人工智能视角下的在线社交网络虚假信息检测、传播与控制研究综述

张志勇[1),2)] 荆军昌[1),2)] 李 斐[3)] 赵长伟[1),2)]

[1)](河南科技大学信息工程学院 河南 洛阳 471023)

[2)](河南省网络空间安全应用国际联合实验室 河南 洛阳 471023)

[3)](广州巨杉软件开发有限公司 广州 510006)

摘 要 随着新一代人工智能技术的发展和应用，在线社交网络（Online Social Networks，OSNs）虚假信息的自动化检测、传播和控制，受到了政府、学术界和工业界人员的广泛关注. 虚假信息检测主要从信息内容和社交上下文辅助信息等方面展开研究，虚假信息传播研究可以追溯到早期复杂网络和小世界网络中的谣言传播动力学模型研究，直到近三年来关于社交自然人和社交机器人的混合型、交互式传播模式研究，虚假信息传播控制主要从传播的节点控制和访问控制/使用控制等方面展开研究. 本文分别从社交客体（虚假信息）和社交主体（社交自然人和社交机器人）两个方面进行深入系统探讨. 首先，回顾了国内外虚假信息检测研究现状，重点论述了虚假信息检测特征和方法. 其次，围绕社交自然人和社交机器人的检测方法和传播模式进行分析和比较，阐述两类社交主体传播虚假信息的一般规律. 然后，对虚假信息传播控制方法进行全面系统的梳理和分析，给出了虚假信息传播的节点控制和使用控制模型，总结了相关数据采集、标注方法和常用的公开数据集等. 最后，提出了社会情境安全和分析框架，以及针对虚假信息在跨平台传播和控制方面，未来研究所面临的问题、挑战及可能的研究方向.

关键词 在线社交网络；虚假信息；社交机器人；人工智能；使用控制；社会情境安全

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2021.02261

Survey on Fake Information Detection, Propagation and Control in Online Social Networks from the Perspective of Artificial Intelligence

ZHANG Zhi-Yong[1),2)] JING Jun-Chang[1),2)] LI Fei[3)] ZHAO Chang-Wei[1),2)]

[1)](*Information Engineering College, Henan University of Science and Technology, Luoyang, Henan* 471023)

[2)](*Henan International Joint Laboratory of Cyberspace Security Applications, Luoyang, Henan* 471023)

[3)](*Guangzhou SequoiaDB Co., Guangzhou* 510006)

Abstract In recent years, with the unceasing development and extensively application of new generation of artificial intelligence technology, the automatic detection, dissemination and control of fake information in online social networks (OSNs) have been widely and generally concerned by the government and regulators, academia and industry. Fake information detection in OSNs is mainly studied and discussed from two different aspects of both information content and social context auxiliary information. Research on fake information propagation on social media dated back to exploring dynamics models on the rumors

收稿日期：2020-01-10；在线发布日期：2020-08-21. 本课题得到国家自然科学基金 (No. 61972133, 61772174)、河南省中原千人计划中原科技创新领军人才项目（No.204200510021）、河南省科技创新杰出人才计划项目（No.174200510011）资助. **张志勇**（**通信作者**），博士，教授，博士生导师，河南省特聘教授，中国计算机学会(CCF)高级会员，主要研究领域为网络空间安全与人工智能、社交大数据分析与挖掘、可信计算与访问控制. E-mail: xidianzzy@126.com. **荆军昌**，博士研究生，中国计算机学会(CCF)会员，主要研究领域为社交网络安全、机器学习与深度学习. **李 斐**，硕士，高级工程师，主要研究领域为人工智能、分布式数据库. **赵长伟**，博士，讲师，中国计算机学会(CCF)会员，主要研究领域为人工智能、网络信息安全.

spreading in complex network and small-word network previously. Even within recent three years, some hybrid and interactive propagation pattern and behavior studies on both social human and social bots have been done. The diffusion control methods of fake information in OSNs mainly focus on the node control and access control/usage control. From two different angles of social object (fake information) and social subjects (social human and social bots), the research of fake information detection, propagation and control are discussed deeply and systematically, the related works are also analyzed and compared in this paper. Firstly, we comprehensively review some important and crucial research works of fake information detection from home and abroad in recent years, and especially focus on the unique characterizations (content features, social context features) and existing models (content models, social context models and hybrid models) of fake information detection. Owing to the difference between fake information and rumor in OSNs, we also briefly summarize the characteristics of rumor detection, which include content features, user features, topic features, propagation features, behavioral features and multimedia features. Secondly, research on the dissemination of fake information is primarily divided into two aspects: the dissemination of fake information based on social human and the propagation of fake information based on social bots. Here, three effective methods for the detection of social bots in OSNs are also discussed, which include graph-based approaches, crowdsourcing-based approaches, and machine learning-based approaches. Based on the analysis and comparison of the detection methods and propagation patterns and strategies of both social human and social bots, the general rules of spreading fake information of two kinds of social subjects are represented, respectively. Then, we systematically review and analyze the control methods of fake information dissemination from two levels: node control and usage control, and present a usage control model that applys to the research of fake information dissemination. Furthermore, the methods of data collection and annotation for fake information are systematically introduced, and some public online datasets that used to do research about detection, propagation and control from popular social media platforms, such as, Twitter, Facebook, Sina Weibo, are described. Finally, a novel social situation security and analytics framework that covers five layers (social entity layer, social environment layer, social behavior layer, social intention layer, social goal layer) and six elements (social object, identity, action, desire, environment, target) are proposed, and future research issues, challenges and possible research directions for cross-platform propagation and control of fake information are presented. We hope that social situation security proposed in this paper will provide theoretical basis, technical support and application scenarios for the realization of both virtual social cyberspace security and ecological governance of network information content.

