社交媒体中的谣言识别研究综述

刘雅辉[1),2),3),4)] 靳小龙[1),2),3)] 沈华伟[1),2),3)] 鲍 鹏[5)] 程学旗[1),2),3)]

[1)](中国科学院网络数据科学与技术重点实验室 北京 100190)

[2)](中国科学院计算技术研究所 北京 100190)

[3)](中国科学院大学计算机与控制学院 北京 100049)

[4)](石河子大学 新疆 石河子 832003)

[5)](北京交通大学软件学院 北京 100044)

摘 要 近年来,社交媒体蓬勃发展,Facebook、Twitter、新浪微博、微信等逐渐成为人们获取或分享信息的重要渠道.这种人人可以参与信息发布和传播的方式在给人们的信息共享提供极大便利的同时,也带来一些突出的问题,特别是网络谣言的不断滋生和快速传播,给社交媒体的有效利用和科学管理提出了严峻挑战.因此,如何快速准确地识别谣言是个重要的研究问题,也是抑制谣言传播、降低谣言危害的前提.该文对社交媒体上的谣言识别工作进行综述,首先介绍了谣言研究的发展历程以及分类;然后,介绍了影响谣言识别的关键要素、当前谣言识别的主要方法;最后,对社交媒体中谣言识别存在的问题与发展趋势进行了总结和展望.

关键词 谣言;社交媒体;谣言识别;谣言传播

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2018.01536

A Survey on Rumor Identification over Social Media

LIU Ya-Hui[1),2),3),4)] JIN Xiao-Long[1),2),3)] SHEN Hua-Wei[1),2),3)] BAO Peng[5)] CHENG Xue-Qi[1),2),3)]

[1)](*CAS Key Laboratory of Network Data Science and Technology*, *Beijing* 100190)

[2)](*Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)

[3)](*School of Computer and Control Engineering*, *University of Chinese Academy of Sciences*, *Beijing* 100049)

[4)](*Shihezi University*, *Shihezi*, *Xinjiang* 832003)

[5)](*School of Software Engineering*, *Beijing Jiaotong University*, *Beijing* 100044)

Abstract In recent years, the popularization of Web 2.0 technology spurs the flourish of social media. There emerged a few globally influential social media platforms, such as Facebook, Twitter, Sina Weibo, and WeChat, which have gradually become a very important channel for people to obtain or share information. On social media, everyone can participate in publishing and disseminating information. According to the 39th Statistical Report on Internet Development in China released by China Internet Network Information Center (CNNIC), as of December 2016, China's Internet users reached 731 million, and the Internet penetration rate was 53.2%. The social media platform makes people more closely connected, and has broken the original six degrees separation, now less than five degrees. The way of information dissemination has undergone enormous changes. Especially, the microblogging system has a strong network effect with the optimization of information flow products, the outbreak of video and the steady progress of the vertical. Active users continue to maintain high-speed growth. By the end of 2016, the monthly

收稿日期:2016-08-23;在线出版日期:2017-05-26.本课题得到国家重点研究发展计划(2016YFB1000902)、国家重点基础研究发展计划(2013CB329602)、国家自然科学基金(61572473,61472400,61232010)、中国科学院青年创新促进会优秀会员项目和 CCF 腾讯犀牛鸟基金项目(20160107)、中央高校基本科研专项资金(2015rc031)资助.**刘雅辉**,女,1979 年生,博士研究生,讲师,中国计算机学会(CCF)会员,主要研究领域为社会计算、数据挖掘.E-mail:lyh@shzu.edu.cn.**靳小龙**,男,1976 年生,博士,副研究员,博士生导师,主要研究领域为社会计算、知识图谱、人工智能.**沈华伟**,男,1982 年生,博士,副研究员,博士生导师,主要研究领域为社会网络分析和社交媒体计算.**鲍 鹏**,男,1987 年生,博士,讲师,主要研究领域为社交媒体计算.**程学旗**,男,1971 年生,博士,研究员,博士生导师,主要研究领域为网络科学与社会计算、互联网搜索与挖掘、网络信息安全、分布式系统与大型仿真平台.

active users of Sina Weibo increased to 313 million. Similarly, Twitter also had 319 million daily active users. CNNIC data show that Sina Weibo in the general social applications ranked second, accounting for 43.5%. 73.9% of users focus on news or hot topics via Sina Weibo, which means that Sina Weibo has become a public platform, and the main channel for people to understand the current hot information. However, as a side-effect of social media, rumors, commonly defined as unconfirmed and uncertain information posted intentionally or unintentionally by some users, are also widely propagated over social media, which may lead to significant negative effects and even severe social problems, e.g., social panic and even chaos. For example, a hacker released a fake tweet about explosions in the White House through the official Twitter account of the Associated Press (AP) on April 23, 2013. This rumor was widely reposted and caused an immediate social panic: S&P 500 Index instantaneously dropped 14 points, wiping out 136.5 billion in a matter of seconds, and the Dow Jones Industrial Average also dropped about 145 points within three minutes. Therefore, how to control the spread of rumors in social media has become one of the major concerns of social media platforms and academia. However, the premise of controlling rumors is to identify rumors first. Therefore, the research of rumor identification has a wide range of practical significance and value for purifying network space, improving the effective use of social media, suppressing rumor spreading and reducing the harm of rumor. This paper presents a survey on the identification of rumors in social media. It first introduces the definition and classification of rumors as well as its research history. It then reviews the three key factors affecting rumor identification, and the current methods for rumor identification based on classification and model. Finally, the paper summarizes existing problems of rumor identification over social media and prospects development trends.

Keywords　rumor; social media; rumor identification; rumor spreading

1　引　言

随着计算机和互联网技术的不断发展,人类社会进入了信息互联和人的互联高度融合的时代.尤其是 Web2.0 技术出现之后,社交网站(如 Facebook、人人网)、博客(如 tumblr、新浪博客、网易博客)、微博(如 Twitter、新浪微博)等社交媒体蓬勃发展[1].社交媒体具有信息传播网络化、信息内容碎片化、线上线下交融等特点,人们可以在网络上自由地发布、传播和获取信息,信息的传播方式发生了前所未有的变化.

社交媒体为谣言传播提供了新的媒介.社交媒体中,谣言具有传播速度快、影响范围广、监测难度大、危害程度深等特点.谣言的产生和传播不仅妨碍了人们对社交媒体的有效利用,而且可能造成民众的误解或引发民众的负面情绪,甚至提高网络犯罪活动发生的可能性[2],影响社会稳定.例如,"2012 年 12 月 21 日是世界末日"的传言引起了很多人的担忧,微博与论坛上出现了大量相关讨论.2013 年 4 月 23 日黑客入侵了美联社的 Twitter 账号并发布"白宫发生两起爆炸,贝拉克·奥巴马受伤"的假消息,立即引起美国资本市场的震动,标准普尔 500 种股票指数 5s 内下跌 14 点,市值蒸发 1365 亿美元①.2015 年中国科协列出的十大"科学流言"②以及由北京市网信办、北京科学协会联合部分网站列出的 10 大生活谣言③也都曾给人们造成过不同程度的负面影响.

鉴于谣言的影响和危害,近几年,社交媒体中的谣言研究引发了广泛关注.工业界针对网络谣言建立了谣言查询网站、辟谣网站以及谣言监测系统等.譬如,在国外,美国有 snopes.com 和 urbanlegends.about.com 等谣言查询网站、威尔斯利学院社会信息学实验室开发的探究谣言源和传播特性的 Twitter-Trails④ 工具[3,4]、哥伦比亚大学的数字新闻研究中

① http://news.hexun.com/2013-04-25/153561212.html
② http://www.ithome.com/html/discovery/199159.htm
③ http://www.bj.xinhuanet.com/2016-01/14/c_1117776818.htm
④ http://twittertrails.com/

心推出的实时谣言跟踪网站 Emergent①、法国的反谣言网站 hoaxbuster 和谣言搜索网站 hoaxkiller.fr②、英国谢菲尔德大学研发的 Twitter 测谎仪 Pheme③[5]等.在国内,有新浪微博采用众包技术建立的微博社区管理中心④、微博辟谣账户,果壳网成立的谣言粉碎机小组以及建在新浪微博上的谣言粉碎机账户,辟谣联盟与四月网合作的辟谣百科以及一些地区建立的辟谣平台.

工业界的这些努力对控制谣言的传播、降低谣言的危害起到了一定作用.然而,这些网站或系统在识别谣言时大部分使用了大众举报和人工验证的方式,不但需要耗费大量的人力和财力,而且在谣言识别上存在较大的时间滞后.谣言发布后,如果没有得到及时控制而被广泛传播,便可能对人们或社会造成不良影响.因此,亟需研制出谣言自动识别系统,使能够在谣言散布后尽可能短的时间内识别出谣言,并采取有效措施最小化谣言产生的影响或危害.为此,学术界针对社交媒体中的谣言展开了广泛的研究.黄少宽等人[6]统计了从 2000 年 1 月 1 日到 2014 年 9 月 30 日"中国期刊全文数据库"(CNKI)中所有以"网络谣言"为主题词的研究文献,共 2922 篇.他们发现,从 2008 年起,"网络谣言"研究文献的增长速度开始加快,2013 年发表的论文数量呈现爆炸式的增长,数量高达 1113 篇.这一统计表明,谣言研究正在引发学术界的研究热潮.这些研究囊括了谣言的各个方面,如谣言溯源、谣言传播过程建模、谣言传播的心理因素分析、谣言识别以及谣言控制等,其中谣言识别在这些研究中占据着重要地位.社交媒体中的谣言识别旨在探索谣言自身及其传播过程中具有的本质特征和存活规律,并利用发现的特征和规律结合有效的模型或方法及时准确地识别出谣言.

社交媒体中,谣言识别研究具有重要的意义和价值.第一,谣言识别不仅为限制谣言传播提供了前提,也有助于发现现实生活中人们所关心的问题、不易发现的社会现象或社会问题.第二,谣言识别的结果对新闻记者、金融市场、紧急事务处理以及社交媒体上的信息质量等都有积极作用.第三,谣言的生命在于传播,谣言传播规律的研究是谣言识别研究的重要组成部分,谣言识别研究有助于促进信息传播研究的发展.

本文第 2 节介绍谣言研究的发展历程和谣言的分类;第 3 节从谣言内容、谣言用户以及谣言传播三个方面介绍影响谣言识别的特征与规律;第 4 节从基于分类和基于模型两个方面介绍谣言的识别方法;第 5 节探讨谣言识别存在的问题并对将来的谣言识别研究进行总结和展望;第 6 节对文章进行总结.

2 谣言及其研究历程

如何定义谣言一直倍受各界关注[7-10].随着信息传播媒介的演变,谣言的内涵也在不断地发展.因此,谣言至今仍然没有一个公认的定义.本文所说的谣言是指在社交媒体中传播的与事实不符或捏造的信息,如错误信息、虚假信息[11]、假故事、恶作剧⑤[12]、阴谋论[13]等.本节首先介绍谣言研究的发展历程,然后介绍谣言在社交媒体中的分类.

2.1 谣言研究的发展历程

谣言古已有之.既有"大楚兴,陈胜王"这种谣传王朝更替的"大谣",也有"曾参杀人"、"三人成虎"等成语故事背后那种市井小民恶作剧式的"小谣".不过,对谣言的学术研究起步很晚,起始于二战时期.

二战期间,大量谣言的爆发不仅影响了民众的斗志和信心,在某些情况下也成为敌人宣传的有力武器,引起了社会学家与心理学家研究谣言的兴趣.早期的研究试图从社会学和心理学的角度解释谣言产生的机制以及谣言传播背后隐藏的各种因素.1944 年 Knapp[14]发表的"谣言心理学"一文,开启了谣言研究的大门.谣言研究的开创性工作是 Allport 和 Postman[15]在 1947 年出版的《谣言心理学》一书,该书提出了谣言的两个重要因素:重要性 I(Importance)和模糊性 A(Ambiguity).

随着二战的结束,20 世纪 50 年代对谣言的研究出现了短暂的间歇期.60 年代中后期谣言研究又迎来了新的热潮,基本上包含了谣言研究的各个方面.Daley 和 Kendall[16]提出谣言传播的数学模型;Buckner[17]提出了谣言的传播理论;社会学家 Shibutani[18]出版了《即兴新闻:谣言的社会学研究》一书,指出谣言是在一群人议论过程中产生的即兴新闻,明确了谣言是一种集体行为;Kapferer[19]在其《谣言:世界最古老的传媒》一书中探讨了人们信谣传谣背后所隐藏的文化和社会背景.

20 世纪 90 年代万维网出现以后,网络数据明确记录了谣言传播的轨迹,为谣言研究提供了前所

① http://www.emergent.info/
② http://www.hoaxkiller.fr
③ https://www.pheme.eu/
④ http://service account.weibo.com/
⑤ http://cogprints.org/7336/

未有的宝贵机遇.因此,研究者们开始研究网络上的谣言,试图了解谣言在网络上是如何产生和传播的.譬如,Bordia 和 Rosnow[20]认为网络谣言是人们面临持续的焦虑与不确定性时为了解决问题而进行的互动;Zanette[21]研究了谣言在小世界网络中的动态传播;Buchegger 等人[22]研究了谣言在移动对等网络中的传播;Moreno 等人[23]研究了谣言在复杂网络中的动态传播.

Web2.0 的出现加速了社交媒体的流行,社交媒体的开放性、高交互性以及传播迅速等特性在给人们的信息交流带来方便的同时也为谣言传播插上了翅膀,使得谣言可以在短时间内广泛地传播开来.为此,研究人员的注意力也开始转移到新的社交媒体上,尤其是影响较大、传播更快的微博系统(如 Twitter,新浪微博等).此时的研究主要探索谣言在社交媒体中的传播机制、识别谣言的方法以及控制谣言的策略等.表 1 简要总结了谣言研究的发展历程.

