Journal of Computer Applications
计算机应用, 2020, 40(6): 1565 - 1573
ISSN 1001-9081
CODEN JYIIDU
2020-06-10
http://www.joca.cn

文章编号:1001-9081(2020)06-1565-09
DOI:10.11772/j.issn.1001-9081.2019101871

基于网络社交媒体的子话题检测技术综述

理姗姗[1],杨文忠[2,3*],王　婷[1],王丽花[1]
(1. 新疆大学 软件学院,乌鲁木齐 830046; 2. 新疆大学 信息科学与工程学院,乌鲁木齐 830046;
3. 社会安全风险感知与防控大数据应用国家工程实验室(中国电子科学研究院),乌鲁木齐 830000)
(*通信作者电子邮箱 ywz_xy@163.com)

摘　要:在当前多种平台崛起的互联网背景下,与传统媒体相比,网络社交媒体中的数据具有传递速度快、用户参与度高、内容覆盖全等特点,其中存在着人们关注并发布评论的众多话题,而一个话题的相关信息中可能存在更深层次、更细粒度的子话题,针对该问题进行基于网络社交媒体的子话题检测技术的研究,这是一个新兴且不断发展的研究领域。通过社交媒体获取话题及子话题信息并参与讨论,这一方式正全方位、深层次改变着人们的生活,但是该领域技术还不成熟,且相关研究在国内尚处于起步阶段。首先,简述网络社交媒体中子话题检测的发展背景和基本概念;其次,将子话题检测技术分为七大类,对每类方法均加以介绍、对比和总结;然后,将子话题检测方式分为在线检测和离线检测两种方式,并将这两种方式进行对比,列举通用技术及两种方式下的常用技术;最后,概括了该领域当前不足及未来发展趋势。

关键词:子话题;话题检测和追踪;网络社交媒体;话题层次;子事件
中图分类号:TP181; TP391　　**文献标志码**:A

Survey of sub-topic detection technology based on internet social media

LI Shanshan[1], YANG Wenzhong[2,3*], WANG Ting[1], WANG Lihua[1]
(1. *College of Software, Xinjiang University, Urumqi Xinjiang* 830046, *China*;
2. *College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang* 830046, *China*;
3. *National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (China Academy of Electronics and Information Technology), Urumqi Xinjiang* 830000, *China*)

Abstract: The data in internet social media has the characteristics of fast transmission, high user participation and complete coverage compared with traditional media under the background of the rise of various platforms on the internet. There are various topics that people pay attention to and publish comments in, and there may exist deeper and more fine-grained sub-topics in the related information of one topic. A survey of sub-topic detection based on internet social media, as a newly emerging and developing research field, was proposed. The method of obtaining topic and sub-topic information through social media and participating in the discussion is changing people's lives in an all-round way. However, the technologies in this field are not mature at present, and the researches are still in the initial stage in China. Firstly, the development background and basic concept of the sub-topic detection in internet social media were described. Secondly, the sub-topic detection technologies were divided into seven categories, each of which was introduced, compared and summarized. Thirdly, the methods of sub-topic detection were divided into online and offline methods, and the two methods were compared, then the general technologies and the frequently used technologies of the two methods were listed. Finally, the current shortages and future development trends of this field were summarized.

Key words: sub-topic; Topic Detection and Tracking (TDT); internet social media; topic hierarchy; sub-event

0　引言

随着互联网媒体技术的飞速发展,众多社交媒体平台随之兴起,例如新浪微博、推特等平台,这些网络平台反映了社会、政治、经济和文化等各领域的热点话题,成为继广播、电视之后最普遍的信息传输方式。其中很多平台都有话题专栏合集,但一般只停留在对话题检测这一层面,却忽略了话题下更细粒度、更深层次、更全角度、更多侧面的内涵。作为随话题检测的发展演变而生的一个新的研究方向,子话题检测旨在解决上述不合理问题,进一步分析同一话题下的组成结构、演化过程和内部关系。在内容上,它有利于全面分析某个话题所包含的不同内容及其关系;在结构上,它有利于建立话题的演化模型,研究发展趋势,更加清晰地掌握网络中话题信息的

收稿日期:2019-11-01;**修回日期**:2019-12-12;**录用日期**:2019-12-17。　**基金项目**:国家重点研发计划项目(2017YFC0820702-3);国家自然科学基金资助项目(U1603115,U1435215);社会安全风险感知与防控大数据应用国家工程实验室主任基金资助项目。

作者简介:理姗姗(1996—),女,河南周口人,硕士研究生,主要研究方向:自然语言处理、文本数据挖掘、信息安全; 杨文忠(1971—),男,河南南阳人,副教授,博士,CCF会员,主要研究方向:网络舆情、情报分析、信息安全、无线传感器网络; 王婷(1996—),女,新疆阿克苏人,硕士研究生,主要研究方向:自然语言处理、文本情感分析、信息安全; 王丽花(1995—),女,河北邯郸人,硕士研究生,主要研究方向:自然语言处理、文本意图检测。

构成情况。自孕育期(1997—2006年)之后,子话题检测技术经历了概念提出期(2007—2009年)、受到关注期(2010—2012年)和兴起与发展时期(2013年以后),在不同阶段具有不同的发展动因及发展结果。

20世纪90年代,随着计算机软件、硬件的发展及互联网技术的完善,舆情监控部门将信息来源转向网络,与传统媒体(电视、广播、报纸、杂志等)相比,网络社交媒体的内容具有规模庞大、形式多样、传播迅速等特点,使得收集并组织相关信息变得愈发困难,话题检测的需求由此而生,其概念源于话题检测与跟踪(Topic Detection and Tracking, TDT)项目[1],此时机器学习(Machine Learning, ML)已成为新的学科并应用于数据分析与挖掘,从而为子话题检测的孕育产生打下基础。进入21世纪之后,互联网中社交媒体逐渐丰富,话题检测技术取得了长足的进步,互联网舆情及信息安全领域的应用需求也随之不断发展,一些学者为全面了解话题各个方面,开始对话题进行细粒度探索,“话题层次”“子话题”等概念被提出(在国外,Nallapati等[2]首次给出在新闻话题内进行事件检测与关系发现的概念;在国内,李军等[3]率先提出“子话题”的概念),并采用聚类、主题模型等技术进行子话题检测。

进入2010年以后,众多社交媒体平台涌现,话题检测成为研究热点,此时子话题检测相关研究也随之进入研究者的视线并受到关注,机器学习、人工智能学术活动空前活跃,自然语言处理技术初步完善。研究者们利用统计分析、知识发现等手段分析数据进行子话题检测,除初始技术外,分类、基于图模型的方法、基于突发状况的方法等技术也很普遍。自2012年进入大数据时代以来,数据挖掘、自然语言处理、多媒体学习等相关领域不断发展,话题检测得到深入研究并取得了丰硕成果,但随着网络社交媒体平台中信息日益变化,它面对着更新的挑战:实时获取热点话题不同方面的内容、全面掌握网络舆情趋势等需求空前旺盛,对话题检测及演化分析的要求也不断提高。现有的话题粒度及层次已无法满足多方面的需求,因此子话题检测技术进入了兴起与发展阶段,研究者们不断探索创新,提出了基于多模态的方法、基于多种技术结合的方法等技术进行子话题检测。

