# 在线社会网络谣言检测综述

陈燕方[1] 李志宇[2] 梁 循[2] 齐金山[2]

[1](中国人民大学信息资源管理学院 北京 100872)

[2](中国人民大学信息学院计算机系 北京 100872)

**摘 要** 大数据环境下,在线社会网络与人们的生活、娱乐以及工作逐渐融为一体.然而"信息过载"和"信息污染"已成为在线社会网络诸多应用发展面临的主要瓶颈之一,并同时造成了用户的"信息焦虑"和"信息迷航"等一系列问题,因此在线社会网络谣言检测是改善在线社会网络信息生态环境质量、提升用户体验的有效手段.在线社会网络谣言检测隶属于信息可信度检测研究范畴,但谣言的不确定性、较强的时效性、主观性和关联性等特征又使得其与虚假信息检测有着本质区别.基于以上,该文从在线社会网络谣言的基本概念和特征研究出发,分别基于目标、对象和时间三个属性,分析了在线社会网络谣言检测研究基本问题的形式化定义,并介绍了研究中数据采集和标注的不同方法.然后,分别对不同类别和应用场景的在线社会网络谣言检测方法和谣言源检测方法进行了分析和总结.最后,该文讨论了在线社会网络谣言检测技术未来发展面临的若干挑战以及可能的研究方向.

**关键词** 在线社会网络;谣言;虚假信息;谣言检测;谣言源检测;网络结构分析

**中图法分类号** TP311 **DOI 号** 10.11897/SP.J.1016.2018.01648

## Review on Rumor Detection of Online Social Networks

CHEN Yan-Fang[1] LI Zhi-Yu[2] LIANG Xun[2] QI Jin-Shan[2]

[1](*School of Information Resource Management, Renmin University of China, Beijing* 100872)

[2](*Department of Computer Science, School of Information, Renmin University of China, Beijing* 100872)

**Abstract** Online Social Networks (OSN) are integrated into people's life, entertainments, and works with the development of big data. However, the issues of information overload and information pollution have become one of the most serious problems hindering the improvements of many OSN applications, decreasing the cost of misinformation diffusion, and resulting in the widely spreading of rumor information, which may also cause information anxiety and information disorientation among people. Thus, OSN rumor detection is an effective way to improve the quality of OSN information ecology environment and user experience. To some extent, rumor detection belongs to the research area of information credibility evaluation; however, the unique features, including uncertainty, time-effective, subjectivity, and relevance, et al., of rumor detection make a big difference from misinformation detection. Currently, there are three main problems in rumor detection, namely timeliness, extendibility and scalability, which are great challenges for all the researchers in this field to tackle. In this paper, we firstly summarized some basic concepts (rumor, online rumor and online social network rumor) and the main features of OSN rumor on five perspectives (who, say what, in which channel, to whom, with what effect). And then, we

收稿日期:2017-02-20;在线出版日期:2017-11-29.本课题得到国家自然科学基金(71531012,71271211)、北京市自然科学基金(4172032)、中国人民大学科学研究基金项目(10XNI029)、中国人民大学 2017 年度拔尖创新人才培育资助计划成果之一资助.**陈燕方**,女,1992 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为社会网络、信息组织、数据挖掘、知识发现.E-mail:cyf@ruc.edu.cn.**李志宇**,男,1991 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为社会计算、网络挖掘、自然语言处理.**梁 循**(通信作者),男,1965 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为神经网络、支持向量机、社会计算.E-mail:xliang@ruc.edu.cn.**齐金山**,男,1977 年生,博士研究生,主要研究方向为社会计算、数据挖掘.

discussed the problem statement of its hot research topic based on different perspectives including detection objective, detection object and detection time to answer the three 3W questions in the process of rumor detection, namely what, who and when. In addition, we showed different ways of rumor data collection (platform, content, quantity, proportion, and method) and data annotation (manual annotation and machine annotation), and some open online rumor datasets. Next, we compared and demonstrated the various performances, strengths, weakness, and applications for both OSN rumor detection based on content and rumor source detection based on network structure techniques. For OSN rumor detection, content classification and content comparison are most common methods currently, while for OSN rumor source detection, model of observing snapshot of sub-graph diffusion and model of observing monitor nodes are the mainstream methods. Finally, we presented six key challenges and future works of OSN rumor detection problem, namely dynamic identification, large-scale distributed detection, unbalanced data set, the heterogeneity of rumor information diffusion, the relevance of rumor information diffusion, and multimedia rumor information detection. OSN rumor detection techniques, as the hot issue in the research areas in social network, information diffusion, et al, has been improved a lot with the development of OSN new topics detection method, new topics diffusion prediction and information credibility prediction. As the current research trend showed, OSN rumor detection techniques have changed gradually from static, time-delay, small-scale and single-source to dynamic, real-time, large-scale and multi-source. This transition requires the following OSN rumor detection model be more efficient and extensible. Of course, the OSN rumor detection model should also keep a flexible balance between efficiency and accuracy based on the specific application scenarios, meanwhile, a certain amount of human factors should be introduced in specific environment, thus further increasing the accuracy and timeliness. Based on reviewing the OSN rumor detection research, this paper hopes to give an effective reference to the following researchers in the relevant research areas, thus co-promoting the development and optimization of the OSN rumors detection model.

**Keywords**　online social networks; rumor; misinformation; rumor detection; rumor source detection; network structure analysis



## 1　引　言

Web2.0 以及移动互联网技术的发展使得在线社会网络(Online Social Networks,OSN),这一重要的信息传播平台(或称载体)逐渐与人们的生活、工作和学习融为一体.在这一平台中,人们不再仅仅是信息的接受者,同时也成为了内容(User Generated Contents,UGC)的创造者[1].其具体形态包括在线社交网站(如 Facebook、人人网等)、在线网络社区(如猫扑)、在线社交媒体(如新浪微博、Twitter 等)、在线论坛、视频分享(如 YouTube、优酷等)以及图片分享播客网站(如 Flickr)等.凭借参与、公开、交流、对话、社区化的特性,在线社会网络大大加速了人与人之间信息交流的速度与深度,但同时我们还需要看到:它在为人们提供便利的信息交流、互动的同时,也降低了不实信息的传播成本,进而逐渐发展成为网络谣言肆意滋生的温床,如著名的"谣盐"恐慌事件.该事件通过社会网络大范围传播,进而引发商户恶意囤积、哄抬价格、扰乱市场等不法现象,并导致民众抢购并囤积碘盐,从而造成各地盐荒.

在线社会网络中"谣言泛滥"的现象严重影响了网络信息生态环境的健康状态,使得人们难以在种类纷繁复杂、质量参差不齐的信息中找到自己所需的可用信息,从而降低人们所获取信息的质量.尤其是遇见突发性事件时(如自然灾害、社会突发事故、卫生安全事故等),在线社会网络的实时性与便利性极易增加谣言传播的速度和广度,从而进一步引发公众的恐慌和焦虑,由此严重影响正常的社会秩序.

由此可见,在线社会网络中有关谣言检测的研究不仅能有效地净化网络信息生态环境,还能帮助公众迅速地甄别有效的信息.同时,各类社会网络平台的管理者也能通过增强平台的信息可信度来提升平台的用户粘度和活跃度.

目前,在线社会网络中的谣言检测研究多被划分为在线社会网络信息可信度的研究范畴.然而,从本质上来说,谣言检测和信息可信度评估是有本质区别的:首先,信息可信度评估的核心是鉴别信息的真伪,但谣言不一定是假的,谣言是那些人们感兴趣或觉得重要的,事实未经证实的阐述(详见 2.2 节),由此,二者在检测的目标上有显著差异.其次,信息可信度检测通常是基于单条信息的,而谣言检测一般基于话题或事件,虽然当前仍有不少谣言检测的研究以单文本信息为对象.以微博平台为例,一则 140 字以内的短文本所包含的信息量极为有限;而同一事件的微博群则汇集了一个群体的智慧,检测结果也更有说服力且更符合真实场景.在传统媒体中,专家也是基于信息集合来判定该事件信息是否为谣言[2].因此,二者在检测对象上也存在差异.总体来说,目前在线社会网络谣言检测研究有以下 3 个核心的问题需要解决:

(1)时效性

谣言的传播可粗略的划分为潜伏期,爆发期与消停期三个阶段.谣言在传播过程中,其检测的延时时间和谣言的生命周期呈近似的线性关系[3],且谣言传播的时间越长,越有可能被媒体或用户发现(尤其在爆发期及之后)[2],但此时谣言可能已经造成了一定的破坏性.因此,在谣言潜伏期实现早期识别是十分重要且必要的,苛刻的说,能够达到实时检测(通常指将检测误差时间控制在一定小的时域内)是最为理想的.然而,由于谣言在潜伏期所提供的信号特征是有限的,因此目前 OSN 谣言检测的研究通常以静态数据集的回溯性检测模型(Static Model Retrospectively)居多,这类模型无法随着新微博的增长而实时更新;因此,针对早期检测与实时检测技术的研究是当前与今后研究的重点方向之一,我们将这类问题统称为谣言检测的冷启动问题.

(2)可扩展性

在全球不同国家不同地区,人们在不同社会网络上针对某一事件的评论方式以及信息传播的特征都是有一定差异的.如将国外的 Twitter 和国内的新浪微博进行对比可以发现:相较于 Twitter 对用户数据保护的严谨性,新浪微博上的用户信息则较为开放,因而相对更好获取.此外,同一时间空间内,针对不同话题的传播特征、文本特征以及用户传播行为特征也同样是有差异的,如旧话题谣言与新话题谣言、突发性话题谣言与蔓延型话题谣言等.这种基于特定地域或特定领域的差异将会导致训练数据集有一定的偏差,从而降低了检测器在实际应用中的性能.因此,在 OSN 谣言检测过程中,对不同数据集的采集与测试,对独立于话题特征的选择是当前研究中数据和特征在获取和选择过程中的重点问题.

(3)可伸缩性

由于谣言检测的类型、方法或者目的存在一定的差异,在具体检测过程中可获取的数据量大小也会有较大的区别.如突发性事件谣言或重要事件谣言能在短时间内产生大量的数据用于检测实验,而一般性话题谣言数据则相对较少.此外,谣言传播过程中检测时间越早,数据越有限.特别的,在大规模数据的训练条件下,模型是否拥有良好的计算性能,在保证检测准确性的同时,不失检测速度,即检测模型的可伸缩性也是 OSN 谣言检测研究的核心问题之一.最后,关于可伸缩性的研究,我们不仅仅需要从数据量级上进行考虑,还需要从不同数据类型的融合性问题上出发,设计符合当前环境下的可伸缩性谣言检测模型或方法.

**行文结构**:如图 1 所示,为本文后续章节行文的基本框架.本文第 2 节对社会网络谣言检测研究中所涉及的基础概念进行总结和辨析;第 3 节分别从谣言检测的目标属性、对象属性与时间属性对谣言检测问题进行形式化的定义;第 4 节对当前谣言检测所涉及的数据采集与标注问题进行分析;第 5 节与第 6 节则分别对基于内容分类的、对比的谣言检测方法以及基于网络结构的谣言源检测方法进行讨论与总结,给出不同方法的主要特征、相互关系以及应用方向等核心内容;接着,本文在第 7 节对当前在线谣言检测研究所存在的挑战进一步进行总结,并给出未来的研究展望;最后,我们在第 8 节对全文进行总结性概述.

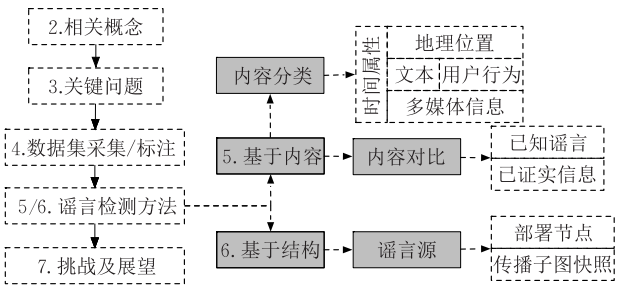


图 1 论文综述结构

## 2　相关概念

### 2.1　在线社会网络

社会网络(Social Networks)是一个典型的复杂网络[4]，它包含了一定范围内人与人之间的关系.其中个人抽象为网络中的节点，关系则抽象为节点之间的链接.在线社会网络(Online Social Networks，OSN)是随着互联网的诞生而出现并逐渐发展的，其最早的形式可以追溯到人们通过 Email 交换信息所构建的关系网络.因此，OSN 本质上是现实中的社会网络在虚拟网络空间中的映射，是基于在线社交网站、在线社区、在线社交媒体等构建的复杂网络[5].与 Web 网络强调内容不同，OSN 是由用户组成的，参与者加入网络后可以发布任何内容并与任何和他们有联系的其他用户产生链接[6].OSN 中的基本要素分别是：Users(用户，即参与到在线社会网络中的每一个人)、Links(链接，即每个用户账号之间建立的链接关系)、Groups(群，用户根据特定兴趣聚成的群组).用户与用户、群组与群组、用户与群组之间都可能产生链接，且这种链接具备双向属性[7]，图 2 是一个典型的社会网络结构[8].

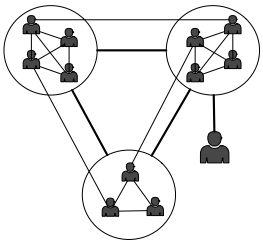

图 2　一个简单的小型社会网络结构图例

当前对 OSN 的研究主要集中在 OSN 结构分析、OSN 演化分析以及 OSN 中的用户行为分析三个方面.其中，结构分析主要是采用平均路径长度、聚类系数、度分布等指数对 OSN 的静态网络结构特征进行度量.演化分析则是从动态的角度研究 OSN 的基本拓扑参数、节点关系以及组关系随时间变化的规律及其内在的联系.而 OSN 中的用户行为分析则包括对用户的会话、点击流等基本行为分析、用户在社交网站应用上的行为分析以及用户行为活动所表现的结果[9].

### 2.2　谣言与在线社会网络谣言

#### 2.2.1　谣　言

谣言是一种特殊的信息，也是一种古老的社会现象，长久以来一直是人们关注的热点研究问题[10].谣言作为一个学术概念，系统的研究始于二战时期，此后大致经历了三次转变.早期，谣言研究主要源于个体心理，以 Knapp[11]、Allport 等人[12]和 Postman 等人[13]为代表的谣言心理学奠基人将谣言分别定义为："旨在使人相信的说法，它与当前时事有关，在未经官方证实的情况下广泛流传."、"是一种为了让人相信，关于当时事件的陈述，常以口头形式在人们中传播，但又没有可靠证明标准."、"一种在人们之间私底下流传的，对公众感兴趣的事物、事件或问题的未经证实的阐述或诠释."自 1960 年开始，随着研究的深入，越来越多的学者开始从社会学角度考察谣言的定义："谣言是在环境模糊时，群体根据已有的信息碎片，对事件意义和解释的建构，是在一群人讨论过程中产生的即兴新闻[14]."、"在社会中出现并流传的未经官方公开证实或者已经被官方辟谣的信息[15]"、"在社会群体中流传的有关当前时事的信息，完全通过口传、往往没有任何事实依据[16]"以及"体现人们对现实世界如何运转的假设的公共传播[17]".进入 21 世纪后，谣言的侧重点开始聚焦于谣言的社会功能，其中该时期谣言研究领域的集大成者 DiFonzo 等人[18]将谣言定义为："谣言是在模糊或危险情境下产生的未经证实却在广泛流传的说法，它能帮助人们弄清事实并控制风险".在中国，谣言的系统研究最早可追溯到 1939 年陈雪屏编写的《谣言的心理》[19]，他指出：谣言是一种复杂的心理社会现象(psycho-social phenomenon)，它与其它一切语言文字的报告或陈述(如新闻、传说等)在实质上是可以相通的，同是追述过去的事实，但谣言所根据的事实较少，主观的补充与改造较多.对于谣言的产生和传播动机的研究也主要集中在心理学[17,20-21]和社会学[13-14,22]两个领域.从个体心理需求层面，认为谣言的产生是个体处于不确定环境中为了排除内在的焦虑和恐惧进行的尝试.从群体层面出发，认为是群体为解决问题、并以获得社会认知为目的而进行的一种集体交易(collective transaction).

综上所述，关于谣言的定义及其产生动机研究主要从心理学和社会学两个视角展开，其侧重点逐步经历了从谣言产生语境到谣言传播特征，最后聚焦到谣言所产生的社会影响上来.虽然从不同时期、不同视角出发所得出的结论存在一定的偏差，但总结下来谣言主要有以下特点：(1)谣言是关于社会上某事某问题的特殊陈述信息，一般来说是关于某事件的断言(Assertion)；(2)谣言会在人群中广泛传播.一般以话题的形式通过口传媒介在人群中流传.当然这与研究者所处的时代相关，显然，在当前环境下，谣言传播的主要媒介以网络媒介为主导；

(3)谣言内容的真伪是未经证实且不确定的,其传播内容最终可能鉴定为真,可能鉴定为假,还可能仍处于不确定状态;(4)谣言一般是公众觉得与自身相关且感觉重要的或是他们感兴趣的内容,同时,这种重要性越高,谣言的流行度也越广;(5)谣言通常包含谣言制造者的个人主观意见较多,而客观事实则相对较少;(6)谣言总是"非官方的",不会在政府或权威的信息机构或个人之间传播,而通常产生于非正式的话语空间,通过非正式渠道进行传播;(7)谣言的传播具有一定的目的性,从个体层面是为了克服内心的恐惧,从群体层面是为了获得社会认知或解决某一问题;(8)谣言同时具备时效性和反复性,因此一旦辟谣成功,便不复存在.