表 1　谣言研究的发展历程

阶段	描述
第一阶段 20 世纪 40 年代	二战引发了社会心理学家对谣言以及谣言控制的研究
第二阶段 20 世纪 60 年代 ～80 年代	这一时期,学术界特别是在社会学和心理学等领域又兴起了研究谣言的热潮,发表了大量文章和书籍
第三阶段 20 世纪 90 年代 ～21 世纪初	万维网的出现使得谣言能够在网络上传播,而且万维网能够保留传播的文本.因此,学术界开始研究网络中谣言的传播
第四阶段 21 世纪 00 年代 中期至今	Web2.0 技术的出现,产生了社交网络、微博等社交媒体.社交媒体自身的特征使其成为谣言滋生的温床,为谣言的传播插上了翅膀.如何识别谣言并控制社交媒体中的谣言传播成为各界研究的热点

2.2　谣言的分类

谣言的产生动机、类型及其生命周期等各方面都存在着差异,这导致研究者对谣言的划分角度各不相同,分类方法也各有所取.本节依据谣言与事实的相关性将谣言分为两大类:源于虚构的谣言和源于事实的谣言.

源于虚构的谣言是造谣者无中生有或有意杜撰的以欺骗他人为目的的虚假信息.造谣者为了提高谣言的影响力和可信度常常假借权威人士或权威媒体的名义发布谣言信息.这类谣言中有些是针对他人或企业实施的攻击或报复,有些是希望引起人们对某件事情的关注,有些则纯属娱乐或恶搞.如“美国航天局报告说未来 10 个月地球将升温 4°”、“安徽牌照专修楼房漏水的面包车在各地偷孩子”及“某明星死亡”等谣言均属此类.

源于事实的谣言是基于某一事实而散布的错误信息.这些错误信息是由于人们对事实的了解或理解中混杂了想象、评价或主观臆断使原本真实的信息发生了变异[24]而产生的.这类谣言的可信度较高,具有很强的迷惑性,很难辨别,甚至有可能通过大众传媒(如电视、广播、报纸等)广泛传播.这类谣言主要包括丢失细节偏离事实、添油加醋捏造细节、蓄意放大夸大事实、过度解读扩展事实、掌握不全曲解事实(如曲解图片、新闻、节目等)、移花接木篡改事实等.它们有些是为了表达个人对事实的一种期望或焦虑,有些是对官方或媒体缺乏事实信息更新的一种补充.如“发现失踪的 MH370”、人们对海关总署修订进境物品税表中关于“进境 200 元奶粉须交 10%的税”的误读等谣言成为网民关注的焦点.

尽管两类谣言来源不同,但无论哪种谣言,只要广泛传播都可能给个人、公众、企业甚至社会秩序造成不同程度的影响.因此,准确而及时地识别谣言对于减少谣言产生的影响具有重要作用,而要识别谣言首先在于研究影响谣言识别的关键要素.

3　谣言识别的要素

在社交媒体上,“关注”功能使人们之间形成了在线关系网络:被关注者成为关注者的朋友(Followees),而关注者是被关注者的粉丝(Followers).被关注者发布或转发消息后,他/她的所有粉丝都可以看到.在 Twitter 或新浪微博上,用户可以通过各种基于 Web 应用的客户端或移动客户端以 140 字以内的文字或结合多媒体(包括图片、视频和音频)的形式发布消息,并实现即时分享.譬如,用户 A 发布了一条消息 m_i,他的粉丝 B 看到他发布的消息后进行直接转发或添加对消息的相关评论内容 c_k 并转发,转发后的消息记为 r_j.微博(Twitter)上直接转发的形式为 B://@A:m_i(B: RT @A:m_i),添加评论并转发的形式为 B:c_k//@A:m_i(B:c_k RT @A:m_i).从上述过程可以得出,消息传播过程经历消息的发布和转发两个阶段.在消息的发布阶段,用户 A 被称为发布用户或原发用户,消息 m_i 被称为原消息;在消息的转发阶段,用户 B 被称为转发用户或者响应用户,消息 r_j 被称为转发消息,c_k 被称为转发评论.消息的这种转发模式决定了消息发布阶段的发布用户和原消息内容以及消息转发阶段的转发用户和转发评论内容会影响消息的传播.

为方便起见,我们把与谣言相对的真实消息称为非谣言.谣言的内容、发布用户以及它的传播有别于非谣言,从而构成识别谣言和非谣言的关键要素.本节主要从这三个方面介绍影响谣言识别的关键要素:首先是谣言原消息内容即谣言内容,其包含的哪些元素能够体现出与非谣言内容的差异.其次是谣言发布者即谣言用户,其具有的哪些特征使他们更容易发布谣言.最后是谣言传播,在谣言传播过程中哪些用户扩大了谣言的传播;人们在转发谣言时发表了怎样的评论,这些对谣言传播产生了什么影响.这些问题的研究和探索可以为谣言识别提供重要的识别依据.此外,在社交媒体上,人们往往转发他们认为可信的消息内容或者转发可信的用户发布或转发的消息[25].因此,可信度(Credibility)[26]研究中挖掘的内容以及用户特征也可以用来辅助识别谣言.

3.1 谣言内容

在社交媒体上,谣言内容涉及到人们生活的各个方面以及当前社会的热点话题或事件.它包括两种形式:第一种是谣言原发者即造谣者发布的消息内容;第二种是一些用户通过直接复制造谣者发布的消息内容而发布的消息,即原消息的副本.这两种消息在内容上没有差异,因此,研究者们通常把它们都看作是谣言的原消息内容,主要研究它们所包含的词语、符号以及所表达的不确定性等特征.本节主要从这三个方面综述谣言和非谣言在内容特征上存在的差异.

3.1.1 内容中包含的词语

研究者们在比较微博中的谣言和非谣言的消息内容时发现,谣言和非谣言的内容在用词上存在着明显差异,如情感词、动词、代词等.

谣言内容中经常包含人们的主观情感,这种情感主要通过情感词来体现.情感词是情感倾向性判断的重要依据,包括积极情感词和消极情感词.研究者们根据情感词典(如中文有知网①、大连理工大学信息检索研究室的情感词汇本体库②等)对谣言和非谣言中的情感词进行提取,通过分析发现谣言内容中往往包含更多的消极情感词,而非谣言内容中则包含更多的积极情感词[27].此外,谣言和非谣言相比会使用更多的动词、代词(如第一人称代词)、脏话[28]以及非标准语法[29]等.

3.1.2 内容中包含的符号

用户经常在消息内容中加入各种符号辅助或加强语义的表达,如摹因(Memes)[30]、情感符号、问号、感叹号、非标准标点符号等.

摹因[30]主要包括 hashtag、URL 以及用户提及(使用@符号:@用户名).Hashtag 指主题标签,受到很多社交媒体的青睐,包括新浪微博、Twitter、Google+、Pinterest、Instagram 等,如在新浪微博(Twitter)中表示为"#话题#"("#话题").Hashtag 方便用户对某一话题进行集中讨论,起到了聚合和归类信息以及用户通过搜索快速聚焦到话题的作用.URL 的作用是对外部消息源的引用或者克服微博的字数限制,缺点是用户不能直接获得 URL 所指向的页面内容.用户提及所起的作用是提醒被提及的用户注意查看消息.其中,URL 和 hashtag 中包含了更丰富的延伸信息,可以增加消息的可信度,经常被用在非谣言的消息中;用户提及经常被造谣者用来吸引被提及用户的注意力,从而诱导被提及的用户转发谣言消息[28].

情感符号类似于情感词,也分为积极的情感符号和消极的情感符号,如在新浪微博上,😄符号可以转换成文字形式"[哈哈]".我们统计了所有新浪微博上的情感符号,分别提取了积极和消极情感符号如表 2 所示,其中带括号的情感符号表示精确匹配,没有带括号的可以进行模糊匹配.模糊匹配可以匹配出同类的情感符号,如"哈哈"可以匹配[哈哈]、[笑哈哈]、[cc 哈哈]、[哈哈哈哈]等积极的情感符号.研究者们[27]发现谣言的消息内容中会包含更多的"笑脸"符号.此外,谣言的消息内容中还经常使用问号、感叹号以及非标准标点等符号来表达惊讶等情感,以达到提醒或吸引人们注意的目的.

3.1.3 内容表达的不确定性

在某些事件中,当一些用户不能从正式渠道(如官方或新闻媒体)获取相关消息时,就会通过制造或传播内容具有较大不确定性[31]的谣言来表达对当前环境中某些话题的担忧.其他用户发现这样的谣言时会根据自己的经验或求助于其它消息源对谣言提供新的证据、表达意见或做出解释,如增加消息的表述、观点的表述、情感的表述以及意会的表述.这些表述减少了谣言内容的不确定性,最终达到集体的一致共识[32-34],但是,用户在寻求消除谣言不确定性的同时也传播了谣言.因此,人们对相关话题内容的不确定性表达可以作为谣言识别的一种途径[35].然而,不确定性是谣言本身的主要特性之一,

① http://www.keenage.com/html/e_index.html
② http://ir.dlut.edu.cn/EmotionOntologyDownload

在实际中捕获这种不确定性面临很大的挑战.这方面的工作还需要研究者们进一步探索,如通过自然语言处理和深度神经网络学习谣言内容的不确定性表达,然后再结合机器学习等方法实现谣言的自动识别.

3.2　谣言用户

用户是社会网络中的主体,他们之间的联系和互动形成了用户的关系网络以及行为等特征,结合用户的自身特征,本节主要从三个方面介绍谣言用户与非谣言用户的差异.

3.2.1　用户的基本特征

研究者们发现可以通过用户资料中使用的信息判断用户的可信度,如用户选择的头像、使用的用户名、性别等.用户的可信度越高发布谣言的可能性越小,反之可能性越大.下面主要介绍谣言与非谣言用户资料中的五点差异.

(1)造谣者为了隐藏自己的真实身份一般不会选用自己的真实照片作为头像,而通常选择系统默认图像以及卡通图片或头像;(2)用户使用的用户名大概有三种方式:话题型的用户名、传统风格的用户名以及网络用户名.这三种方式相比,话题型的用户名更能增加用户的可信度,它代表用户可能是某个领域的专业人士或者知名人士,如"LR 机器学习"、"计算机视觉"、"全球健身中心"、"微博搞笑排行榜"等.然而,也有一些特定意图的用户使用话题型的用户名发布谣言[36]来增加谣言的可信度.使用传统风格或网络风格作为用户名的用户有相对较低的可信度,他们发布谣言消息的可能性较大[29].在社交媒体上,用户使用的头像和用户名是人们对他与他所发布消息内容是否可信的最直观的判断;(3)识别谣言的能力也与性别有关①,女性相对于男性发布的消息更可能成为谣言,尤其是政治类的消息;(4)用户所在地的差异以及受文化背景影响的差异也决定了他们识别谣言的能力,如来自自由地区的人比来自保守地区的人更可能发布谣言[36];(5)造谣者为了逃避相关责任,很少人使用认证账户发布谣言消息;而许多正常用户(如名人、媒体等)却使用认证账号来提高他们的名誉.

3.2.2　用户的网络特征

用户所在的关系网络可以作为用户是否可能发表谣言的判断依据.研究者们发现,朋友数大于粉丝数的用户更可能发布谣言消息[29].当用户的朋友中有多人发布或转发了谣言消息时,他发布或转发谣言消息的概率也会增大.一些有蓄谋的造谣者为了更快更广泛地传播谣言,会通过关注功能吸引所关注用户的注意或寻找一些时机来增加粉丝,直到粉丝数达到一定值时才发布谣言[37].

3.2.3　用户的行为特征

用户在社交媒体上发表过的消息代表了他们的活跃程度和历史.用户越活跃、注册时间越长,他们辨识谣言的能力越强.因此,已经发表过很多消息的用户发布谣言的可能性越小[27].除考虑用户已经发布的消息数量还需考虑这些消息的质量,如用户 u_i 转发了用户 u_j 的消息 m,表示为 u_i: //@$u_j m$,m 有两种可能成为谣言:第一种,u_j 在过去有发布和转发谣言的历史;第二种,u_i 已经发布或转发过谣言[11].因此,如果用户有发布或转发谣言的历史,那么他有可能再次发布或转发谣言.

3.3　谣言传播

消息的转发过程,即消息的传播过程.用户读到不同的消息时,反应也会有很大差异,往往通过发表自己的观点或意见来评论消息.随之也带动关注他们的大量粉丝进行转发.Mendoza 等人[38]已经发现在 Twitter 上谣言和新闻的传播存在着很大的不同.本节主要从转发的评论内容以及转发用户两方面介绍相关研究工作.

3.3.1　转发的评论内容

评论内容包括人们对谣言表达的观点或意见等信息,如支持、质疑、反驳以及中立等.针对转发的评论内容,研究者们重点研究人们是如何参与谣言的讨论过程以及他们对谣言传播产生怎样的影响[39].造谣者在编造谣言时会尽量使谣言看起来更真实更可信,然而,即使一些可信度比较高、形象丰满的谣言,也会留下很多令人质疑的隐患,如图片模糊、链接打不开以及广告太假等[40].因此,很多人在转发谣言时会根据他们识别的一些线索或具备的知识对谣言提出质疑或反驳[38].图 1 标注了转发者对"国宝大熊猫被大规模捕杀"这条谣言提出质疑和反驳的转发评论.研究者们通常把对谣言反驳的转发消息称为更正(Corrections)[41],也称为批判[42]或辟谣[43].为了便于对转发中出现的质疑和更正词的理解,我们分析并标注了新浪微博中 3286 条谣言的转发,抽取了出现频率较高的质疑和更正词,如表 2 所示.