目前,对有关社交媒体子话题的检测技术并没有系统的阐述和介绍。已有的国内外相关综述性文献,如文献[4-9]仅涉及对话题检测技术进行总结,只完成了话题检测层面综述性工作,而对于话题下子话题的概念没有统一的定义,对子话题检测相关技术也没有全面的分类说明。故本文概括当前研究中子话题相关概念,将网络社交媒体子话题检测技术分为基于突发状况的技术、基于分类的技术、基于聚类的技术、基于主题模型的技术、基于结构图的技术、基于多模态特征的技术和其他检测技术,逐类给出了详细介绍,并进行总结和比较;同时,将子话题检测方式按实时性要求分为在线检测和离线检测,对二者作出对比并列举通用技术及不同方式的常用技术;最后,提出了当前研究的不足及对未来发展的展望。

1 研究背景及相关概念

1.1 研究背景

当前互联网数据增长飞快。截至2019年6月,我国网民规模达8.54亿人,较2018年年底增长2 598万人;互联网普及率达61.2%,较2018年年底提升了1.6个百分点[10]。越来越多的人将互联网视为获取知识、传递信息、发表评论和交流看法的最佳媒介,通过各类社交媒体平台,人们可以在网络上实时获取新闻资讯和各种相关报道,发表相应的评论来对自己感兴趣的内容提出见解,并由此形成层出不穷的种种话题。随着用户竞相参与,话题的热度在不断飙升,其相关信息的数量也在激增。

网络社交媒体中的大多数内容都是由用户自发创造的,包括文字、图片、音频和视频等多种表现形式。针对这些数据,传统的话题检测多把话题当作一个整体,往往忽略了话题下子话题的存在及子话题关系演变的刻画。同一个话题下的数据是复杂多样的,包含很多隐藏内容,且可能含有多个联系紧密、相似性很高的子话题,如何更有效地从海量数据中找到用户感兴趣的话题,并挖掘相关话题下的子话题,帮助用户全面准确地了解话题详细内容及演化的各个方面,是研究领域内一个新的难题。

1.2 相关概念

话题、事件、子话题和子事件这几个术语将贯穿全文。在最初研究阶段,“话题”和“事件”含义相同[11],一个“话题”指由某些原因、条件引起,发生在特定时间、地点,涉及特定的参与者,且可能产生一些必然后果的一个事件。对于“子话题”这一术语,在国内已有的研究中使用的比较多,例如文献[3,12-13]等,而在国外通常以“子事件”来进行描述,但实质上是表达同一概念,例如文献[14-16]等。下面对国内外研究中所提出的相关概念加以陈述和总结。

1.2.1 话题检测和追踪相关概念

话题检测和追踪(TDT)是美国国防高级研究计划局(Defense Advanced Research Projects Agency, DARPA)于1996年开展的研究项目,其目标为实现按话题查找、组织和利用来自多种新闻媒体的多语言信息[2],包括新闻报道的切分、新事件识别、报道关系识别、话题识别、话题跟踪和层次话题检测等子任务[17]。其中的“话题”概念不再等同于信息检索中的“主题”,并非某一个“领域”,而是表示一个相对具体的“事件”[12]。某些情况下“话题”与“事件”可以通用,不作严格的区分[18]。为了区别于语言学上的概念,TDT评测会议对相关要素进行了定义[19],陈述如下:

1)话题(topic):由一个种子事件或活动,和全部与之直接关联的后续事件或活动构成。

2)报道(story):新闻片断,包含两个以上独立陈述某个事件的子句,与话题关系紧密。

3)事件(event):由特定原因、条件引起,发生在某些特殊时间、地点,并可能伴随特定后果的特例。

1.2.2 子话题检测相关概念

互联网社交媒体信息的话题内容具有多元性、演化性等特点,大多数关于话题检测的研究仅集中于静态地识别信息数据中存在的话题,却忽略了一个主话题下可能存在的子话题层次,或是忽略了随着时间的推移话题内容可能产生的扩充和演变。目前中外相关领域对子话题的研究中,对“子话题”或是“子事件”的概念没有一个统一的定义,下面描述几种具有代表性的定义。

在中文领域,李军等[3]定义“子话题”是话题内一组相关事件或活动的集合。洪宇等[5]定义“话题”由一个种子事件以及后续直接相关的事件或活动组成,“子话题”是针对其中某一事件的相关描述,“事件”则定义为发生于特定时间和特定地点的事情。吕楠等[20]定义“子话题”为话题的一个方面:话题T在i时刻的状态T_i由若干个子话题组成,记为$T_i=\{T_{1i},T_{2i},\cdots,T_{ni}\}$,每个子话题$T_{ji}$代表话题在$i$时刻的某一个方面。程葳等[21]提出了“子话题”概念:关于同一事件或活动的相似报道集合称为“子话题”;一个子话题可以包含多篇报道,

但唯一从属于一个话题；一个话题可以包含多个子话题。王巍[22]定义“话题”由一个种子子话题和其他相关子话题构成，认为“子话题”等同于“子事件”。代翔等[13]把“话题”“子话题”“事件”划为三个层次，定义“子话题”作为衔接“话题”与“事件”的桥梁，能够相对清晰地呈现某一类具体事情。

在国外研究领域，Nolasco等[14]定义“事件”是受时间和相关位置限制的一个重要事情，“子事件”是通过组合关联关系与另一个事件关联的事件，事件包括两个或多个子事件。Srijith等[23]在文献中给出“子话题”的定义为：子报道检测将与同一现实事件相关的推文分成与不同子事件相关的类，与这些子事件相关的讨论话题称为子话题。Abhik等[24]提出：“子事件”是在某特定事件中按时间或位置分隔的较小事件。Panem等[25]在文献中定义“话题”与现实世界中的重大事件相关，“子话题”则是此类事件细粒度的一个方面。Wu等[26]定义“子事件”是由事件的演变产生的，假设存在关系rel，若关系rel(事件，事件*)为真，则表示“事件*”是“事件”的子事件。

2　子话题检测技术

本文概括了常用的网络社交媒体子话题检测技术，将其分为基于突发状况的技术、基于分类的技术、基于聚类的技术、基于主题模型的技术、基于结构图的技术、基于多模态特征的技术和其他检测技术七大类别，并做出总结和对比。子话题检测技术分类如图1所示。

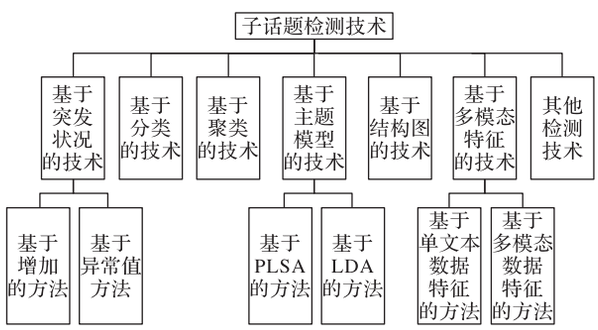

图1　子话题检测技术的分类

Fig. 1　Classification of sub-topic detection technologies

2.1　基于突发状况的技术

基于突发状况的检测方法是用于社交媒体平台子话题检测的一种常见技术，最初被应用于突发事件检测，后被应用至话题检测领域，并不断深入探索，现在也被用于进行子话题检测。该方法有两种思路，即基于增加的方法和基于异常值的方法，研究者们首先提出了基于增加的方法，随后针对其存在的不足提出了基于异常值的改进思路。