Keywords Online Social Networks; Fake Information; Social Bots; Artificial Intelligence; Usage Control; Social Situation Security

1 引 言

随着移动互联网技术和 Web 4.0 的产生，以各类在线网络社区、在线社交工具、平台和服务为代表的社交网络生态系统得到飞速发展，数以亿计、十亿计的社交应用已经渗透到整个社会生产生活，成为人们获取海量信息资源的重要渠道[1-5]. 社交用户在获取信息的同时，也产生了大量的虚假信息，例如阴谋论（Conspiracy theories）、标题党（Clickbait）、伪科学（Peudo science），甚至捏造的“假新闻”（Fake news）等. 虚假信息[6-7]是指制造者故意误导读者，并能够通过一些其它来源证实其结果为假的信息，通常具有故意性(Intent)和可证实性（Verifiability）. 此外，2018 年 Vosoughi 等人[8]在 Science 杂志上也指出了虚假信息具有上述两个特性. 进一步地讲，虚假信息具有两层含义：1）信息具有一定的表面或片面陈述，不能客观反应出事物的本质. 例如，研究人员使用数据库检索系统，查

出表面相关而本质不相关的文献. 2）由于双方（多方）之间存在利益或竞争关系，为了达到一定目的而人为制造的不准确信息. 例如，宣传广告中故意夸大内容、股市领域的人为陷阱等. 由于传播媒介是虚假信息存在的必要条件，因此随着媒介的不断进化，在线社交平台对虚假信息的产生、传播和影响起到关键性作用，我们将此类信息定义为 OSNs 虚假信息. 谣言作为信息的一种存在形式，从谣言内容的真伪上分析，它是指尚未及时得到官方的证实，最终传播内容可判定为真、假和不确定. 综上所述，在线社交网络中的虚假信息与谣言具有一定联系，但又存在本质区别. 虚假信息具有故意性和可证实性特征，而谣言具有不确定性、时效性、主观性和关联性等[9]. 由于虚假信息具有一定的诱惑性，能够快速吸引用户的眼球，因此，随着社交平台的不断推广，虚假信息的传播变得愈演愈烈，已成为虚拟网络空间安全应用所普遍关注的热点和难点问题[10-13]. 大量虚假信息的大量传播，被认为是全球存在的一个重大风险，不仅影响了社交用户之间正常的信息共享和交流，也影响到经济社会发展，甚至国家安全和政治生活. 因此，当前研究社交网络虚假信息检测、传播和控制，对虚拟网络空间安全和治理意义重大，亟待深入开展. 我们通过调研文献发现，众多国内外大学和研究机构都对此进行深入研究，如麻省理工学院，印第安纳大学，亚利桑那州立大学，中国人民大学，国防科技大学等，其研究成果发表在 Nature、Science、ACM Transactions 系列和 IEEE Transactions 系列等国际顶级期刊和会议上.

当前，人工智能技术已上升至国家重要战略地位，大数据驱动下的机器学习、深度学习和神经网络等成为人工智能的核心技术，从弱人工智能到新一代的通用人工智能和强人工智能技术也持续不断地渗透到人们生产和生活的各个领域，研究人员正企图全面地了解智能本质，生产出一种新的能用与人类智能相似的方式做出反应的智能机器[14]. 尤其在社交网络领域，智能化的社交机器人（Socialbots）是由自动化程序控制的社交账号，它能够根据人为设定的程序，自动化执行相应的操作，从而模仿人类行为参与一系列 OSNs 活动[15]. 随着用户间频繁地通过 OSNs 进行互动与交流，大量社交机器人用户生成的海量内容充斥着在线社交平台[16-17]. 如图 1 所示，根据互联网公司 GlobalDots 公布的《2018 年机器人流量报告》，正常人在 2018 年在线流量占 62.10%，机器人流量已经达到 37.90%，其中恶意机器人占 20.40%，恶意机器人相对于正常机器人所占的比例较大[18]. Davis 等研究人员指出，社交网络平台中也存在着大量的恶意社交机器人，它们通过模仿正常用户的操作行为，获取用户个人的隐私信息，传播恶意虚假信息，甚至干扰全球政治选举活动，影响金融股市交易等，对网络平台的安全与稳定带来了极其严重的影响[19-21]. 例如，2016 年全世界目睹了社交机器人在美国总统选举期间传播虚假新闻的风暴[22]，2017 年德国联邦总统选举期间，社交机器人传播大量垃圾新闻的事件[23]，2017 年法国总统选举之前，社交机器人在 Twitter[24]上传播关于马克龙泄密的竞选文件等虚假新闻[25].

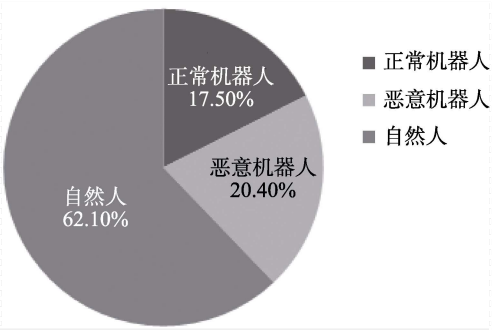

图 1　2018 年机器人流量报告[18]

OSNs 虚假信息的大量传播，不仅损害了网络媒体的公信力，也给社交用户的生产和生活带来了严重的影响. 因此，国家政府、学术界和工业界的研究人员对此广泛关注，他们分别从 OSNs 虚假信息检测、社交自然人和社交机器人传播模式等方面展开研究. 在虚假信息检测方面，早期传统的虚假信息检测方法主要是根据信息内容的真实程度进行判断，而随着 OSNs 不断发展，社交上下文辅助信息也成为研究人员关注的焦点. 针对虚假信息传播研究，可以追溯到早期复杂网络和小世界网络传播. 起初研究人员以谣言为研究对象，基于经典的流行病传播机理，研究谣言传播的动力学模型. 近三年来，随着大量社交网络平台应用和社交机器人的出现，研究人员开始转向对虚假信息传播主体的研究，即对社交用户和社交机器人的传播特征和传播模式研究[8,26-28]，其中一些研究学者对动态传播虚假信息的两类主体的行为特征，进行提取、分析，利用有效的分类或聚类算法将这两类主体进行区分，从而进一步研究社交用户虚假信息传播的潜在规律. 另外一些研究学者主要针对社交用户主体，利用社交网络统计性质，例如基尼系数、帕尔马比例等指标来研究社交用户虚假信息传播模式. 以上这些研究

工作使我们对 OSNs 虚假信息检测和传播的研究，有了进一步的认识和理解.