(1)更正对谣言传播的影响

研究者们发现更正对谣言识别或抑制谣言传播起着重要作用[44,45].Starbird等人[41]通过分析谣言

① http://www.pishu.cn/zxzx/xwdt/377463.shtml

图 1 转发评论中的质疑和反驳

表 2 新浪微博上的情感符号以及评论观点词或短语

	类型	主要表达词
情感符号极性	积极情感符号	哈哈、呵呵、嘻嘻、偷乐、笑、开心、鼓掌、赞、ok、耶、给力、给劲、拍手、V5、爱、亲亲、害羞、鬼脸、兴奋、顶、膜拜、得意、抱抱、钱、蜡烛、话筒、熊猫、兔子、奥特曼、礼物、v5、江南 style、[haha]、[挤眼]、[阳光]、[太阳]、[good]、[威武]、[酷]、[心]、[握手]、[多云转晴]、[hold 住]
	消极情感符号	泪、抓狂、惊、害怕、鄙视、悲伤、生病、感冒、委屈、伤心、失望、衰、无语、崩溃、阴险、恶心、怒、吐、悲催、凌乱、晕、黑线、囧、汗、鼻、疑问、思考、哼、最差、可怜、最差、猪头、围观、神马、浮云、困、睡觉、嘘、不要、拳头、[懒得理你]、[猥琐]、[小人得志]、[闭嘴]、[草泥马]、[弱]
评论的观点	质疑	真的吗、真滴吗、真得吗、真事吗、真实吗、是真的、真实的吗、真的假的、真假的、求证、求真相、真吗、属实吗、可能吗、可信吗、能信吗、信吗、大家信吗、有此事吗、有这事吗、有依据吗、有吗、靠谱吗、科学吗、假的吗、假吗、假新闻吗、辟谣了吗、辟谣的吗、辟谣吗、谣言吗、猜忌、疑团、迟疑不决、疑忌、猜疑、顿生疑窦、疑问、靠不住、疑虑重重、疑心、生疑、打问号、迟疑、狐疑、半信半疑、犯嘀咕、可疑、吃不准、质疑、疑惑、怀疑、置疑、蹊跷、岂非、问号、起疑、不见得
	更正	谣言、造谣、谣传、辟谣、假冒、诽谤、伪造、假的、闹剧、误读、误解、尚未、消息不实、不实信息、不属实、不真实、内容失实、信息失实、无此报道、网传、传闻、传言、流言、假消息、假信息、假新闻、假图、作出澄清、经查、未收到、未接到、未发布、未发现、未发生、未经证实、无此事、暂停发布、吸费电话、理解有误、说法有误、虚构

传播过程中人们对谣言的支持与更正消息随时间变化的情况发现，有的更正对谣言传播完全没有影响，有的有些影响，有的甚至会抑制谣言的传播[46]. 更正产生这样的效果可能与人们辨别谣言的难易程度有关. 对于易于辨别的谣言，更正能阻止谣言的传播，如谣言中提到的人或组织发现有消息对他们造谣时，他们会立刻发布更正来抑制谣言继续传播[47]. 然而，对于比较难辨别的谣言，更正并不能起到抑制谣言传播的作用[48]. 当然，更正的作用效果还受其它一些因素的影响，如信息传播所依托的信息共享平台对消息的显示方式[42]. 如果人们在转发谣言前先看到其它转发中的更正信息或者包含的更正链接（如链接到 snopes. com）[43]，他们很可能放弃对谣言的转发. 然而，有些谣言尽管包含了上百条的更正消息，仍然能继续传播，其主要原因是谣言与其更正不能同时显示，一些人在转发谣言时并没有看到相关的更正消息.

（2）更正的传播速度

Zeng 等人[49]通过对谣言原消息到所有支持谣言的转发消息所获得的时间间隔与原消息到所有更正所获得的转发时间间隔的对比发现，支持谣言的转发消息具有更长的时间间隔. 这意味着更正有更快的传播速度，大概比谣言的传播速度快 2 倍[47]. 导致这个现象的原因是，与支持某个观点相比，人们对挑战该观点更感兴趣. 由此可见，如果有计划的安排一些传播用户（如意见领袖[50]）发布更正消息可能会有效抑制谣言的传播.

从上述研究可以看出，更正是识别谣言或抑制谣言传播的显著特征. 然而，从更正数量与支持数量随时间的变化发现，更正大多出现在谣言传播的中后期[13,41]. 虽然更正对谣言能够产生一定的抑制效果，但是更正出现时谣言已经产生了较大影响. 此外，并不是所有的谣言消息都会有用户更正谣言，如比较难验证的科技谣言. 因此，谣言识别工作中不宜把更正作为重点建模目标，而应该把它作为有效的辅助特征并结合其它特征或方法来识别谣言，同时也会在谣言识别上赢得更多的时间.

3.3.2 传播用户

传播用户的内在因素、特定的目的以及其权威性、影响力与行为特性等决定了谣言传播的广度、深度以及谣言的爆发点. 研究者们在探索和研究参与谣言和非谣言传播的用户差异时发现，这种差异能够为谣言识别提供重要依据.

（1）用户的传播行为

人们为什么发布或传播谣言?研究发现,人们传播谣言是由诸多因素驱动的,包括传播用户的内在因素、特定目的、社会因素以及环境因素等.

传播用户的内在因素(如性格特征)以及他们的特定目的影响了他们在社交媒体上传播消息的行为. Chen[51]研究发现,传播用户性格中的神经性(Neuroticism)和开放性(Openness)对谣言消息的传播具有显著的影响:性格具有神经质的人往往不喜欢随意传递或转发无关的消息,因此他们很少传播谣言;性格具有开放性的人可能传播更多的谣言消息来探究他们的非常规想法.此外,传播用户的娱乐(Entertainment)、社交(Socializing)与寻求地位(Status seeking)的动机对谣言消息的传播也有重要影响,其中娱乐目的对谣言消息的传播影响最大,其次是寻求地位.

谣言传播除了与传播用户的内在因素有关,也与传播用户的社会因素和环境因素有关.如在某些事件中,用户在物理位置或情感上与事件的接近度对他们是否发布或转发消息有很大影响[52].当人们有物理位置接近于事件发生地或情感上与他们有关联的人或地(如朋友、家人或以前居住地等)可能受到事件的影响时,他们会传播更多的消息,有时也想通过传播消息帮助他人.然而,这些消息中可能包含大量的谣言,他们的转发扩大了谣言的传播范围.

(2)起主导作用的传播用户

①有些谣言会被大量转发,给人们的生活和社会秩序带来很大的影响.是否所有的转发用户都起到了主导作用呢?

新浪微博包含的用户认证类型有普通用户、名人、政府、企业、媒体、校园、网站、应用、团体(机构)、达人等.不同认证类型的用户有不同的用户关注度和影响力.信誉好或有影响力的人或媒体参与到谣言的传播过程中会使谣言发生质的变化,即成为准新闻,促使潜在的信谣者大幅增加[19]. Liao 和 Shi[33]研究名人、认证用户、大众媒体、组织、网站、达人以及普通用户在谣言传播过程中的贡献和影响力时发现,在传播过程中起主要作用的用户是达人、名人与大众媒体.这三类用户在谣言传播过程中形成了三阶段模式的主导作用:达人首先发布谣言,使得谣言在一定范围内扩散,然后名人参与传播导致更大范围的转发,最后大众媒体发挥他们的优势提供大量报道或调查引爆谣言的传播.

虽然媒体扩大了谣言的影响,扮演了"推手"的作用,但是当他们知道错误地报道了谣言时,也会转而宣传辟谣消息来阻止谣言的传播.因此,媒体在谣言传播中扮演了推手和辟谣的双重角色[53].此外,在危机事件中,官方账号,如新闻媒体、紧急事件的响应者与组织机构,也扮演着形成谣言的讨论以及发布更正减缓或停止谣言传播的双重角色. Andrews 等人[54]发现在危机事件中谣言传播中的大部分转发是由少量主流媒体和新闻账号所引发,但随着谣言的传播,一些账号会发布更正来揭露谣言,尤其是官方账号的更正会激励一些参与谣言传播的账号去更正自己.也有研究者发现大概有 14.7%的人更正他们已经转发的谣言消息[47].即便有些传谣的用户没有进行更正,也会减少其继续传播谣言的可能性,从而起到对谣言的抑制作用.

由此可见,只有少数起主导作用的用户引发了谣言的大量转发,如 Gupta 等人[55]研究转发高峰期前后两个小时中参与转发的用户发现,0.3%的用户导致了 90%的转发.从传播用户主导的先后顺序看,谣言这种由小众到大众的传播模式和正常的新闻消息有明显差异,因此,这种传播模式上的差异可以作为谣言识别研究的一个很好的切入点.

②大部分在传播谣言过程中起主导作用的用户具有一定的权威性和影响力,最直接的衡量是他们拥有大量的粉丝.用户的粉丝越多,他发布或转发的消息就会被越多人看到.因此,大部分消息产生的直接转发主要来自用户的粉丝.那么,是否拥有大量粉丝的用户在发布或转发消息后都会引起大量的转发,反之不会呢?

事实上,粉丝数与转发数之间的相关性不强[48],不是粉丝数越多转发数就越多.然而,用户的粉丝数越多,越可能获得大量转发. Arif 等人[56]从转发量、曝光量(Exposure,通常用粉丝数来衡量)以及内容生成(Content production)三个互补视角,通过定性、定量以及可视化的方法研究了曝光量对谣言传播的影响.他们发现谣言衍生内容量的大小(直接或非常接近于原内容的拷贝)和曝光量的大小两两组合可以产生四种效应:巨效应、雪球效应、闪灭效应以及轻效应.名人、大众媒体、官方媒体以及紧急事件响应者等可能触发巨效应;如果一个用户只有少量粉丝,要想引起大量转发,就需要通过设计内容来造成雪球效应.闪灭效应和轻效应都引发很少的转发量,前者一般发生在谣言生命周期快要结束时或人们支持和反驳谣言的交互期间,许多人看到了谣言但是并不进行转发;后者,消息源有很少的粉丝限制了消息被其他人看到且消息的内容也不足引

起大量转发.由此可见,在考虑一条谣言消息的影响范围时,应该综合考虑消息的内容与曝光量,但是拥有大量粉丝为消息的广泛传播提供了重要前提.

3.4 小 结

研究者们探索谣言传播的关键要素是为了发现谣言的显著特征或内在规律,从而为谣言的自动识别或控制提供依据和线索.前述特征的总结如图 2 所示.其中,已经有部分特征应用到了谣言识别中,如词汇特征、情感特征、更正特征等.已有研究还存在以下不足:

(1)从谣言内容和转发评论方面,大多数研究把对内容包含的词(如正/负情感词数、更正数等)、符号(如问号数、提及数等)等的统计数据作为特征,并没有挖掘内容中包含的语义等深层特征.从谣言用户和传播用户方面,只考虑了用户的基本特征(如性别、用户名等)、网络特征(如朋友数、粉丝数)、历史行为特征(如已发布的消息数)和传播过程中的转发行为特征(如转发数、评论数等),而没有考虑用户的影响力以及行为等的深层特征.

(2)从特征的应用方面,一些被发现的显著特征或规律还需要将来在谣言识别中进一步证实,或探究它们如何更好地应用到谣言识别研究中,如传播用户的特征、谣言的可信度特征等.

(3)从研究方式方面,有些研究只考虑谣言是如何传播,这种方式由于没有对比性而面临着很大挑战,因此在将来的研究中,研究者们可以考虑把同种类型的谣言传播和非谣言传播进行对比研究,会更容易发现谣言传播的独特之处,同时也符合社交媒体所具有的从大量信息中识别出谣言的场景.

(4)在分析方法方面,主要采用社会学理论、定量分析、定性分析以及可视化[57]等方法进行的实证研究,实验的数据量较小,得出的结论还需在更大的数据集上进一步验证.