2.1.1　基于增加的方法

基于增加的方法其基本思想为：新发生事件迅速吸引了人们的注意，使得社交媒体中与之相关的发文和讨论内容突然增加，因此在子话题的检测中可以考虑评论发布数量或相关词汇频率的增加。在外文领域早期的研究中，研究者们通过比较当前时间片推文数据量与前一时刻的数据量，认为若当前时间片数据量突然增加，这一现象可能反映一个重要子话题的产生，从而来识别自然灾害[27]、政治事务[28]和体育赛事[29]等话题中的子话题。

2.1.2　基于异常值的方法

基于增加的方法可能会出现漏检错误：在网络社交媒体中，由于不同的用户对话题的关注点和关注度均不相同，即使发生了子事件，可能仅有部分用户参与讨论，其数据量并未大幅增加，此时基于增加的方法就无法检测出实际存在的部分子话题。基于异常值的方法观察相关话题下当前时间片与所有历史时间片的推文数据量并统计比较，认为与常规数据相比，当前推文速率是一个异常值时，就产生了子话题。

Chen等[30]使用推特流数据在线检测子事件，使用卡尔曼滤波器、高斯过程和概率主成分分析三种统计方法，将子事件识别过程定义为异常检测问题。Zubiaga等[31]对足球赛事实时总结时，将当前的发文速率与历史所有发文速率进行比较，采取基于异常值的方法进行子事件的检测。实验结果证明，该方法平均覆盖了84%的子事件和100%的关键子事件类型。与基于增加的方法相比，其优点是考虑到特定话题有特定的受众，且在推文速率保持不变时也能检测到赛事中存在的连续子事件。

2.2　基于分类的技术

分类是一种有监督的学习方法，其任务是在预先给定的类别标记集合下，根据文本内容判定它的类别[32]。基于分类的子话题检测算法用分类器来判断文档是否属于特定子话题，基本思想是按照某种规则给样本贴标签，通过学习得到分类器，再对未知类别的样本进行区分归类。常用的分类算法有决策树算法、贝叶斯算法、神经网络算法、逻辑回归算法、支持向量机等。

Sakaki等[33]提出了一种监控推文和检测目标事件的算法，基于推文中的时间、空间特征(关键字、单词数量以及上下文等)来设计推文分类器，可以估计灾难事件位置的中心和轨迹。Badgett等[34]提出了一种自动提取子事件的两阶段方法：在第一个阶段用一个引导人工神经网络来识别可能包含子事件短语的句子；在第二阶段识别出符合预定连词模式的短语，完成子事件提取。Bekoulis等[35]利用推特流的时间顺序并考虑其序列性质，将社交媒体流中的子事件检测问题构造为序列标记任务，本质上是对线性序列中每个元素根据上下文内容进行分类的问题。Chierichetti等[36]使用一个逻辑回归分类器，以推文和转发率为特征进行研究完成子事件检测；随后Araki等[37]对其进行了改进，提出了一个多分类逻辑回归模型，并使用一组丰富的特征识别子事件及确定子事件的关系。Aldawsari等[38]在2019年提出通过有监督的逻辑回归模型来自动识别子事件，并融入了一些语言和叙事特征，以及少量的特征修改。

为训练一个无偏的子事件分类器，需要丰富的先验知识，必须提供大量的样本，并断定所有的待分类样本都一定对应一个类别。但这并不符合实际要求，尤其是面对海量数据时，若想通过数据预处理来满足分类算法的要求，代价会很大，此时可以考虑使用聚类算法。

2.3　基于聚类的技术

聚类是一种非常重要的非监督学习技术，其任务是按照某种标准或数据的内在性质及规律，将目标样本分成若干个簇，保证每个簇内的样本相似性尽可能大，且不同簇间的样本相似度尽可能小。聚类技术被广泛应用于数据挖掘、统计学、机器学习等领域，且在子话题检测领域的最初阶段就被纳入采用。随着子话题检测研究的发展不断完善，常见的聚类算法有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密度的聚类算法等。

张小明等[39]通过引入子话题的方法提高话题检测的准确率，使用基于增量聚类的算法进行自动话题检测，实验表明该方法的召回率为0.80、准确率为0.84、F1值为0.84，能迅速检测话题，且以较小的误差(小于10%)检测出话题数量。代翔

等[13]为解决主题建模分类结果粒度过粗的问题，在主题建模之后通过层次聚类算法对话题下的文本进行二次聚类，得到话题下的子话题，通过实验表明，与Single-Pass算法和*K*-means聚类算法相比，基于层次聚类得到的结果更具有真实性。

2.4 基于主题模型的技术

在子话题检测领域，早期的研究中多使用向量空间模型（Vector Space Model，VSM），但其在语义探索和表示上有许多欠缺之处，因此研究者们提出了潜在语义分析（Latent Semantic Analysis，LSA）模型。随后，引入概率统计方法对其改进，提出了概率潜在语义分析（Probabilitistic Latent Semantic Analysis，PLSA）模型。但是PLSA模型并非完整的概率生成模型，针对这一不足，相关领域专家们又提出了沿用广泛的文档主题生成LDA（Latent Dirichlet Allocation）模型，LDA模型是统计主题模型的典型代表，在文本建模上具有独特的优越之处，因此已成为自然语言处理领域内新的研究热点。

2.4.1 基于向量空间模型的方法

向量空间模型（VSM）是用空间向量表示文本信息的数学模型，可通过计算向量之间的相似性来度量文档间的相似性，其最常用的词权重设置方法是词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）赋权。向量空间模型在文本检索、信息过滤、信息提取、文件索引、分类、聚类等问题中都得到了广泛应用。

由于同一话题内的事件往往非常相似，导致话题内的事件检测精确度较差。为了克服这一缺陷，张阔等[40]使用向量空间模型，根据词频赋权，用层次聚类算法挖掘每个事件的核心词元，利用核心词元完成话题内事件检测与关系发现。针对在新闻话题中报道突发、热点相似且子话题层次丰富的现象，周学广等[41]提出了基于依存连接权VSM的子话题检测与跟踪方法，使用关联词邻接图方法改进VSM，引入词语之间的连接权值，通过依存树分析构造有向节点，在外部引入领域命名实体词典并放大相应权值，从而完成子话题检测与跟踪。该方法能迅速地在特定领域信息范围中检测热点话题，但需要外在的领域词典，因而应用场景过于局限。

2.4.2 基于PLSA及其改进方法

向量空间模型应用十分普遍，但其没有能力探究隐藏在字、词背后的涵义，无法处理一词多义和一义多词问题，而潜在语义分析（LSA）方法的引入能减轻类似的问题。LSA基于奇异值分解（Singular Value Decomposition，SVD），能将高维度的词汇-文档共现矩阵映射到低维度的潜在语义空间，使得表面毫不相关的词体现出深层次的联系[42]，但LSA缺乏严谨的数理统计基础，而且SVD非常耗时。

为此Hofmann[43-44]提出了基于概率统计的PLSA模型，并用期望最大化算法（Expectation-Maximization algorithm，EM）学习模型参数，通过一个生成模型来为LSA赋予了概率意义上的解释。该模型假设：每一篇文档都包含一系列潜在话题，文档中每一个单词都不是凭空产生，而是在这些潜在话题的指引下通过一定概率生成的。图2为PLSA模型，其中：d、Z、W分别表示文档、主题和词语；M和N分别表示文档数和词数。