在虚假信息传播控制研究中，依据社交网络拓扑结构特性和语义情感分析等方面，主要从节点控制和访问控制入手展开研究. 虚假信息传播的节点控制主要分为基于时间戳的节点控制方法和基于影响力的节点控制方法，但这些方法最终都是采用反应式地删帖、封号或切断信息源头等补救措施. 然而，这样仍然存在虚假信息已在小范围内传播的可能，甚至会造成不可挽回的负面影响. 基于传统的访问控制技术，主要是从基于角色的访问控制模型、基于属性的访问控制模型和基于关系的访问控制模型展开，但是通常缺少对支持事前（传播前）、事中（传播中）的使用控制方法研究.

本文的其余部分组织如下. 第 2 节给出 OSNs 虚假信息检测关键问题描述；第 3 节分别从社交自然人和社交机器人两个角度，对虚假信息传播研究进行总结和比较；第 4 节对虚假信息传播控制相关研究进行综述；第 5 节分别从虚假信息数据集的采集和标注以及虚假信息公开数据集进行综述；接着，本文第 6 节对当前 OSNs 虚假信息传播和控制所存在的挑战进行总结，并给出未来的发展趋势；最后，我们在第 7 节对全文进行总结.

2 OSNs 虚假信息检测关键问题描述

虚假信息检测属于信息可信度检测研究范畴，它是研究虚假信息传播和控制的基础，旨在帮助社交媒体用户及时发现伪造不实信息，进一步提高社交媒体承载信息的生态环境质量. 在人工智能视角下，如何实现一套自动化程度高、鲁棒性强、可靠和高效的 OSNs 虚假信息检测的可计算方法，成为描述虚假信息检测研究的关键问题.

2.1 虚假信息检测问题定义

定义 1. $I=\{i_1,i_2,\cdots,i_{|I|}\}$ 是 OSNs 平台中待检测的虚假信息集合，类标签集合 $C=\{C_F,C_N\}$，C_F 表示虚假信息（Fake information）集合，C_N 表示正常信息（Normal information）集合. OSNs 虚假信息检测目的是判断待测信息 i_j 是否属于虚假信息集合 C_F，其决策函数为

$$\varphi(i_j,c_k):I\times C\to\{-1,+1\}(1\leqslant j\leqslant|I|,k\in\{F,N\}).$$

其中，

$$\varphi(i_j,c_k)=\begin{cases}+1,i_j\in C_F\\-1,i_k\in C_N.\end{cases}$$

定义 2. 待测虚假信息的特征向量可表示为 $\boldsymbol{a}^{(i)}=(b_1^{(i)},b_2^{(i)},\cdots,b_n^{(i)},\mathrm{c}^{(i)})$，其中 i 表示信息的 ID，n 表示特征数量，$b_j^{(i)}$ 表示关于信息 i 的第 j 个特征，$c^{(i)}$ 是信息 i 的类型（虚假信息或正常信息）.

2.2 虚假信息检测方法

虚假信息检测分为特征提取和模型构建两个阶段. 特征提取阶段是以形式化的数学结构来表示信息内容和社交上下文相关辅助信息. 模型构建阶段是进一步构建基于特征表示的信息内容模型、社交上下文模型和混合模型，来更好地检测虚假信息和真实信息.

2.2.1 特征提取

在 OSNs 虚假信息检测过程中，有效地提取虚假信息检测的关键特征，直接影响到后期模型构建的效果. 本节我们主要从虚假信息检测的内容特征和社交上下文特征两个方面进行阐述.

（1）内容特征

内容特征是由文本中提取的信息组成，包括信息的发布者、标题、正文内容中的文本、图片和视频等. 研究人员通常将内容特征分为基于语言的特征和基于视觉的特征[29]. 1）基于语言特征由词汇特征和句法特征构成. 词汇特征包括字符层面和词层面的特征，比如词的总数、每个词的字符数、词的频率和独特词汇等. 句法特征主要是指信息内容中句子层面的特征，例如，虚词和短语的频率、标点符号和词性标注等. 2）基于视觉的特征是从视觉元素（例如图像和视频）中提取虚假信息特征，包括清晰度得分、一致性得分、相似性分布直方图、多样性得分和聚类得分等.

（2）社交上下文特征

社交上下文特征主要由用户、用户发布（转发）的帖子和网络三个方面的特征构成. 1）用户特征分为个人层面用户特征和群组层面用户特征. 个人层面用户特征主要是通过运用用户各个方面的统计资料（例如注册时间、关注者数量/粉丝数量、已发布的推文数量等）来推断一个用户的可信度和可靠性. 群组层面用户特征是由个人层面用户特征通过平均和加权等方法计算得到. 2）用户发布（转发）帖子的特征主要从帖子层面、群组层面和时间层面进行考虑. 帖子层面的特征是对每一个帖子生成相应的特征值，群组层面的特征指在通过使用众包技术来聚合对于特定信息的所有相关帖子的特征值. 时间层面的特征是指帖子层面特征值随时间的变化. 3）基于网络的特征是从发布相关社交帖子的用户中，通过构建立场网络（stance network）、共存网络

（occurrence network）、朋友关系网络（friendship network）和扩散网络（diffusion network）等特定的网络进行特征提取[29-31].