对于谣言的传播规律,Allport 和 Postman[12]提出了谣言传播公式:$R=i\times a$. 其中,$R$(Rumor)为谣言传播的流行性,$i$(importance)为谣言所涉及事件的重要性,$a$(ambiguity)为该事件的模糊度. 其中,关于二者的乘积关系有两点需要说明:(1)事件越重要或越模糊,则谣言的影响程度越大;(2)当重要程度或模糊性为 0 时,谣言会停止传播. 在该公式中,谣言传播是无意识主体做出的反应,对此 Crouse[23]在上述基础上加入了人的影响因素,从而将公式进一步修改为 $R=i\times a\times c$,其中 $c$(critical ability)为公众批判能力,即公众的批判能力越强,则谣言产生的可能性就越小. 同时,Rosnow[17]则认为还应考虑个体的心理指标,如个人的焦虑和担忧. 此外,还有学者认为除了心理指标变量,还应考虑社会环境要素. 如国内学者胡钰[23]提出事件的反常态度应该是谣言产生的主要因素之一,反常性越大的事件越容易引起人们的兴趣,并定义谣言的传播公式为:$R=a\times a'\times a''$,其中 $R$ 为谣言,$a$ 为关注度(attention),$a'$表示模糊度(ambiguity),$a''$表示反常度(abnormality),该公式从谣言的传播规律出发,更贴近谣言传播的实质,即越是戏剧性的谣言越容易引起人们的注意,传播的速度也越快.

对谣言传播建模研究始于 20 世纪 60 年代,Sudbury[24]根据谣言传播与疾病传播的相似性,将基于数学统计的宏观数学模型:传染病模型,用来对谣言的传播进行建模. 如图 3 所示,模型中,种群个体被抽象为三种类型(状态):易感染(Susceptible,S)、感染(Infected,I)、治愈(Recovered,R). 根据状态之间的转换关系,可构建 SI[25]、SIS[26]和 SIR[27]三种基础模型. 其中,SI 模型仅考虑易感染人群转变成感染人群且不会治愈的状况. SIS 模型考虑到了易感染人群转变成感染人群以及感染人群又变成易感染状态的状况. 而 SIR 模型则在 SI 模型上增加了感染人群转换到治愈状态且再也不会被感染的情况. 此后,学者们根据谣言的感染群体[28-29]、感染方式[30-31]及感染程度[32]的类型对谣言传播的传染病模型进行了进一步的修正和完善. 此外也有学者从微观数学出发,提出了借助随机过程方法建立的 D-K 谣言数学模型[33]以及基于离散数学的原细胞自动机模型[34].

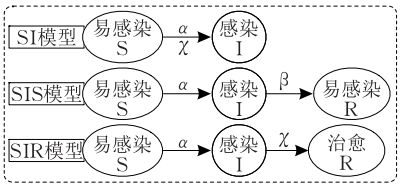


图 3 一种基于"传染病模型"的谣言传播模型(其中,S/I/R 分别对应于易感染(Susceptible)、感染(Infected)以及治愈(Recovered)类型,指示箭头上的系数表示为感染率)

### 2.2.2 在线社会网络谣言

媒介每一次的进化都意味着谣言传播载体的丰富与补充,并逐渐出现以新代旧的趋势. 从口耳传播、大众媒体传播、网络传播再到如今新兴的在线社会网络传播. 信息传播技术以及传播环境的转变只是改变了谣言传播的形态,而并未改变谣言的实质内容. 当然,不同传播形态的谣言也产生于相应的社会背景,同时反映了一定的社会问题,从而显示出了一系列新的特征. OSN 谣言产生于 OSN 的兴起,属于网络谣言的范畴,而网络谣言也通常作为谣言研究的一个分支,并未发展成为一个独立的研究领域. 目前对 OSN 谣言的理论研究多基于网络谣言研究,并结合微博、Twitter 等活跃的 OSN 平台展开一系列的实践研究. 因此,下文将通过剖析网络谣言的定义、特点来分析 OSN 谣言的内涵和特性.

对于网络谣言的定义,大多数学者多从媒介论的角度出发,即认为网络谣言只是传播媒介由传统的人际媒介或大众媒介转向了网络媒介. 如 Bordia 和 Rosnow[35]指出网络谣言和传统谣言在本质上是相同的(如特征、内容与情感等),但网络谣言的传播方式、传播范围、传播节奏和传统谣言是有差异的. 对此,周裕琼[36]认为网络谣言除了强调网络的传播属性外,还应关注网络谣言作为谣言分支的特殊属性,即其广泛的社会传播可能造成的社会影响. 因此,她将网络谣言定义为那些互联网在其产生、传播、影响的某阶段或全过程中起到过关键作用的谣言,它们的产生、传播、影响与社会密切

相连，它们关乎社会问题，隶属社会传播，具有社会意义.

综合以上分析，本文将在线社会网络谣言定义为在线社会网络谣言是指在线社会网络在其产生、传播、影响的某阶段或全过程起到过关键作用的，内容未经证实的，且造成了一定社会舆论影响的阐释或阐述.

OSN 谣言继承了 OSN 和谣言的双重特点，同时也有区别于传统谣言的个性化特征. 下面，本文将基于信息传播的 5W 要素[37]，即 who、say what、in which channel、to whom、with what effect，对传统谣言与 OSN 谣言进行对比，如表 1 所示.

**表 1　在线社会网络谣言与传统谣言比较**

| 类别 | 传统谣言 | 在线社会网络谣言 |
|---|---|---|
| 传谣群体特征 | 精英化、信息垄断 | 去中心化、草根性、匿名化 |
| 谣言内容特征 | 事实与数据不足，内容较易更改 | “真实的谣言”、内容和形式较为固定、可再创造 |
| 传播渠道特征 | 时空限制、无反馈、难以溯源 | 突破时空限制、有反馈、可溯源、成本降低 |
| 受众群体特征 | 被动性 | 双重性、个性化、虚拟性 |
| 传播影响特征 | 范围局限 | 范围广、难以控制 |

(1) 传谣群体特征

除了网络带来的虚拟化特征，OSN 传谣群体最大的特征表现为去中心化. 传统谣言的来源主要是信息生产者(内容提供商)等精英化或意见领袖角色，如大众媒体. 但在 OSN 环境中，由于信息发布者的草根性，传谣群体脱离了传统媒体的信息垄断地位，呈现离散的去中心化状态，同时 OSN 也综合了人际媒介的点对点和大众媒介的点对面的双重传播特点. 此外，随着传谣背后相关利益输送下“产业链”的不断壮大，一个新兴的群体——社会网络水军，也应运而生. 社会网络水军具备普通网络水军的共性特点：如目标明确(大多都是为了获取经济利益或造成一定的网络影响)[38]，内容特征异常(如存在大量广告、链接，情感极性较强等)，行为模式异常(如不平衡关注-粉丝比)、关系模式异常(如朋友间缺乏链接关系)[39]等. 但相较于其它网络水军(如电子商务、邮件、论坛等领域的网络水军)而言，社会网络水军隐蔽性更高，更趋同于正常用户，异常模式识别难度也更大，通常需要结合内容、行为和关系等综合特征来进行识别，这也是谣言检测过程中的难点之一.

(2) 谣言内容特征

在谣言内容特征上，传统谣言与 OSN 谣言的区别主要体现在三个方面：第一，OSN 谣言内容丰富且逼真. 相较于传统谣言，OSN 谣言包含了音频、视频等大量多媒体信息以及丰富的事实和数据等细节，在形式上进一步逼近真实消息. 而传统谣言口口(Word of Mouth)传播的方式通常是传谣者对受众思想或是心灵的感染[40]，没有 OSN 谣言丰富的事实依据. 同样的事实在可信度研究中也有共识，即电视媒体信息比纸媒信息更让人觉得可信[41]. 第二，OSN 谣言的内容在传播过程中较为固定，较少被修改，即通常呈现出直接转发，或复制重发等现象. 传统谣言的内容在传播过程中极易由于人们记忆、表述、个人主观情感的偏差而产生误传，而 OSN 谣言内容在传播过程中，用户只需复制粘贴或转发，其内容是可以追溯的. 第三，OSN 谣言内容可再创造. 由于多数 OSN 平台的转发加评论功能使得 OSN 传谣者可以在原谣言内容不变的基础上继续发挥，在评论中加入个人的主观意见进一步渲染谣言的舆论氛围.

(3) 传播渠道特征

在谣言的传播渠道上，OSN 谣言主要体现出四点特征：第一，高速实时的跨地域传播. 由于 OSN 的虚拟性特征，谣言的传播突破了时空限制，在网络上实现了无需面对面的实时交流模式. 第二，可溯源传播. 在传统环境中谣言是短暂的现象，很难留下痕迹，因此，同一谣言很容易在不同时期反复传播[22]. 但在在线社会网络中，信息是可以记录并追溯的，因此一旦辟谣成功后，谣言很难产生二次广泛影响. 第三，低成本传播. 在互联网中，发布任何虚假的、具有误导性的谣言都变得十分容易[42]. OSN 平台的虚拟性、匿名性、高用户活跃度等特点都使得谣言广泛传播的成本大大降低了. 第四，实时的辟谣反馈. OSN 中用户可以对他人发布的消息进行点评，因此在谣言的传播过程中，随着时间的推移，评论中越来越多的质疑使得谣言的传播与辟谣常常是同时发生的.

(4) 受众群体特征

在 OSN 环境下，受众群体通常由传统的被动接受转变为主动参与，且具有双重可能身份. 而在传统环境中，谣言的接受者只能被动接受信息，难以将自己的观点公之于众. 但在 OSN 环境下，用户可以对接收的谣言进行评论并表明自身的立场，甚至继

而成为新的谣言传播者.因此在 OSN 中传谣群体和受众群体是密不可分的,同时也是相互影响的.

(5)传播影响特征

OSN 谣言传播的影响特征主要体现在两个方面:第一,较强破坏力的强势舆论.传统谣言的传播受到时空限制,其负面影响相对较小.由于 OSN 平台的活跃度较高,用户群体较为密集,信息传输所需的平均路径较短,使得 OSN 谣言一经传播,将以分秒级的速度波及世界各地,并形成较强的社会舆论,其所辐射的人群之广,影响范围之大,将带来更加难以预测的潜在危害.第二,复杂的传播结构难以控制.不同于传统谣言的线型传播结构、传统大众媒体的扇型传播结构以及网络谣言的网状传播结构,OSN 谣言裂变式的传播模式会呈现出链状、树状、放射状、漩涡状等多种传播结构,各种传播模式之间的特征既存在同质性,也存在较大的差异性,因此需要针对不同的传播模式设计有效的检测方法.

综上所述,OSN 谣言与传统谣言以及一般的网络谣言在各类特征上都存在较大的差别.因此,需要针对性的对 OSN 谣言检测进行研究和分析,从而设计出特定的检测模型,达到能够对 OSN 谣言传播进行高效的管理与控制的目的.接下来,本文将对 OSN 谣言所涉及的问题进行形式化的描述,以期给后续研究者提供有效的分类参考.

## 3 OSN 谣言检测的关键问题描述

谣言检测问题描述中需要分析的 3W,分别是检测目标(What)、检测对象(Who)以及检测时间(When).依据对上述三要素的不同排列组合,从而形成适用于不同场景的研究问题.为了方便对 OSN 谣言检测问题进行形式化表达,本文在表 2 中给出了文中所涉及的通用实体及其描述说明.

**表 2 符号说明**

| 表达 | 描述 |
|---|---|
| $G=(V,E)$ | $G$ 指 OSN,$V$ 指 OSN 中的节点(用户),$E$ 指 OSN 中的边(关系) |
| $S$ | 在线社会网络中的信息流 |
| $e_i \in E\{e_1,e_2,\cdots,e_n\}$ | OSN 中的流行话题事件集合 |
| $d_i \in D\{d_1,d_2,\cdots,d_n\}$ | OSN 中的单条文本信息集合 |
| $u_i \in U\{u_1,u_2,\cdots,u_n\}$ | OSN 中的用户集合 |
| $f_i \in F\{f_1,f_2,\cdots,f_n\}$ | 待检测对象的特征集合 |
| $t_i \in T\{t_1,t_2,\cdots,t_n\}$ | 传播时间序列 |
| $r_i \in R\{r_1,r_2,\cdots,r_n\}$ | 检测输出的谣言结果集合,其中 $r \in \{R, NR\}$,$R$ 为 Rumor,$NR$ 为 No Rumor |

### 3.1 基于目标属性的问题定义

OSN 谣言检测目标的选择取决于研究者对谣言的定义以及其检测的目的与需求.目前谣言的检测目标大致可分为三大类:对谣言的检测、对错误谣言的检测以及对虚假信息的检测.

(1)OSN 谣言检测问题

第一类研究目标也是大部分研究者的选择,他们认为谣言虽然无统一定义,但其核心特征为"未经证实性"(结果为真、假或仍无法确定).因此,他们的检测目标即为那些事实尚未确定的特殊阐述.此时,谣言被定义为真实性未经确定的信息,并将未经官方新闻证实过的信息视为待检测谣言的候选项[43].

(2)OSN 错误谣言检测问题

第二类研究目标以文献[44-47]等为代表,认为与被传播成为谣言的真实信息相比,那些错误的谣言信息危害更大.此时,研究者多采用"false rumor"、"misinformation"和"disinformation"等词与"rumor"一词区分开来,从而特指那些最后被证实为虚假的谣言信息.如文献[48]将错误的谣言信息与新闻信息作为一对概念,他们认为来自官方渠道的新闻往往都是可信的,以此通过二者对比来检测错误谣言信息.

(3)OSN 虚假信息检测问题

第三类研究目标则以文献[49-51]等为代表.他们并未将谣言按照其最终证实的结果进行分类,而是将谣言一词等同于那些错误虚假的信息.同时并不区分"rumor"和"misinformation"、"disinformation"等词存在的差异.该类研究直接将谣言检测转化为虚假信息、错误信息的识别问题或信息可信度排序问题[52].例如,文献[53]直接将与官方新闻不匹配的信息视为谣言,即虚假信息.

综上所述,对研究目标的确定主要存在以下几个问题:第一,对谣言的检测首先需要厘清谣言的概念与本质,严格区分谣言与虚假信息的区别,详细分析谣言检测与虚假信息识别、信息可信度识别的差异.第二,虽然谣言中那些错误信息破坏性较大,但是对于那些最终正确和仍无法证实真伪性的信息,它们的危害性仍不容小觑.辟谣是为了探求真相,以防非法信息扰乱公众视野,从而构建良好的信息生态环境.第三,以上三种分类中,均有研究者将谣言/错误谣言/虚假信息和新闻作为一对概念进行比较.然而当前网络媒体的公信力在逐渐削弱,为博得大众眼球,新闻媒体常常缩短消息证实环节,不实报道随处可见.因此,官方新闻的可靠性同样是值得

怀疑的.

### 3.2　基于对象属性的问题定义

谣言是一则人们感兴趣或觉得重要的，且事实尚未确定的陈述或阐述.它通常产生于一个或多个谣言制造者，并在人群中广泛流传最后形成关于该陈述或阐述的谣言信息群.可见，OSN 谣言的检测对象从检测粒度上可分为基于单文本信息的细粒度检测对象和基于多文本信息的粗粒度检测对象，同时还包括信息源角度的谣言源检测对象.

(1) 单文本信息 OSN 谣言检测问题

单文本信息的 OSN 谣言检测类似 OSN 虚假信息或垃圾信息检测问题，均以单则网络消息为检测对象.其检测目标简单明确，同时特征较易抽取.然而，这类问题的弊端也显而易见：首先在 OSN 中，一则消息所包含的信息量是极为有限的.因此在检测模型的设计中仍需考虑文本集群中有价值的信号特征，如整体传播特征、网络结构特征等，以此来弥补单一文本中有限的信号特征.此外，针对单则消息的检测通常来讲并不符合真实场景的需求.辟谣工作的对象一般是某一话题陈述中的断言(Assertion)，即围绕某一事件或人物的判断性陈述语句.用户所关心的也是该陈述的断言而非单则消息.一旦确定为谣言，与这一陈述相符的信息均将被用户视为谣言，不予采信.

(2) 多文本信息 OSN 谣言检测问题

基于多文本信息的 OSN 谣言检测将对象看成一个包含了多条文本信息的话题或事件[2].通常信息源用户发表一则陈述信息后，相关用户进行了一系列的转发或点评，而以上文本信息集合均是围绕该原始陈述信息的.OSN 谣言检测模型通过对传输文本信息之间所包含的内容特征、用户特征、网络特征以及传播特征等进行特征抽取后，建立特征向量，从而进行分类识别.Sun 等人[54]首次提出了基于事件的谣言检测问题，并基于内容、基于用户和基于多媒体三个维度共提取 15 个特征，将基于事件的谣言检测转化为一个分类问题.

基于多文本信息的 OSN 谣言检测在具体应用中一般可分为两种：针对普通事件的日常谣言以及针对突发事件的特殊谣言.通常而言，突发事件产生后(如地震，海啸等)，会围绕该事件产生关于不同话题的谣言，因此针对突发事件的多文本信息 OSN 谣言检测多将突发事件作为一个整体，然后识别出该事件中的各类谣言.但该类检测往往依赖于事件的话题信息，不具备检测的普适性.

(3) 谣言源检测问题

谣言源检测是指通过将 OSN 谣言传播网络抽象为某一信息传播模型(如传染病传播模型 SI、SIS、SIR)，然后根据谣言传播特征找出谣言传播的原始节点，即谣言源.一般认为，谣言总是从小部分群体处产生的，因而对谣言源的检测又分为单谣言源检测和多谣言源检测.在具体研究中，为了简化检测问题通常仅考虑单谣言源检测问题.

