深度学习缺乏应用的问题。深度学习在话题检测与追踪领域缺乏有效的应用，我们认为有两方面的原因：一方面，深度学习从 2006 年取得突破性进展后，最先应用于图像和语音领域，而在自然语言处理领域的应用则是近些年才开始的，这导致话题检测与追踪领域目前还没有较为成熟的模型。另一方面，话题检测与追踪数据的时效性很强，而深度学习方法非常依赖对数据的训练和学习，所以频繁的数据变化也是导致深度学习方法没有广泛应用到话题检测与追踪任务的原因之一。

参考文献

[1] Fiscus J G, Doddington G R. Topic detection and tracking evaluation overview[M]. Topic detection and tracking. Springer US, 2002: 17-31.

[2] 赵华，赵铁军，于浩，等. 面向动态演化的话题检测研究[J]. 高技术通讯，2006，16(12)：1230-1235.

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(Jan): 993-1022.

[4] Mcauliffe J D, Blei D M. Supervised topic models[C]//Proceedings of the 20th International Conference on Neural Information Processing Systems, 2008: 121-128.

[5] Zhang Z, He Q, Gao J, et al. A deep learning approach for detecting traffic accidents from social media data[J]. Transportation Research Part C: Emerging Technologies, 2018, 86: 580-596.

[6] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. Association for Computational Linguistics, 2009: 248-256.

[7] 周先琳. 基于动态 Labeled-LDA 模型的微博主题挖掘[D]. 合肥：合肥工业大学硕士学位论文，2015.

[8] 江雨燕，李平，王清. 用于多标签分类的改进 Labeled LDA 模型[J]. 南京大学学报(自然科学版)，2013，49(04)：425-432.

[9] 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(02)：102-107.

[10] 李倩. Single-Pass 聚类算法的改进及其在微博话题检测中的应用研究[D]. 济南：山东师范大学硕士学位论文，2016.

[11] 叶施仁，杨英，杨长春，等. 孤立点预处理和 Single-Pass 聚类结合的微博话题检测方法[J]. 计算机应用研究，2016，33(08)：2294-2297.