此外，文献[32]在文献[33-35]提出的虚假新闻检测特征基础上，通过将内容特征和社交上下文特征相结合，进一步将虚假新闻的检测特征分为文本特征（例如，语言处理技术）、新闻源特征（例如，可靠性和可信度）和环境特征（例如，OSNs 结构）. 新闻源特征由新闻文章发布者的信息组成. 环境特征包括用户参与度（例如，点赞、分享、评论等交互行为）统计量和时间模式，如表 1 所示.

综上所述，我们对国内外研究人员常用的虚假信息检测特征进行系统总结. 由于虚假信息检测和谣言检测在特征选取上具有值得借鉴之处. 本小节最后，我们也对谣言检测特征进行简要总结，为今后研究学者更加精准选取虚假信息的检测特征提供参考[36-39]. 谣言检测常用的特征包括内容特征、用户特征、话题特征、传播特征、行为特征和多媒体特征等，具体如图 2 所示.

表 1　虚假新闻检测的常用特征列表[32]

类别	特征
文本特征	1）语言特征
	2）词汇特征
	3）心理语言学特征
	4）语义特征
	5）主观性特征
新闻来源特征	1）偏见特征
	2）可靠性和可行性特征
	3）域位置特征
环境特征	1）参与度
	2）时间模式

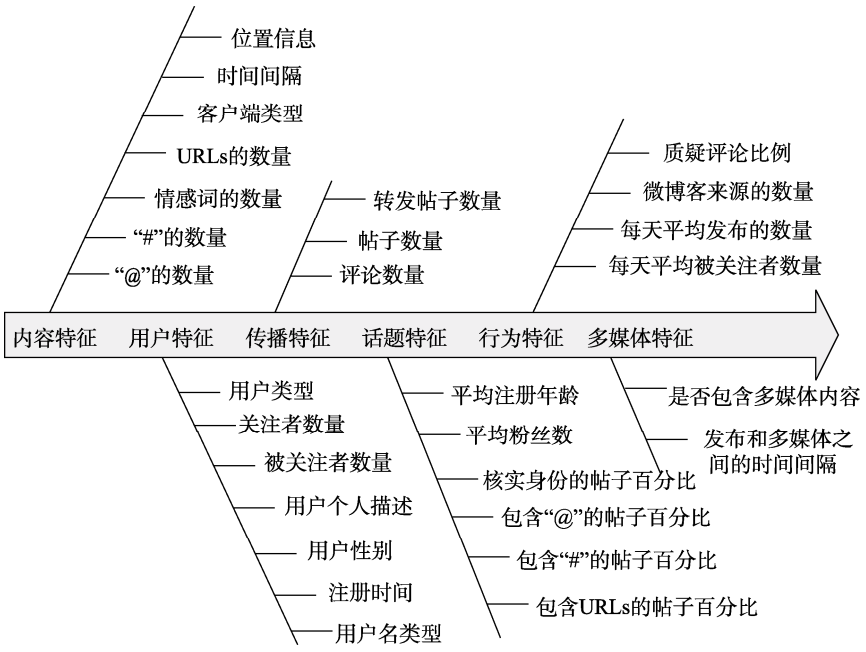

图 2　谣言检测常用的特征归类

2.2.2　模型构建

虚假信息检测的模型构建建立在内容特征和社交上下文特征的基础上，主要分为信息内容模型、社交上下文模型和混合模型.

（1）信息内容模型

现有的信息内容模型主要分为基于知识(Knowledge-based)的方法和基于风格（Style-based）的方法. 1)基于知识的方法旨在使用外部来源来核实信息内容的真实性. 现有的事实核查方法主要有面向专家(Expert-oriented)、面向众包(Crowdsourcing-oriented)和面向计算（Computational-oriented）的方法. 2）基于风格的方法是通过捕获操作者在信息内容中的写作风格来检测虚假信息，主要分为以欺骗性为中心（Deception-oriented）和以客观性为中心（Objectivity-oriented）的方法. 以文献[40-45]为代表的研究人员，通常采用群体智慧的众包方法来标注训练数据集，用于虚假信息的内容模型检测分析. 例如，Canini 等研究人员通过使用 MTurk 来建立充分真实的数据集(Ground truth data)，并强调收集真实的数据集是评估信息可信度的重要环节. 但是由于部分参与者可能会缺少一些相关技术知识，因此依靠群体智慧的众包方法可能会存在一些

偏见，从而影响信息可信度的判定[44-45]. 此外，Kumar 等人[7]使用认知心理学中相关内容来检测 OSNs 中虚假信息（Misinformation\Disinformation）和宣传（Propaganda）的传播，利用社交网络的协同过滤特性，提出了一种有效检测 OSNs 虚假信息恶意传播的算法，该算法对信息来源的可信度和新闻内容的质量进行检测，所提方法的有效性已在 Twitter 上进行验证.

（2）社交上下文模型

社交上下文模型主要基于用户的社交参与度进行构建，具体可分为基于立场（Stance-based）的方法和基于传播（Propagation-based）的方法[46-49]. ①基于立场的方法是利用用户对相关信息内容的观点，来推断用户发布信息内容的准确性. 通常，用户发布信息的立场主要分为显式和隐式两种，显式的立场是情感观点的直接表达，比如在社交网络平台上表达“赞成”、“反对”和“中立”等，隐式立场是从 OSNs 用户发布的信息中自动提取. 立场的检测是从用户发布信息的内容中自动判断用户是赞成、中立还是反对某个事件或观点. ②基于传播的虚假信息检测方法主要是通过分析用户在社交网络平台上发布相关内容的相互关系，建立同构和异构的可信度网络来预测信息的真伪.