通过在传统的PLSA基础上引入背景语言模型，能降低背景词对子话题的干扰[45]，周楠等[46]在此基础上发现子话题关键词，结合外部知识库生成事件子话题的标签。通过实验表明，该算法相较*K*-means和LDA等方法具有更好的性能，通过其生成的子话题标签可以发现事件共性，反映子话题热度的趋势，比传统方法具有更好的准确性和概括性。该算法在发现子话题时，能有效克服同一话题下文档的相似性问题，但是采用监督的方法生成子话题标签，当处理庞大的数据量时可能面临计算量巨大、计算复杂度高、时间开销大等问题。

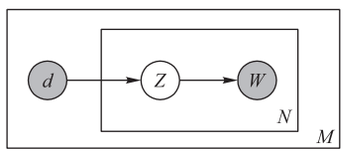

图2 PLSA模型

Fig. 2 PLSA model

PLSA作为LSA的变形，具有更坚实的数学基础及易于利用的数据生成模型，可用于信息检索和自然语言处理等许多领域。但是PLSA并非完备的概率模型，当文档数量增加时，PLSA模型也会线性增加，变得十分庞大；其中EM算法反复迭代，计算开销很大。为了克服PLSA的不足，领域专家们又提出了一些其他的主题模型，其中包括应用最为广泛的LDA主题模型。

2.4.3 基于LDA及其改进方法

LDA由Blei等[47]在PLSA的基础上提出，是一个完整的概率生成模型。图3表示LDA模型，其中：$\boldsymbol{\theta}$代表文本-主题概率分布，$\boldsymbol{\varphi}$代表主题-词概率分布，$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$分别表示$\boldsymbol{\theta}$和$\boldsymbol{\varphi}$的超参数，W表示词语，M、N和K分别表示文本数、词数和主题数。

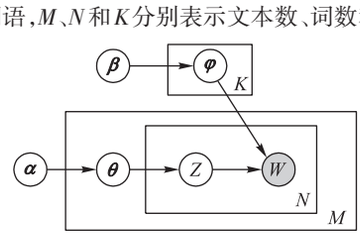

图3 LDA模型

Fig. 3 LDA model

李军等[3]基于LDA模型进行子话题划分，证明LDA模型较VSM模型性能更优越，但是在聚类时忽略了子话题之间的联系。对此，楚克明等[48]提出了基于LDA模型新闻话题的演化方法。通过话题抽取、话题过滤和话题关联三个步骤找到一对多或多对多的子话题之间的演化关系，体现了话题在内容上的变化。Nolasco等[14]在社交网络中收集数据，使用LDA算法和主题标注方法从原始文本进行子事件挖掘，该方法可以自动检测给定事件的子事件，并为其分配代表性标签来进行描述。

LDA主题建模是大批量文本数据中进行话题检测最有效的方法之一，但也存在一定局限性，例如结果过于笼统、可读性差等。为解决上述问题，研究者们对LDA模型做出了改进。Huang等[49]率先探讨了词分配对LDA检测话题能力的影响。Ge等[50]提出了基于关键词的LDA（Key Phrase LDA，KPLDA）模型的中文新闻热点子话题发现和推荐方法，采用关键短语代替独立词作为特征，基于KPLDA模型对语料库进行训练，得到主题短语分布，通过聚类完成子话题发现。实验证明：KPLDA训练时间消耗多于KPLDA，但KPLDA模型的热点子话题发现质量和准确性均优于LDA。苏婧琼等[51]针对LDA建模结果较泛化及传统相似度计算方法语义性欠缺、无法精确区分各个子话题的问题，提出了一种基于LDA和诱导划分（Derived Partition，DP）的子话题划分方法，采用LDA对

文档集建模,综合考虑全覆盖模型在表征文档时的描述能力,使用诱导划分实现子话题划分。该方法对子话题划分的效果很好,但诱导划分的时间复杂度和空间复杂度都很高。李湘东等[52]提出一种基于LDA和知网语义词典(HowNet)相结合的多粒度子话题划分方法,用LDA模型对不同源的新闻集进行初划分,结合HowNet来计算新闻文档之间的相似度,通过增量聚类算法实现子话题划分。

胡艳丽等[53]基于在线主题模型(Online LDA, OLDA)利用先验知识抽取网络信息中隐含的子话题,定义子话题演化类型,根据语义相似度和时序关系建立子话题关联。Srijith等[23]提出分层主题模型HDP(Hierarchical Dirichlet Process),能实时准确检测话题下多数子话题,非常适合子话题检测任务。李静远等[54]提出了抑制背景噪声的LDA子话题挖掘算法,预先抽取专题文档集合的共同背景知识,有效解决了共同背景下专题文章集合的子话题挖掘难题。Banu等[55]提出了一种前景动态主题建模方法,提取噪声内容并从语料库中提取前景推文,在其上构建模型,通过避免噪声数据检测子话题,随时间的推移抽取生成子话题的摘要。

2.4.4　各种主题模型的比较

对各类主题模型的比较如表1所示。

表1　不同主题模型比较

Tab. 1　Comparison of different topic models

模型	优点	缺点
VSM	1)表示简单、可操作性和可计算性强 2)向量维度意义明确 3)实现了语言问题向数学问题的转化	1)向量矩阵高度稀疏 2)未考虑特征项之间的联系,造成语义的丢失 3)新数据加入时需重新计算特征权值,维护成本高
LSA	1)将词和文档映射到潜在语义空间,提高了信息检索的精确度 2)可以解决一词多义和一义多词的问题	1)缺乏严谨的数理统计基础 2)SVD非常耗时
PLSA	1)为LSA赋予了概率意义上的解释 2)拥有更坚实的数学基础,能为信息提取提供更好的词汇匹配	1)无法生成未知的文档 2)随着文档数量增加,模型越来越复杂 3)容易出现过拟合问题
LDA	1)完整的概率生成模型 2)对于每一个主题均可找出一些词语来描述	1)话题数量需要指定 2)检测结果可理解性差 3)对短文本处理效果不好

2.5　基于结构图的技术

随着子话题研究的深入,研究者们发现传统的文本表示方法将词语单独考虑,缺少结构化信息,随后有专家提出使用结构图的方法来表示社交媒体信息:图的点代表词,边表示词与词之间的语义关系,通过构建图来识别文本信息的关联关系,完成话题及子话题的检测。

Liu等[56]构建了话题相关的事件结构图,通过对事件图划分形成子话题。Katragadda等[57]使用时间演化图从推特数据流中检测子事件,定义了用于识别两个图簇关系的度量,并引入事件生命周期模型来映射所识别的关系来检测子事件。Meladianos等[58-59]在推文数据集中检测演化事件中的子事件,将较短时间间隔内连续的推文表示为一个加权的单词图,使用图退化的概念来识别子事件,实验证明基于结构图的方法可以有效地捕捉子事件。

2.6　基于多模态特征的技术

2.6.1　基于单文本数据特征的方法

王巍[22]根据搜索引擎的某个话题结果进行子话题划分,提出了基于关键词和基于时间信息的两种子话题聚类方法。但是在基于关键词的划分方法中并未分析子话题的内容特征;在基于时间信息的划分方法中并未考虑相同时间可能涌现多个子话题的情况。为此,仲兆满等[60]提出融合内容和时间特征对中文新闻子话题聚类的方法,重点分析了子话题内容特征的表现规律,研究了子话题特征词的权重计算和降维方法。