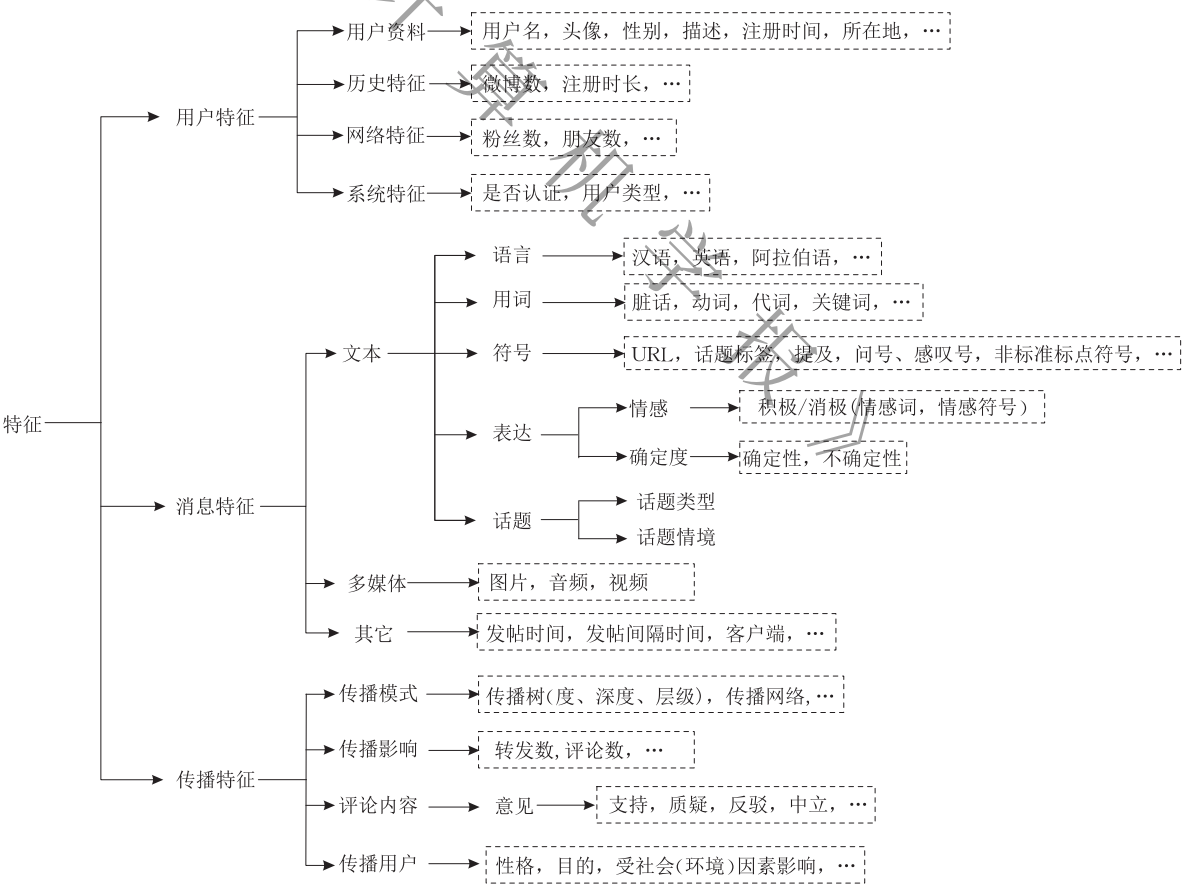

图 2 谣言识别特征总结

4 谣言识别方法

在社交媒体中,谣言识别是谣言跟踪以及采取措施控制谣言传播、降低谣言影响的关键前提.因此,谣言研究具有重要的研究意义和应用价值.当前,大部分社交媒体平台主要采用用户举报和人工验证的方法识别谣言,时间上存在很大的延迟,识别

效率不高.因此,亟需自动或半自动的辅助方法来进行谣言识别,以便尽早发现谣言,也提高谣言识别效率.现有的自动谣言识别主要采用分类的方法,也有少量基于模型的识别方法.本节对上述两类方法进行详细介绍.

4.1　基于分类的谣言识别方法

基于分类的谣言识别借助自然语言处理、社会网络分析、数据挖掘以及机器学习等技术和方法,从谣言的内容、用户以及传播中提取或挖掘一些显著的特征,并将它们应用到谣言识别中.研究者们通常把谣言识别看成二分类问题,使用不同的分类算法,如支持向量机(SVM)、决策树、随机森林以及朴素贝叶斯等对谣言和非谣言进行分类.由此可见,基于分类方法识别谣言的关键是挖掘出谣言和非谣言具有显著差异的特征集.通常,数据中每个消息的特征集表示成特征向量的形式

$$x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})$$

x_i表示第 i 条消息,$x_i^{(j)}$ 表示 x_i 的第 j 个特征,可能是消息的内容、用户或者传播特征,n 表示消息的特征个数.分类的类别标签 y_i表示为

$$y_i = \begin{cases} 1, & \text{消息 } x_i \text{ 为谣言} \\ -1, & \text{消息 } x_i \text{ 为非谣言} \end{cases}$$

谣言分类问题包括学习和分类两个过程.在学习过程中,根据已知的训练数据集利用有效的学习方法学习一个分类模型;在分类过程中,利用学习得到的分类模型对给定的消息 x_k预测其是否为谣言.分类算法可以选择已有的机器学习库或工具包,如 SVM 可使用台湾大学林智仁教授开发的 LibSVM①,朴素贝叶斯、决策树、随机森林可使用 Weka②、R③、scikit-learn④ 或 Matlab 等软件的工具包.

表 3　混淆矩阵

		预测类别	
		1	−1
实际类别	1	*TP*	*FN*
	−1	*FP*	*TN*

分类结果使用的评价指标是根据表 3 的混淆矩阵计算的准确率(Accuracy)=$(TP+TN)/(TP+FP+TN+FN)$、精确度(Precision)=$TP/(TP+FP)$、召回率(Recall)=$TP/(TP+FN)$以及 $F1=2\times Precision\times Recall/(Precision+Recall)$.精确度和召回率在某些情况下是矛盾的,即一个值很高的时候,另一个值会很低.在谣言识别中,谣言往往给人们或社会带来一些消极影响,因此,希望分类器尽量识别出所有的谣言,即召回率要高.当然,也不应该把很多非谣言识别为谣言而增大谣言的验证和排查工作量,即谣言识别的精确度也不能太低.为此,谣言识别通常采用精确度与召回率的调和平均数 $F1$ 来反映两个值的均衡情况.$F1$ 越高说明分类器识别谣言的总体效果越好.

近几年,社交媒体(如 Twitter 和新浪微博)蓬勃发展,它具有的传播特性使得社交媒体上的谣言识别研究引发广泛关注,产生了一系列谣言识别工作.这些工作通常要比较所提方法的谣言识别效果,因此,我们首先介绍两个经常被其它工作作为 baseline 进行比较的研究工作.这两个工作分别使用 Twitter 和新浪微博数据集提取谣言识别的基本特征,并使用分类器进行谣言识别.两个工作实施谣言识别的具体方法如下.

第一个工作来自 Castillo 等人[27].该工作主要评估 Twitter 上的新闻话题的可信度,提取的 68 个特征分为 4 类:基于消息的特征、基于用户的特征、基于话题的特征(每个话题包含多条相关的消息)以及基于传播的特征,并使用最优优先选择方法选择最好的 15 个特征用于分类.他们使用基于 J48 决策树的分类器,在预测新闻话题是否可信上获得了 86%的准确率.最好的 15 个特征中基于话题的特征包括消息的平均情感分数、有积极情感分数的消息比例、有消极情感分数的消息比例、有 URL 的消息比例、不同的短 URL 数、最频繁发帖作者发布的消息比例、包含用户提及的消息比例、包含问号的消息比例、包含笑脸符号的消息比例以及包含第一人称的消息比例,这 10 个话题特征中有一半的特征是获取用户所表达的情感;用户特征包括用户的平均注册年龄、平均粉丝数、平均朋友数以及平均发布的消息数,这 4 个用户特征反应了所有发布某一话题的用户的名誉,它们的值越小,发布谣言的可能性越大;传播特征只包含传播树中除根节点的孩子,层包含的最多节点数,这个特征反应除根节点外是否还有其它的节点引发更多的转发.

第二个工作来自 Yang 等人[58].他们使用新浪微博上的谣言数据,从消息中抽取了两个新特征:用户发布消息使用的客户端和消息内容中提及事件的

① http://www.csie.ntu.edu.tw/~cjlin/libsvm/
② http://www.cs.waikato.ac.nz/ml/weka/
③ https://www.r-project.org/
④ http://scikit-learn.org/stable/

发生位置.他们发现用户发布消息时使用非移动客户端程序和移动客户端程序两种形式,大概 71.8% 的谣言使用非移动客户端发帖;谣言消息内容中所提及的事件位置大概有 56.1% 发生在国外.为了验证这两个特征识别谣言的有效性,他们结合已有研究提出的内容、用户或传播特征[11,27,29],使用 SVM 分类器自动识别谣言和非谣言.内容特征包括是否包含多媒体、积极与消极情感符号数、是否包含 URL 以及发帖时间与用户注册时间之间的时间间隔,结合两个新特征获得了 78% 的谣言识别准确率;用户特征包括用户的注册时间、粉丝数、朋友数、消息数、是否认证、有无描述、性别、用户头像类型、用户名类型以及注册地点,这些用户特征在一定程度上反应了用户的活跃程度、影响力以及权威性,结合两个新特征获得了 77.4% 的谣言识别准确率;传播特征只考虑消息是否是转发消息、评论数以及转发数,结合两个新特征获得了 78.7% 的谣言识别准确率.

Castillo 等人把同一主题的一组消息看作一个话题,把这组消息各项特征的均值、比例或最值作为该话题的特征值.如果某个话题是谣言,我们把这种谣言称为基于话题的谣言.而 Yang 等人把每一条消息当作一个谣言,特征取值只针对一条消息,我们把这种谣言称为基于消息的谣言.从前面的介绍我们可以看到,Castillo 等人和 Yang 等人所提的特征在谣言的识别效果上还有待提高.因此,后续研究工作在他们所提特征的基础上从内容、用户或传播的角度进一步探索新的特征或方法来提高谣言识别效果.尽管这些工作同时考虑了内容、用户与传播等不同角度的特征,但在这三个角度的特征上各有侧重.因此,本节分为三小节分别介绍以内容、用户与传播为主要特征的研究工作,最后分析并总结各项工作的分类效果.

4.1.1 以内容特征为主的分类方法

用户发布的消息的内容主要包含两类:文本和多媒体.这两类内容只要其中之一被编造或篡改都会使消息成为谣言.下面介绍以这两类内容为主要特征的谣言识别方法.

(1)基于文本内容特征的方法

贺刚等人[59]认为 Yang 等人抽取的浅层文本特征难以有效地区分谣言和非谣言,需要挖掘文本内容的深层特征以获得更好的谣言识别效果.因此,他们提取了文本的符号、链接、关键词分布以及时间差四类新特征.其中,符号特征包括 hashtag、@以及标点符号(如!、?、...、…等),并以它们在消息中出现的频次作为特征值;链接特征衡量消息内容与 URL 链接所指向的网页内容的相似度,其值使用 Jaccard 系数计算获得;关键词分布特征用来区分谣言和非谣言消息中包含的关键词在频次分布上的差异,特征值取消息中包含的所有关键词在谣言与非谣言中频次差的总和再取均值;时间差特征是消息发布时间与谣言原消息发布的时间差,时间差越大,消息是谣言的可能性越小.这些特征结合 Yang 等人[58]提出的文本特征,使用 SVM 分类器获得了 81.2% 的准确率.实验表明,在这四类特征中,关键词分布特征对分类的结果影响最大,而链接特征并没有提升分类的效果,是因为消息中存在的 URL 链接有一部分是位置链接,这些链接对分析消息与外部资源的相似度所起的作用不大.

Zhang 等人[60]也认为浅层特征在很多情况下不能很好地区分谣言与非谣言.因此,他们抽取了 4 个基于内容的隐式特征:流行度取向、内外一致性、情感极性、评论观点.流行度取向特征可以获得消息内容和当前社会的热点话题或事件的相关性,使用 Jaccard 系数计算获得特征值;内外一致性衡量消息内容与相关的外部网页内容的关联性,关联性越少,消息成为谣言的可能性越大;情感极性包括积极、消极以及中立三种类型,通过捕获消息中使用的情感词或情感符号,然后基于不同类型的情感词典计算词的权重;评论观点捕获转发者对消息内容的接受程度,使用支持的评论数与不支持的评论数的比值作为特征值.他们使用 SVM 分类器获得了 72.4% 的精确度和 58.6% 的召回率.为了验证所提出的隐式内容特征的有效性,他们和 Yang 等人提取的内容特征所获得的分类效果做了对比,发现分别在精确度和召回率上提升了 10.5% 和 4.7%.

还有一些研究根据相关的语料库对于消息中使用的短语进行提取,并用于识别谣言.Qazvinian 等人[11]试图进一步区分某种类型反复出现的短语之间的差异,如对谣言的认可、反驳、质疑以及简单谈论等类型的短语,并使用计算的方法来识别谣言.他们提取的内容特征包括基于单个词/两个词的词汇模式、基于单个词/两个词的词性模式以及摹因(包括 hashtag 和 URL).每个特征对应一个贝叶斯分类器,以每个消息中该特征的对数似然比作为特征值.URL 的特征值通过计算消息中 URL 所指向的内容中单个词/两个词的对数似然比来获得.通过检测特征的分类效果发现内容特征中的语言特征(前

4 个特征)获得了很高的精确度,大概 95%左右.然而,因为很多谣言中都没有 hashtag 和 URL,所以摹因特征虽然获得了高的精确度,但召回率低.

其它特征,如文本内容的话题类型对谣言的识别结果也有重要影响.因为谣言往往集中在少数大家比较关心的敏感话题上,如食品安全、儿童救助或丢失等话题.

(2)兼顾文本内容与多媒体特征的方法

有些谣言只包含了文本内容,也有一些谣言除了文本内容还包含了图片,即用户在发布或转发消息时加入与原消息内容不符或伪造的图片使其成为谣言.这种图文并茂的谣言比只使用文本形式的谣言往往具有更高的可信度.

① 图文不符是指文本表达的含义和图片不符.图片往往是真实的,已经在网络上被发布过,即为过时图片,而文本内容针对过时图片却表达得张冠李戴.Sun 等人[61]发现 80%的事件谣言带有图片,而且它们中大多数是图文不符.因此,他们抽取了 4 个新的文本特征以及 1 个多媒体特征来识别这类谣言.其中,文本特征包括描述事件的动词数、包含事件动词的消息比例、是否包含强消极词以及包含强消极词的消息比例.大多数事件是消极的社会事件,在事件期间产生的谣言通常包含更多的事件动词和强消极词,并且组合过时图片来增加谣言的可信度.作者还计算了包含事件动词的消息比例和包含强消极词的消息比例特征来衡量用户在事件谣言中使用的动词和消极情感词与以往发布消息的差异,使用包含事件动词或强消极词的消息数与用户发布过的所有消息数的比例作为特征值.这两个值越高,消息成为谣言的可能性越大.多媒体特征主要获取消息发布时间与图片最初发布时间之间的跨度.在已有研究所提特征基础上加入这 5 个新特征,使用决策树获得了 75%的精确度和 66.3%的召回率,比不使用新特征分别提升了 19.7%和 46.1%,说明了新提出的特征对基于图文不符的事件谣言识别的有效性.