（3）混合模型

混合模型首先是将选取的内容特征和社交上下文特征进行有效地组合，然后运用自然语言处理（Natural Language Processing，NLP）、机器学习和深度学习等先进技术，来预测待测信息是否为虚假信息. 随着研究人员对该领域的不断深入，混合模型的检测方法越来越受到青睐. 以文献[32,50-55]为代表的研究人员，进一步地将上述两类特征划分为文本特征、新闻来源特征和环境特征等，然后运用逻辑斯蒂回归（Logistic Regression, LR）、支持向量机（Support Vector Machine, SVM）、朴素贝叶斯法（Naïve Bayes, NB）、决策树（Decision Tree，DT）、K 近邻法（K-Nearest Neighbor, KNN）、随机森林（Random Forests, RF）和极端梯度提升（XGBoost, XGB）等经典机器学习模型和循环神经网络(Recurrent Neural Network, RNN)、卷积神经网络（Convolution Neural Network, CNN）以及图神经网络（Graph Neural Networks, GNN）等深度学习模型. 最后通过选取精度（Precision）、召回率（Recall）、ROC 曲线下的面积（AUC）和 F-score 等评估指标，进一步比较分类器的泛化能力.

在上述的混合模型中，研究人员主要使用有监督和半监督机器学习算法进行模型构建. 然而，无监督的机器学习算法还鲜有报道. 无监督学习算法包括 K-means 算法、模糊 C-means 算法和隐马尔可夫模型等. 在无监督学习方法中，Abbasi 等人[56]在无监督层次聚类算法的基础上，提出了一种 CredRank 算法. 首先根据式（1）和（2）计算用户的行为相似性，将相似用户聚为一个簇. 其次计算每一个簇中各个用户的权重，并依据权重的大小来判断每一个用户发布信息的可信度.

$$Sim(u_i,u_j)=\frac{1}{t_n-t_0}\sum_{t=t_0}^{t_n}\sigma(B(u_i,t),B(u_j,t)) \tag{1}$$

$$\sigma(B_i,B_j)=\frac{\left|B_i\cap B_j\right|}{\left|B_i\cup B_j\right|} \tag{2}$$

其中，$B(u_i,t)$是用户 u_i 在 t 时刻的行为，$\sigma(B(u_i,t), B(u_j,t))$分别表示用户 u_i和 u_j在 t 时刻的行为相似度.

综上所述，我们分别从特征选择和模型构建两个方面对虚假信息检测问题进行了详细的描述. 通过分析发现，信息内容模型、社交上下文模型和混合模型在虚假信息检测方面已经取得了比较理想的效果. 但是上述模型对虚假信息传播的早期检测并不完全适用，由于时间因素的限制，模型所需的内容特征和社交上下文特征选取不够充分，会导致检测虚假信息准确率比较低. Liu 等研究人员[57]通过对新闻传播路径进行分类，提出一种虚假新闻早期检测模型. 首先对每一条虚假信息建立关于多变量时间序列的传播路径，其次通过递归神经网络和卷积神经网络建立关于时间序列的分类器，分别捕获传播路径上用户特征的全局变量和局部变量. 实验结果显示，在 Twitter 和微博社交平台中，从发布到传播的前 5 分钟，检测准确率分别为 85%和 92%.

3 OSNs 虚假信息传播关键问题描述

随着人们逐渐将社交媒体作为获取信息的主要渠道，社交媒体中虚假信息的传播极大地影响了 OSNs 信息生态系统的质量和用户体验效果. 当前研究人员重点围绕社交自然人和社交机器人两类传播主体展开研究. 在传播模型的研究中，主要针对传播动力学模型、独立级联模型和线性阈值模型等. 在传播行为模式的研究中，社交主体主要是通过发布、转发、提及、评论等多种混合式行为方式进行虚假信息传播[26,58-61]. 图 3 分别从虚假信息传播主体、传播客体、传播媒介和传播模型等方面，展现出早期复杂网络和小世界网络以及近 5 年来关于社交用户和社交机器人的混合型、交互式传播现状.

时间轴	早期研究工作 (20世纪60年代-90年代)	近5年研究工作 (2015年至今)
传播主体	社交自然人	• 社交自然人 • 社交机器人 • 半社交机器人
传播客体	谣言	• 谣言 • 虚假信息和真实信息 • 不同可信度等级的新闻来源信息
传播媒介	• 口耳相传 • 传统大众媒体	• 单一的社交平台 • 跨平台(两个或多个社交平台)
传播模型	基于SI、SIS和SIR等传染病模型	• 基于SI、SIS和SIR等传染病模型 • 基于社交网络统计性质传播模型 • 跨平台虚假信息传播模型

图 3　虚假信息传播研究进展和分类图

3.1　基于社交自然人的虚假信息传播

社交自然人是 OSNs 虚假信息传播主体的重要组成部分. 社交自然人虚假信息传播研究可以追溯到早期复杂网络和小世界网络中的谣言传播动力学模型研究[62-63]. 在 OSNs 中，虚假信息传播与社交自然人的兴趣、身份、工作和生活密不可分. 由于社交自然人相互之间关系的复杂性、灵活性和多变性以及虚假信息之间存在关联性等特征，造成以社交自然人为主体的虚假信息传播研究面临着若干挑战. 如何运用复杂的人工智能技术和统计学分析方法来研究社交自然人虚假信息传播模式成为研究的关键问题. 表 2 分别从社交自然人传播虚假信息类型、采用的模型（方法）和评估指标等方面进行对比分析.