Abhik等[24]通过使用社交媒体数据的多个特征分两步进行子事件检测:首先,将每个特征单独考虑,形成聚类并对其赋权;然后,将所形成的聚类解以主加权方式组合,得到最终的聚类结果。张瑞琦[61]将整合去重后的关键特征映射到话题空间上形成初始话题;然后,对初始话题进行聚类得到子话题,并进行子话题关键特征的抽取。

2.6.2　基于多模态数据特征的方法

除文本数据以外,网络社交媒体中图像、音频或视频形式的数据也蕴含大量信息,在进行子话题检测时值得纳入考虑。Pohl等[62-64]对社交媒体中的文本、图像、视频等多模态数据,采取聚类算法识别与危机相关的子事件,证明了对多模态数据使用聚类技术检测子事件的可行性。

多用户网络社交媒体平台信息中存在数据异构和时间不同步等问题,因此跨媒体库的子事件检测任务准确率不高。Zaharieva等[65]提出多用户图像集中的媒体同步与子事件检测,利用上下文中时间、位置信息以及图像内容来挖掘多模态数据集,研究在数据未同步情况下使用聚类算法检测子事件的可行性。Qian等[66]提出一种基于社交媒体的事件汇总方法,使用用户文本图像共同聚类的方法,从多种媒体类型(用户、文本和图像)的微博中共同发现子事件,通过实验证明,与单一文本聚类方法相比,该方法具有优越性。

2.7　其他检测技术

上述方法并不互相独立,为提升子话题检测的准确率和实用性,许多研究者对各种方法进行探索、结合与改进,提出了一些诸如结合在线和离线的方法、结合概率论和数理统计的方法、结合有监督和无监督的方法等其他检测方法。

Panem等[67]结合离线方法和在线方法,提出了基于推特子话题检测的实体实时跟踪方法,探索了基于语义和基于概念空间表示来解决动态聚类问题的方法。在离线阶段,通过训练数据获得种子集群,然后在在线阶段使用种子集群来对推文进行集群测试,定期清理团簇以保持其纯度,从而提高子话题检测的准确性,保证系统的高效性和实用性。

另一些研究者引入概率论和数理统计的方法进行子话题检测,魏明川等[68]提出一种基于吸收马尔可夫链的子话题发现方法,该方法将聚类生成的话题关键词组合生成子话题,用吸收马尔可夫链对子话题进行吸收衍化,重排序生成最终子

话题。实验结果表明,该方法能同时保证生成子话题的重要性和多样性。Khurdiya等[69]使用条件随机场模型从推文中识别、提取和构建围绕大型热点事件的小型子事件结构图。

针对话题检测多停留在二维平面集合操作,而忽略了话题及子话题可能存在的层次关系问题,韩冰等[70]结合生物学知识,引入觅食基础上改进的蚁群算法,通过改进相似度度量方法以及状态转换函数来改进现有的蚁群算法,并利用改进的蚁群聚类算法实现新闻话题的子话题自动划分。

Chen等[71]结合有监督和无监督技术,首先提出了用于子事件检测的无监督深度神经网络,使用一种新的编码器-存储器-解码器框架进行社交媒体子事件检测,该模型以数据驱动的方式学习,通过为每条推文选择最合适的子事件表示来完成子事件检测,从而最大限度提高文本重建概率。

3 子话题检测方式

在社交媒体子话题检测的任务中,本文按照对实时性的要求将其分为离线检测和在线检测两种方式。话题检测最初是应用在离线的静态文本上的,在进行子话题检测时,离线检测是指先将相关数据获取到本地,随后再对其进行处理和检测。随着社交媒体的发展,更多用户希望不仅能检索历史事件,还能实时获取最新的热点事件和焦点话题,同时进一步了解相关话题下的不同子话题,这产生了在线子话题检测的应用需求。离线检测和在线检测两种方式的通用技术包括基于聚类的方法、基于主题模型的方法、基于结构图的方法等,但面向不同场景,二者所采用的技术也有不同,在3.1节和3.2节中对它们分别作详细介绍,其对比如表2所示。

表2 两种子话题检测方式对比

Tab. 2 Comparison of two sub-topic detection methods

检测方式	优点	缺点	常用技术
在线检测	具有及时性、可对数据实时分析	当前时间片信息不全面、检测准确率低下	基于突发状况的方法
离线检测	准确率高、可分析子话题演化过程建立全局认知	无法完成实时检测任务	基于聚类的方法、基于主题模型的方法

3.1 在线子话题检测

社交媒体的信息具有实时性、更迭迅速的特点。在线子话题检测常用的技术主要是基于突发状况的方法,常使用无监督或半监督学习方法检测数据的显著变化。此外,还有一些研究者使用滑动窗口技术对在线数据流处理的方法、基于聚类的方法、基于主题模型的方法、基于结构图的方法等技术来完成在线子话题检测任务。

程葳等[21]针对互联网新闻的特点提出了在线话题检测算法,提出子话题概念,建立具有子话题层和话题层的双层检测结构和基于滑动窗口的跟踪策略,解决信息冗余、议题发散和话题漂移等问题,实验表明该方法的最小错误代价为0.138 8,远低于传统single-pass算法的0.371 9。Saravanou等[72]提出一种对通用文本流以在线方式进行子话题检测和描述的方法DeLi(Detection and deLineation),结合社交网络的结构与内容属性,通过跟踪用户节点和内容节点的连通图检测事件和子事件,并选择最中心的内容节点来表示这些子事件,实验表示该方法在精度(0.15)、召回率(0.49)、F-Score(0.22)和运行时间(139 s)上都表现优越。Tokarchuk等[73]通过实时微博监控框架进行子事件检测:首先,使用自适应微博爬虫爬取数据;然后,采用可以实时完成的流划分方法,通过突发检测算法来分析时间特征;最后,从每个划分的流中提取内容特征并重新组合以提供子事件的最终概括。实验证明该框架能更全面准确地识别子事件,在召回率(44.44%)和精确度(9.57%)上均有良好的表现。

Gonçalves等[74]模拟在线子话题检测的实验,比较*K*-means、非负矩阵分解(Nonnegative Matrix Factorization, NMF)、LDA和动态主题模型(Dynamic Topic Model, DTM)几个算法的性能,把归一化互信息(Normalized Mutual Information, NMI)、调整兰德系数(Adjusted Rand Index, ARI)和归一化折损累积增益(Normalized Discounted Cummulative Gain, NDCG)@10作为评估指标,并得出结论:NMF为最优聚类方法(0.740,0.421,0.576 7);其次是*K*-means算法,其中使用余弦距离的*K*-means方法(0.726,0.387,0.577 1)优于使用JS散度(Jensen-Shannon divergence)的*K*-Means方法(0.736,0.406,0.577 1);由于缺乏文本元素和训练模型的文档,LDA(0.613,0.228,0.442 6)和DTM(0.658,0.267,0.468 8)表现较差。

3.2 离线子话题检测

通常网络社交媒体平台中的话题具有生命周期,且同一话题在热度降低以后隔一段时间可能随其他话题再次被提起,因此话题和子话题可能阶段性地分布在不同时间段的社交媒体数据中,且每次出现都伴随着大量的相关信息。基于这种特性,在线子话题检测往往只能局限地识别出当前的子事件,而不能检测出构成话题的全部子话题内容以及话题演化的过程。因此在没有时效性和及时性的要求时,为提高检测准确率,全面而具体地反映话题及子话题的内容和演化历程,当前大多数的子话题检测均采用离线的处理方式进行。