综上所述，基于对象属性问题的选择，本文认为谣言是关于事件的陈述，应基于多文本信息的特征来检测某陈述是否为谣言信息.虽然单文本信息检测将谣言检测简化为单文本信息可信度检测问题，且目前研究较为成熟，但这仅仅只能作为谣言检测的基础，而非等价问题.

### 3.3　基于时间属性的问题定义

按照 OSN 谣言检测的时间属性划分，可分为回溯性 OSN 谣言检测(Retrospective Rumor Detection)、早期 OSN 检测谣言(Early Rumor Detection)和实时 OSN 谣言检测(Real-time Rumor Detection)三类核心问题.

(1) 回溯性 OSN 谣言检测问题

回溯性 OSN 谣言检测主要针对 OSN 上的旧谣言或在谣言的爆发期及其后期对历史的、静态的谣言数据集进行检测，即模型的训练和测试都是基于谣言爆发期或消停期之前的历史数据，这也是目前研究最为广泛的模型.回溯性检测的优势在于可训练的数据集较多，从而使得可提取的信号特征也更多，尤其是传播特征的数量(如转发数、点评数等).但回溯性检测方法并不适用于真实的 OSN 场景，当突发事件爆发时，急需尽早地识别谣言并抑制其传播，以尽量减轻谣言所带来的负面影响.

(2) 早期 OSN 谣言检测问题

早期 OSN 谣言检测是指在谣言尚处于潜伏期时对其进行检测.早期 OSN 谣言检测一般基于事件层次，即多文本的信息检测.随着时间的推移，该事件阐述所包含的文本信息不断增加，因此在检测时该事件阐述至少包含一条文本信息；同时早期检测旨在尽量缩短关于该事件阐述的第一条文本信息与检测时刻之间的时间差.如文献[2]中，假设对于一系列新闻事件 $E=\{e_1,e_2,\cdots,e_n\}$，每个事件 $e_i$ 相应地包含了文本集 $D_i=\{d_{i,1},d_{i,2},\cdots,d_{i,m_i}\}$，随着时间的推移，某一事件所包含的文本集 $m_i$ 会不断增多.对此，将某事件在时间 $t$ 的文本集定义为 $D(t)=\{(d_i,t_i),\cdots,(d_m,t_m)\}$，其中 $t_i<t_{i+1},\cdots,t_m\leq t$.因

此,早期谣言检测矩阵函数 $\boldsymbol{M}(t)$将某一事件文本集 $D(t)$特征矩阵定义为一个在实数域 $\mathbb{R}^{m\times f}\rightarrow\mathbb{R}^{f}\rightarrow\{1,0\}$上,在时间 $t$ 的 $f$ 维的矩阵向量,最后映射为一个在时间 $t$ 上的可信度二分类问题:即判别为真实(0)或者谣言(1).

(3)实时 OSN 谣言检测问题

实时检测即基于谣言实时数据流的检测,通常针对单文本信息,即信息一经发布后就开始检测该信息是否为谣言或谣言候选项.如文献[43]认为文档一旦发表后未来就有可能成为谣言,将信息流 $S$ 中在 $t$ 时刻发布的文档 $d$ 记为 $S\{d_0,d_1,d_2,\cdots,d_n\}$,$d_t$的相关特征向量记为 $\boldsymbol{f}_{d,t}$,权重为 $w$,谣言记为 $RS_{d,t}=\boldsymbol{w}^{\mathrm{T}}\times\boldsymbol{f}_{d,t}$,谣言的预测则基于一个固定的阈值 $\theta$,若 $RS_{d,t}>0$,则消息 $d_t$为谣言,其中参数 $w$ 与 $\theta$ 的最优值由训练集学习得到.此外,也有针对事件的实时检测,实质上是将基于多文本信息的早期 OSN 谣言检测与基于单文本信息的实时 OSN 谣言检测结合.例如,在文献[55]中,将早期 OSN 谣言检测问题形式化表示如下:假设 OSN 中存在信息流输出为 $D=\{(d_1,t_1),(d_2,t_2),\cdots\}$,其中 $d_i$为 $t_i$时刻发布的文本信息,$i\in\{1,2,\cdots\}$.实时检测的结果即为自 $t$ 时刻起,每隔 $\Delta t$ 输出结果序列 $R_t=\{R_{t,1},R_{t,2},\cdots,R_{t,l}\}$,其中 $R_{t,j}$为每个聚类陈述主题 $s_{t,j}$的检测结果,且 $\Delta t$ 内至少产生了关于该阐述 $s$ 的一条信息.

综上所述,对于谣言检测时间属性特征可以概括为以下三点:第一,回溯性检测、早期检测和实时检测的根本区别为是否引入时间参数 $t$,由于谣言具有较强的时效性,因此早期检测与实时检测更符合现实需求,能有效地降低谣言可能产生的不良影响.第二,早期检测、实时检测相对于回溯性检测而言可利用的谣言信号特征较少,因而挑战性较大.第三,就谣言检测的准确度而言,一般很难实现 100% 的准确检测.相较之下,回溯性检测方法由于可获取信号特征最多,检测准确度最好,一般可达到 90% 左右[56-57];其次是早期检测,再次是实时检测.一般来讲,谣言检测准确率和谣言检测时间属性之间存在直接的关系,检测时间越早,检测效用越大(即尽早地抑制谣言可能产生的危害),但检测准确率相对不高.反之,检测时间越延后,检测效用相对越小,但检测准确率不断提升.因此,综上可知,尽量缩短检测时延的早期检测能较好的平衡检测准确率与时间延迟之间的舍取,且此时谣言产生的不良影响还未放大,因而更适合 OSN 谣言检测的时间属性选择.

# 4 在线社会网络谣言数据采集与标注

## 4.1 OSN 谣言数据集采集方法

数据收集和注释在谣言检测中非常重要,直接影响到后期检测模型的精确度.数据采集的目标需要确定数据的采集平台、采集内容、采集数量和比例以及关键的数据采集方法等几方面的内容.

对于数据采集平台而言,国外研究基本集中在 Twitter 上,而 Yang 等人[49]首次在新浪微博平台上提出了谣言分析与检测的问题后,作为中国最大社交媒体平台的新浪微博也开始成为国内外研究人员的重点研究对象.与 Twitter 相比,新浪微博对用户的隐私保护不够完善,但这也使得研究人员可以获取更多微博用户的历史数据和实时动态数据,进一步增强模型预测的精准性.

对于数据话题的选择,针对 Twitter 的相关研究,以文献[49]为代表的研究者通常利用 Snopes 网站(国外专门供用户核查并揭穿谣言和传闻的网站)公布的谣言话题作为"种子"构造检索式在 Twitter 上收集谣言信息.而国内有关新浪微博的研究,则利用"@微博辟谣"、"@谣言粉碎机"等官方辟谣账号发布的谣言话题作为"种子"搜集相关谣言数据.与此同时,其对应的真实信息则通常以新华社等官方新闻门户的信息为来源作为对比.

在数据采集的数量和比例方面:数据量目标的确定需要明确谣言声明(即谣言话题)的数量,谣言消息(即谣言微博)的数量以及谣言与非谣言之间的比例.基于单文本谣言信息检测的研究通常仅选取一到两个谣言话题进行研究,如关于地震、台风等自然灾害谣言检测研究.对于多文本谣言信息检测的研究,研究者们通常搜集几十到上百不等的谣言与非谣言话题,进而收集一定数据量的谣言消息.在谣言与非谣言消息数量的比例上,以文献[43,46,49]为代表的大多数研究都采取了 1∶1 的分布策略,但是这实际上是不符合真实情况的,毕竟在 OSN 中,正常信息远大于谣言信息.对此文献[58]通过微博官方辟谣账号获取 400 条谣言微博,然后通过新浪官方热门话题推荐服务获取了 3600 条真实信息微博,大致以 1∶9 分布.

在 OSN 谣言数据的具体采集方法上,大多数研究者都是首先确定感兴趣的谣言,然后通过使用与谣言微博相关的关键字进行过滤收集.其中直接调用微博 API 接口是较为常见的方法.如 Qazvinian 等

人[59]使用 Twitter 的搜索 API 和正则表达式查询收集 2009 年到 2010 年期间的数据.然而,这种基于特定关键词检索方式的弊端也显而易见.

第一,该方法是基于先验的.例如,对于在 2014 年马航失联中产生的微博,研究者可以通过搜索“马航失联”这一关键词来收集与谣言相关的微博.但是这要求所讨论的谣言是先验已知的,并且将无法搜集到之前没有定义特定关键字的相关谣言.为了克服这一问题,文献[60]引入了动态时间序列来增强基于事件的谣言收集的召回率.作者提出首先邀请新闻专家给出当前可能产生谣言的相关事件,然后利用相关标签和关键词实时收集与事件相关的所有数据,并通过转发推文的追溯找到更早期发布的消息源.接着将以上数据按照时间轴排序并请专家标注,该方法能有效地克服基于特定关键词的谣言信息检索所产生的数据不全的问题.

第二,该方法无法获取微博用户之间的对话信息以及用户互动中上下文之间的关联.为了解决这个问题,文献[61]提出了自动获取对话结构树的数据收集方法.由于 Snopes 网站有提供专业的谣言库和验证服务.当 Twitter 用户不确定某一信息的真实性时,会在 Snopes 上恳求其他用户验证他们的阐述或提供与该阐述相关的已知事实.该文献基于 Snopes 的提问功能,利用 snopes,#snopes,snopes.com 和 @snopes 等关键词从 Twitter 信息流中收集谣言对话.利用新浪 API 的搜索功能,将每一个 Snopes 提问都作为一则种子微博来重建对话结构.类似的,文献[62]则针对微博中的转发和评论文本设计了如下的四步数据获取策略:(1)通过以 Twitter 服务器可以接受的最短时间间隔(一小时)发送请求,获取包含该时间间隔的所有推文;(2)由于社交活动通常以丰富的关键词作为主题,因此通过发送多个关键词获得完整的数据;(3)通过设置搜索范围,得到 $U$、$V$、$W$ 三个不同集合,其中 $U$ 代表所有的原始和转发推文集合;$V$ 代表原始推文集合,是 $U$ 的子集;$W$ 是由经过认证的 VIP 用户发出的原始推文集合,且是 $V$ 的子集;(4)基于以上策略,通过 Twitter API 同时获得每条推文的转发和评论.由上可知,谣言数据的获取不仅仅包括浅显的微博文本内容信息和基本用户信息,还可从用户的转发文本、评论文本、评论中的对话文本等渠道收集具有较高价值的特征信息.

### 4.2　谣言数据集标注方法

对于 OSN 谣言数据集的标注,目前的主流方法分为人工标注与机器自动标注两种形式.人工标注法即对搜集来的数据请几个专家按照(0/1 或 −1/1)进行标注,主要依赖于标注员的个人经验和知识.一般而言,研究人员会求助于资深的新闻记者来协助信息标注.但随着网络众包模式的兴起,亚马逊土耳其机器人(Amazon Mechanical Turk,AMT)为谣言标注问题开辟了新的解决途径.AMT 是一个 Web 服务应用程序接口(API),开发商通过它可以将人的智能整合到远程过程调用(RPC).研究人员通过发出标注请求,应用程序就会将这一请求发送给执行任务的自然人,该方法能有效地协调使用人类智慧来执行计算机无法做到的任务.对于人工标注结果,一般采用 Kappa 系数对标注结果进行一致性检验,如式(1):

$$k=\frac{\Pr(a)-\Pr(e)}{1-\Pr(e)} \tag{1}$$

式中,$\Pr(a)$为实际一致率,$\Pr(e)$为理论一致率.

机器标注法则通常利用外部信息资源(如官方网站发布的新闻)作为参考依据进行标注.如文献[63]通过两个官方微博辟谣账号发布的辟谣信息获取谣言源微博后,从爬取的相关微博中筛选出社会新闻微博.接着挑选出未标记的微博,并通过 TF-IDF 提取微博关键词用于构造搜索关键词,然后在新浪和搜狐中搜索该关键词,最后利用式(2)计算未标记微博和返回结果之间的相似性,即基于新闻是可信的这一假设,来进行对比标注,同时这类方法仍然会采用人工标记进行辅助,并使用 Kappa 系数进行一致性检验.

$$sim(m_i,m_j)=\frac{m_i.keywords\cap n_j.keywords}{m_i.keywords\cup n_j.keywords} \tag{2}$$

对于标注的结果形式,除了常见的 0/1 二分类标注法(根据研究检测目标的不同,见 3.1 节,0 和 1 表示的含义也有一定的差异,通常 0 表示非谣言/非虚假谣言/真实信息,1 表示谣言/虚假谣言/虚假信息),还有将谣言进一步细分的多分类标注方法,如文献[59]中,谣言数据被标注为三类:0(微博与谣言无关)、11(微博赞成谣言)、12(微博否认或质疑谣言,或保持中立态度).

### 4.3　公开数据资源

上节主要对 OSN 谣言数据的搜集和标注方式进行了分析和论述,本节将介绍可用于 OSN 谣言检测研究的公开数据资源.当前,可用于 OSN 谣言检测研究的公开数据资源还比较少.同时,研究平台主要集中于 Twitter 和新浪微博.依据研究对象的不同,分为基于消息和基于事件的两种类型.如表 3

所示，为当前 OSN 谣言常用数据集. 基于消息的数据集仅针对单文本数据进行收集并标注(谣言或非谣言)；基于事件的数据集则以话题所包含的多文本进行收集，并针对话题进行标注.

**表 3 公开 OSN 谣言数据集**

| 类型 | 平台 | 微博数 | 话题/消息数量(R/NR) | 用户 | 数据链接 |
|---|---|---|---|---|---|
| 基于消息 | Twitter | 7507 | 2695/4812 | 7507 | https://ndownloader.figshare.com/files/4988998 |
| | Twitter | 5802 | 1972/3830 | / | https://www.pheme.eu/software-downloads/ |
| | 新浪微博 | 5137 | 2601/2536 | 5137 | http://adapt.seiee.sjtu.edu.cn/~kzhu/rumor/ |
| 基于事件 | Twitter/新浪微博 | 110.2 万(T)<br>380.5 万(W) | 498/499(T)<br>2313/2351(W) | 49 万(T)<br>274 万(W) | http://alt.qcri.org/~wgao/data/rumdect.zip |
| | Twitter | | 51/60 | | https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FBFGAVZ |
| | 新浪微博 | 11 万 | 50/274 | 8.8 万 | https://www.dropbox.com/sh/9lmy4veobd2oknk/AABEcn77PRHwKJcNJitm7d0Ma?dl=0 |

基于消息的公开数据集包括 PHEME(欧盟委员会为期 3 年的研究项目基金，Zubiaga 等人[64])公开数据集和 Wu 等人[65]分别针对 Twitter 和新浪微博收集的实验数据，数据量均不大(1 万以下). 基于事件的公开数据集的典型代表是 Ma 等人[66]同时针对 Twitter 和新浪微博收集的大规模数据集，包括 110.2 万推文和 380.5 万微博，其中涉及 498 个 Twitter 谣言话题和 2313 个微博谣言话题以及 49 万 Twitter 用户和 274 万微博用户. 相比之下，Kwon 等人[67]针对 Twitter 和 Jin 等人[68]针对新浪微博收集谣言话题数据集相对较少.

针对新浪微博谣言实验，较知名和有效的公开实验数据集还包括清华大学自然语言处理与社会人文计算实验室孙茂松、刘知远等建立的中文谣言数据库(http://rumor.thunlp.org)[69]. 数据的来源主要包括三个途径：中文社交媒体谣言数据抓取、中文社交媒体谣言自动识别以及用户提交. 下载的谣言数据集时间跨度为 2012 年 5 月到 2016 年 6 月，共计 30038 条谣言信息，目前仍在不断更新.

## 5 基于内容的谣言检测方法

基于内容的谣言检测方法主要是从微观的角度，通过观察谣言传播过程中文本内容、用户内容、传播内容的特征以及随时间变化的趋势来识别谣言. 目前常见基于内容的谣言检测方法，根据处理方式的不同可分为基于内容分类的机器学习方法和基于内容对比的方法. 依据处理内容的不同又可分为基于文本内容特征的识别方法和基于非文本特征(即用户特征、传播特征等)的识别方法；依据时间属性又可分为动态检测方法和非动态检测方法. 下文将从基于内容分类的机器学习方法和基于内容对比的方法两个维度对当前的相关研究文献进行综述.

### 5.1 基于内容分类的机器学习方法

基于内容的分类模型将 OSN 谣言检测看成针对信息流 S 中的单文本或多文本的机器学习分类问题. 在机器学习中，分类问题多采用基于监督或半监督的学习算法，即从已标记的训练数据中训练预测函数的任务. 如图 4 所示，由已标记的 OSN 谣言(即数据集输入对象匹配期望输出的对象)组成训练数据，提取特征向量，然后通过相关算法分析该数据，并推断出可应用于判别标记新输入对象的预测模型函数，最后得出期望的预测结果.