3.1.1　基于 SIR 等传染病传播模型研究

基于 SIR、SIS 等传染病的传播动力学模型是早期研究虚假信息传播的主要模型. 研究人员通常以单一类型信息作为研究对象，分别从易染状态（Susceptible, S）、感染状态（Infected, I）、免疫状态或恢复状态（Recovered, R）等维度来刻画虚假信息的传播过程，并对模型中的参数进行分析. 随着信息传播种类的丰富，一些正面信息（例如，官方发布的新闻和观点等）和负面信息（例如，谣言、流言蜚语等）的混合式传播，不断影响着社交平台

表 2　社交自然人信息传播研究对比分析

文献	信息类型	模型/方法	评价指标
[60]	信息	基于用户属性、社交关系和微博内容等特征，预测用户的转发行为 基于概率级联模型，预测微博转发路径	查全率和查准率 F1 度量和预测精度
[64][65]	谣言	基于引力学思想提出一种谣言传播分析模型 GRPModel 基于随机过程的传播动力学 SIS 模型	用户节点影响力 谣言的影响力 用户对谣言的接触率
[59][66]	正面信息（新闻、观点等）和负面信息（谣言、流言蜚语）	提出正面信息和负面信息混合式传播动力学模型	传播信息的延迟时间 判断用户接收信息时状态（乐观悲观）的准确度 判断用户偏好和拒绝倾向的准确度
[57][67][68]	真实信息和虚假信息	利用递归神经网络和卷积神经网络建立信息传播路径模型	信息传播速度； 虚假信息检测速度； 虚假信息检测的准确率、精度、召回率和 F1 度量
[8][26][69][70]	不同可信度等级的新闻来源信息	分别从人口群体和人口子群体两个方面，评估分析传播不同可信度新闻源的用户分布均匀程度、用户分布数量、用户分享不同来源信息的快慢程度和分享的用户特征等	洛伦兹曲线和基尼系数 密度和边节点比 信息级联的互补累计分布函数等
[71]	舆情事件	提出关于信息传播者集合 P、信息接收者集合 R、信息内容集合 C、传播媒介集合 M 和信息传播效果评价 A 的 PRCMA 多元信息传播模型	传播过程的质量 传播者信息反馈

信息传播的质量. 由于正面信息和负面信息通常包含与人相关的因素，因此这两类信息同时传播，不能认为是两个独立传播过程的叠加. 针对正面信息和负面信息的混合式传播，研究学者主要运用独立级联模型（Independent cascade model，ICM）和线性阈值模型（Linear threshold model，LTM）展开研究，但这两类模型都属于仿真模型，不能考虑时间动力学因素对传播模型的影响，因此不适用于真实的社交网络平台.

针对上述模型存在的不足之处，Wen 等人[59]提出了一种正面信息和负面信息混合式传播的分析模型. 该模型既呈现了传播动力学特性，也呈现了人们在接收到这两种信息时做出选择的行为. 在分析模型的基础上，进一步研究了参数对传播动力学模型的影响. 该研究结果证明了通过传播正面信息来抑制负面信息是抑制虚假信息传播的一种有效策略. 文献[65]利用传播速度、易受骗性、验证为恶作剧谣言的概率和忘记当前信任的概率 4 个因素，建立一种基于随机过程的传播动力学 SIS 模型，来研究恶作剧的传播情况.

3.1.2 基于社交网络统计性质的虚假信息传播研究

由于传统信息来源的真实性和可靠性不断受到人们的质疑，一些研究人员依据信息源的质量，将信息源划分为不同的类别[26, 67-70]. Glenski 等人[26]将新闻信息的来源分为可信信息(Trusted)、敲击诱饵(Clickbait)、阴谋论(Conspiracy theories)、宣传(Propaganda)和虚假信息(Disinformation)等 5 类. 针对 Twitter 上的 1100 万条帖子，分别从人口群体和人口子群体两个方面，利用洛伦兹曲线、基尼系数、密度、边节点比等社交网络统计性质，评估分析传播上述 5 类不同可信度新闻源的用户分布均匀程度、用户分布数量、用户分享不同信息来源的快慢程度和分享的用户特征等. 分析结果发现一小部分高度活跃的用户负责大部分虚假信息的传播，年收入和受教育程度较低的用户相对于其他用户会分享更多的虚假信息，年龄大的用户比年轻用户分享可信新闻来源更快，但对于可疑新闻来源内容的分享，年龄大的用户会比年轻用户在时间上分享的更迟，分享敲击诱饵和阴谋论新闻来源的用户更有可能分享宣传新闻来源的用户分享的内容. Vosoughi 等研究人员[8]在 Science 期刊上发表论文声称，通过研究 2006 年至 2017 年在 Twitter 上发布的所有已证实真实和虚假的新闻内容传播情况，发现在所有的信息种类中，虚假新闻比真实新闻传播得更远、更快、更深和更广泛. 对于虚假的政治新闻，其影响比有关恐怖主义、自然灾害、科学、城市传说或金融信息等更为明显. 由于虚假新闻比真实新闻更新颖，人们更愿意分享新颖的虚假信息. 在传播过程中，虚假新闻激发了人们的恐惧、厌恶和惊讶，真实新闻激发了人们的期待、悲伤、喜悦和信任. 与传统的观点相反，社交机器人以同样的速度加速了真实和虚假新闻的传播，这意味着虚假新闻比真实新闻传播得更多，因为人（而不是机器人）更有可能传播虚假新闻. 文献[68]通过分析 Twitter 上 2016 年美国总统选举的 40 条新闻（20 条真实新闻和 20 条虚假新闻）在传播模式上的差异，发现随着时间的推移，虚假新闻的传播数量不断增加，真实新闻的传播数量急剧下降. 这充分表明虚假新闻的传播是可持续的，可以达到更广泛的传播范围，但该文献没有对传播结果进行详细的理论解释. Glenski 等人[69]针对社交机器人和社交自然人对不同可信度新闻源信息做出的反应，发现社交机器人对不同可信度的新闻来源的反应存在差异，通过使用细粒度模型分析社交机器人和社交自然人对新闻来源的响应，并标记为回答、欣赏、同意、不同意、阐述、幽默或消极反应. 验证发现对可信的新闻来源做出反应时，社交自然人占绝大多数. 当社交机器人对可信度为宣传新闻来源信息做出反应时，反应延迟比人要更短. 文献[70]使用 Twitter 和 Reddit 两个社交媒体平台来研究用户对可信性和欺骗性的新闻来源反应，提出了一个基于内容和语言的神经网络模型，将用户对新闻来源的反应分为九种类型，例如回答、阐述和怀疑等，通过使用 1080 万条 Twitter 帖子和 620 万条 Reddit 评论来测量用户对可信性和欺骗性新闻来源的反应速度和反应类型. 研究结果发现 Twitter 平台上的用户对可信和虚假新闻来源的反应速度和类型上存在显著差异，但在 Reddit 平台上的差异要小得多.