在离线处理中,数据是已知的,可以统计从最近几个月到一两年甚至更久的已有信息,进行数据的建模和历史信息的统计分析。通过相关的数据元信息(例如地理位置、文本内容、关键词等),可以对话题的局部或全局做出明确的认知。离线子话题检测方式以存储在本地的离线数据为基础,结合机器学习、数据挖掘、文本分析等进行子话题的检测,常用技术有基于聚类的方法、基于主题模型的方法,也可以采用基于分类的方法、基于特征的方法、基于结构图的方法等。

4 结语

通过对以上研究成果的分析,本文得出结论:将多个不同子话题作为同一话题下的纵向挖掘深入拓展,通常能更为有效地描述某一话题的不同侧面,反映同一事件中不同子事件的发酵和演化过程,以便全面掌握全局信息。当前,对于社交媒体平台中子话题检测研究已有了一些初步进展,但还存在以下几个方面的问题。

1)话题检测粒度过于粗糙:对话题检测的研究较多,而针对某一特定话题下的子话题检测研究较少。当前的研究多是把话题当成一个整体,而忽略了内部结构和其联系,欠缺对子话题层次的深入细化。

2)同一话题下子话题之间的相似性:在子话题检测任务中,各个子话题同属于一个主话题,拥有相同的背景,因而具有很强的相似性,当前普通的话题挖掘方法对于具有相同背景的子话题数据集合检测效果不好,检测结果区分度受限。

3)在线检测技术的不足:网络社交媒体的用户规模和信

息量持续增长,相应产生实时变化的数据流和海量的数据集,以在线方式快速准确检测子话题仍值得关注。

4)文本特征选择问题:社交媒体平台中的数据多为短文本,存在特征稀疏性问题,需要充分挖掘短文本中更多特征及重要性关系,提高处理效率和结果的准确率。

5)数据多模态问题:当前子话题检测研究主要围绕文本数据,但网络社交媒体中图像、音频或视频等形式的数据也蕴含大量信息,在子话题检测任务中能发挥重要作用。

6)跨平台检测问题:大部分子话题检测研究都是针对某一平台的单源数据。现阶段网络中各社交平台相互紧密关联(例如可以将知乎中的帖子分享至微博),话题传播途径也全面覆盖多种平台。因此,子话题检测过程中应增加更多的数据来源,以便全面反映网络中的话题及子话题内容。

7)子话题的呈现问题:当前广泛应用的各类方法(例如聚类、主题模型)检测出的话题一般用无序的词语或短语表示,语义理解性较差。所以,提供语义清晰、逻辑通顺的高质量子话题呈现成为备受关注的一个用户需求,可以考虑结合领域词汇集或外部知识库扩充主题词,或抽取相关语句进行描述,抑或是利用可视化技术实现直观呈现。

8)评估指标问题:在话题检测中常用准确率、召回率和F值等作为评估指标,许多学者将其借鉴用于子话题检测任务中,虽能一定程度衡量系统的准确性,但仍有局限。而针对不同的子话题检测方法,相应地也涌现出不同的评估方法和指标,如聚类方法常用NMI、ARI作为评价指标,基于主题模型的方法常用困惑度作为评价指标等,这些指标在一定程度上完成了对不同方法的评估与比较,但是目前还没有一个可以普遍适用的完整而成熟的评估系统。

参考文献 (References)

[1] ALLAN J, CARBONELL J, DODDINGTON G, et al. Topic detection and tracking pilot study final report [EB/OL]. [2019-02-12]. http://nyc. lti. cs. cmu. edu/yiming/Publications/allan-tdt1-final-report. pdf.

[2] NALLAPATI R, FENG A, FU C, et al. Event threading within news topics [C]// Proceedings of the 13th ACM Conference on Information and Knowledge Management. New York: ACM, 2004: 446-453.

[3] 李军,李涓子. 新闻专题内子话题划分[C]// 第四届全国信息检索与内容安全学术会议论文集(上). 北京:中国中文信息学会, 2008: 442-451. (LI J, LI J Z. Subtopic division in news special [C]// Proceedings of the 4th National Conference on Information Retrieval and Information Content security (I). Beijing: Chinese Information Processing Society of China, 2008: 442-451.)

[4] 张晓艳,王挺. 话题发现与追踪技术研究[J]. 计算机科学与探索, 2009, 3(4): 347-357. (ZHANG X Y, WANG T. Research of technologies on topic detection and tracking [J]. Journal of Frontiers of Computer Science and Technology, 2009, 3(4): 347-357.)

[5] 洪宇,张宇,刘挺,等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报, 2007, 21(6): 71-87. (HONG Y, ZHANG Y, LIU T, et al. Topic detection and tracking review [J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87.)

[6] 王卫姣. 话题追踪技术研究综述[J]. 软件导刊, 2013, 12(4): 147-149. (WANG W J. Research status of topic tracking technology [J]. Software Guide, 2013, 12(4): 147-149.)

[7] 孙国梓,黄斯琪,张禹森,等. 基于数据挖掘的微博话题检测方法研究进展[J]. 金陵科技学院学报, 2014(1): 15-20. (SUN G Z, HUANG S Q, ZHANG Y S, et al. Research on Mircoblog's topic detection based on data mining [J]. Journal of Jinling Institute of Technology, 2014(1): 15-20.)

[8] ATEFEH F, KHREICH W. A survey of techniques for event detection in Twitter [J]. Computational Intelligence, 2015, 31(1): 132-164.

[9] 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. (PENG M, GUAN C Y, ZHU J H, et al. A survey on topic detection and tracking in social media text [J]. Journal of Wuhan University (Natural Science Edition), 2016, 62(3): 197-217.)

[10] 中国互联网络信息中心. 第44次中国互联网络发展状况统计报告[EB/OL]. [2019-08-30]. http://www. cnnic. net. cn/hlwfzyj/hlwxzbg/hlwtjbg/201908/P020190830356787490958. pdf. (China Internet Network Information Center. The 44th China statistical report on Internet development [EB/OL]. [2019-08-30]. http://www. cnnic. net. cn/hlwfzyj/hlwxzbg/hlwtjbg/201908/P020190830356787490958. pdf.

[11] FISCUS J G, DODDINGTON G R. Topic detection and tracking evaluation overview [M]// ALLAN J. Topic Detection and Tracking: Event-based Information Organization, INRE 12. Boston: Springer, 2002: 17-31.

[12] 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. (ZHANG Y S, DUAN Y X, HUANG G J, et al. A survey on topic detection and tracking methods in social media [J]. Journal of Chinese Information Processing, 2019, 33(7): 1-10, 30.)

[13] 代翔,黄细凤,唐瑞,等. 基于层次聚类的子话题检测算法[J]. 华南理工大学学报(自然科学版), 2019, 47(8): 84-95. (DAI X, HUANG X F, TANG R, et al. Subtopic detection algorithm based on hierarchical clustering [J]. Journal of South China University of Technology (Natural Science Edition), 2019, 47(8): 84-95.)

[14] NOLASCO D, OLIVEIRA J. Subevents detection through topic modeling in social media posts [J]. Future Generation Computer Systems, 2019, 93: 290-303.

[15] SARAVANOU A, KATAKIS I, VALKANAS G, et al. Detection and delineation of events and sub-events in social networks [C]// Proceedings of the IEEE 34th International Conference on Data Engineering. Piscataway: IEEE, 2018: 1348-1351.