[12] 周雪梅，闫用杰，程山英，等. 基于文本重构的网络话题检测模型研究[J]. 南昌航空大学学报(自然科学

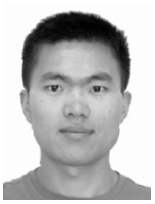
版),2015,29(03):32-37.
[13] Chen P, Zhang N L, Liu T, et al. Latent tree models for hierarchical topic detection[J]. Artificial Intelligence, 2017, 250: 105-124.
[14] Geng X, Zhang Y, Jiao Y, et al. A novel hybrid clustering algorithm for microblog topic detection [G]. AIP Conference Proceedings. Melville: AIP Publishing LLC., 2017, 1890(1): 040074.
[15] Dong G, Yang W, Zhu F, et al. Discovering burst patterns of burst topic in Twitter[J]. Computers & Electrical Engineering, 2017, 58: 551-559.
[16] Ghoorchian K, Girdzijauskas S, Rahimian F. DeGPar: Large scale topic detection using node-cut partitioning on dense weighted graphs[C]//Proceedings of the 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017: 775-785.
[17] Zhang C, Wang H, Cao L, et al. A hybrid term-term relations analysis approach for topic detection[J]. Knowledge-Based Systems, 2016, 93: 109-120.
[18] Torres-Tramón P, Hromic H, Heravi B R. Topic Detection in Twitter using topology data analysis[G]. LNCS9396: Proc of 2015 International Conference on Web Engineering. Berlin: Springer, 2015: 186-197.
[19] 黄贤英,陈红阳,刘英涛.短文本相似度研究及其在微博话题检测中的应用[J].计算机工程与设计,2015,36(11):3128-3133.
[20] 金镇晟. 基于改进的TF-IDF算法的中文微博话题检测与研究[D].北京:北京理工大学硕士学位论文,2015.
[21] 刘志雄. 面向用户兴趣与社区关系的微博话题检测方法[D].北京:北京交通大学硕士学位论文,2017.
[22] 刘玉新. Web 2.0互联网在线话题发现和热度评估[D].广州:华南理工大学硕士学位论文,2013.
[23] 李正. 基于地理位置信息的中文微博突发话题检测技术研究[D].哈尔滨:哈尔滨工程大学硕士学位论文,2016.
[24] Zhang W, Chen T, Li G, et al. Fusing cross-media for topic detection by dense keyword groups[J]. Neurocomputing, 2015, 169: 169-179.
[25] 万越,隋杰.基于用户行为影响的微博突发话题检测方法[J].中国科学技术大学学报,2017,47(04):328-335.
[26] 贺敏,刘玮,刘悦,等.基于特征驱动的微博话题检测方法[J].中文信息学报,2017,31(03):101-108,124.
[27] 王征,王林森,赵磊.基于信息密度的微博突发话题检测模型研究[J].情报理论与实践,2016,39(03):125-129.
[28] Fang Y, Zhang H, Ye Y, et al. Detecting hot topics from Twitter: Amultiview approach[J]. Journal of Information Science, 2014, 40(5): 578-593.
[29] 席耀一,林琛,李弼程,等.基于语义相似度的论坛话题追踪方法[J].计算机应用,2011,31(1):93-96.
[30] Chen H, Lu J, Wang F, et al. A new method of topic tracking for Micro Blog texts based on semantic Relevance[C]//Proceedings of the 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). IEEE, 2017, 2: 349-353.
[31] 唐晓波,王中勤,钟林霞.基于维基语义扩展的微博话题追踪模型研究[J].情报科学,2017,35(02):80-85.
[32] Papka R. On-line new event detection, clustering, and tracking[R]. University of Massachuseffs Amherst, MA, USA, 1999.
[33] J Allan, J Carbonell, G Doddington, et al. Topic detection and tracking pilot study: Finalreport[C]// Proceedings of the DARPA BroadcastNews Transcription and Understanding Workshop, Virginia: Lansdowne, 1998, 194-218.
[34] Lewis DD, Schapire R E, Callan J P, et al. Training algorithms for linear text classifiers[C]//Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1996: 298-306.
[35] Huang J, Peng M, Wang H, et al. A probabilistic method for emerging topic tracking in Microblog stream[J]. World Wide Web, 2017, 20(2): 325-350.
[36] Khandelwal V, Gupta R, Allan J. An evaluation corpus for temporal summarization[C]//Proceedings of the 1st International Conference on Human Language Technology Research. Association for Computational Linguistics, 2001: 1-5.
[37] Y Lo, J L Gauvain. The LIMSI topic tracking system for TDT 2002 [C]//Proceedings of Topic Detection Tracking Workshop. Gaithersburg, USA, 2002.
[38] 张辉,周敬民,王亮,等.基于三维文档向量的自适应话题追踪器模型[J].中文信息学报,2010,24(05):70-76.
[39] 郑燕. 基于增量学习的自适应话题追踪技术研究[D].济南:山东师范大学硕士学位论文,2013.
[40] 刘彦伟. 微博话题追踪系统的研究与实现[D].北京:北京交通大学硕士学位论文,2013.
[41] 柏文言,张闯,徐克付,等.一种融合用户关系的自适应微博话题跟踪方法[J].电子学报,2017,45(06):1375-1381.
[42] Hu F, Wu G, Zhao C. Research on topic tracking based on event-time relation model[J]. Intelligent Computer and Applications, 2016, 1: 008.

(下转第30页)

[18] 王荀，李素建，王宇昕. 内容标签和关系标签相结合的汉语篇章标注规范[J]. 中文信息学报，2015，29(03)：65-70.

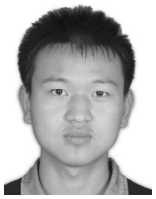
[19] Shi Y. The establishment of modern Chinese grammar：The formation of the resultative construction and its effects[M]. John Benjamins Publishing，2002.

[20] Joty S，Carenini G，Ng R，et al. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics，2013，1：486-496.



侯圣峦(1989—)，通信作者，博士研究生，主要研究领域为自然语言处理、文本摘要、机器学习。
E-mail：houshengluan1989@163.com

张书涵(1991—)，博士研究生，主要研究领域为自然语言处理、机器学习。
E-mail：zhangshuhan@ict.ac.cn



费超群(1992—)，博士研究生，主要研究领域为自然语言处理、机器学习。
E-mail：feichaoqun15@mails.ucas.ac.cn

(上接第 10 页)

[43] 彭敏，官宸宇，朱佳晖，等. 面向社交媒体文本的话题检测与追踪技术研究综述[J]. 武汉大学学报(理学版)，2016，62(3)：197-217.

张仰森(1962—)，博士，教授，博士生导师，主要研究领域为中文信息处理、网络内容安全、人工智能。
E-mail：zhangyangsen@163.com

黄改娟(1965—)，高级实验师，主要研究领域为中文信息处理、智能仓储与物流。
E-mail：hgj@bistu.edu.cn

段宇翔(1992—)，硕士研究生，主要研究领域为中文信息处理。
E-mail：tjlgdxdyx@163.com