② 伪造图片是指谣言发布者通过对图片的处理而生成的虚假图片.它扭曲了原图片的含义,并以两种形式出现在消息中,一种是直接显示图片,另一种是显示图片的 URL.Gupta 等人[55]分析 2012 年飓风桑迪期间 Twitter 上传播的虚假图片的特征发现,包含虚假图片 URL 的消息中有 86%是转发消息;发布或转发虚假图片的用户,他们的粉丝网络和转发网络中只有 11%的重叠率,说明只有少量转发源自于用户的粉丝,其它大部分转发是转发用户通过 Twitter 搜索或从热点话题中发现而进行的转发,这部分转发用户并不关心他们是否关注了对方.这种行为说明,用户转发陌生人的消息比转发他们所关注用户的消息更有可能是谣言[47].Gupta 等人为了识别包含虚假图片 URL 的谣言,提取了消息的文本特征,包含符号特征(如?、!、愉快/悲伤的情感符号、hashtag、URL、@)、词汇特征(如词的个数、第一/二/三人称代词、积极/消极情感词数、大写字符数)、消息的长度以及转发数.他们使用决策树分类器获得了 97.7%的准确率,比利用用户特征提升了 44.4%,说明内容特征在预测包含假图片 URL 的消息时具有非常好的效果.

从图片的角度识别谣言的相关研究较少,一方面是由于数据的限制,使用 API 只能得到文本内容,如果要获得对应的图片还需单独抓取;另一方面,图文不符的谣言相对容易识别,而包含伪造图片的谣言识别具有一定的挑战,它需要具备图片识别的相关知识.然而,从上述研究中的统计数字可以看出,谣言中带有图片的比例较大,而且通过图片辅助的方式识别谣言的准确率较高.因此,研究者们可以把消息中的图片作为将来谣言识别的一个重要考虑因素.

综上所述,基于内容特征为主的谣言识别研究主要从消息的文本内容、多媒体、消息与外部内容的关联等方面进行特征提取(如表4),然后利用这些

表 4　内容特征

特征类	特征	说明	效果
词的特征	关键词	谣言和非谣言消息中关键词在频次分布上的差异	Y
	词	单个词、两个词,词的个数	Y
	词性		Y
	动词	动词个数或消息的动词占所有已发布消息动词的比例	Y
	代词	个数或比例	y
	情感词	包括积极、消极和中立.情感分数或情感词个数	Y
符号特征	摹因	包括 hashtag、URL、@符号.是否包含或个数或比例	y
	?	是否包含或包含的个数	y
	!	是否包含或包含的个数	y
	情感符号	包括积极、消极.情感分数或情感符号个数	Y
多媒体特征	图片	是否包含图片	y
	时间间隔	消息发布时间与图片的时间间隔	Y
与外部消息的关联	链接特征	消息内容与 URL 链接所指向的网页内容的相似度情况	n
	流行度取向	消息内容和当前社会的热点话题或事件的相关性	Y
	内外一致性	消息内容与相关的外部网页内容的关联性	Y
消息的整体特征	话题类型		Y
	消息的长度		y

注:Y 表示对识别结果有重要影响,y 表示对识别结果有影响,n 表示识别结果不好.

特征进行分类.从分类的结果可以看出,基于文本内容特征的方法获得了很高准确率,但是对谣言预测的精确度和召回率却不高,因此,内容特征很好地预测了非谣言消息.其中,消息的话题、情感、流行度取向以及内外一致性等特征获得了显著的识别效果.

4.1.2 以用户特征为主的分类方法

用户的历史行为特征犹如用户的一面镜子,可以反射出用户的活跃程度和行为动机.用户的影响力特征在一定程度上反映他们发布消息后能激起的反响力.本节主要从用户的行为特征和影响力特征介绍谣言识别的相关工作.

(1)基于用户行为特征的方法

Liang 等人[37]认为用户的行为特征是谣言识别的隐藏线索.他们使用微博数据,抽取了 5 个新的发帖用户和传播用户的行为特征,包括平均每天关注的朋友数、平均每天的发帖数、可能的消息源数、质疑评论比率和更正数,并结合其它的行为特征(如用户是否认证、粉丝数、转发数、评论数)使用决策树分类器获得了 86.5%的精确度和 85.4%的召回率,与已有研究提出的最好的 11 个用户行为特征相比分别提升了 19.8%和 25.4%.这说明被新提出的用户行为特征比已有研究所提的用户行为特征更能有效地识别出谣言.他们分析用户的行为特征发现,通常用户关注的人越多,他获得的粉丝数越多.为了快速地吸引更多的粉丝,造谣者在非常短的时间内关注很多人.因此,造谣者平均每天关注的朋友数比正常用户要高.造谣者在发布谣言后,为了逃避可能的责任很少或从来不使用发布谣言的账号登录.因此,造谣者平均每天的发帖数远小于正常用户.谣言消息通常起源于一个人或一小部分人,而非谣言消息可能被很多无关的个人目击或发布.因此,谣言的消息源数比非谣言的消息源数要小.从上述对用户的行为特征分析可以看出,谣言发布者的行为不同于正常用户的行为,用户的行为隐藏了谁可能是造谣者或者什么样的消息可能是谣言的信息.

Qazvinian 等人[11]构建了两个用户模型来捕获用户发帖或转发的历史行为(见 3.2.3 小节解释的消息 m 成为谣言的两种可能):第一个是积极的用户模型,即用户没有发布过谣言或者没有被发布或转发过谣言的用户转发过;第二个是消极的用户模型,即用户发布过谣言或者被发布或转发过谣言的用户转发过.他们提出了发布用户和转发用户两个特征,并利用两个用户模型对给定的消息计算对数似然比作为两个特征的特征值.这里对消息的发布用户与转发用户做了区分,这种区分是非常重要的,因为转发用户可能在转发时改变被转发消息的意思.他们利用这两个特征来识别谣言获得了大概 90%的精确度和 80%的召回率.这说明用户的行为是一个很好的谣言指示器.

(2)基于用户影响力特征的方法

Zhang 等人[60]抽取了基于用户的 3 个隐式特征来识别谣言,获得了 65.7%的精确度和 57.2%的召回率,比 Yang 等人的分别提升 3.4%和 0.5%.用户的 3 个隐式特征包括社会影响力、观点的转发影响力以及消息的匹配度.其中,社会影响力通过用户的粉丝数与朋友数的比例来反映用户在社交媒体中的交际能力;观点转发的影响力捕获用户的观点被其他用户的接受程度,用户被他人接受的程度越高,发布谣言的可能性越小,以用户所有消息的转发数总和与消息数量的比例作为特征值;消息的匹配度是指用户的职业和消息内容的匹配程度,谣言的主题通常与用户历来发布消息的内容有很大的差异,因此使用消息内容的话题分布与用户历史内容的话题分布的余弦相似性来表达消息的匹配度.

对上述研究中挖掘的用户特征总结如表 5 所示.从分类的结果可以看出,基于用户行为特征的方法获得了很高的精确度和召回率,说明用户的行为特征对谣言有很好的预测力.基于用户影响力特征的方法虽然也能提升谣言的识别效果,但是获得了较低的召回率.因此,谣言用户的行为特征有别于非谣言用户的行为特征,可以作为谣言识别的重要依据.

表 5 用户特征

特征类	特征	说明	效果
行为特征	平均每天关注的朋友数		Y
	平均每天的发帖数		Y
	可能的消息来源数	发布的原消息数	Y
	是否发布或转发过谣言	如果发布或转发谣言,再次发布或转发的可能性增大	Y
	消息的匹配度	用户的职业和消息内容的匹配程度,用消息的话题分布与用户已发布消息的话题分布的余弦相似性来衡量	y
影响力特征	社会影响力	通过用户的粉丝数与朋友数的比例来反应	y
	观点的转发影响力	用户的观点被其他用户接受的程度,用已发布的所有消息的转发数与消息数的比例来衡量	y

注:Y 表示对识别结果有重要影响,y 表示对识别结果有影响.

4.1.3 以传播特征为主的分类方法

研究者们使用最频繁的传播特征是转发数和评论数,它们可以被看作是用户对消息的一种响应行为,同时也是消息流行度的一种反映.除此之外,还有转发的评论内容特征、时序特征以及结构特征.

(1)基于转发评论特征的方法

在社交媒体上,用户基于不同的需求转发消息,其中包括带有质疑地去转发有争议的消息.这种转发为识别该消息为谣言提供了重要线索.Zhao 等人[13]依据这样的线索,首先识别带有质疑和更正的信号消息,然后对相似的消息进行聚类,并抽取每一类的描述,再根据描述收集其它相关的消息,组成谣言候选集,最后基于统计特征按每类消息成为谣言的可能性大小进行排序.其中,统计特征是对每类中带有信号的消息与全部消息分别进行统计而获得的特征量,具体包括信号消息的百分比、词频分布的熵的比率、消息的长度、转发消息的百分比以及摹因特征(包括 hashtag、URL、用户提及)等.他们发现,在波士顿事件中,只使用更正信号在候选集的 Top 10 中获得谣言的精确度为 46.6%,这说明在社会网络中并不是每个基于话题的谣言都可以被大众揭露;而更正和质疑同时使用在候选集的 Top 10 中获得了 52.1%的精确度.实验结果说明单独使用更正信号发现谣言并不能获得很好的谣言识别效果.相对于更正,质疑信号会更多地出现在谣言中,能够提升谣言的识别效果.然而,并不是所有消息中包含质疑或更正的都是谣言.因此,需要结合更有效的方法提高谣言识别的精度.文章中也使用质疑和更正信号验证了谣言的早期发现,与只使用更正信号相比早了 4.3 h.

此外,转发评论中包含的平均情感/情感符号分数、平均质疑分数以及平均感叹分数等特征[62]也能明显提升谣言的识别结果.

(2)基于转发时序特征的方法

谣言传播过程中主要经历发布初期、爆发期以及消退期.在这个动态的传播过程中,谣言的很多特征随时间的变化不同于非谣言.

①转发数量随时间的变化.消息在发布后,随着时间的推移,转发量会大幅增加,当达到某个时间点时会发生爆发现象,在图示上表现为峰的形式.消息在生命周期中可能没有爆发的峰、可能有一个或多个爆发的峰.其中,多个爆发的峰受到很多因素的影响.例如,受昼夜交替或某些转发用户的影响,谣言可能在生命周期内出现多峰现象;此外,谣言即使在某个时间被官方辟谣,受一些因素的影响可能在某个时间出现"死灰复燃"的复发现象,如"小龙虾是一种处理过尸体的虫子,外国人从不吃",这也可以导致谣言在多个时段上出现峰的现象.Kwon 等人[63]观察到谣言的多峰现象有别于非谣言通常有一个显著峰的现象,并且在信息传播模型 SpikeM[64]的基础上构建了周期性的外部震动模型来捕获谣言的这种周期性爆发的特性,然后结合结构和语言特征使用随机森林分类器获得了 93.5%的精确度和 89.2%的召回率.然而,Takahashi 等人[47]发现谣言在传播过程中爆发可能是不同类型的转发消息所引发,如支持谣言的消息或更正消息.如果区分这两种爆发,并不是所有支持谣言的转发都会出现爆发的多峰现象.这种情况下使用爆发这个特征并不能很好地区分谣言与非谣言.

②部分用户、内容以及传播特征也随时间发生变化.一些特征量,如转发数、更正数,在谣言传播过程中是动态更新的,当谣言传播一定时间后它们能提供明显的信号来区分谣言与非谣言.

Ma 等人[65]认为一些研究者在谣言识别上只考虑了用户特征、内容特征以及传播特征的统计量,忽略了它们随时间的动态变化,如第一人称代词会频繁出现在谣言传播的早期、问号的数量会更多的出现在谣言传播的晚期等.虽然 Kwon 等人[63]试图捕获转发量随时间的变化,但是在他们的模型中只使用了 3 个参数来刻画这种变化,这可能导致重要信息的丢失.因此,Ma 等人建立了动态时间序列结构模型(Dynamic Series-Time Structure,DSTS)来捕获在文献[27,58,62]中所提的部分用户、内容与传播特征在消息生命周期中的动态变化.基于 DSTS 的 SVM 分类器(SVM_{all}^{DSTS}),在 Twitter 数据集上获得了 88%的精确度和 90.9%的召回率;在新浪微博数据集上获得了 86.1%的精确度和 85.4%的召回率.同时,他们也验证了 SVM_{all}^{DSTS} 早期谣言识别的效果.在初始阶段,SVM_{all}^{DSTS} 和其它分类算法(如决策树、随机森林、SVM)获得了相似的准确率,但是随着时间的推移,在 Twitter 数据集上 SVM_{all}^{DSTS} 大概用 15 h 达到了随机森林 25 h 达到的准确率 86%左右,在新浪微博数据集上 SVM_{all}^{DSTS} 大概用 20 h 到达了随机森林 70 h 达到的准确率 82%左右.

Liu 等人[66]利用人们对基于话题的谣言表达的支持、反驳或质疑等信念特征作为线索,结合传统新闻调查的语言、用户、传播以及元特征实时地揭露谣言.与实时谣言跟踪网站(如 snopes.com,emergent.com)揭露谣言的时间相比,75%的谣言揭露时间比它们早,准确率在 85%以上;与新闻媒体揭露的时间相比,仍然有 75%的谣言揭露时间比它们早,准确率在 80%以上.在识别谣言过程中,用户特征始终起到重要的作用;随着时间的推移人们逐渐

认识到谣言的真实性,信念特征会起到越来越重要的作用.