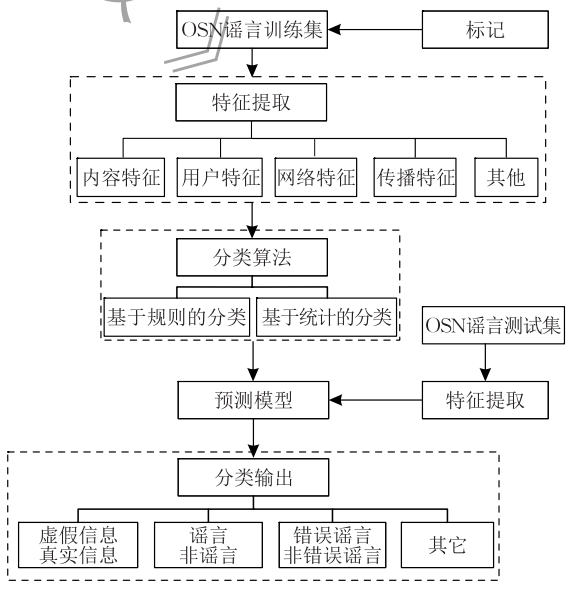


图 4 分类模型预测流程图

在以往研究中，关于 OSN 谣言检测的分类模型的差异主要包括以下几个关键点：输入文本粒度、

数据集特征选取、分类算法选取以及分类输出对象类型.(1)输入文本粒度,即 3.2 小节对检测对象属性问题定义所阐述的内容,分为基于粗粒度的多文本输入对象(基于事件)和基于细粒度的单文本输入对象(基于消息);(2)数据集特征选取,即找出谣言与正常信息的特征差异从而构建分类器输入的特征向量.目前常见的特征向量选择包括内容特征、用户特征、网络特征、传播特征及其相互组合特征等;(3)分类算法的选择,目前常见的分类模型可以分为两大类:基于规则的分类方法(决策树、关联规则和粗糙集等)和基于统计的分类方法(朴素贝叶斯、支持向量机等);(4)分类输出对象,即 3.1 小节关于检测目标属性问题定义所阐述的内容.对二分类问题,输出对象组合包括虚假信息/真实信息,谣言/非谣言,错误谣言/非错误谣言等.此外,还有多分类结果的输出.

在分类模型的各步骤中,分类器的精度和效度很大程度上依赖于输入特征的选择.对此,Castillo 等人[44]在对信息可信度的预测中提出了基于消息的、基于用户的、基于话题的和基于传播的详细特征.Qazvinian 等人[59]提出了基于内容的、基于网络的和基于特定微博的特征来识别谣言.在后续的谣言分类预测研究中,大多数研究者都通过在此特征提取基础上进行舍取和创新,从而提高分类器的准确性.具体而言主要包括文本特征的方法和非文本特征的方法.

5.1.1　文本特征的方法

文本特征是指消息文本内容以语法为主的显性特征、以语义、情感为主的隐性特征以及随着时间变化表现出的新特征.在文献[60,57,70,71]等中,实验结果均表明微博消息或事件所包含微博消息集合的内容特征在分类性能上明显优于其它基于用户、基于网络或基于传播的特征.

(1)以语法为主的显性特征

以语法为主的显性特征分析主要包括消息文本内容的词语特征、符号特征、简单的情感特征[72]、模因[73](Memes)特征等,如表 4 所示.其中词语特征、符号特征以及模因特征主要通过简单的人工统计进行分析(如是否包含表情符号及其个数).而简单的情感特征分析则是比较文本信息中积极情感词汇和消极情感词汇的数量,从而最终确定文本的情感倾向.其中情感词汇集合来源 WordNet[74](英文领域)和 HowNet[75](中文领域).

**表 4　基于语法的显性特征分类**

| 特征类 | 具体内容 |
|---|---|
| 词语特征 | 字符串长度、词性、大小写字母等 |
| 符号特征 | 标点符号特征、表情符号特征等 |
| 简单情感特征 | 积极和消极情感词汇数量等 |
| 模因特征 | 是否包含标签、是否包含链接等 |

基于显性文本特征分类的机器学习检测方法是早期用于 OSN 谣言检测的常见方法,如谣言检测的开创性研究者 Castillo 等人[44]将文本特征细分为字符串长度、词语个数、是否包含标点符号(问号、感叹号等)、是否包含表情符号(微笑和皱眉等)、人称代词(第一人称、第二人称、第三人称)、大写字母个数、发布时间、积极词汇数量、消极词汇数量、是否含有标签、是否含有链接等.此外,Takahash 等人[76]发现谣言消息的词汇分布与非谣言信息有一定的区别,因此将谣言和非谣言信息的词语分布比率作为检测谣言的文本特征之一.Ratkiewicz 等人[73]则利用 Twitter 上政治谣言内容中包含的标签、链接和提问作为文本特征,建立了 Truthy system 来检测谣言.

由以上可知,早期基于文本特征的 OSN 谣言识别主要以浅显的显性特征为主.这类方法简单便捷,但仅适用于数据规模较小,谣言话题特定(如仅针对政治谣言、食品安全谣言等)时的检测.因为,特定类型的谣言话题更易表现出其独特的显性文本特征.但整个分类模型的准确率和普适性仍不令人满意.对此,文本隐性特征的提取成为 OSN 分类研究中的重点研究方向之一.

(2)以语义为主的隐性特征

基于语义为主的隐性特征分析主要针对消息文本在深层次的语义层面进行特征提取或抽象表示,从而获取消息文本的潜在语义特征、情感特征、消息间相互关联特征以及其它文本特征,如表 5 所示.

**表 5　基于语义的隐性特征分析方法**

| 特征类 | 已有方法 |
|---|---|
| 深层次语义特征 | 词袋模型、神经网络等 |
| 多粒度情感特征 | 稀疏相加生成模型、词向量、情感分类器等 |
| 消息间关联特征 | 语义相似性计算 |

① 深层次语义特征分析

对于深层次语义特征分析,现有方法通常忽略了语法和语序特征,通过语义表示学习来抽象表示消息文本,进而获取其深层次的语义特征.常见的方法包括词袋模型[59,77]、神经网络模型[66,78-80]等.

词袋模型的基本思想是假定对于一个文本,忽

略其词序、语法、句法特征，仅仅将其看作是一些词汇的集合(且文本中每个词汇都是独立的)；进而将每个文档表示为向量空间中的一个单词向量，使具有相似内容的文档具有相似的向量[81]. 该词语集合中的单个单词或多个连续单词又可表示为文本的词袋特征，其中一元特征(unigram)、二元特征(bigram)和三元特征(trigram)为常用的词袋特征. 词袋特征在文本的观点挖掘、情感分析等方面是一个极为有效的特征. 如文献[59]将文本特征具体细分为一元词汇特征、二元词汇特征、一元词性特征和二元词性特征，达到了较好的分类预测效果. 文献[82]使用了词袋模型和神经网络语言模型(Word2vec(100)和 Word2vec(400))两种文本表示模型进行对比测试，在对新浪微博 10000 个帖子的测试中发现，词袋模型的最佳分类精确度超过超过 90%，神经网络语言模型的最佳精确度达到了 60%以上.

神经网络曾在机器学习领域名声大噪，但由于其容易过拟合且参数训练速度较慢，且传统的人工神经网络是一个浅层的结构，和人们预期的人工智能相距甚远，故而逐渐淡出人们视野[83]. 然而，随着计算机处理速度和存储能力的提高，深层神经网络的实现逐渐成为可能，如循环神经网络(RNN)/卷积神经网络(CNN)等不同形式. 深层神经网络的提出是对传统特征选择和提取模式的突破，因而常用于自然语言处理中文本的语义特征抽象. 如文献[66]是首次采用深度学习模型检测微博谣言的研究. 基于循环神经网络(RNN)的模型相较于基于手动爬取特征的先机学习算法有较大的改进，通过三个广泛使用的递归单元：tanh 函数、长短期记忆(LSTM)和门限递归单元(GRU)，学习消息间的隐层表达，比现有的检测方法达到了更高的准确度. Chen 等人[78]在此基础上提出了一种基于保护机制的新型循环神经网络模型，通过从时序文本中获取潜在语境变化特征来实现早期的谣言检测. Li 等[84]则以词向量作为输入，运用卷积神经网络模型(CNN)对评论文本进行语义特征抽象，从而用于虚假评论的识别研究. 深度神经网络建模方法采用连续化向量表示文本，克服了特征稀疏的问题，且在一定程度上模拟了人脑的思维模式，准确率往往较高；但是该模型参数较多，学习收敛较慢且需要大量的语料库进行计算.

② 情感分析

情感分析也隶属于文本语义分析的一种方法，即从情感极性的角度分析消息文本的语义特征. 由于谣言在传播过程中很大程度上受心理和情绪因素的驱动，且相比正面情绪，煽动负面情绪更易加速谣言传播，因此文本的情感特征也作为重要特征之一被许多研究者纳入谣言检测的文本特征进行研究. 在自然语言处理中，情感分析也可称为观点挖掘，依据挖掘的粒度一般包括篇章层级、句子层级和词语层级的情感分析. 在具体研究方法上，相似性计算、机器学习等方法仍是主流，如文献[85]利用词向量计算文本词汇与情感词典词汇的相似度，文献[86]基于机器学习的情感分类器计算谣言文本的情感倾向.

在分析对象上，相比微博主体内容的情感特征，OSN 群体的转发内容和评论内容的文本情感特征也颇受关注[86-87]. 如文献[86]在遵循前人将检测任务作为分类问题求解的框架基础之上，重点关注微博评论中的情感反馈，提出将评论的总体情感正负倾向(不考虑强度，仅分为正向、中立和负向)作为一项新的特征，用于 OSN 谣言检测的分类任务中. 具体而言，作者基于谣言微博的评论总体上比普通微博更倾向于负面情感的假设，提出基于词频特征的分类器对单条评论进行情感倾向的有效识别，从而得到总体情感倾向特征值. 最后基于微博语料库进行实验验证，表明所提出的新特征在现有特征基础上对分类结果有可观的提升.

③ 隐性关联分析

对于微博信息之间的隐性关联研究则主要指消息与消息之间的语义关联、情感关联、情境关联等分析方法. 对于语义关联分析，语义相似性计算是较为常见的方法，如文献[88]假设可观察到的和丢失词语的语义空间能组成句子完整的语义特征，然后使用由 Guo 等人[89]提出的语义文本相似性(Semantic Textual Similarity，STS)模型提取句子的潜在语义，从而创建每条微博的潜在向量表示(Tweet Latent Vector，TLV). STS 是通过标记和补充来对每个短文本进行预处理，然后删除不常用的单词，并采用 TF-IDF 赋权，最后提取潜在语义，并形成 100 维的潜在语义特征向量. 文献[90]提取了热门话题的相关性、内部和外部的一致性(即消息内容和其 URL 链接的话题内容的语义相似性，越相似，则越不可能是谣言)、情感极性以及用户评论中的接受程度四个隐性特征. 并将基于以上隐性特征构建的分类模型和基于显性特征(是否有 URL，是否有多媒体信息，是否有@其它用户特征等)构建的分类模型相比，准确性提高了 10.5%，召回率提高了 4.7%.

对于情感特征关联分析以及情景特征关联分析.简单的分析方法包括借助外部辅助信息,如标签或链接关系.文献[61]提出了内容中的两个隐性链接(标签链接和 Web 链接)来分析具有有潜在关联的文本信息,即同时含有同一标签链接或 Web 链接的微博文本具有一定的相似性.更为深层次的挖掘方法则采用对谣言事件传播网络建模的方式,刻画各个话题谣言以及各个消息层级以及层内的潜在关联.

文献[91]以如图 5 的虚假新闻实例构建了一个如图 6 的基于三层次的可信度网络传播模型:消息层(Message Layer),子事件层(Sub-event Layer)和事件层(Event-Layer).该三层可信度网络又形成了四种类型的网络链路以反映网络节点之间的隐性关系,即图 6 虚线表示的层内链接(消息到消息,子事件到子事件)和实线表示的层间链接(消息到子事件,子事件到事件).

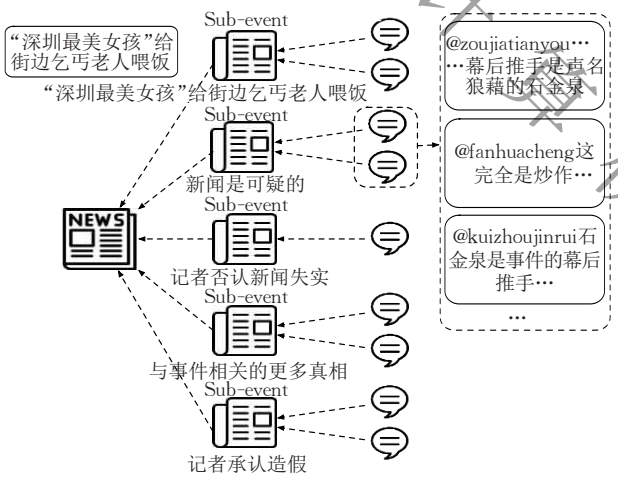


图 5　一个基于三层可信网络实例,以 2013 年虚假新闻"'深圳最美女孩'给街边乞丐老人喂饭"为例

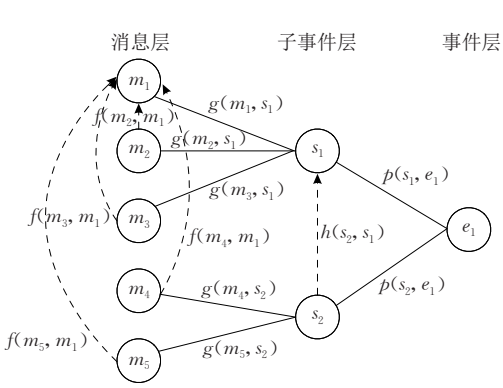


图 6　基于三层的分层可信网络抽象,$m$ 代表消息,$s$ 代表子事件,$e$ 代表事件,为了简化抽象图省略了部分链接

其中层内链接反映同一层级内实体之间的关系,而层间链接则反映了层级与层级之间的关系.对于给定新闻事件及其相关微博,通过聚类算法生成子事件,并利用子事件层捕获链路间的隐性语义信息.在实验过程中,该文分别基于内容特征、基于传播特征和基于用户特征生成了三个独立的模型 $S_{\text{content}}$、$S_{\text{propa}}$、$S_{\text{user}}$,同时还采用逻辑回归来综合以上三个独立模型得到一个综合模型,最后构建可视化的结果.实验结果表明 $S_{\text{content}}$ 优于其它两个独立模型,且综合模型性能最好.

综上所述,对消息文本语义特征分析通过挖掘文本的深层次语义特征、情感特征以及关联特征,在预测的准确率上相比基于语法的显性特征提取有较大的提升.以神经网络模型为代表,通过向量的形式对文本进行语义抽象和特征提取,是未来基于深度学习的谣言检测的新兴研究方法.但目前对文本内容的研究,多以消息文本本身的内容为主,对消息的评论文本、转发文本的语义分析以及关联分析不足,这也是未来需要进一步完善的地方.

5.1.2　多媒体信息特征方法

在 OSN 谣言传播过程中,多媒体内容(如图片、音频和视频等等)往往比单独的文字信息更容易引人注意,具有更强的感染力和说服力,Sun 等人[54]在研究中指出 80%的事件谣言都包含有图片.由此可见在 OSN 谣言检测中多媒体内容特征是对文本内容特征的有效补充.然而,在当前已有研究中,对多媒体信息内容特征挖掘深度不足,大多数仅将有无(0/1)多媒体作为文本特征的补充[85,86,92],而针对多媒体信息本身作为谣言媒介的检测研究则较少(尤其是针对视频或音频信息的深度分析),例如,仅有少量针对图片特征的提取分析[54,93],且均是利用图片的内容标签或外部知识(如搜索引擎)来印证图片内容,而未挖掘图片本身的内容特征.如 Sun 等人[54]主要研究了图文不符的谣言类型,并指出 80%的事件谣言都包含有图片,且需要图片是有限的,其中 86%的虚假图片都是通过转发得到的.在具体研究中,他们将谣言划分为四种类型:虚构消息、过时消息、篡改消息、图文不符消息,并利用多媒体特征中的图片特征来检测图文不匹配类的谣言.具体实现过程中,作者将微博的图片提交到外部搜索引擎来检索图片出处,如无返回结果则表示该微博图文相符,否则将根据网站的可信度和原始图片的发布时间对记录结果降序排序,然后爬取排名靠前的网站内容,并采用 Jaccard 系数计算微博内容与爬取网站内容的相似性,如果相似则证明该微博为非谣言,否则即为图文不匹配谣言.Gupta 等

人[93]则是直接对 Sandy 飓风期间 Twitter 上传播的虚假图片进行检测,但该检测方法仍是通过分别提取消息的文本特征和用户特征,然后使用决策树分类器进行对比检测,研究结果发现内容特征在分类预测虚假图片推文的效果明显优于用户特征.进一步地,文献[94]提出了从图片的视觉特征和统计特征两个角度识别图片类虚假新闻.其中,视觉特征包括图像的清晰度、相似性、一致性、多样性、聚类数五个特征;统计特征包括事件新闻中图像个数、包含图像与不包含图像微博的比率、图像数量与推文数量比例等七个特征.实验表明,与常见的基准特征相比,图像特征在各类分类器上都获得了更好的检测效果.

由以上分析可知,多媒体内容已逐渐成为谣言传播过程中不可获取的一部分,然而当前基于多媒体特征检测谣言的研究却十分匮乏,且主要体现在两方面:首先,当前研究虽然考虑到了图片外部标签所蕴含的丰富信息,但真正利用的有效信息严重不足.事实上,一张用户上传的图片是包含有大量元数据信息的:照片标题、ID、拍摄时间、拍摄地点以及拍摄者的相关个人信息等[95].其次,当前研究中基于多媒体内容的深度挖掘较为罕见,即未使用相关的多媒体内容处理技术识别其内在语义特征.因此,在未来研究中,需要进一步充分地将多媒体标签内容以及多媒体语义挖掘技术应用到谣言检测研究中.

5.1.3 用户行为特征方法

基于用户行为特征的 OSN 谣言检测方法一般包括信息发布者行为特征分析、信息传递者行为特征分析、信息接收者行为特征分析以及各个用户之间的交互行为特征分析四种类型.