[16] ZHAO S, ZHONG L, WICKRAMASURIYA J, et al. Human as real-time sensors of social and physical events: a case study of Twitter and sports games [EB/OL]. [2019-03-20]. https://arxiv.org/ftp/arxiv/papers/1106/1106.4300.pdf.

[17] WAYNE C L. Multilingual topic detection and tracking: successful research enabled by corpora and evaluation [C]// Proceedings of the 2nd International Conference on Language Resources and Evaluation Conference. Stroudsburg: ACL, 2000: 1487-1493.

[18] 陈儒华. 中文微博子话题构建技术研究与实现[D]. 长沙:国防科技大学, 2013: 3-6. (CHENG R H. Research and implementation on building subtopics for Chinese Microblog [D]. Changsha: National University of Defense Technology, 2013: 3-6.)

[19] JAMES A. Topic Detection and Tracking: Event-Based Information Organization [M]. Norwell: Kluwer Academic Publisher, 2002: 1-16.

[20] 吕楠,罗军勇,刘尧,等. 基于话题三层结构模型的话题演化分析算法[J]. 计算机工程, 2009, 35(23): 71-72, 75. (LYU N, LUO J Y, LIU Y, et al. Topic three layer model based topic evolution analysis algorithm [J]. Computer Engineering, 2009, 35(23): 71-72, 75.)

[21] 程葳,龙志祎. 面向互联网新闻的在线话题检测算法[J]. 计算机工程, 2009, 35(18): 28-30. (CHENG W, LONG Z Y. Online topic detection algorithm for Internet news [J]. Computer Engi-

neering, 2009, 35(18): 28-30.)

[22] 王巍. 基于关键词和时间点的网络话题演化分析[D]. 上海:复旦大学,2009:7-8.(WANG W. Evolution analysis of Internet topics based on keywords and time points [D]. Shanghai: Fudan University, 2009: 7-8.)

[23] SRIJITH P K, HEPPLE M, BONTCHEVA K, et al. Sub-story detection in Twitter with hierarchical Dirichlet processes [J]. Information Processing and Management, 2017, 53(4): 989-1003.

[24] ABHIK D, TOSHNIWAL D. Sub-event detection during natural hazards using features of social media data [C]// Proceedings of the 22nd International Conference on World Wide Web. New York: ACM, 2013: 783-788.

[25] PANEM S, BANSAL R, GUPTA M, et al. Entity tracking in real-time using sub-topic detection on Twitter [C]// Proceedings of the 2014 European Conference on Information Retrieval, LNCS 8416. Cham: Springer, 2014: 528-533.

[26] WU Q, MA S, LIU Y. Sub-event discovery and retrieval during natural hazards on social media data [J]. World Wide Web, 2016, 19(2): 277-297.

[27] EARLE P S, BOWDEN D C, GUY M. Twitter earthquake detection: earthquake monitoring in a social world [J]. Annals of Geophysics, 2011, 54(6): 708-715.

[28] EOM Y H, PULIGA M, SMAILOVIĆ J, et al. Twitter-based analysis of the dynamics of collective attention to political parties [J]. PLoS One, 2015, 10(7): Article No. e0131184.

[29] ZHAO S, ZHONG L, WICKRAMASURIYA J, et al. SportSense: real-time detection of NFL game events from Twitter [EB/OL]. [2019-03-20]. http://arxiv.org/abs/1205.3212.

[30] CHEN C, TEREJANU G. Sub-event detection on Twitter network [C]// Proceedings of the 2018 IFIP International Conference on Artificial Intelligence Applications and Innovations, IFIPAICT 519. Cham: Springer, 2018: 50-60.

[31] ZUBIAGA A, SPINA D, AMIGÓ E, et al. Towards real-time summarization of scheduled events from Twitter streams [J]. Proceedings of the 23rd ACM Conference on Hypertext and Social Media. New York: ACM, 2012: 319-320.

[32] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859.(SU J S, ZHANG B F, XU X. Advances in machine learning based text categorization [J]. Journal of Software, 2006, 17(9): 1848-1859.)

[33] SAKAKI T, OKAZAKI M, MATSUO Y. Earthquake shakes Twitter users: real-time event detection by social sensors [C]// Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010: 851-860.

[34] BADGETT A, HUANG R. Extracting subevents via an effective two-phase approach [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2016: 906-911.

[35] BEKOULIS G, DELEU J, DEMEESTER T, et al. Sub-event detection from Twitter streams as a sequence labeling problem [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2019:745-750.

[36] CHIERICHETTI F, KLEINBERG J, KUMAR R, et al. Event detection via communication pattern analysis [C]// Proceedings of the 8th International Conference on Weblogs and Social Media. Menlo Park: AAAI Press, 2014: 51-60.

[37] ARAKI J, LIU Z, HOVY E, et al. Detecting subevent structure for event coreference resolution [C]// Proceedings of the 9th International Conference on Language Resources and Evaluation. Paris: ELRA, 2014: 4553-4558.

[38] ALDAWSARI M, FINLAYSON M A. Detecting subevents using discourse and narrative features [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 4780-4790.

[39] 张小明,李舟军,巢文涵. 基于增量型聚类的自动话题检测研究[J]. 软件学报,2012,23(6):1578-1587.(ZHANG X M, LI Z J, CHAO W H. Research of automatic topic detection based on incremental clustering [J]. Journal of Software, 2012, 23(6): 1578-1587.)

[40] 张阔,李涓子,吴刚,等. 基于关键词元的话题内事件检测[J]. 计算机研究与发展,2009,46(2):245-252.(ZHANG K, LI J Z, WU G, et al. Term-committee-based event identification within topics [J]. Journal of Computer Research and Development, 2009, 46(2): 245-252.)

[41] 周学广,高飞,孙艳,等. 基于依存连接权VSM的子话题检测与跟踪方法[J]. 通信学报,2013,34(8):1-9.(ZHOU X G, GAO F, SUN Y, et al. Sub-topic detection and tracking based on dependency connection weights for vector space model [J]. Journal on Communications, 2013, 34(8): 1-9.)

[42] 石晶,戴国忠. 基于PLSA模型的文本分割[J]. 计算机研究与发展,2007,44(2):242-248.(SHI J, DAI G Z. Text segmentation based on PLSA model [J]. Journal of Computer Research and Development, 2007, 44(2): 242-248.)

[43] HOFMANN T. Probabilistic latent semantic indexing [C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999:50-57.

[44] HOFMANN T. Probabilistic latent semantic analysis [C]//Proceedings of the15th Annual Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1999: 289-296.

[45] LU Y, MEI Q, ZHAI C. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA [J]. Information Retrieval, 2011, 14(2): 178-203.

[46] 周楠,杜攀,靳小龙,等. 面向舆情事件的子话题标签生成模型ET-TAG[J]. 计算机学报,2018,41(7):1490-1503.(ZHOU N, DU P, JIN X L, et al. ET-TAG: a tag generation model for the sub-topics of public opinion events [J]. Chinese Journal of Computers, 2018, 41(7): 1490-1503.)

[47] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[48] 楚克明,李芳. 基于LDA模型的新闻话题的演化[J]. 计算机应用与软件,2011,28(4):4-7,26.(CHU K M, LI F. LDA model-based news topic evolution [J]. Computer Applications and Software, 2011, 28(4): 4-7, 26.)

[49] HUANG C M, WU C Y. Effects of word assignment in LDA for news topic discovery [C]// Proceedings of the 2015 IEEE International Congress on Big Data. Piscataway: IEEE, 2015: 374-380.