(3)基于转发结构特征的方法

在 Facebook、Twitter 以及微博上,一条消息的所有转发路径形成了树状结构.这种结构通常被称为传播树[27,38],树的根节点表示原发用户,其它节点表示转发用户,节点间的连接表明用户间的转发关系.研究者们对谣言与非谣言在传播树结构上存在的差异做了相关的探索.

①传播树中节点的用户类型.不同类型的用户拥有的粉丝数可能差距几个量级,如普通用户的粉丝数从几十、几百到几千不等,而名人或媒体的粉丝数达到几十万甚至上千万.这些用户在发布或者转发消息时,由于粉丝数的差异,引起的转发量也存在显著差异,尤其是他们发布或转发人们所关心或担忧的消息内容.Wu 等人[62]发现"谣言通常被普通用户发布,然后被一些意见领袖(粉丝数与朋友数的比例达到某一阈值)转发,最后被大量普通用户转发,而非谣言则是被意见领袖发布,直接被大量普通用户转发"的传播模式.因此,他们把每个消息的传播过程构建成传播树,对每个节点标记相应的用户类型并结合用户对原消息的意见和情感信息,使用随机游走图核计算传播树之间的相似性来区分谣言与非谣言传播模式的差异.同时,他们结合基于内容、用户和传播特征的径向基核函数构成混合 SVM 分类器预测给定的消息是否为谣言,混合 SVM 分类器获得了 90.5%的精确度和 92.2%的召回率.他们也通过混合 SVM 实施早期的谣言发现,在传播开始时获得了 72%的准确率,24 h 后获得了 88%的准确率.一些研究[63]使用低度到高度节点的比例来衡量谣言的这种传播模式,和 Wu 的方法相比丢失了一些信息.

②传播树的基本结构.传播树的基本结构包括传播树的宽度与深度.宽度表明了消息的延展度,深度表明了消息的渗透力[67].在不同情况下,消息传播树的结构存在很大差异.例如,Nadamoto 等人[48]发现,在正常形势下,谣言传播树的深度有多层,而在灾难形势下,谣言传播树的深度只有 2 到 3 层.刘雅辉等人[68,69]研究正常形势下谣言与非谣言传播树的结构发现,大部分谣言传播树的深度比非谣言深,传播树的宽度比非谣言窄.这说明谣言的渗透力更深;而非谣言的延展度更广,传播得更集中.

上述两个工作都是研究基于消息的谣言传播结构.对于基于话题的谣言,同一个话题中所有消息的转发树构成了一个森林结构.在结构中,没有获得转发的消息以单点的形式存在,Kwon 等人[63]发现基于话题的谣言比非谣言包含了更多单点.

综上所述,传播特征总结如表 6 所示.从各方法对谣言的识别结果看,它们在精确度和召回率上都获得了很高的值.因此,传播特征对谣言有很好的预测效果,尤其是时间特征和结构特征.转发评论中人们对原消息的质疑或反驳也为谣言识别提供了重要线索.

表 6 传播特征

特征类	特征	说明	效果
转发评论	质疑	消息被转发时对其提出的质疑	Y
	更正	消息被转发时对其提出的反驳	Y
	情感词特征	积极和消极情感词的分数或个数	Y
	符号特征	情感符号的分数,?、!的分数或比例	Y
时序特征	转发数量	谣言和非谣言有不同的爆发次数	Y
	用户、内容以及传播特征	谣言和非谣言的这些特征随时间的变化有一定的差异	Y
	支持、质疑、反驳	作为实时谣言发现线索	Y
结构特征	传播树的节点类型模式	谣言从普通用户到意见领袖,非谣言从意见领袖到普通用户	Y
	传播树的基本结构	传播树的宽度、深度	Y
整体转发特征	转发数	所有转发的数量	y
	评论数	所有评论的数量	y

注:Y 表示对识别结果有重要影响,y 表示对识别结果有影响.

大部分以传播特征为主的方法是基于谣言的完整传播过程,一些特征取值也依赖传播过程结束才能获得,如转发数、评论数等特征.然而,谣言识别的最终目的是能在实际中得到应用,即在大量的信息流中,谣言能够被实时地、尽可能早地识别出来,以降低谣言产生的各种影响.流数据环境中包含了大量的噪声,使得实时谣言识别面临着巨大的挑战.一些研究已经尝试使用特征随时间的变化进行实时谣言识别(如表 7 所示).他们为实时谣言识别研究奠定了一定的基础.然而,从表 7 的比较可以看出,实时谣言识别效果并不理想,且缺少评价标准.研究者们可以考虑把谣言的爆发点作为实时谣言识别效果的评价标准,即在爆发点之前谣言被识别出来的准

表 7 实时谣言识别效果比较

动态信号	比较	实时识别效果
信念特征	与谣言跟踪网站比	75%的谣言揭露比他们早,准确率在 85%以上
	与新闻媒体揭露的时间比	75%的谣言揭露比他们早,准确率在 80%以上
质疑与更正	与只使用更正比	平均早 4.3 h
传播模式	与其它分类方法比	24 h 时获得 88%的准确率,比其它方法高 5%~11%
内容、用户、传播特征	与其它分类方法比	15 h 时获得 36%的准确率,比其它方法高 4%左右

确率.这个评价标准更合理,也能够很好地反映实时的谣言识别效果.

综合以内容、用户以及传播特征为主的方法,它们在验证特征对谣言和非谣言的识别效果时都使用了均衡数据集,但在真实的环境中,谣言与非谣言消息存在着较大的不均衡性.这可能导致有些特征在均衡的数据条件下表现明显,而在不均衡的条件下识别效果会变差的问题.因此,在以后的研究工作中还需要进一步探索,并且考虑在数据不均衡的情况下谣言的识别效果,挖掘出更具有鲁棒性的显著特征.

4.2　基于模型的谣言识别方法

基于分类的方法需要提取大量的特征,虽然部分特征能够很好地提升谣言的识别效果,但需要花费大量时间和人力.因此,有研究者采用基于模型的方法来识别谣言,如传染病模型、信息传播模型、隐马尔可夫模型以及神经网络模型.

4.2.1　基于传染病模型

研究者们在研究建模信息的传播模型时通常把用户是否参与传播作为主要考虑因素,如典型的传染病模型:SI 模型与 SIR 模型[70].一些研究者认为谣言的传播与传染病的传播具有很高的相似性,均可看成是在多个个体组成的一个封闭且同质群体中进行的传播过程.1964 年 Daley 和 Kendall[16] 借鉴传染病的 SIR 模型首次提出了谣言传播的数学模型,被称为 DK 模型.以此工作为基础产生了大量谣言研究工作①[71-73].这些研究工作试图通过建模谣言在社交媒体中的传播模式来研究谣言的演变、传播机制、识别以及传播控制等问题.

Jin 等人[74]首次尝试了通过建模谣言在 Twitter 上的传播模型来识别基于话题的谣言.他们首先使用传染病模型的变体 SEIZ(Susceptible-Exposed-Infected-Skeptic)[75] 建模谣言和新闻的传播,然后从模型中获得用户状态间的转换参数,最后使用这些参数组成的比率形式来识别谣言和新闻.虽然这个比率对一些特殊的话题并不能较好地区分,但是这个工作为谣言识别提供了另一种思路,即可以结合谣言的传播模型以及谣言的内容和用户模型或特征来提高谣言的识别效果.目前已经有一些研究使用传染病模型的变体或物理模型[76] 等来建模谣言的传播过程,他们已经证实谣言的传播模型和非谣言存在差异.然而,谣言传播过程的建模涉及到很多因素,传染病模型是否能建模谣言的传播过程还有待研究者们对谣言的演变、传播机制进一步深入的研究和证实.

4.2.2　基于信息传播模型

有些研究工作把所有发布、接受、转发同一消息的用户看成同质的用户群体,他们忽略了用户之间的个体差异,如用户的年龄、教育背景以及在线朋友对他们的影响等.如果考虑用户之间的这些差异,谣言的传播模式和非谣言的传播模式有着明显的不同.这种传播模式通常是一条消息的传播网络,即发布用户、接受用户以及转发用户组成的网络.

Liu 等人[77] 受 Jin 等人用传染病模型识别谣言的启发,构建基于不同用户行为的信息传播模型来自动识别谣言和可信的消息.他们构建的信息传播模型具有两种模式,包括谣言模式和可信的消息模式.他们进一步提取了 15 个用户特征,并使用逻辑函数计算用户转发消息的可能性.最后,他们用所构建信息模型识别谣言的方法如下:给定一条消息,首先抽取消息的传播网络,然后在两种模式下分别计算产生这样的消息并传播网络的可能性;如果计算的可能性在谣言模式下大于可信的消息模式,证明这条消息是谣言.Liu 等人的模型取得了 81.2% 的精确度和 79.3% 的召回率.

4.2.3　基于隐马尔可夫模型

Vosoughi[78] 认为特征的时序性对于谣言的识别非常重要.因此,他选择隐马尔可夫模型来建模特征的时序动态性,并针对谣言数据和非谣言数据训练了两个隐马尔可夫模型.对于给定的消息集分别计算在两个隐马尔可夫模型上的可能性.如果使用谣言数据训练的模型获得的可能性大,则为谣言,反之为非谣言.在官方或者新闻媒体公布谣言之前,模型能获得 75% 的准确率.他们使用的特征包括语言特征、用户特征以及传播特征.其中,语言特征包括否定的消息比例、消息内容的正式度和熟练度均值、包含意见和洞察力的消息比例、推断和试探性的消息比例;用户特征包括用户的争议性、原创性、可信性、影响力、角色以及活跃性;传播特征包括高影响力用户转发低影响力用户的比例、最大联通分量节点数的比例、深度与转发数的比例均值、新用户的比例、原帖的比例、包含外部链接的消息比例以及孤立节点的比例.这 3 类特征中,传播特征在模型中的谣言识别效果表现最好,其次是用户特征.

① http://infoscience.epfl.ch/record/176326/files/project-report.pdf

4.2.4 基于神经网络模型

神经网络起源于 20 世纪 40 年代. 近十几年,神经网络研究取得了引人注目的进展,从而激起了学术界和工业界研究者们的巨大热情和浓厚兴趣. 先后出现了神经网络的几种典型的结构形式,如深度神经网络(DNN)、卷积神经网络(CNN)、循环神经网络(RNN)等. 这些结构应用到不同领域,如计算机视觉以及自然语言处理等领域,均获得了非常好的效果.

传播用户在转发评论中留下了各种讨论谣言真实性的线索. Ma 等人[79]首次利用循环神经网络及其变体长短期记忆(LSTM)和门限递归单元(GRU)学习基于话题的消息的转发评论随时间变化的隐层表达,并利用隐层表达预测某话题消息是否为谣言. GRU 获得了最好的预测结果,在 Twitter 和新浪微博数据集上分别获得了 85.1%的精确度、95%的召回率和 87.6%的精确度、95.6%的召回率. 在谣言的早期发现上,谣言传播到 12 h 时在 Twitter 和新浪微博数据集上分别获得了 83.9%和 89%的准确率. 实验结果表明神经网络能够很好地捕获转发评论中能区分谣言与非谣言的隐藏线索,与最新使用特征工程的分类算法相比有明显的提升. 因此,在将来的研究中可以考虑使用神经网络的相关方法学习谣言的传播结构、用户等能够很好区分谣言与非谣言的要素.

综上所述,基于模型的方法避免了特征工程的缺陷,获得了很好的谣言识别效果. 然而,由于它们比较复杂,有大量参数,训练时需要大量谣言数据. 因此,如何获得大量且完整的谣言数据集成为应用这些方法的重要提前.

4.3 小 结

对本节主要的方法总结如表 8 所示. 针对基于分类的谣言识别方法分析如下:

(1)分类器. 各方法主要使用的分类器包括朴素贝叶斯、决策树、随机森林以及 SVM,其中 73%的方法选择了 SVM 的径向基核函数(RBF)来分类谣言和非谣言. 一些方法使用相同的特征集和数据集而选择不同的分类器进行谣言识别,如随机森林、决策树、SVM. 通过比较,随机森林的识别效果最好,决策树与 SVM 在不同方法上表现效果不同,相差大概 2%~10%. 由此可见,在基于分类的谣言识别方法中,分类器的选择对谣言的识别效果也有很大影响.

(2)特征和评价指标. 分类器主要基于内容特征、用户特征与传播特征中的一种或几种,不同的方法提取的特征侧重也不同. 分类效果的评价指标主要使用了准确率、精确度、召回率以及 $F1$. 表 8 中主

表 8 各研究工作对谣言识别的效果比较

研究工作	方法	特征			评价指标		数据集		比较的 baseline	
		内容 C	用户 U	传播 P	准确率	$F1$	Twitter	微博	Castillo	Yang
Yang 等人[58]	SVM	y	y	y*	78.7%			✓		
Castillo 等人[27]	决策树	y	y	y	86.0%	84.9%	✓			
贺刚等人[59]	SVM	y&**E**			81.2%			✓		↑10.4%
Sun 等人[61]	决策树	y&**E**(**M**)	y			70.4%		✓		
Gupta 等人[55]	决策树	y**&E**	y		96.7%		✓			
	朴素贝叶斯	y**&E**	y		91.5%		✓			
Liang 等人[37]	决策树		**E**			85.9%		✓		
	朴素贝叶斯		**E**			77.8%		✓		
	SVM		**E**			76.9%		✓		
Zhang 等人[60]	SVM	y&**E**	y&**E**			74.4%		✓		↑6.7%
Qazvinian 等人[11]	SVM	**E**	**E**			89.7%	✓			
Kwon 等人[63]	SVM	y		y&**E**(**T**(**V**))	87.3%	86.7%	✓		↑7.9%	
	决策树	y		y&**E**(**T**(**V**))	82.1%	82.2%	✓		↑13.7%	
	随机森林	y		y&**E**(**T**(**V**))	89.7%	87.8%	✓		↑11.6%	
Ma 等人[65]	SVM	y	y	y&**E**(**T**(**C&U&P**))	89.6%	89.4%	✓		↑10.9%	↑12.8%
	SVM	y	y	y&**E**(**T**(**C&U&P**))	84.6%	85.7%		✓	↑5.7%	↑5.2%
Wu 等人[62]	SVM	y	y	y&**E**(**S**)	91.3%	91.3%		✓	↑5.9%	↑13.9%
Liu 等人[77]	信息传播模型		**E**			81.3%		✓		
Vosoughi[78]	隐马尔可夫模型	y	y	y	75.0%		✓		↑11.0%	
Ma 等人[79]	循环神经网络	**E**			88.1%	89.8%	✓		↑15.8%	↑24.7%
	循环神经网络	**E**			91.0%	91.4%		✓	↑8.3%	↑9.7%

注:y 用了基本特征,y* 获得了最好评价的特征类,M 多媒体,T 时序,S 结构,V 转发数量,& 并且,E 强调了某类特征,如 E(T(V))强调了转发数的时序变化,↑提升,是与 $F1$ 比,没有 $F1$ 的与准确率比.