对于信息发布者行为特征分析,通常以搜集用户近段时间内的发文量、转发量、关注数量、粉丝数量以及发文异常模式等特征作为判别依据,这也是已有研究中最常关注的用户行为特征:即谣言发布者的行为特征.如文献[96]通过识别极端用户来检测 Twitter 上的政治谣言,主要包括用户介绍是否包含极端关键词、用户最新发表的 100 篇推文中是否包含极端关键词以及用户粉丝等其它常见用户特征.该方法的局限在于人工设置参数的依赖性较强.实验表明:不同数据集、不同类型谣言的最佳规则是不同的.因此对于给定的新数据集可能无法得知哪些规则是最好的.其次,使用极端关键词识别极端用户在该文的政治谣言数据检测中效果较好,但在其它类型谣言中极端用户不一定使用极端关键词.

对于信息传递者行为特征分析,一般以转发用户的行为特定为依据,如转发用户的属性特征(粉丝量、关注数等)、转发用户的异常模式等.通常基于传递者的行为特征分析也被称为基于传播特征分析,即分析基于转发路径形成的传播结构.文献[46]通过对谣言传播规律的分析,明确指出了谣言和非谣言在传播过程中,转发者模式的区别:谣言通常由普通用户发布,然后被意见领袖转发,最后再被大量普通用户转发.而非谣言则通常由意见领袖发布,然后直接被普通用户转发.在具体研究中,作者将转发者行为特征和信息发布者行为特征、消息内容特征相结合,利用混合 SVM 分类器来预测谣言.

对于信息接收者行为特征分析,一般以用户的评论行为为切入点来提取重要特征进行判别.如文献[62]聚焦于紧急情况下用户的回应行为来识别新浪微博谣言.通过提取用户的转发和评论文本特征,然后进行聚类分析来验证特征集选择的有效性,最后采用机器学习分类技术实现谣言检测.实验发现,在过去被视为噪声的停用词、标点符号以及一些表达人群响应情感的词语或符号在检测中发挥了重要作用.

对于用户之间的交互行为特征分析.事实上,更多的用户行为特征分析是提取用户行为模式和交互模式的综合特征.如文献[97]依据微博发布者和微博阅读者双方用户的特征建立了谣言检测分类器.提出了区别于以往研究的五个新特征,具体包括基于微博发布者的 3 个行为特征:平均每天关注数量(总关注数量/注册天数)、平均每天发文量、发布与特定微博相似的人群数量(文章假设谣言总是从少数甚至一个人传出来的,而正常信息则可能有多个消息源),和基于微博接收的 2 个行为特征:质疑评论数量占比、纠正评论数量(Shirai 等人[98]指出 14.7% 的人或机构在发现谣言后会及时的发布辟谣信息),实验结果表明新选取特征在谣言检测中效果显著.

谣言的传播是一个双向的过程,在整个谣言信息传播生态链上存在:谣言生产者、谣言传递者(转发者)、谣言接受者、谣言分解者(辟谣者)等不同用户.然而在已有研究中主要关注点在谣言生产者行为特征上,而事实上,谣言传递者行为、谣言接受者行为都蕴含有大量有效的潜在信息用于谣言检测.如上文提到的谣言转发者行为特点、谣言接受者的回应行为特点以及二者的交互行为特点等.因此,在未来研究中,对谣言传播中用户行为的研究需要进

一步综合生产者、传递者和接受者三者的行为特点及其交互特点.

5.1.4　时间属性特征的方法

随着时间的推移，谣言的传播模式以及有效的检测信号特征都可能发生变化[99]. 而已有的研究大多数都是基于一个任意设置的单个观察窗口，从一个固定观察点获得的结果往往难以表示一般的谣言传播模式. 对此，文献[67]将用户，结构，语言和时间特征全面结合，分析了从谣言传播的前三天到近两个月，在不同时间窗口上，谣言分类检测的表现水平. 统计分析发现，结构和时间特征在长期窗口中区分效果较好，但在初始传播阶段不是很理想. 相比之下，用户和语言特征在传播初始阶段容易获得，也是谣言初始传播阶段检测的良好指标. 由此可见，时间属性特征对谣言检测的重要性. 事实上，动态预测是当前 OSN 谣言检测面临的重要挑战之一，具体到实际研究问题中，OSN 谣言检测的动态性主要有两大途径：① 对实时的流数据进行检测，② 对时变信息、用户或网络结构模型进行检测. 针对途径①，苛刻的来看，真正意义上基于动态流数据的谣言检测较少采用基于内容分类的机器学习方法，因为该方法需要一定数量的初始数据进行训练方能启动测试. 对此，文献[2]通过加入信息源可靠性和用户态度两大类新特征，分别对获取不同数量实时数据(5 条、100 条和 400 条)以及不同时间延迟后(1 小时后，12 小时后，72 小时后)的分类准确率进行对比，研究发现，随着时间的推移，可获得的初始数据越多，分类的准确率越高.

针对途径②，Kwon 等人[100]首次指出了谣言传播过程中时间属性的重要性，并提出了推文数量随时间变化的时间序列拟合模型，实验显示能获取更好的检测效果. 此后，Ma 等人[101]在 Kwon 等人研究的基础上进一步扩展了随时间变化的特征集合，利用动态时间序列来观察社交情景特征随时间的变化，具体包括随时间变化的文本特征、用户特征和传播特征，最后利用基于动态时间序列的 SVM 分类器分别在 Twitter 数据集和新浪微博数据集上获得了较好的检测结果.

由于谣言的检测往往需要依赖于一定的话题、网络模型等外在因素，真正意义上普适性的模型很难实现. 因此在使用机器学习分类方法时，为有效识别谣言，需要一定量的训练集来训练模型，而这也成为谣言实时检测的弊端之一. 由此可见，基于机器分类的谣言检测用于回溯性检测的效果是最好的，通过加入一定的时间属性改进后的模型也可用于早期检测(如文献[101]使用的基于动态时间序列的 SVM 分类模型)，但仍旧很难实现实时检测. 对实时谣言的检测还需依赖下文讨论的基于内容对比的方法.

5.1.5　地理位置特征的方法

一般的，在谣言传播过程中存在三种类型的地理位置特征：传播网络所在地点、谣言事件发生地点以及信息发布者地点(或信息传播者发布消息的地点). 在全球不同地区的 OSN 网络(如 Twitter 和新浪微博)中，消息的传播特征以及人们的信息行为方式都有一定的差异，因此谣言传播网络空间也会影响检测性能. 对此，文献[102]选取了来自全世界不同地方发生的九个谣言组成的实验数据集，构建了一个与主题无关的谣言分类器，然而分类效果并不理想. 实验显示：谣言识别的大多数特征分布是基于特定环境的，而事件发生的地点(不同国家、不同社会群体、不同文化特征)起到了关键性作用. 因此，在未来谣言检测模型中，应在不同网络环境中、不同类型的数据集上进行测试，以保证模型的现实可用性.

同时，谣言事件发生地的距离也会影响谣言的传播. 如文献[49]提出了基于事件发生位置以及消息发布客户端的两个新非文本特征. 基于事件发生位置的特征是指消息提及事件的发生地点，文中分为国内和国外两种类型. 基于客户端的特征是指用户发布消息时所使用的客户端，分为智能客户端和非智能客户端. 实验发现：谣言信息更倾向于使用网页发布，且谣言信息所涉及事件的地点多在国内. 最后，作者将位置特征和客户端特征结合基于消息、账户和传播的三类特征(共计 19 个)训练 RBF 核函数的 SVM 分类器来识别谣言，取得了较好的效果.

对于信息发布者位置特征，随着 Web2.0 和位置感知技术的发展，社会网络允许人们额外的添加时间戳和地理位置作为标签，使上传的文件或发布的内容更容易根据时空信息进行共享和搜索[103]. 例如，Flickr 上有超过 4000 万张照片、Panoramio 上有超过 400 万张照片都是都是有地理位置标记的，这些可以公开访问的照片不仅包含丰富的地理信息，而且还传达了各种人口背景的人的情感和观点[104]. 这些标签极具个性化，除了常见的地名定位(如北京、长城等)，还可能包含有日期、天气、相机参数，甚至用户个人的心情等. 应用在谣言检测中，信息发布者当前的区域位置或其发布消息的具体位置以及发布者位置与谣言事件位置的距离[2]都是值得

关注的重要信息.对此,文献[105]发现用户发布者的位置信息在政治事件方面对其它用户的可信度感知有较大的影响,但是具体如何影响,文章尚未证明.

地理位置特征为 OSN 谣言检测提供了一个新的有效信号特征,尤其针对突发性自然灾害或社会重大事故产生的谣言,信息发布者位置和事件发生位置成为谣言识别的重要特征.目前对位置信息的挖掘多局限于基于 LBS 技术的外部标签定位信息,事实上,用户上传的图片属性信息也会包含其消息发布的位置信息,但目前尚未有研究充分利用这一有效信息,由此进一步说明多媒体信息对谣言检测的重要性.

### 5.2 基于内容对比的检测方法

基于内容对比的检测方法通过选取一个可信度已知的对象作为单条谣言信息或谣言陈述的比照对象,从而达到对原始信息进行判别的目的.具体对比方法包括与已知的谣言信息进行对比或与已知的正常信息进行对比.

#### 5.2.1 与已知谣言信息对比的方法

与已知谣言信息进行对比的方法需选取与待检测文本话题相关的已知谣言信息,包括根据官方辟谣机构发布的谣言信息建立的谣言库或已知谣言的内容模式或用户行为模式,并根据该模式作为参照对象进行比对.

(1)与已知谣言语料库匹配对比.该方法通常依赖于手工或半自动建立的谣言语料库,需要人员不断地更新与维护[106].如文献[107]通过与 Snopes 上已公布的谣言信息进行匹配,来检测 2016 年美国总统选举期间关于两位候选人的谣言.在具体研究中采用了最常见的几种文本匹配算法进行对比测试:基于 TF-IDF 的文档相似度计算、BM25 算法、深度学习中的词向量模型(基于单词的 Word2Vec 模型和基于段落的 Doc2Vec 模型)以及词典匹配方法,其中前四种通过将文本转化为数值向量进行匹配,最后基于词典匹配的方法真是基于关键词匹配的模式,最后结果显示 BM25 算法效果最优.然而,基于已知谣言文本匹配的方法有两个明显的弊端:覆盖面有限且存在一定的时滞性.一般谣言只有在传播了一段时间且产生了一定影响后才会引起相关辟谣机构的注意,然后进行验证和检测.因此,模式匹配的对比方法更为常见,主要从谣言消息或非谣言消息的消息内容文本模式、消息传播模式以及用户行为模式等方面入手,探究二者之间的差异性,并以此为依据进行对比匹配.

(2)消息文本模式匹配对比.通过分析并提取出谣言文本区别于正常文本的显著特征.研究最多的谣言文本模式是由 Marcelo 等人[50]提出的质疑、警告等关键词特征,作者指出谣言消息比正常消息通常更容易受到人们的质疑,因此对于消息文本中出现的与质疑相关的特征词是检测谣言的重要信号特征.基于以上思想,文献[108]利用基于警告关键词匹配的方式提出了"Dematter"系统来检测错误谣言.该系统的检测过程分为三个阶段:首先,它采用 MeCab(日文形态分析器)从待检测的文本中提取关键词;然后检索包含该关键词的推文并将其存储为待验证推文信息,如果检索结果不够,将通过减少关键词来扩大检索策略直到获取足够的相关推文为止.最后,Dematter 系统通过计算待验证推文信息中含有警告推文的比例来判断被检测文本是否为谣言,如"demagoguery"、"false information"、"gaseous"等,该方法的缺陷在于无法检测到非热门谣言.特别的,当没有足够数量的相关推文时,系统将输出"不可判定".

此外,这种简单的关键词匹配存在语义缺陷,没有考虑同义词等语义相似性问题.对此,可以采用共现、本体或聚类等方式解决语义匹配问题.与文本信号特征模式对比类似,文献[55]将微博信息中包含用户疑问和纠正的信息作为谣言的信号特征,并设计了更为全面的正则表达式来进行匹配(如表 6 所示).具体匹配过程如图 7:对于收集得到的数据,按信号特征分为 signal tweets 和 non-signal tweets 两个集合.通过将 signal tweets 聚类并提取出相关的陈述集合,然后再将之与 non-signal tweets 进行匹配,最后将匹配成功的 non-signal tweets 归入相应的陈述集合,并对这些陈述集合排序作为潜在谣言集合.但是,并非所有存在质疑的信息都是谣言,因此该文并未给出最终的谣言检测结果,只是给出了潜在的谣言集合作为供参考的候选项.该类基于谣言文本重要信号特征(如警告词汇)的检测方法的精确度较高,但是召回率通常较低,因为还存在不少不包含有如警告词汇等文本信号特征的谣言.

**表 6 谣言信号特征正则表达式**

| 正则表达式 | 类型 |
|---|---|
| Is(that\|this\|it)true | 质疑 |
| Wh[a]*t[?!][?1]* | 质疑 |
| (real?\|really?\|unconfirmed) | 质疑 |
| (rumor\|debunk) | 纠正 |
| (that\|this\|it) is not true | 纠正 |

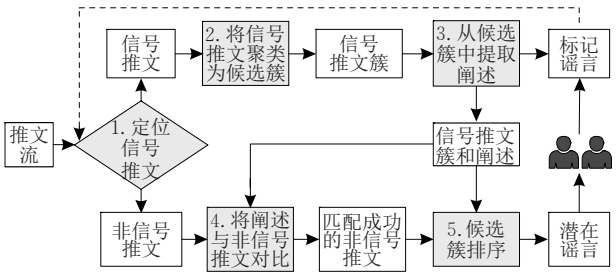


图 7　信号推文与非信号推文匹配方法

(3)消息传播模式匹配对比. 通过抽象谣言信息传播网络和正常信息传播网络进行对比分析,从而识别出谣言. 文献[58]将谣言检测转化为一个谣言信息传播网络与真实信息传播网络匹配对比的问题,并将谣言定义为包含了关于某事件或事实的错误信息的流行微博. 在模型构建中,着重考虑用户的转发行为,并抽象出 spreader 和 stifler 两类用户 spreader 指那些从关注者处收到消息后进行转发的用户,stifler 则指那些从关注者处收到消息后不进行转发的用户. 作者假设用户接受微博信息后转发与否的概率不仅依赖文本特征,还依赖用户特定的属性,如年龄、教育背景等. 基于以上假设,作者针对谣言检测,采用异质用户表征建立了真实信息传播模型和谣言信息传播模型. 真实信息传播模型有两个核心因素:第一,用户转发微博倾向与用户过去三个月转发微博数量及其占比相关. 第二,用户关注人群的类型对消息真实性有影响. 谣言信息传播模型:除了以往研究中常用的通用特征,还采用表 7 的特征描述用户鉴别信息的能力. 在具体实验过程中,对于每一个 spreader 或 stifler 用户,首先解析它的传播网络,然后分别计算其与真实信息传播模型和谣言信息传播模型的相似性 $Pr$ 和 $Pc$,接着比较二者的大小从而判断该信息为谣言还是真实信息.

**表 7　文献[58]用户特征选取**

| 编号 | 特征 | 编号 | 特征 | 编号 | 特征 |
|---|---|---|---|---|---|
| $f_1$ | 年龄 | $f_6$ | 微博等级 | $f_{11}$ | 是否为微博专家 |
| $f_2$ | 注册时间 | $f_7$ | 活跃天数 | $f_{12}$ | 教育背景 |
| $f_3$ | 粉丝数量 | $f_8$ | 信誉得分 | $f_{13}$ | 是否有个人网站 |
| $f_4$ | 关注数量 | $f_9$ | 标签数量 | $f_{14}$ | 工作描述 |
| $f_5$ | 好友数量 | $f_{10}$ | 是否为 VIP | $f_{15}$ | 个人描述 |

(4)用户行为模式匹配对比. 文献[48]将用户态度和传播结构模式结合,针对实时数据流的匹配来检测谣言. 根据文献[109]和文献[110]提出的常见微博拓扑结构特征(如图 8),指出除了基本的两点拓扑结构,T2(S3)和 T4(S4)是最常见的串级结构,并将这两个基本结构作为消息流的子图结构. 同时,作者将用户的态度作为标签赋予传播结构中的每个节点并按照图 9 格式存储. 在研究过程中,输入对象为 $ES=\{e_1,e_1,e_3,\cdots\}$ 和 $PT=\{p_1,p_1,p_3,\cdots\}$. 其中每条边 $e$ 包含两个节点以及传播的时间并定义为 $e=\langle n_{\text{star}},n_{\text{end}},time\rangle$,子图模式 $p$ 分为星型(S3)和线型(S4)两种模式,以图 10 为例的两个子图可定义如下:$p=\{'Star',n_{\text{root}}.label=DENY,\ n_{\text{left}}.label=SUPPORT,n_{\text{right}}.label=QUESTION\}$ 和 $p=\{'Path',n_{\text{root}}.label=SUPPORT,\ n_{\text{up}}.label=SUPPORT,\ n_{\text{down}}.label=DENY\}$. 通过输入和匹配不断更新补充信息传播模式,最后基于 TF-IDF 提取出谣言信息流中最常见的传播模式,从而用于匹配检测谣言.