[50] GE B, HU C, HU S, et al. Chinese news hot subtopic discovery and recommendation method based on key phrase and the LDA model [C]// Proceedings of the 2018 International Conference on Electrical, Control, Automation and Robotics. Paris: Atlantis Press,2018: 349-358.

[51] 苏婧琼,刘建霞,谢珺,等. 面向新闻文档的子话题划分方法研究[J]. 小型微型计算机系统,2017,38(8):1850-1855.(SU J Q, LIU J X, XIE J, et al. Research of sub-topic division method in news documents [J]. Journal of Chinese Computer Systems, 2017, 38(8): 1850-1855.)

[52] 李湘东,巴志超,黄莉. 基于LDA模型和HowNet的多粒度子话题划分方法[J]. 计算机应用研究,2015,32(6):1625-1629.(LI

X D, BA Z C, HUANG L. Multi-granularity subtopic division based on LDA and HowNet [J]. Application Research of Computers, 2015, 32(6):1625-1629.)

[53] 胡艳丽,白亮,张维明. 一种话题演化建模与分析方法[J]. 自动化学报,2012,38(10):1690-1697. (HU Y L, BAI L, ZHANG W M. Modeling and analyzing topic evolution [J]. Acta Automatica Sinica, 2012, 38(10): 1690-1697.)

[54] 李静远,丘志杰,刘悦,等. 抑制背景噪声的LDA子话题挖掘算法[J]. 华南理工大学学报(自然科学版),2017,45(3):54-60. (LI J Y, QIU Z J, LIU Y, et al. LDA subtopic detection algorithm with background noise restraint [J]. Journal of South China University of Technology (Natural Science Edition), 2017, 45(3): 54-60.)

[55] BANU S H, CHITRAKALA S. Trending topic analysis using novel sub topic detection model [C]// Proceedings of the 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics. Piscataway: IEEE, 2016: 157-161.

[56] LIU W, WANG D, XU W, et al. A sub-topic partition method based on event network [C]// Proceedings of the 7th International Conference on Internet and Web Applications and Services. Washington, DC: IEEE Computer Society, 2012: 194-199.

[57] KATRAGADDA S, BENTON R, RAGHAVAN V. Sub-event detection from tweets [C]// Proceedings of the 2017 International Joint Conference on Neural Networks. Piscataway: IEEE, 2017: 2128-2135.

[58] MELADIANOS P, NIKOLENTZOS G, ROUSSEAU F, et al. Degeneracy-based real-time sub-event detection in Twitter stream [C]// Proceedings of the 8th International Conference on Web and Social Media. Menlo Park: AAAI Press, 2015: 248-257.

[59] MELADIANOS P, XYPOLOPOULOS C, NIKOLENTZOS G, et al. An optimization approach for sub-event detection and summarization in Twitter [C]// Proceedings of the 2018 European Conference on Information Retrieval, LNCS 10772. Cham: Springer, 2018: 481-493.

[60] 仲兆满,李存华,戴红伟,等. 融合内容与时间特征的中文新闻子话题聚类[J]. 计算机科学与探索,2013,7(4):368-376. (ZHONG Z M, LI C H, DAI H W, et al. Clustering Chinese news subtopic integrating content and time features [J]. Journal of Frontiers of Computer Science and Technology, 2013, 7(4): 368-376.)

[61] 张瑞琦. 基于关键特征聚类的Top *N*热点话题检测方法研究[D]. 北京:北京理工大学,2015:4-5. (ZHANG R Q. Research on Top *N* hot topics detection method based on key features clustering[D]. Beijing: Beijing Institute of Technology, 2015: 4-5.)

[62] POHL D, BOUCHACHIA A, HELLWAGNER H. Automatic sub-event detection in emergency management using social media [C]// Proceedings of the 21st International Conference on World Wide Web. New York: ACM, 2012: 683-686.

[63] POHL D, BOUCHACHIA A, HELLWAGNER H. Supporting crisis management via detection of sub-events in social networks [J]. International Journal of Information Systems for Crisis Response and Management, 2013, 5(3): 20-36.

[64] POHL D, BOUCHACHIA A, HELLWAGNER H. Online indexing and clustering of social media data for emergency management [J]. Neurocomputing, 2016, 172: 168-179.

[65] ZAHARIEVA M, RIEGLER M. Media synchronization and sub-event detection in multi-user image collections [C]// Proceedings of the 2nd ACM International Workshop on Human-centered Event Understanding from Multimedia. New York: ACM, 2015: 13-18.

[66] QIAN X, LI M, REN Y, et al. Social media based event summarization by user-text-image co-clustering [J]. Knowledge-Based Systems, 2019, 164: 107-121.

[67] PANEM S, BANSAL R, GUPTA M, et al. Entity tracking in real-time using sub-topic detection on Twitter [C]// Proceedings of the 2014 European Conference on Information Retrieval, LNCS 8416. Cham: Springer, 2014: 528-533.

[68] 魏明川,朱俊杰,张瑾,等. 基于吸收马尔可夫链的子话题发现方法[J]. 中文信息学报,2014,28(1):41-46,55. (WEI M C, ZHU J J, ZHANG J, et al. An algorithm for subtopic detecting based on absorbing Markov chain [J]. Journal of Chinese Information Processing, 2014, 28(1): 41-46, 55.)

[69] KHURDIYA A, DEY L, MAHAJAN D, et al. Extraction and compilation of events and sub-events from Twitter [C]// Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Piscataway: IEEE, 2012:504-508.

[70] 韩冰,汪波. 一种基于改进蚁群算法的子话题划分方法[J]. 济南大学学报(自然科学版),2016,30(6):473-478. (HAN B, WANG B. A subtopic partition based on improved ant colony algorithm [J]. Journal of University of Jinan (Science and Technology), 2016, 30(6): 473-478.)

[71] CHEN G, XU N, MAO W, et al. An encoder-memory-decoder framework for sub-event detection in social media [C]// Proceedings of the 27th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2018: 1575-1578.

[72] SARAVANOU A, KATAKIS I, VALKANAS G, et al. Detection and delineation of events and sub-events in social networks [C]// Proceedings of the IEEE 34th International Conference on Data Engineering. Piscataway: IEEE, 2018: 1348-1351.

[73] TOKARCHUK L, WANG X, POSLAD S. Piecing together the puzzle: Improving event content coverage for real-time sub-event detection using adaptive microblog crawling [J]. PloS One, 2017, 12(11): Article No. e0187401.

[74] GONÇALVES G, MARTINS F, MAGALHÃES J. Analysis of subtopic discovery algorithms for real-time information summarization [C]// Proceedings of the Web Conference 2018. New York: ACM, 2018: 1855-1856.

This work is partially supported by the National Key Research and Development Program of China (2017YFC0820702-3), the National Natural Science Foundation of China (U1603115, U1435215), the Laboratory Director Foundation of National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data.

LI Shanshan, born in 1996, M. S. candidate. Her research interests include natural language processing, text data mining, information security.

YANG Wenzhong, born in 1971, Ph. D., associate professor. His research interests include Internet public opinion, intelligence analysis, information security, wireless sensor network.

WANG Ting, born in 1996, M. S. candidate. Her research interests include natural language processing, text emotional analysis, information security.

WANG Lihua, born in 1995, M. S. candidate. Her research interests include natural language processing, text intention detection.