要列出了准确率和 $F1$ 两个指标,大部分方法都使用了 $F1$.从 $F1$ 的值可以看出,使用内容特征的分类器获得了较低的 $F1$ 值,用户特征获得的 $F1$ 比内容特征的 $F1$ 值略高,而使用传播特征的分类器获得了更好的 $F1$ 值.综合前面每部分的分析,我们可以得出用户特征和传播特征更易于发现谣言消息,而内容特征更易于发现非谣言消息的结论.因此,只通过分析消息的内容并不能很好辨别消息是否为谣言,需要结合用户以及传播特征才能达到更好的谣言识别效果.

(3)数据集.各方法所使用的谣言数据来自 Twitter 或新浪微博,它们选择一个或同时选取两种数据进行特征分类性能的验证.两种平台最明显的差异特征是新浪微博上有辟谣账号公布谣言信息,相对于 Twitter 能获得更高质量标注的谣言数据.从相同的方法在两个数据集上获得的准确率和 $F1$ 值看,在微博上获得的谣言识别效果比在 Twitter 上好,这可能是因为微博中的所有转发消息能够集中获得,转发消息之间的转发关系也比较容易提取,而 Twitter 上获得转发消息和转发关系相对于微博可能会丢失一些信息.

(4)方法比较的 baseline.各方法比较的 baseline 主要包括 Castillo 等人及 Yang 等人的工作.表中提升的效果主要是针对评价指标 $F1$.与 baseline 进行比较的工作中,大多数侧重从传播特征的角度识别谣言,$F1$ 提升得比较明显,证明传播特征能够很好地识别谣言.从同时和这两个 baseline 比较的方法的提升结果看,Castillo 的谣言识别效果比 Yang 的要好.这说明基于话题的谣言比基于消息的谣言包含更多的谣言信号,因为一些信号在单条消息中表现不明显,但在整个话题中会有明显体现.

此外,表中也总结了基于模型进行谣言识别的方法,它们主要基于传播过程进行建模,也获得了很好的识别效果.然而,谣言的内容和用户特征是谣言识别的基本特征,可以达到大概 70%的识别效果.这两种特征可以作为实时谣言识别开始时所依赖的基本特征.因此,在将来的研究中可以考虑将谣言内容和用户中包含的基本特征与谣言传播模型相结合进行谣言识别.

5　谣言识别研究存在的问题和展望

基于前面几节对国内外社交媒体中谣言识别研究工作的综述,本节总结谣言识别面临的问题并探讨未来的研究方向.

5.1　谣言识别研究存在的问题

(1)谣言数据难以获得

数据是谣言识别研究赖以进行的基础.现有的大部分研究工作都没有公布所用的数据集,譬如 Kwon 等人[63]由于保护数据中个人隐私[80]的原因,只给出了文章所使用的部分数据①.因此,到目前为止还没有公开且完整的谣言数据集,科研人员不得不通过社交媒体平台开放的 API 接口获取数据.然而,社交媒体平台也通常因为保护用户隐私之故设有一些限制,加之大数据时代数据处理和分析面临的数据复杂性、计算复杂性和系统复杂性挑战[81],所以研究人员只能获得少量且不完全的谣言数据,这对谣言的研究产生了很大限制.

(2)研究背景受限

大多已有研究工作以某些灾难或事件中传播的谣言为研究对象,这使得一些结论带有很强的事件依赖性.其实,在日常生活中也会产生大量谣言,如 Nadamoto 等人[48]发现事件背景下的谣言和日常生活中的谣言具有不同的传播模式.因此,研究者除了考虑灾难或事件中的谣言之外,也应该关注日常生活中的谣言,探索和挖掘它们有别于非谣言的特征或传播规律,为普通谣言的识别提供有效依据和重要线索.

5.2　谣言识别研究展望

社交媒体的广泛使用使得谣言识别研究得到了更多的关注.虽然已经取得了一定的进展,但是达到谣言的自动并实时识别,仍然有很多研究问题有待将来进行更深入的探究.

(1)谣言的传播规律发现

谣言也是一种信息,受很多因素的影响,因此,研究者们要发现谣言有别于非谣言的显著特征和本质规律面临着巨大挑战.存在的研究已从内容、用户以及传播方面进行了浅层和深层特征挖掘和规律发现.基于消息的整个传播过程进行谣言识别已经获得了很好的效果,准确率达到了 80%以上.因此,未来的研究可以探究传播过程中还未发现的显著特征和基本规律,如谣言传播模式随时间的动态变化,为谣言识别提供更好的依据.

(2)实时谣言识别

实时谣言识别的最终目标是在谣言传播的早期就能识别出谣言,即最好的情况是在谣言爆发的高

① http://mia.kaist.ac.kr/publications/rumor/

峰期之前识别出谣言,以减少谣言产生的不利影响.目前,针对谣言实时识别的研究工作较少.然而,实时谣言识别是将来应用和研究的重点.在社交媒体环境下,实时谣言识别涉及各方面的因素,如人的因素、社会因素和环境因素等,加上这些因素随时间的动态变化以及争取时间降低谣言影响的目的,识别方法获得较高的谣言识别准确率面临很大的困难.因此,识别方法可以给出可能是谣言的消息候选集,然后对候选集进行排序,也可以结合消息的流行度预测[82-84]对排序的消息进行预测,找出可能在未来某个时间内流行且有很大可能成为谣言的消息进行监测或人工排查.这样不仅能更早地甄别出谣言,而且还为制定有效的措施抑制谣言的传播赢得宝贵的时间.

(3)面向话题类型的谣言识别

话题类型包括政治类、经济类、环境类、娱乐类、生活类、健康类、科技类等,不同的话题类型,人们关心的可信程度不同[29],产生谣言的比率也不同[85].因此,很难从某一事件产生的谣言得出适用于所有话题类型谣言的共性传播规律,而且,并不是所有话题类型的谣言都能对人们产生很大的影响.譬如,娱乐类谣言通常不会产生太大的社会影响.此外,对于部分谣言,社交媒体平台也具有一定的自我净化能力.因此,未来的研究工作应该更多关注那些发布谣言后会对人们或社会产生很大影响且比较流行的话题类型,如生活类的话题、健康类的话题、科技类的话题等.研究这些类话题中谣言传播所独有的规律,建立适用不同话题类型的谣言识别模型.这样不仅可以降低谣言实时识别的处理量,而且也能提高谣言识别的准确率,缩短识别所用的时间.

(4)跨学科的谣言识别研究

谣言识别涉及到了心理学、社会学、新闻传播学、数据挖掘、机器学习、自然语言处理、社会网络分析等多个领域的理论、技术和方法,因此加强多学科合作对谣言识别研究具有重要意义.

(5)构建评测平台

谣言识别的分类方法中,各研究使用不同的数据集和特征集,虽然有一些研究工作同其它的工作进行了比较,但是被比较的工作在两个数据集上的识别效果存在着很大的差异[61].因为一些特征在其它数据集中不可用或者作用不明显,导致相同的方法在不同数据集上识别准确度不同.因此,将来可以考虑构建统一的评测平台弥补这一缺陷,从而促进谣言识别研究的进一步发展.

6 总 结

社交媒体中的谣言传播涉及信息、人和网络等多个方面的因素,既体现了信息传播的网络效应,又融合了人的社会心理因素.谣言在社交媒体中的传播会对人们的生活甚至整个社会造成不同程度的影响,因此谣言研究的最终目的是要抑制甚至阻止谣言的传播.然而,抑制谣言传播的前提是能够及时识别出谣言,因此,谣言识别成为社交媒体领域的重要研究问题之一.本文在充分调研和分析已有工作的基础上,对该研究领域的工作进行了综述.首先,介绍了谣言及其研究发展历程,分析了社交媒体中谣言识别的关键挑战和研究价值;然后,从影响谣言识别的关键要素与谣言识别方法两个方面详细介绍已有的研究工作,具体介绍了各个研究点的研究现状并进行了分析和总结;最后,本文总结了谣言识别存在的主要问题并探讨了未来的发展方向.

参 考 文 献

[1] Cao Bo-Lin. Social media: Definition, history of development, features and future—The ambiguous cognition of social media. Journal of Hunan Radio & Television University, 2011, 2011(3): 65-69(in Chinese)
(曹博林. 社交媒体:概念、发展历程、特征与未来——兼谈当下对社交媒体认识的模糊之处. 湖南广播电视大学学报, 2011, 2011(3): 65-69)

[2] Budak C, Agrawal D, El Abbadi A. Limiting the spread of misinformation in social networks//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 665-674

[3] Metaxas P T, Finn S, Mustafaraj E. Using TwitterTrails.com to investigate rumor propagation//Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing. Vancouver, Canada, 2015: 69-72

[4] Finn S, Metaxas P T, Mustafaraj E, et al. Trails: A system for monitoring the propagation of rumors on twitter//Proceedings of the Computation and Journalism Symposium. New York, USA, 2014: 1-6

[5] Zubiaga A, Liakata M, Procter R, et al. Towards detecting rumours in social media//Proceedings of the AAAI 2015 Workshop on Artificial Intelligence for Cities. Austin, USA, 2015: 35-41

[6] Huang Shao-Kuan, Huang Xiao-Bin, Li Bo. Research review of Internet rumors information governance. E-Government, 2015, 146(02): 54-63(in Chinese)

(黄少宽，黄晓斌，李波. 网络谣言信息治理研究综述. 电子政务，2015，146(02)：54-63)

[7] Wang Run. Theoretical concept of gossip and rumors in the Internet environment. Tribune of Social Sciences, 2015, 2015(03): 238-243(in Chinese)
(王润. 互联网环境下流言与谣言概念刍议——基于 Rumor 词源的分析. 社会科学论坛，2015，2015 (03)：238-243)

[8] Min Qing-Fei, Liu Xiao-Dan. A review of rumor study based on media evolution. Journal of Intelligence, 2015, 34(04): 104-109(in Chinese)
(闵庆飞，刘晓丹. 谣言研究综述:基于媒介演变的视角. 情报杂志，2015，34(04)：104-109)

[9] Difonzo N, Bordia P. Rumor, gossip and urban legends. Diogenes, 2007, 54(1): 19-35

[10] Difonzo N, Bordia P. Rumor Psychology: Social and Organizational Approaches. Washington, US: American Psychological Association, 2007

[11] Qazvinian V, Rosengren E, Radev D R, et al. Rumor has it: Identifying misinformation in microblogs//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 1589-1599

[12] Kumar S, West R, Leskovec J. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes//Proceedings of the 25th International Conference on World Wide Web. Montreal, Canada, 2016: 591-602

[13] Zhao Z, Resnick P, Mei Q. Enquiring minds: Early detection of rumors in social media from enquiry posts//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 1395-1405

[14] Knapp R H. A psychology of rumor. Public Opinion Quarterly, 1944, 8(1): 22-37

[15] Allport G W, Postman L. The Psychology of Rumor. Oxford, UK: Henry Holt, 1947

[16] Daley D J, Kendall D G. Epidemics and rumours. Nature, 1964, 204(4963): 1118-1118

[17] Buckner H T. A theory of rumor transmission. Public Opinion Quarterly, 1965, 29(1): 54-70

[18] Shibutani T. Improvised News: A Sociological Study of Rumor. New York, USA: Bobbs-Merrill, 1966

[19] Kapferer J N. Rumors: World's Oldest Media. Paris, France: Le Seuil Editions, 1987

[20] Bordia P, Rosnow R L. Rumor rest stops on the information highway transmission patterns in a computer-mediated rumor chain. Human Communication Research, 1998, 25(2): 163-179

[21] Zanette D H. Dynamics of rumor propagation on small-world networks. Physical review E, 2002, 65(4): 041908

[22] Buchegger S, Le Boudec J. The effect of rumor spreading in reputation systems for mobile ad-hoc networks//Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks. Sophia-Antipolis, France, 2003: 1-10

[23] Moreno Y, Nekovee M, Pacheco A F. Dynamics of rumor spreading in complex networks. Physical Review E, 2004, 69(6): 066130

[24] Difonzo N. Rumour research can douse digital wildfires. Nature, 2013, 493(493): 135

[25] Rosnow R L, Yost J H, Esposito J L. Belief in rumor and likelihood of rumor transmission. Language & Communication, 1986, 6(3): 189-194

[26] Fogg B, Tseng H. The elements of computer credibility//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Pittsburgh, USA, 1999: 80-87

[27] Castillo C, Mendoza M, Poblete B. Information credibility on twitter//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 675-684