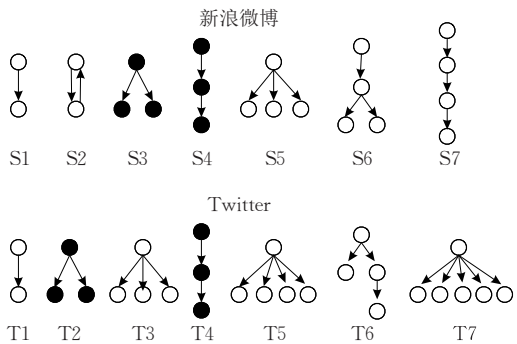


图 8　微博传播的常见拓扑结构

| NodeID | Support_in | Support_out | Deny_in | Deny_out | Question_in | Question_out |
|---|---|---|---|---|---|---|

图 9　节点信息存储格式

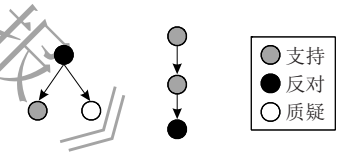


图 10　两个谣言模式的例子

5.2.2　与已知的正常信息对比的方法

与已知的官方信息进行匹配需选取与待检测文本话题相关或包含该话题的已证实信息. 一般而言,研究者会选取官方网站(如新浪新闻、搜狐新闻等网站)的新闻信息作为可信信息,然后通过主题抽取进行匹配检测. 如文献[53]提出假设:Twitter 上经验证的官方新闻账号的信息比未经验证的普通用户的信息更可信. 作者基于这一假设采取了如图 11 所示的匹配流程:首先使用主题标签将实时热点话题下的推文聚类,然后按照推文的来源(经验证的官方新闻账户或未经验证的账户)分为两类,接着采用语义和情感计算来比较包含同一主题,来自不同类型账户推文的匹配情况,最后将不匹配率值高于阈值的话题标注为谣言. 该方法实现过程简单,能有效地识别出潜在谣言,但是精准性识别效果不理想.

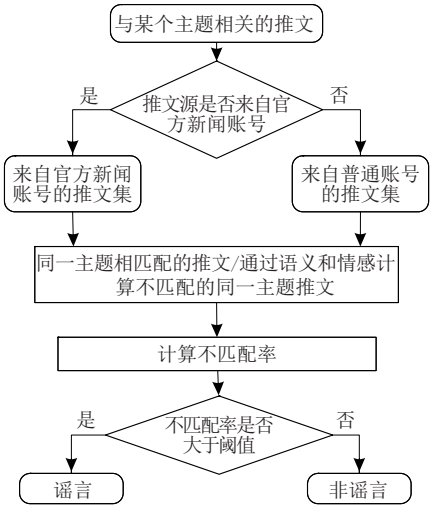


图 11 与已证实消息对比的谣言检测流程图

### 5.3 小 结

综上所述,基于分类机器学习方法能够面向谣言数据集的多个特征,具有较高的精确度,但这些特征大多数只能在被用户大量转发后的谣言爆发期或消停期才能获取.因此,在实际情况中,这种方法无法及时有效地遏制谣言可能产生的不良社会影响.基于内容对比的方法能有效地解决检测时效性的问题,且相对分类方法可操作性更强,但该方法多基于单一或少量的属性特征,在检测的精准度上值得商榷,其检测的结果更多的是嫌疑谣言消息.

## 6 基于网络结构的谣言源检测方法

谣言检测的最终目的是及时并有效地阻断未经证实消息的传播,防止其可能产生的不良社会影响.其中,谣言源的识别与控制至关重要,它能有效地找到谣言传播的根结所在,并能最高效地控制其进一步传播.基于网络结构的谣言检测聚焦于谣言源的检测,且隶属于信息源推断问题,它以图的形式抽象地描述社会网络拓扑结构,同时抽象出信息在社会网络中的传播模型,然后依据感染传播子图快照构建谣言源节点估计器,从而使估计的准确率最大.

在基于网络结构的谣言源检测模型中,两个首先需要厘清的基本模型分别是谣言信息传播模型和信息网络模型.对于谣言信息传播模型,除了本文 2.2.1 节指出 SI、SIS 和 SIR 三大传染病模型,还有常见的独立级联模型(Independent Cascade Model)[111] 和线性阈值模型(Linear Threshold Model)[112].对于信息网络模型,具有代表性的模型主要包括规则网络、随机网络、小世界网络、无标度网络和实际网络,其中规则网络是研究者们首选的信息网络模型.

在具体识别对象上,谣言源的检测分为单谣言源检测和多谣言源检测.单谣言源检测是谣言源检测问题的简化形式,也是研究者们关注的焦点.而多谣言源的检测也通常可以转化为对多个单谣言源的检测,如文献[113]中提出在实际应用中,谣言总是爆发自一个节点集群,由多个谣言源发出.其中主级谣言源是一类主要的谣言源节点的节点,因为其具有巨大的影响力或丰富的连接性而成为最重要的源节点,因此多源检测问题可以转化为对主级谣言源节点的检测.在具体的检测方法上,基于网络结构的谣言源检测主要分为基于传播子图快照的检测方法和基于部署观察点的检测方法.

### 6.1 基于传播子图快照的检测方法

基于传播子图快照的检测方法是指通过一次或多次获取全部或部分节点是否收到谣言消息(即是否成为感染节点)的状态子图,然后对某一网络拓扑属性度量特征估计,再推算出网络中最大可能成为谣言源的节点,这也是目前关于谣言源检测最常见的方法.其中传播子图的网络拓扑属性度量特征以 Shah 和 Zaman 等人[114] 提出的谣言中心性(rumor centrality)最为常见,即基于组合数最大似然估计的源点估计量.此外还包括传播临界边缘概率、模拟感染路径等其它传播子图检测方法.

对于谣言中心性的提出,奠基性研究成果始于 Shah 和 Zaman 等人[115] 的工作.他们假设每个节点是谣言源的概率相同,然后根据网络中观察到的感染节点组成的传播子图来计算从当前节点到其它已感染节点的路径条数以及路径出现的概率,从而建立似然函数.最后,分别计算每个节点的最大似然估计值,其中最大似然估计值最大的节点即为网络的信息源.在此过程中,他们针对谣言源检测提出了一个著名的拓扑属性度量指标:谣言中心性.谣言中心性是一个"图分数"函数,他将网络 $G(V,E)$($G$ 为社会网络,$V$ 代表网络中的节点,$E$ 代表网络中的边)作为输入,并假设每个节点为源节点时,由源节点到最后感染节点组成的路径序列的数量为每个节点的得分值即谣言中心性,其中得分最高的节点即为谣言源,又称为谣言中心.谣言中心性的公式化形式如下式(3):

$$R(u,G)=\frac{|V|!}{\prod_{w\in V}T_w^u} \tag{3}$$

其中 $T_w^u$ 表示子树 $T_w^u$ 中易感染节点的个数，如图 12 所示，令 $u$ 为节点 1，则 $|V|=5$，子树的规模分别为 $T_1^1=5, T_2^1=3, T_3^1=T_4^1=T_5^1=1$. 由此可知节点 1 的谣言中心性：$R(1,G)=8$，同理 $R(2,G)=12$，$R(3,G)=2, R(4,G)=3, R(5,G)=3$. 因此，节点 2 为所推断出的谣言源. 在实际计算过程中，对于一个具有 $n$ 个节点的一般树 $G_n$，可以使用一个消息传递算法计算 $G_n$ 中所有节点的谣言中心性 $\{RC(s,G_n), S\in G_n\}$，且复杂度为 $O(n)$，谣言中心性最大的节点即为谣言中心.

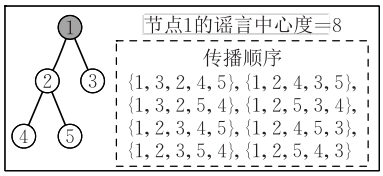


图 12　谣言中心性计算实例

此外，Shah 等人还根据网络拓扑结构对网络结构进行划分，分别建立了在正则树、一般树和一般图中的似然函数，使所提出的算法更具普适性[116]. 但该方法的缺陷也显而易见，主要包括：① 研究假设可以一次观察到整个网络中感染节点的传播结构图，但在现实情况中，我们很难一次观察到整个网络的状态，且一次性观察到的网络感染状态不一定是真实的感染状态，存在部分节点被感染但未被表达出来的情况；② 研究假设每个节点是谣言源的概率是相同的，然而事实上只有部分节点才可能是谣言源；③ 该方法是基于静态网络检测的，未考虑时间属性特征；④ 该方法假设谣言传播仅存在单一的谣言源，未考虑多谣言源的情况；⑤ 该方法通常假设底层的网络传播模型是固定也已知的，大大地限制了其应用范围. 以上问题也是谣言源检测研究过程中几个关键性的核心问题，在后续的研究中，在逐步的探索与改进.

针对问题①. 目前已有两种解决方案：第一，转化为基于多次独立信息源传播过程观察的单个谣言源检测问题，并采用了一个基于联合谣言中心性(union rumor centrality)的统一推理框架[117]. 第二，假设每个感染节点都以 p 的概率被观察到(即考虑到节点被感染但没表达出来的情况)，然后利用部分可观察到的节点信息来推测谣言源[118]. 以上两种方案通过引入多样本观察和概率的方式充分利用了所有的自由度，避免了单一样本(即一次性观察)难以观察到整个网络状态的弊端，但仍存在一定的局限性. 对于方法一，虽然检测准确度随着观测样本数和网络节点的度增加而增加，即丰富的连通性和多样性都可以增加检测的性能. 但随着观测次数的增多和网络节点度的增加，计算的时间和空间开销也随之增加. 对于方法二，研究发现对于规则树，当 $p$ 大于阈值时，谣言中心性估计器检测性能趋向于已知完整感染节点信息情况下的最优检测性能；对于几何树，在任何 $p>0$ 的情况下，估计器基本能实现 $p=1$ 情况下的检测性能；但是阈值的优化仍是值得今后继续探讨的问题.

针对问题②. 一般来说在无先验知识的情况下假设每个节点成为谣言源的概率相同的假设能较好地简化模型，但实际情况中，我们往往能通过一定的先验知识确定部分节点更可能是谣言源. 比如社会网络中信誉较低的节点更可能传播谣言. 对此，文献[119]提出了局部谣言中心性(local rumor centrality)的概念(即在度为 $\delta$ 的常规树的有限和渐近区域中观察 $n$ 个感染节点的源估计器的正确检测概率)，来研究在 SI 模型下当有嫌疑人节点集合的先验知识和受感染节点快照情况的下单信息谣言源检测问题. 由此可知，有关嫌疑节点的先验知识为谣言源检测问题引入了新的思路，进一步减小了检测过程中的计算开销；然而，关于嫌疑节点的先验知识往往因不同情况而异，多数情况下仍旧依赖于专家的主观经验，对检测结果有一定的影响. 此外，嫌疑节点的不同连接模式也会影响到检测的性能.

针对问题③. 目前已有的谣言源检测方法通常需要网络拓扑结构(如树形结构中)，节点与节点之间的固定连接，以便检测者可以沿着确定的链接追溯到谣言源(即树的根部)，以上过程充分依赖于静态网络. 由于动态实时网络观察过程的复杂性和不确定性，为减少检测所需的计算开销，目前已有的解决方案有两类：第一，间接法. 将感染节点的时序信息作为一个新的自由度来提供更准确检测谣言源的可能，即采用感染节点的先后顺序来代替感染时刻，并把这些可以反馈感染顺序的已感染节点称之为“锚节点”[113]. 但在不同的实际情况中，可获取感染时序信息的“锚节点”在已感染节点中的占比以及“锚节点”之间的链接关系都将直接影响到谣言源正确检测的概率. 第二，直接法. 由于直接检测连续性动态网络特征挑战较大，对此，文献[120]通过引入时间积分窗口将时变网络转化为一系列静态网络，然后采用反向传播策略来指定一组真实谣言来源的嫌疑犯，从而进一步缩小检测范围和计算开销. 但该方法是基于离散时间的整合窗口，在某种意义上不

算绝对的时变网络检测.因此,未来对连续时间窗口中的谣言源检测将会是一个新的思路和方向.由以上可以看出,基于传播子图快照的检测方法即使改进后,仍较难真正满足真实实时动态网络中的谣言源检测,这也是该方法的缺陷之一.

针对问题④提出的多谣言源情况.在实际情况中,谣言的产生往往有多个源头,如有关机密信息可能从不同的来源被泄露了,随之而产生的谣言也可能出自不同的来源.对于多谣言源检测,大多数研究都在单一谣言源检测的基础上利用分区的思想来检测多谣言源,且假设多个谣言源都是在同一时间开始扩散的.即首先对感染网络进行分区,然后在每个区内分别检测一个谣言源[121-122].但该分区方法的弊端也显而易见:首先,来自不同谣言源的被感染节点可能会合并,并随之在不同的分区独立地感染扩散,且基于分区的思想很可能导致该情况被检测为两个独立的谣言源.其次,不同的谣言源可能在不同的时间开始感染扩散,在检测过程中应考虑这种实际状况的差异性.考虑到各个谣言源在不同时间开始感染的多谣言源检测问题,文献[123]利用覆盖的思想代替了分区的思想,允许分配给不同感染源的感染区域之间的重叠重叠(如图 13 所示),并引入了重中心(heavy center)的概念来描述当感染是以未知的速率确定时,被单个感染源感染的子图.

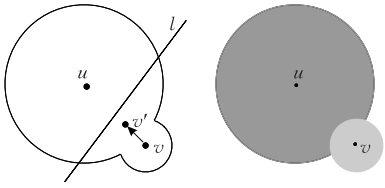


图 13 分区思想和覆盖思想之间的区别示意图(假设 $u$ 和 $v$ 是谣言源.分区思想(左)给出由行 $l$ 分隔的两个区域;然而,明显被 $u$ 感染的部分在检测中被分配给 $v$.如果在 $v$ 的区域中应用单源检测算法,则源 $v$ 的估计位置可以从其真实位置移动到 $v'$.而覆盖思想(右)允许重叠,同时确保区域被"正确地"分配给两个节点 $u$(深灰色区域)和 $v$(浅灰色区域))

针对问题⑤,当前基于传播子图快照的检测方法通常假设网络底层的传播模型是已知也固定的,如常使用的 SI 模型、SIR 模型等,或分别考虑不同传播模型下的检测方法.但在实际情况中,识别正确的传播模型总是需要一定的先验知识,例如,很难为新的谣言选择适当的传播模型,此外,很难在预先选择的底层传播模型中获取参数的真实值.以上都大大地限制了该方法的应用范围,因此,在不清楚基础传播模型的情况下检测谣言源是十分必要也有一定挑战性的问题.对此,对文献[124]提出了源头突出(source prominence)的思想(即被较大比例受感染节点包围的节点更有可能是谣言源),一方面,在受感染地区的边缘,节点往往感染较少的邻居.另一方面,在感染区域的中心,节点往往有更多的感染邻居.接着,作者提出了基于标签传播的源标识(LPSI)的多源检测方法,并在现实数据集的广泛实验上证明了该方法的有效性和高效性.虽然该源突出的思想仍有一定的缺陷性,如不同时间开始感染的谣言源,后扩散源节点周围受感染的节点也可能较少.但通过以上分析我们仍可以看出,跳出固定传播模型制约的框架,提出更具普适应的多源检测方法是未来谣言源检测的重要趋势之一.

除了谣言中心性结合最大似然估计来定位谣言源的方法,其它方法,例如基于图形中心性的(graph-centrality)度量方法也有相关研究,如文献[125]提出了距离中心性(Distance Centrality)的概念,认为在信息传播子图中,到其它节点的距离之和最小的节点是网络的信息源点.文献[126]则提出了易接入性(Accessibility)的概念,将易接入性最大的节点定位为网络的信息源.文献[127]定义了谣言传播边界概率,在感染传播子图的快照中,通过寻找传播边界边缘概率最大的节点来定位谣言源点.但这类基于图形中心性度量的检测方法往往需要完整的传播子图,这对时变社会网络较难实现.同时该方法在多源检测方面也存在一定的局限性,不利于实际需求的满足.

### 6.2 基于部署节点的检测方法

在 6.1 节基于传播子图快照的检测方法中已提及由于 OSN 网络的规模巨大且复杂,传播图中节点的真实感染状态是很难完整获取的.同时,在 OSN 网络中,由于每个节点的重要程度是不同的,如果对网络中的所有节点都进行研究势必会增加无谓的计算开销.因此,对于给定的网络,在不了解网络节点感染状态和节点关系的情况下,可以选取适量数量和重要位置的节点作为整个网络的观察点进行研究.基于部署节点的检测方法正是依据这一思想,在社会网络中部署少量的观察点(observer nodes)或称监测器(monitor nodes)来记录它们首次收到邻居节点发送来的消息时的时间和方向,然后通过统计计算推断出当前网络的谣言源.

例如,在如图 14 所示的谣言传播网络 $G$ 中,在

未知时间 $t=t^*$ 时，信息源 $s^*$ 开始感染扩散，其中实线表示已感染边缘. 图中存在 $o_1$、$o_2$、$o_3$ 三个观察点，其主要作用是观察在何时从哪个邻居节点处接收信息并进行记录，从而依据这些记录的统计信息估计网络 $G$ 中哪个节点是信息源.

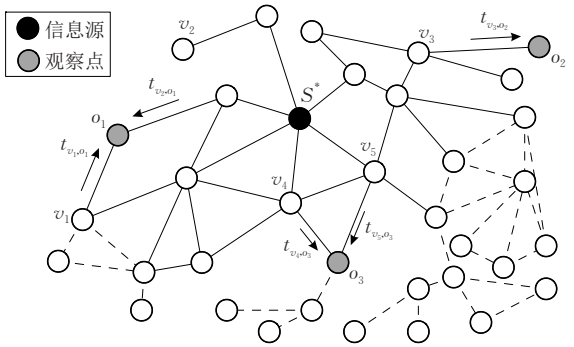


图 14　基于部署观察点的任意图上的信息源估计

基于部署节点的检测方法在具体操作过程中主要包含三个步骤：部署观察点，估算信息源，最后将估计值最大的候选源作为当前网络的谣言源[128]. 具体流程如图 15 所示.