3.1.3 基于跨媒介的虚假信息传播研究

针对网络舆情信息跨媒介的传播研究，Rao 等人[71]分别从信息传播者集合 P、信息接收者集合 R、信息内容集合 C、传播媒介集合 M、信息传播效果评价 A 等 5 个要素，提出了 PRCMA 多元信息传播模型，并将该模型形式化地表示为 IP={P, R, C, M, A}，图 4 具体给出 PRCMA 元组的特征及其关系.

3.2 基于社交机器人的虚假信息传播

社交机器人的一个重要功能是对社交平台信息的交流和共享，随着社交平台用户数量的持续增多和智能化社交机器人行为复杂性增加，社交机器人在信息传播方面扮演着不可替代的角色，同时也吸

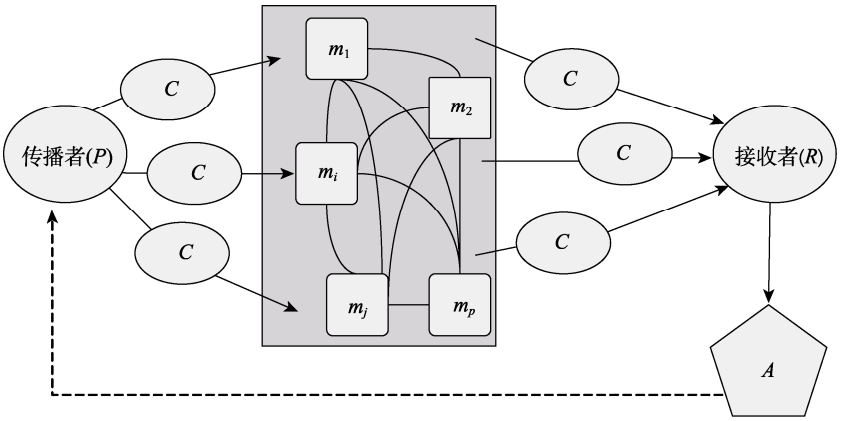

图 4　跨媒介舆情信息传播 PRCMA 多要素模型[71]

引了众多研究人员对社交机器人传播虚假信息的机制进行深入系统研究，为社交网络平台上虚假信息传播控制提供了有效的保障. 社交机器人检测是研究社交机器人传播规律的基础，只有充分准确地检测出社交平台上已有的社交机器人账号，才能够进一步地研究社交机器人传播虚假信息的规律. 因此，本节余下部分，分别从社交机器人检测方法、社交机器人虚假信息传播特征的研究现状和发展趋势进行总结分析.

在社交媒体网络中，区分社交机器人和社交自然人是社交机器人检测的目的之一，研究精准有效的社交机器人检测方法，有助于及时控制社交机器人传播大量的虚假信息. 目前，社交机器人检测的常用方法主要分为基于图的方法、基于众包的方法和基于机器学习的方法.

（1）基于图的社交机器人检测方法

在社交网络平台中，社交网络图直观反应了社交自然人之间的关系. 研究人员通常采用基于信任传播、图聚类和图的一些度量和性质等方法来检测社交机器人账号. 其中基于信任传播的方法主要是通过评估两个社交图信任关系的强弱来进行判定，图聚类方法主要是利用用户之间的距离等相似特性，对社交图的相关节点进行分类，图的度量和性质主要包括概率分布、无标度图结构和中心性等[72-75]. 文献[73]提出了一种基于随机游走的 SybilWalk 方法进行恶意社交机器人检测，分别使用标记有正常用户和恶意社交机器人标签的两个额外节点来扩充社交图，依据随机游走的概率来判断账户为恶意社交机器人可能性. Mehrotra 等研究人员[74]通过选取社交图中节点中心性的六个特征，运用人工神经网络、决策树和随机森林三种算法来检测社交平台中虚假的关注者，实验结果表明随机森林算法具有最好的泛化能力，准确率达到 95%.

（2）基于众包的社交机器人检测方法

基于众包的社交机器人检测方法主要是通过选取相关技术人员查看社交平台中给定账户的个人资料以及分享内容，来区分社交自然人和社交机器人[76-78]. Alarifi 等研究人员[76]使用人工标注的方法，通过招募一批志愿者，对 2000 个随机账户进行评级和标注，并评估标记数据的真实性和可靠性，结果表明在标记过程中准确率达到 96%. 虽然众包方法在检测社交机器人准确率方面已经比较理想，但仍存在一些缺陷. 首先，当数据集中样本比较大时，需要雇佣大量的相关技术人员，从而提高了检测的成本；其次，在标注过程中用户的一些个人隐私信息，可能会暴露给外部的工作人员，从而造成个人隐私泄露；最后，由于众包用户在执行任务过程中采用匿名的方式，对众包工作者的相关技术水平和能力进行审核时缺少明确的标准，从而可能会出现一部分用户为获得报酬而不认真完成任务，影响检测结果的准确性[78].