[28] Gupta A, Kumaraguru P. Credibility ranking of tweets during high impact events//Proceedings of the 1st Workshop on Privacy and Security in Online Social Media. Lyon, France, 2012: 2-8

[29] Morris M R, Counts S, Roseway A, et al. Tweeting is believing?: Understanding microblog credibility perceptions//Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. Seattle, USA, 2012: 441-450

[30] Ratkiewicz J, Conover M, Meiss M, et al. Detecting and tracking the spread of AstroTurf memes in microblog streams. Computing Research Repository, 2010, 2010(10): 249-252

[31] Difonzo N, Bordia P. A tale of two corporations: Managing uncertainty during organizational change. Human Resource Management, 1998, 37(3-4): 295

[32] Lewandowsky S, Ecker U K, Seifert C M, et al. Misinformation and its correction continued influence and successful debiasing. Psychological Science in the Public Interest, 2012, 13(3): 106-131

[33] Liao Q, Shi L. She gets a sports car from our donation: Rumor transmission in a chinese microblogging community//Proceedings of the 2013 Conference on Computer Supported Cooperative Work. San Antonio, USA, 2013: 587-598

[34] Bordia P, Difonzo N. Problem solving in social interactions on the Internet: Rumor as social cognition. Social Psychology Quarterly, 2004, 67(1): 33-49

[35] Starbird K, Spiro E, Arif A, et al. Expressed uncertainty and denials as signals of online rumoring//Proceedings of the Collective Intelligence. Santa Clara, USA, 2015: 1-4

[36] Yang J, Counts S, Morris M R, et al. Microblog credibility perceptions: Comparing the USA and China//Proceedings of the 2013 Conference on Computer Supported Cooperative Work. San Antonio, USA, 2013: 575-586

[37] Liang G, He W, Xu C, et al. Rumor identification in microblogging systems based on users' behavior. IEEE Transactions on Computational Social Systems, 2015, 2(3): 99-108

[38] Mendoza M, Poblete B, Castillo C. Twitter under crisis: Can we trust what we RT?//Proceedings of the 1st Workshop on Social Media Analytics. Washington, USA, 2010: 71-79

[39] Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter//Proceedings of the 43rd Hawaii International Conference on Systems Science. Hawaii, USA, 2010: 1-10

[40] Zhou Yu-Qiong. Will QQ group communication lead participants to believe rumors more? An experiment of control of four Olympic game rumors. Journalism & Communication, 2010, 2010(2): 76-87+111(in Chinese)
(周裕琼. QQ群聊会让人更相信谣言吗?——关于四则奥运谣言的控制实验. 新闻与传播研究, 2010, 2010(2): 76-87+111)

[41] Starbird K, Maddock J, Orand M, et al. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston marathon bombing//Proceedings of the iConference. Berlin, Germany, 2014: 654-662

[42] Tanaka Y, Sakamoto Y, Matsuka T. Toward a social-technological system that inactivates false rumors through the critical thinking of crowds//Proceedings of the 46th Hawaii International Conference on System Sciences. Hawaii, USA, 2013: 649-658

[43] Friggeri A, Adamic L A, Eckles D, et al. Rumor Cascades//Proceedings of the 8th International Conference on Weblogs and Social Media. Michigan, USA, 2014: 1-10

[44] Bordia P, Difonzo N, Schulz C A. Source characteristics in denying rumors of organizational closure: Honesty is the best policy. Journal of Applied Social Psychology, 2000, 30(11): 2309-2321

[45] Bordia P, Difonzo N, Haines R, et al. Rumors Denials as Persuasive Messages: Effects of Personal Relevance, Source, and Message Characteristics. Journal of Applied Social Psychology, 2005, 35(6): 1301-1331

[46] Maddock J, Starbird K, Al-Hassani H J, et al. Characterizing online rumoring behavior using multi-dimensional signatures//Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. Vancouver, Canada, 2015: 228-241

[47] Takahashi T, Igata N. Rumor detection on twitter//Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems (SCIS), and The 13th International Symposium on Advanced Intelligence Systems (ISIS). Kobe, Japan, 2012: 452-457

[48] Nadamoto A, Miyabe M, Aramaki E. Analysis of microblog rumors and correction texts for disaster situations//Proceedings of the 15th International Conference on Information Integration and Web-Based Applications & Services. Vienna, Austria, 2013: 44

[49] Zeng L, Starbird K, Spiro E S. Rumors at the speed of light? modeling the rate of rumor transmission during crisis//Proceedings of the 49th Hawaii International Conference on System Sciences. Hawaii, USA, 2016: 1969-1978

[50] Sharara H, Getoor L, Norton M. Active surveying: A probabilistic approach for identifying key opinion leaders//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Catalonia, Spain, 2011: 1485-1490

[51] Chen X. The influences of personality and motivation on the sharing of misinformation on social media//Proceedings of the iConference. Philadelphia, USA, 2016: 1-11

[52] Huang Y L, Starbird K, Orand M, et al. Connected through crisis: Emotional proximity and the spread of misinformation online//Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. Vancouver, Canada, 2015: 969-980

[53] Sun Yan. Rumor storm: Research on the crisis of network opinion after disasters. Journalism & Communication, 2011, 2015(5): 52-62+111(in Chinese)
(孙燕. 谣言风暴:灾难事件后的网络舆论危机现象研究. 新闻与传播研究, 2011, 2015(5): 52-62+111)

[54] Andrews C, Fichet E, Ding Y, et al. Keeping up with the Tweet-dashians: The impact of 'official' accounts on online rumoring//Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. San Francisco, USA, 2016: 451-464

[55] Gupta A, Lamba H, Kumaraguru P, et al. Faking sandy: Characterizing and identifying fake images on Twitter during hurricane sandy//Proceedings of the 22nd International Conference on World Wide Web companion. Rio de Janeiro, Brazil, 2013: 729-736

[56] Arif A, Shanahan K, Chou F-J, et al. How information snowballs: Exploring the role of exposure in online rumor propagation//Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. San Francisco, USA: 465-476

[57] Wang Yuan-Zhuo, Jin Xiao-Long, Cheng Xue-Qi. Network Big Data: Present and future. Chinese Journal of Computers, 2013, 36(6): 1125-1138(in Chinese)
(王元卓, 靳小龙, 程学旗. 网络大数据:现状与展望. 计算机学报, 2013, 36(6): 1125-1138)

[58] Yang F, Liu Y, Yu X, et al. Automatic detection of rumor on Sina Weibo//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. Beijing, China, 2012: 13-20

[59] He Gang, Lv Xue-Qiang, Li Zhuo, Xu Li-Ping. Automatic Rumor Identification on Microblog. Library and Information Service, 2013(23): 114-120(in Chinese)
(贺刚, 吕学强, 李卓等人. 微博谣言识别研究. 图书情报工作, 2013(23): 114-120)

[60] Zhang Q, Zhang S, Dong J, et al. Automatic detection of rumor on social network//Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing. Nanchang, China, 2015: 113-122

[61] Sun S, Liu H, He J, et al. Detecting event rumors on Sina Weibo automatically//Proceedings of the Web Technologies and Applications—15th Asia-Pacific Web Conference, APWeb 2013. Sydney, Australia, 2013: 120-131

[62] Wu K, Yang S, Zhu K Q. False rumors detection on Sina Weibo by propagation structures//Proceedings of the 31st IEEE International Conference on Data Engineering. Seoul, South Korea, 2015: 651-662

[63] Kwon S, Cha M, Jung K, et al. Prominent features of rumor propagation in online social media//Proceedings of the 13th International Conference on Data Mining Dallas, USA, 2013: 1103-1108

[64] Matsubara Y, Sakurai Y, Prakash B A, et al. Rise and fall patterns of information diffusion: Model and implications//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 6-14

[65] Ma J, Gao W, Wei Z, et al. Detect rumors using time series of social context information on microblogging websites//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 1751-1754

[66] Liu X, Nourbakhsh A, Li Q, et al. Real-time rumor debunking on Twitter//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 1867-1870

[67] Bao P, Shen H-W, Chen W, et al. Cumulative effect in information diffusion: Empirical study on a microblogging network. PloS One, 2013, 8(10): e76027

[68] Liu Ya-Hui, Jin Xiao-Long, Shen Hua-Wei, et al. Rumor identification research and its development trend in social media. Chinese Association for Artificial Intelligence, 2016, 2016(3): 18-22(in Chinese)
(刘雅辉, 靳小龙, 沈华伟等. 社交媒体中的谣言识别研究及其发展趋势. 中国人工智能学会通讯, 2016, 2016(3): 18-22)

[69] Liu Y, Jin X, Shen H, et al. Do rumors diffuse differently from non-rumors? A systematically empirical analysis in Sina Weibo for rumor identification//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Jeju, South Korea, 2017: 407-420

[70] Kermack W, Mckendrick A. A contributions to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1927, 115(772): 700-721

[71] Bao Y, Yi C, Xue Y, et al. A new rumor propagation model and control strategy on social networks//Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara, Canada, 2013: 1472-1473

[72] Zhao L, Wang J, Chen Y, et al. SIHR rumor spreading model in social networks. Physica A: Statistical Mechanics and its Applications, 2012, 391(7): 2444-2453

[73] Zhao L, Wang Q, Cheng J, et al. Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal. Physica A: Statistical Mechanics and Its Applications, 2011, 390(13): 2619-2625

[74] Jin F, Dougherty E, Saraf P, et al. Epidemiological modeling of news and rumors on Twitter//Proceedings of the 7th Workshop on Social Network Mining and Analysis. Chicago, USA, 2013: 8:1-8:9

[75] Bettencourt L, Cintron-Arias A, Kaiser D I, et al. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. Physica A: Statistical Mechanics and its Applications, 2006, 364(364): 513-536

[76] Han S, Zhuang F, He Q, et al. Energy model for rumor propagation on social networks. Physica A: Statistical Mechanics and its Applications, 2014, 394(2): 99-109

[77] Liu Y, Xu S, Tourassi G. Detecting rumors through modeling information propagation networks in a social media environment. Social Computing, Behavioral-Cultural Modeling, and Prediction. Berlin, Germany: Springer. 2015: 274-278

[78] Vosoughi S. Automatic detection and verification of rumors on Twitter [Ph. D. Dissertation]. Cambridge, Massachusetts, USA: Massachusetts Institute of Technology, 2015

[79] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 3818-3824

[80] Liu Ya-Hui, Zhang Tie-Ying, Jin Xiao-Long, et al. Personal privacy protection in the Era of Big Data. Journal of Computer Research and Development, 2015, 52(01): 229-247(in Chinese)
(刘雅辉, 张铁赢, 靳小龙等人. 大数据时代的个人隐私保护. 计算机研究与发展, 2015, 52(01): 229-247)

[81] Cheng Xue-Qi, Jin Xiao-Long, Wang Yuan-Zhuo, et al. Survey on Big Data system and analytic technology. Journal of Software, 2014, 25(9): 1889-1908(in Chinese)
(程学旗, 靳小龙, 王元卓等人. 大数据系统和分析技术综述. 软件学报, 2014, 25(9): 1889-1908)

[82] Bao P, Shen H-W, Huang J, et al. Popularity prediction in microblogging network: A case study on sina weibo//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 177-178

[83] Shen H-W, Wang D, Song C, et al. Modeling and predicting popularity dynamics via reinforced Poisson processes//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec, Canada, 2014: 291-297

[84] Bao P, Shen H-W, Jin X, et al. Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 9-10

[85] Wang Q, Lu T, Ding X, et al. Think twice before reposting it: Promoting accountable behavior on Sina Weibo//Proceedings of the 18th International Conference on Computer Supported Cooperative Work in Design. Taiwan, China, 2014: 463-468

LIU Ya-Hui, born in 1979, Ph. D. candidate, lecturer. Her research interests include social media computing and data mining.

JIN Xiao-Long, born in 1976, Ph. D., associate professor, Ph. D. supervisor. His research interests include social computing, knowledge graph, and artificial intelligence.

SHEN Hua-Wei, born in 1982, Ph.D., associate professor, Ph. D. supervisor. His research interests include social network analysis and social media computing.

BAO Peng, born in 1987, Ph. D., lecturer. His research interest is social media computing.

CHENG Xue-Qi, born in 1971, Ph. D., professor, Ph. D. supervisor. His research interests include network science and social computing, web search and mining, network and information security, distributed systems and large-scale simulation platform.

Background

In recent years, the vigorous development of social media, such as Facebook, Twitter, Sina Weibo, WeChat etc. which have gradually become an important channel for people to get or share information. It is convenient for everyone to participate in the release and dissemination of information. However, it also brings some problems. The rapid growing and spread of Internet rumors have brought serious challenges to the effective use of social media and scientific management. How to identify the rumor quickly and accurately is one of the main problems in the field of industry and academia. It is also the key premise of the rumor tracking and taking measures to control the spread of rumors and reducing the impact of rumors. Therefore, it is of great significance and application value to study the rumor identification. This paper surveys and summarizes the latest research result of rumor identification from the perspective of the elements and methods of rumor identification.

This work was funded by the National Key Research and Development Program of China under Grant No. 2016YFB1000902, the National Basic Research Program of China (973 Program) under Grant No. 2013CB329602, the National Natural Science Foundation of China under Grant Nos. 61572473, 61472400, 61232010. H. W. Shen is also funded by Youth Innovation Promotion Association CAS and the CCF-Tencent RAGR (No. 20160107). P. Bao is also funded by the Fundamental Research Funds for the Central Universities under Grant No. 2015RC031 and the State Scholarship Fund from the CSC. These projects focus on rumor identification, rumor control, information diffusion on online social media. Our group has been working on these projects for several years and published a number of related papers. This paper is a survey of the issues of rumor identification.