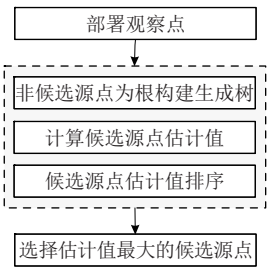


图 15　基于部署节点属性度量的谣言源定位步骤

该方法始于对信息源的推断研究，文献[129]通过在网络中稀疏地部署少量的观察点，获取观察点记录的传播信息，然后计算网络中各节点是真实信息源的概率，从而实现信息源的定位. 在对谣言源检测的实际应用中，文献[130]对比了各类中心性度量的观察点选取方法，并将观察点分为积极观察点(Positive Observer)和消极观察点(Negative Observer)两类，然后分别计算各节点到这两类观察点的可及性(Reachability)和距离(Distance)，进而形成两个指标：贪婪源集合大小(Greedy Source Set Size，GSSS)和贪婪信息传播的最大距离(Maximal Distance of Greedy Information Propagation，MDGIP)，最后结合这两个指标，通过贪婪算法来定位谣言源. 而文献[131]则采用独立级联模型来刻画社会网络中的信息传播，并定义了谣言量词(rumor quantifier)，即一个基于概率的数值，用于排列节点将成为谣言源的可能性. 此时，作者假设：级联更可能从谣言源传播到积极的观察点(即接收到消息的感染点)，而不太可能传播到消极的观察点. 实验表明，当有合理数量的观察点时，作者通过可扩展的 RSD 算法来计算每个节点的谣言量词，能有效地识别出谣言源. 文献[132]从每个观察点来反推信息到达其它节点的时间，然后计算多个观察点反推所得出时间的均方差，最小均方差值对应的节点即为信息源.

由以上综述可知，基于部署观察点的方法仅需要获取少量观察点反馈的传播信息，而无需窥探整个网络传播状态，在实际应用中的可行性较高. 但是，该方法检测的准确度和算法开销取决于观察点在网络中所部署的位置和数量. 对此，研究者们达成的共识是在网络中对信息传播影响力越大的节点，其在信息传播过程中接收谣言信息的可能性越大，记录的谣言传播信息也越有效. 在目前的相关研究中，中心性是网络节点重要程度的一个有效度量指标[133]，具体包括度中心性度量(Degree Centrality)，介数中心性度量(Betweenness Centrality)和紧密性中心性度量(Closeness Centrality). 节点的度中心性指该节点的邻居节点个数，主要描述节点在网络中产生的直接影响. 紧密中心性指该节点与网络中其它节点距离的反比，值越大，则表明该节点与其它节点越近. 介数中心性指网络中两两节点间的最短路径经过该节点的次数，主要描述节点在网络信息流动中的重要性.

文献[134]在观察点对信息源定位准确率的影响研究中，基于节点的中心性度量分析了 6 种常见的部署策略对定位准确性的影响，即度、介数、聚类系数、特征向量、$k$-核及随机策略，发现这几种策略对定位准确率的影响基本相同；最后作者提出了基于 $r$ 覆盖率的观察点部署策略，以观察点结合的 $r$ 覆盖率为目标函数，实现了 $r$ 覆盖率优先观察点选取算法. 在目前基于观察点部署的谣言源检测的常见策略中，主要以度和介数两个度量指标居多[130,133,135]. 如文献[130]分别比较了随机性的、观察点的相互距离(Inter-Monitor Distance，Dist)、节点入边数量(Number of Incoming Edges，NI)、NI 和 Dist 的组合，介数中心性(Betweenness Centrality，BC)，BC 和 Dist 的组合，共六种中心性比较方法选择部署的观察点.

### 6.3　小　结

整体上，基于传播子图快照的检测方法和基于

部署观察点的检测方法优劣各异.基于传播子图快照的检测方法虽能有效地定位谣言源,但其往往需要谣言感染过程中大量、动态的感染网络拓扑结构状态图(如整个网络的感染状态、多次观察的感染网络快照、节点感染状态可观察的概率等),且该方法往往是基于静态的网络拓扑图,检测的时效性和动态性较差.因此在当前大规模动态的社会网络实际应用中难度较大.另一方面,基于部署观察点属性度量的检测方法,无需掌握整个网络的感染状态,仅需稀疏部署的观察点获取其反馈信息,因此网络资源计算开销较小,但是观察点数量和位置的合理部署仍是当前研究中的核心挑战.

## 7 研究挑战以及展望

随着在线社会网络的迅速发展,用户规模不断扩大,传统 OSN 谣言检测技术面临着一些新的挑战,同时也为 OSN 谣言检测技术的发展创造了新的条件与机遇.对此,下文将介绍 OSN 谣言检测技术未来面临的挑战以及值得进一步关注或探讨的问题.

(1)动态识别

由于谣言阐述的不确定性会随着时间的推移而改变,此刻不确定的事实很可能在一段时间后得到证实.由此可见谣言并非一成不变的,某一时间段的谣言检测结果也不能盖棺定论.因此对于反复性谣言,不能完全依赖与谣言知识库对比的检测方法.同时,突发性事件(如自然灾害、社会重大事故)谣言的潜在破坏力大,影响广,故而需要及时有效的检测手段.过去的研究中,多是基于历史谣言数据集或是静态网络拓扑关系的研究,尤其在谣言源检测研究中,对于根据时间变化的时变网络下的谣言源推断问题研究不足.因此,如何利用在线机器学习、时间序列模型等动态预测模型处理 OSN 谣言传播过程中的演变;如何利用流数据对 OSN 谣言用户行为建模;如何研究时变网络拓扑结构关系下 OSN 谣言传播模型都是值得今后继续挖掘的问题.

(2)大规模分布式检测

目前 OSN 谣言检测主要聚焦在模型的准确性上,实验数据多以几千几万级别的小数据为主;但现实社会网络产生的数据规模和用户规模十分巨大,海量的数据给现有的计算机体系结构和算法带来了巨大的挑战,现有模型算法在大规模计算下的时间和空间开销已然成为严峻的问题.因此,未来谣言检测算法在大规模计算环境下的效率有待进一步提升.

(3)不平衡数据集

现有研究中,实验数据集中谣言与非谣言数据集很多为 1∶1,或非谣言数据稍多于谣言数据,这是不符合实际 OSN 谣言传播状况的.现实 OSN 中,谣言信息远少于正常信息,谣言传播群体数量也远少于正常用户数量,是典型的不平衡数据集分类问题[136].对此,文献[137]研究表明,不平衡数据集的不平衡程度、概念复杂度以及训练数据规模等因素会对传统分类算法性能产生较大的影响.因此已有的许多机器学习模型,在实际检测过程中,很可能由于谣言信息空间中混杂着大量的正常信息,从而模糊其分类边界,导致分类准确率较低.由此可见,在未来基于分类的谣言检测方法研究中,可利用 SMOTE[138]等采样方法调整谣言数据集中的类间分布,或将代价信息加入到传统分类算法中,从而将已有分类算法修改为代价敏感学习算法(Cost Sensitive Learning, CSL)[139],即在模型构建过程中考虑每个类别的错误分类代价是不同的,从而提高谣言类的识别准确率.

(4)谣言信息传播的异构性

谣言传播过程中的异构性体现在三个方面.首先谣言信息具有异构性,多条信息可能来自不同的谣言源,并且不同的来源之间一般存在竞争或合作的关系[140],而目前研究以单谣言源研究居多.第二,谣言传播模型也具有异构性,不同的谣言信息可能采取不同的传播模型和机制;同一谣言信息的传播机制也可能随着时间动态变化[141],而目前研究以树形传播结构居多.第三,谣言传播网络具有异构性,信息可能会在不同的网络或者平台中同时进行传播,例如某条信息可能在新浪微博、腾讯微博、微信朋友圈等不同平台同时传播.因此,未来研究中,需要将这些不同角度的异构性综合考虑,进行深入的研究.

(5)谣言信息传播的关联性

作为 OSN 舆情传播的一种,OSN 谣言在传播过程中并非孤立的,谣言之间的关联现象普遍存在.某一则谣言兴起时,往往会通过共同主体或共同主题关联其它谣言,如玉树地震谣言与汶川地震谣言.这一关联既包括单一热点事件中的谣言集,也包括多个热点事件之间的谣言集.而在目前研究中,每一

则谣言传播网络、每一个节点都是孤立存在的，没能考虑谣言传播过程中的耦合性、关联性特征以及网络中相邻节点和多个节点簇之间的合作与分享. 因此如何定量的分析谣言传播之间的关联性、节点之间的合作与共享，进而更好地指导谣言检测方法的实用性，是一个有待继续深入的问题.

(6) 多媒体谣言信息检测

多媒体信息(图片、视频、音频)已成为 OSN 谣言传播过程中重要的补充信息，甚至成为谣言主体内容. 当前，不少谣言来自于对图像的 PS、音频或视频的剪辑与合成. 目前研究中主要以谣言文本信息为主，少量研究者利用图像的标签和外部知识来辅助谣言检测，而对视频、音频类信息的挖掘十分罕见. 因此，将图片、音频与视频处理技术合理地运用到谣言检测中是未来研究的又一趋势.

## 8　结束语

在线社会网络谣言检测是社会网络、信息传播等研究领域的热点问题，随着在线社会网络新兴话题检测、传播预测和信息可信度预测技术的不断发展，该问题一直受到研究人员的广泛关注. 针对这一研究热点，本文首先梳理了在线社会网络谣言的基本概念与本质特点，然后分析了在线社会网络谣言检测研究中目标、对象和时间属性三个维度上的差异性，并阐述了在线社会网络谣言数据采集和标注的不同方法. 接着，本文对不同类别的在线社会网络谣言信息检测方法和在线社会网络谣言源检测方法进行深入分析和比较，并分别指出了它们的优势、不足及其适用场景. 最后，本文讨论了当前在线社会网络谣言检测领域面临的挑战及未来可能的研究方向.

从当前的研究趋势来看，在线社会网络谣言检测的识别逐步从静态的，时滞的，小规模的以及单源的向动态的、实时的、大规模的和多源的转变. 这种转变要求后续模型的计算效率更高，可扩展性更好，例如可以利用单一模型融合多源谣言数据进而实现精准的识别. 当然，在线社会网络的谣言识别模型还应该根据具体的应用场景灵活调整效率与准确率之间的关系，同时在特定的环境下引入一定量的人为要素，进一步地增加监控的准确性以及时效性. 本文希望在对当前在线社会网络谣言检测方法进行综述的基础上，给后续相关领域的研究者以相应的参考，共同推动在线社会网络谣言检测模型的发展与优化.

## 参　考　文　献


[1] O'reilly T. What is Web 2.0: Design patterns and business models for the next generation of software. Social Science Electronic Publishing, 2007, 97(7): 253-259

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

[3] Tripathy R M, Bagchi A, Mehta S. A study of rumor control strategies on social networks//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada, 2010: 1817-1820

[4] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Oxford, England: Cambridge University Press, 1994

[5] Hu Yan-Li. Research on Key Technologies of Public Opinion Evolution in Online Social Networks[Ph. D. dissertation]. National University of Defense Technology, Changsha, 2011 (in Chinese)
(胡艳丽. 在线社会网络中的舆论演化关键技术研究[博士学位论文]. 国防科技大学，长沙，2011)

[6] Mislove A, Marcon M, Gummadi K P, et al. Measurement and analysis of online social networks//Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. California, USA, 2007: 29-42

[7] Breiger R L. The duality of persons and groups. Social Forces, 1974, 53(2): 181-190

[8] Garton L, Haythornthwaite C, Wellman B. Studying online social networks. Journal of Computer-Mediated Communication, 1997, 3(1): 75-106

[9] Xu Ke, Zhang Sai, Chen Hao, Li Hai-Tao. Measurement and analysis of online social networks. Chinese Journal of Computers, 2014, 37(1): 165-188(in Chinese)
(徐恪，张赛，陈昊，李海涛. 在线社会网络的测量与分析. 计算机学报，2014，37(1)：165-188)

[10] Donovan P. How idle is idle talk? One hundred years of rumor research. Diogenes, 2007, 54(1): 59-82

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

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

[13] Peterson W A, Gist N P. Rumor and public opinion. American Journal of Sociology, 1951, 57(2): 159-167

[14] Shibutani T. Improvised News: A Sociological Study of Rumor. Oxford, UK: Bobbs-Merrill, 1966

[15] Kapferer J N. Rumeurs: Le Plus Vieux Media du Monde (in French). Paris, France: Le Seuil Editions, 1987

[16] Morin E. Rumour in Orleans. New York, USA: Pantheon Books, 1971: 11

[17] Rosnow R L. Rumor as communication: A contextualist approach. Journal of Communication, 1988, 38(1): 12-28

[18] DiFonzo N, Bordia P. Rumor Psychology: Social and Organizational Approaches. Washington, USA: American Psychological Association, 2007

[19] Chen Xue-Ping. Rumors of Psychology. Taiwan, China: Arts and Literature Series Editorial Department, 1939 (in Chinese)
(陈雪屏. 谣言的心理. 台湾, 中国: 艺文丛书编辑部, 1939)

[20] Allport G W, Postman L. An analysis of rumor. Public Opinion Quarterly, 1946, 10(4): 501-517

[21] Jung C G. Contribution to the psychology of rumour. Indiana Association for Health Physical Education Recreation & Dance Journal, 1909, 18(1): 1-26

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

[23] Hu Yu. The Effect of Mass Communication: Problems and Countermeasures. Beijing: Xinhua Publishing House, 2000 (in Chinese)
(胡钰. 大众传播效果: 问题与对策. 北京: 新华出版社, 2000)

[24] Sudbury A. The proportion of the population never hearing a rumour. Journal of Applied Probability, 1985, 22(2): 443-446

[25] Hurley M, Jacobs G, Gilbert M. The basic SI model. New Directions for Teaching and Learning 2006, 106(6): 11-22

[26] Pastor-Satorras R, Vespignani A. Epidemic dynamics and endemic states in complex networks. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2001, 63(6): 138-158

[27] Moreno Y, Pastor-Satorras R, Vespignani A. Epidemic outbreaks in complex heterogeneous networks. The European Physical Journal B, 2002, 26(4): 521-529

[28] Xiong F, Liu Y, Zhang Z, et al. An information diffusion model based on retweeting mechanism for online social media. Physics Letters A, 2012, 376(30): 2103-2108

[29] 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

[30] Kawachi K, Seki M, Yoshida H, et al. A rumor transmission model with various contact interactions. Journal of Theoretical Biology, 2008, 253(1): 55-60

[31] Zhao L, Cui H, Qiu X, et al. SIR rumor spreading model in the new media age. Physica A: Statistical Mechanics and Its Applications, 2013, 392(4): 995-1003

[32] Roshani F, Naimi Y. Effects of degree-biased transmission rate and nonlinear infectivity on rumor spreading in complex social networks. Physical Review E, 2012, 85(3): 036109

[33] Daley D J, Kendall D G. Stochastic rumours. IMA Journal of Applied Mathematics, 1965, 1(1): 42-55

[34] Xuan Hui-Yu, Zhang Fa. Complex System Simulation and Application. Beijing: Tsinghua University Press, 2008 (in Chinese)
(宣慧玉, 张发. 复杂系统仿真及应用. 北京: 清华大学出版社, 2008)

[35] 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

[36] Zhou Yu-Qiong. A Study on the Online Rumors in Contemporary Chinese Society. Beijing: The Commercial Press, 2012 (in Chinese)
(周裕琼. 当代中国社会的网络谣言研究. 北京: 商务印书馆, 2012)

[37] Lasswell H D. The structure and function of communication in society. Communication Theory and Research, 2007, 24 (8): 215-228

[38] Mo Qian, Yang Ke. Overview of Web spammer detection. Journal of Software, 2014, 25(7): 1505-1526(in Chinese)
(莫倩, 杨珂. 网络水军识别研究. 软件学报, 2014, 25(7): 1505-1526)

[39] Murmann A J. Enhancing Spammer Detection in Online Social Networks with Trust-Based Metrics[M. S. dissertation]. San Jose State University, San Jose, USA, 2009

[40] Nekovee M, Moreno Y, Bianconi G, et al. Theory of rumour spreading in complex social networks. Physica A: Statistical Mechanics and Its Applications, 2007, 374(1): 457-470

[41] Metzger M J, Flanagin A J, Eyal K, et al. Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. Annals of the International Communication Association, 2003, 27(1): 293-335

[42] Sunstein C R. On Rumors: How Falsehoods Spread, Why We Believe Them, and What Can be Done. Princeton, USA: Princeton University Press, 2014

[43] Qin Y, Wurzer D, Lavrenko V, et al. Spotting Rumors via Novelty Detection. arXiv preprint arXiv:1611.06322, 2016

[44] 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

[45] Chen W, Yeo C K, Lau C T, et al. Behavior deviation: An anomaly detection view of rumor preemption//Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference. Vancouver, Canada, 2016: 1-7

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

[47] Okazaki N, Nabeshima K, Watanabe K, et al. Extracting and aggregating false information from Microblogs//Proceedings of the Workshop on Language Processing and Crisis Information. Nagoya, Japan, 2013: 36-43

[48] Wang S, Terano T. Detecting rumor patterns in streaming social media///Proceedings of the 2015 IEEE International Conference on Big Data. Santa Clara, Canada, 2015: 2709-2715

[49] 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

[50] 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

[51] Yang Y K, Niu K, He Z Q. Exploiting the topology property of social network for rumor detection//Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering. Songkhla, Thailand, 2015: 41-46

[52] Liang C, Liu Z, Sun M. Expert finding for Microblog misinformation identification//Proceedings of the 24th International Conference on Computational Linguistics. Mumbai, India, 2012: 703-712

[53] Jain S, Sharma V, Kaushal R. Towards automated real-time detection of misinformation on Twitter//Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics. Jaipur, Rajasthan, 2016: 2015-2020

[54] 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. Sydney, Australia, 2013: 120-131

[55] 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

[56] Yang H, Zhong J, Ha D, et al. Rumor propagation detection system in social network services//Proceedings of the International Conference on Computational Social Networks. Ho Chi Minh City, Vietnam, 2016: 86-98

[57] Yang Z, Wang C, Zhang F, et al. Emerging rumor identification for social media with hot topic detection//Proceedings of the 12th Web Information System and Application Conference. Jinan, China, 2015: 53-58

[58] Liu Y, Xu S, Tourassi G. Detecting rumors through modeling information propagation networks in a social media environment//Proceedings of the 8th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Washington, USA, 2015: 121-130

[59] 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

[60] Zubiaga A, Liakata M, Procter R. Learning reporting dynamics during breaking news for rumour detection in social media. arXiv preprint arXiv:1610.07363, 2016

[61] Sampson J, Morstatter F, Wu L, et al. Leveraging the implicit structure within social media for emergent rumor detection//Proceedings of the 25th ACM International Conference on Information and Knowledge Management. Indianapolis, USA, 2016: 2377-2382

[62] Cai G, Wu H, Lv R. Rumors detection in Chinese via crowd responses//Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Beijing, China, 2014: 912-917

[63] Liang G, Yang J, Xu C. Automatic rumors identification on Sina Weibo//Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. Changsha, China, 2016: 1523-1531

[64] Zubiaga A, Liakata M, Procter R, et al. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS One, 2016, 11(3): 1-34

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

[66] 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

[67] Kwon S, Cha M, Jung K. Rumor detection over varying time windows. PLoS One, 2017, 12(1): e0168344

[68] Jin Z, Cao J, Jiang Y G, et al. News credibility evaluation on Microblog with a hierarchical propagation model//Proceedings of the 2014 IEEE International Conference on Data Mining. Shenzhen, China, 2014: 230-239

[69] Liu Zhi-Yuan, Zhang Le, Tu Cun-Chao, Sun Mao-Song. Statistical and semantic analysis of rumors in Chinese social media. Science China Information Sciences, 2015, 45(12): 1536-1546(in Chinese)
(刘知远，张乐，涂存超，孙茂松. 中文社交媒体谣言统计语义分析. 中国科学：信息科学，2015，45(12)：1536-1546 )

[70] Dayani R, Chhabra N, Kadian T, et al. Rumor detection in Twitter: An analysis in retrospect//Proceedings of the 2015 IEEE International Conference on Advanced Networks and Telecommunications Systems. Kolkata, Indian, 2015: 1-3

[71] Hamidian S, Diab M. Rumor detection and classification for Twitter data//Proceedings of the 15th International Conference on Social Media Technologies, Communication, and Informatics. Barcelona, Spain, 2015: 71-77

[72] Andreevskaia A, Bergler S. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses// Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy, 2006: 209-216

[73] 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

[74] Miller G A. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39-41

[75] Dong Z, Dong Q. Hownet and the Computation of Meaning. Singapore: World Scientific Publishing, 2006

[76] Takahashi T, Igata N. Rumor detection on Twitter// Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligence Systems. Kobe, Japan, 2012: 452-457

[77] Ma B, Lin D, Cao D. Content representation for Microblog rumor detection//Proceedings of the Advances in Computational Intelligence Systems. New York, USA, 2017: 245-251

[78] Chen T, Wu L, Li X, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. arXiv preprint arXiv:1704.05973, 2017

[79] Ruchansky N, Seo S, Liu Y. CSI: A hybrid deep model for fake news. arXiv preprint arXiv:1703.06959, 2017

[80] Cheng Liang, Qiu Yun-Fei, Sun Lu, Research on detecting microblogging rumours. Computer Application and Software, 2013, 30(2): 226-228(in Chinese)
(程亮, 邱云飞, 孙鲁. 微博谣言检测方法研究. 计算机应用与软件, 2013, 30(2): 226-228)

[81] Lin D, Lv Y, Cao D. Rumor diffusion purpose analysis from social attribute to social content//Proceedings of the 2015 International Conference on Asian Language Processing (IALP). Suzhou, China, 2015: 107-110

[82] Ma B, Lin D, Cao D. Content representation for microblog rumor detection//Proceedings of the Advances in Computational Intelligence Systems. New York, USA, 2017: 245-251

[83] Jiao Li-Cheng, Yang Shu-Yuan, Liu Fang, et al. Seventy years beyond neural networks: Retrospect and prospect. Chinese Journal of Computers, 2016, 39(8): 1697-1716(in Chinese)
(焦李成, 杨淑媛, 刘芳等. 神经网络七十年: 回顾与展望. 计算机学报, 2016, 39(8): 1697-1716)

[84] Li L, Ren W, Qin B, et al. Learning document representation for deceptive opinion spam detection//Proceedings of the 14th China National Conference, CCL 2015 and Third International Symposium. Guangzhou, China, 2015: 393-404

[85] Mao Er-Song, Chen Gang, Liu Xin, Wang Bo. Research on detecting Microblog rumors based on deep features and ensemble classifier. Application Research of Computers, 2016, 33(11): 3369-3373(in Chinese)
(毛二松, 陈刚, 刘欣, 王波. 基于深层特征和集成分类器的微博谣言检测研究. 计算机应用研究, 2016, 33(11): 3369-3373)

[86] Guo Kai. The research of Microblog rumors detection based on comments sentiment [M. S. dissertation]. Dalian University of Technology, Dalian, Liaoning, 2014(in Chinese)
(郭凯. 基于评论情感的微博谣言检测研究[硕士学位论文]. 大连理工大学, 辽宁,大连, 2014)

[87] Wu Hao. Research on Rumor Detection Based on Crowd's Sentiment Features in Theme Tweet[M. S. dissertation]. Guilin University of Electronic Technology, Guilin, Guangxi, 2015(in Chinese)
(吴昊. 基于主题微博中群体情感特征的谣言检测研究[硕士学位论文]. 桂林电子科技大学, 桂林, 广西, 2015)

[88] Hamidian S, Diab M T. Rumor identification and belief investigation on Twitter//Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 3-8

[89] Guo Weiwei, Diab Mona. Modeling sentences in the latent space//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Jeju Island, Korea, 2012: 864-872

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

[91] Jin Z, Cao J, Jiang Y G, et al. News credibility evaluation on Microblog with a hierarchical propagation model// Proceedings of the 2014 IEEE International Conference on Data Mining. Shenzhen, China, 2014: 230-239

[92] Wang Ze-Hui. Early Detection and Analysis of Rumors in Social Network[M. S. dissertation]. Zhejiang University, Hangzhou, 2016(in Chinese)
(王泽慧. 社交网络谣言的及时检测和分析[硕士学位论文]. 浙江大学, 杭州, 2016)

[93] 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. Rio de Janeiro, Brazil, 2013: 729-736

[94] Jin Z, Cao J, Zhang Y, et al. Novel visual and statistical image features for Microblogs news verification. IEEE Transactions on Multimedia, 2017, 19(3): 598-608

[95] Su S, Wan C, Hu Y, et al. Characterizing geographical preferences of international tourists and the local influential factors in China using geo-tagged photos on social media. Applied Geography, 2016, 73(6): 26-37

[96] Chang C, Zhang Y, Szabo C, et al. Extreme user and political rumor detection on Twitter//Proceedings of the 12th Advanced Data Mining and Applications. Gold Coast, Australia, 2016: 751-763

[97] 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

[98] Shirai T, Sakaki T, Toriumi F, et al. Estimation of false rumor diffusion model and estimation of prevention model of false rumor diffusion on Twitter//Proceedings of the 26th Annual Conference of Japanese Society for Artificial Intelligence. Yamaguchi, Japan, 2012: 1-4

[99] Friggeri A, Adamic L A, Eckles D, Cheng J. Rumor cascades//Proceedings of the 8th International AAAI Conference on Weblogs and Social Media. Michigan, USA, 2014: 101-110

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

[101] 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 on Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 1751-1754

[102] Tolosi L, Tagarev A, Georgiev G. An analysis of event-agnostic features for rumour classification in Twitter//Proceedings of the 10th International AAAI Conference on Web and Social Media. Cologne, Germany, 2016: 151-158

[103] Majid A, Chen L, Mirza H T, et al. A system for mining interesting tourist locations and travel sequences from public geo-tagged photos. Data & Knowledge Engineering, 2015, 95(1): 66-86

[104] Li S, Zhang Y J. Semi-supervised classification of emotional pictures based on feature combination//Proceedings of the SPIE—the International Society for Optical Engineering. Beijing, China, 2011: 27

[105] Aladhadh S, Zhang X, Sanderson M. Tweet author location impacts on Tweet credibility//Proceedings of the 2014 Australasian Document Computing Symposium. Melbourne, Australia, 2014: 73-76

[106] Cheng Zhi, Wang Qi-Dong, Zheng Li-Hui, et al. Design and application of online earthquake rumor monitoring system. South China Journal of Seismology, 2012, 32(2): 79-86(in Chinese)
(程志，王启东，郑黎辉 等. 网络地震谣言监测系统的设计和应用. 华南地震，2012，32(2)：79-86)

[107] Jin Z, Cao J, Guo H, et al. Rumor detection on Twitter pertaining to the 2016 US presidential election. arXiv preprint arXiv:1701.06250, 2017

[108] Toriumi F, Shinoda K, Kaneyama G. Accuracy evaluation of demagogue detection system using social media. IPSJ Digital Practice, 2012, 12(3): 201-208

[109] Zhou Z, Bandari R, Kong J, et al. Information resonance on Twitter: watching iran//Proceedings of the 1st Workshop on Social Media Analytics. Washington, USA, 2010: 123-131

[110] Fan P, Li P, Jiang Z, et al. Measurement and analysis of topology and information propagation on Sina-Microblog//Proceedings of the 2011 IEEE International Conference on Intelligence and Security Informatics. Beijing, China, 2011: 396-401

[111] Tripathy R M, Bagchi A, Mehta S. A study of rumor control strategies on social networks//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada, 2010: 1817-1820

[112] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 137-146

[113] Wang Chao-Xu. Research on Identifying Information Source in Networks[Ph. D. dissertation]. University of Science and Technology of China, Hefei, China, 2016(in Chinese)
(王朝旭. 信息传播网络中信息源推断问题的研究[博士学位论文]. 中国科学技术大学，合肥，2016)

[114] Shah D, Zaman T. Rumor centrality: A universal source detector. ACM SIGMETRICS Performance Evaluation Review, 2012, 40(1): 199-210

[115] Shah D, Zaman T. Rumors in a network: Who's the culprit?. IEEE Transactions on Information Theory, 2011, 57(8): 5163-5181

[116] Shah D, Zaman T. Finding rumor sources on random trees. Operations Research, 2015, 64(3): 736-755

[117] Wang Z, Dong W, Zhang W, et al. Rumor source detection with multiple observations: Fundamental limits and algorithms//Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems. Austin, USA, 2014: 1-13

[118] Karamchandani N, Franceschetti M. Rumor source detection under probabilistic sampling//Proceedings of the 2013 IEEE International Symposium on Information Theory. Istanbul, Turkey, 2013: 2184-2188

[119] Dong W, Zhang W, Tan C W. Rooting out the rumor culprit from suspects//Proceedings of the 2013 IEEE International Symposium on Information Theory. Istanbul, Turkey, 2013: 2671-2675

[120] Jiang J, Sheng W E N, Yu S, et al. Rumor source identification in social networks with time-varying topology. IEEE Transactions on Dependable and Secure Computing, 2016, 99(1): 1-15

[121] Zhang Z, Xu W, Wu W, et al. A novel approach for detecting multiple rumor sources in networks with partial observations. Journal of Combinatorial Optimization, 2017, 33(1): 132-146

[122] Prakash B A, Vreeken J, Faloutsos C. Spotting culprits in epidemics: How many and which ones?//Proceedings of the

2012 IEEE 12th International Conference on Data Mining. Brussels, Belgium, 2012: 11-20

[123] Ji F, Tay W P. An algorithmic framework for estimating rumor sources with different start times. IEEE Transactions on Signal Processing, 2017, 65(10): 2517-2530

[124] Wang Z, Wang C, Pei J, et al. Multiple source detection without knowing the underlying propagation model// Proceedings of the 31th AAAI Conference on Artificial Intelligence. California, USA, 2017: 217-223

[125] Entringer R C, Jackson D E, Snyder D A. Distance in graphs. Czechoslovak Mathematical Journal, 1976, 26(2): 283-296

[126] Comin C H, Costa L F. Identification of starting points in sampling of complex networks. Physics, 2011, 84(5): 1-17

[127] Zheng L, Tan C W. A probabilistic characterization of the rumor graph boundary in rumor source detection//Proceedings of the 2015 IEEE International Conference on Digital Signal, Singapore, 2015: 765-769

[128] Xu Chao. The Research of Social Network Source Location Based on Partial Propagation Paths[M. S. dissertation]. Northeastern University, Shenyang, 2014(in Chinese)
(徐超. 基于部分传播路径的社交网络传播源点定位方法研究[硕士学位论文]. 东北大学, 沈阳, 2014)

[129] Pinto P C, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks. Physical Review Letters, 2012, 109(6): 068702

[130] Seo E, Mohapatra P, Abdelzaher T. Identifying rumors and their sources in social networks//Proceedings of the Society of Photo-Optical Instrumentation Engineers Conference Series. Baltimore, USA, 2012: 1-13

[131] Xu W, Chen H. Scalable rumor source detection under independent cascade model in online social networks// Proceedings of the 2015 11th International Conference on Mobile Ad-hoc and Sensor Networks. Shenzhen, China, 2015: 236-242

[132] Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks//Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media. California, USA, 2009: 361-362

[133] Ghosh R, Lerman K. Predicting influential users in online social networks. arXiv preprint arXiv:1005.4882, 2010

[134] Zhang Yu-Bo, Zhang Xi-Zhe, Zhang Bin. Observer deployment method for locating the information source in social network. Journal of Software, 2014, 25(12): 2837-2851(in Chinese)
(张聿博, 张锡哲, 张斌. 面向社交网络信息源定位的观察点部署方法. 软件学报, 2014, 25(12): 2837-2851)

[135] Pinto P C, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks. Physical Review Letters, 2012, 109(6): 068702

[136] Lu Tong-Qiang. Research on Rumor Detection of Weibo Based on Semi-Supervised Learning[M. S. dissertation]. Shangdong University, Jinan, 2015(in Chinese)
(路同强. 基于半监督学习的微博谣言检测研究[硕士学位论文]. 山东大学, 济南, 2015)

[137] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent Data Analysis, 2002, 6(5): 429-449

[138] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357

[139] Elkan C. The foundations of cost-sensitive learning// Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, USA, 2001: 973-978

[140] Myers S A, Leskovec J. Clash of the contagions: Cooperation and competition in information diffusion//Proceedings of the 2012 IEEE 12th International Conference on Data Mining. Brussels, Belgium, 2012: 539-548

[141] Farajtabar M, Gomez-Rodriguez M, Wang Y, et al. Co-evolutionary dynamics of information diffusion and network structure//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 619-620

**CHEN Yan-Fang**, born in 1992, Ph. D. candidate. Her research interests include social networks, information organization, data mining, and knowledge discovery.

**LI Zhi-Yu**, born in 1991, Ph. D. candidate. His research interests include social computing, Web mining, and natural language processing.

**LIANG Xun**, born in 1965, Ph. D., professor, Ph. D. supervisor. His research interests include neural networks, support vector machine and social computing.

**QI Jin-Shan**, born in 1977, Ph. D. candidate. His research interests include social computing, data mining.

**Background**

Online rumor has great potential bad effect on people's acquiring information, especially in Online Social Network (OSN) which has been already integrated into people's life, entertainment and work. So, it is essential for us to detect rumor information in OSN, thus improving the quality of OSN information ecology environment and user experience. In the era of big data, there are three main problems in rumor detection, namely timeliness, extendibility and scalability, which are great challenges for all the researchers in this field to tackle.

In the previous study, many researchers think rumor detection is a part of information credibility detection, and some ones (especially in Asia) even suppose rumor is similar to misinformation, which treats "rumor" as a negative term. Although many variations of the definition of rumors have been proposed in the literature of sociology and communication studies, there are two key points which explain how it is different from misinformation. First, the true value of rumor is uncertain, which means only "false rumor" (eventually found to be false) belongs to misinformation. Second, rumor is a statement about an object, event, or issue, in other words, it is about a topic. Misinformation, by contrast, is just a piece of information.

Based on discussed above, in this review, firstly, we summery the concept and some basic features of OSN rumor. Then, we conclude and explain different problem statements based on three attribute perspectives (objective, object and time), which have different application scenes. In addition, we also compare and analyze various OSN rumor detection model (classification model based on content and matching model) and OSN rumor source model (model of observing snapshot of diffusion sub-graph and model of observing monitor nodes), including its strengths and weaknesses. In the future, we hope our advices could help a lot in this field.

This work is supported by the National Natural Science Foundation of China (Grant Nos. 71531012, 71271211), the Natural Science Foundation of Beijing (Grant No. 4172032), the Outstanding Innovative Talents Cultivation Funded Programs 2017 of Renmin University of China.