（3）基于机器学习的社交机器人检测方法

基于机器学习的社交机器人检测方法，首先对社交机器人和社交自然人两类主体的传播特征进行分析，然后采用机器学习中的分类和聚类算法开展研究. 表 3 分别从检测特征、模型、评价指标和数据集等方面，系统性地给出国内外研究学者关于社交机器人检测的效果比较. 针对社交机器人的检测特征研究，主要从网络特征，用户特征，交友特征，时间特征，内容特征和情感特征等 6 类静态和动态特征入手展开研究，其中网络特征是从不同维度获取信息的传播模式，通过转发、提及等操作建立网络，并从中提取出节点度分布、聚类系数和中心性等一些统计特征；用户特征包括语言、地理位置和

账号创建时间等；交友特征包括与帐户社会联系相关的描述性统计数据，例如账户的关注者、被关注者和帖子等数量分布的中位数、时间和熵；时间特征包括内容生成和消费的时间模式，如发布推文的速率和连续发布两个推文之间的时间间隔等；内容特征是基于通过自然语言处理得到的语言线索，尤其是词性标注等；情感特征是使用一些通用的和Twitter特有的情感分析算法，包括幸福感、情绪化分数等[15,19,28]. 针对分类和聚类算法，主要采用随机森林、支持向量机、聚类算法、深度学习等机器学习方法，来区分两类社交主体[28,79-85]. 综上所述，我们只有通过选取有效的分类或聚类方法，来准确地区分这两类社交主体，才可以及时删除社交平台中的恶意机器人账号，使得OSNs生态系统更加安全、可信和可控. 图5给出基于机器学习的社交机器人检测框架.

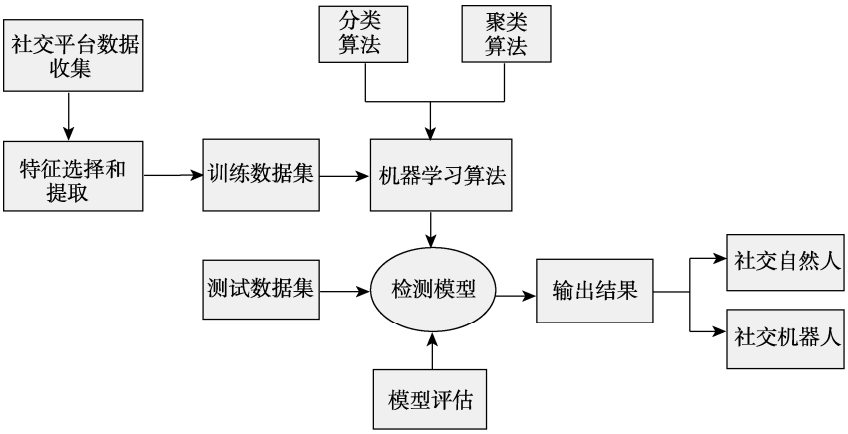

图5 基于机器学习的社交机器人检测框架

表3 社交机器人检测效果比较

作者	模型	特征						评价指标		数据集
		网络	用户	交友	时间	内容	情感	准确率	F1	
Varol 等人[28]	RF	✓	✓	✓	✓	✓	✓	90%		Twitter
Alarifi 等人[76]	SVM	✓	✓	✓	✓	✓		93%	93%	Twitter
Morstatter 等人[81]	AdaBoost	✓	✓			✓		79.76%	75.91%	Twitter
Costa 等人[82]	Act-M				✓			96.5%		Reddit
					✓			94.7%		Twitter
Jr 等人[83]	Wavelet, RF	✓	✓	✓		✓		94.47%		Twitter
Fazil 等人[84]	RF	✓	✓	✓				79.7%	78.91%	Twitter
Shi 等人[85]	K-means[86]	✓	✓		✓			93.1%	95.2%	CyVOD[87]
Cai 等人[88]	BeDM	✓		✓	✓	✓		88.41%	87.32%	Twitter
	BeDM					✓		83.49%	84.11%	Twitter
Ping 等人[89]	CNN, LSTM	✓			✓	✓		98.6%	98.1%	Twitter
Sneha 等人[90]	LSTM	✓	✓	✓				96%	96%	Twitter
Cai 等人[91]	CNN, LSTM	✓	✓			✓		87.58%	88.30%	Twitter
	Boosting		✓	✓		✓		85.23%	84.77%	Twitter
	BoostOR	✓			✓	✓		83.16%	86.10%	Twitter
Clark 等人[92]	NLP	✓	✓			✓		90.32%		Twitter
Walt 等人[93]	RF	✓	✓	✓		✓		87.11%	49.75%	Twitter

注：在特征列表中“✓”表示该文献使用对应的特征，空白表示该文献未使用对应的特征. 在评价指标列表中，若文献对应的F1指标空白则表示该文献未使用该项评价指标.

Morstatter 等人[81]通过选取用户转发推文数量在发布推文数量中所占比例、用户发布推文的平均长度、URL和连续两次转发时间间隔等特征分别如(3)-(6)式，提出一种增加召回率启发式的有监督学习BoostOR模型，该模型通过评估准确率和召回率之间的关系，来检测平台中存在的恶意机器人. 实验结果显示该方法具有较高的精度，以便于研究人员可以从他们的社交媒体数据集中删除更多的社交

机器人账户，从而关注真实用户产生的信息.

$$Retweet(u)=\frac{\left|\left\{x\middle|x\in tweets^{u},x \text{ is reweet}\right\}\right|}{\left|tweets^{u}\right|} \tag{3}$$

$$Length(u)=\frac{\sum_{i}^{\left|tweets^{u}\right|}\left|tweets_{i}^{u}\right|}{\left|tweets^{u}\right|} \tag{4}$$

$$URL(u)=\frac{\left|\left\{x\middle|x\in tweets^{u},\ x\ contains\ URL\right\}\right|}{\left|tweets^{u}\right|} \tag{5}$$

$$Time(u)=\frac{1}{\left|tweets^{u}\right|-1}\sum_{i=2}^{N}(t_i-t_{i-1}) \tag{6}$$

文献[82]基于用户行为活动时间，建立关于时间活动的数学模型 Act-M（Activity Model, Act-M），该模型通过拟合社交媒体用户不同行为的时间间隔分布，从而更加准确地检测社交媒体中的恶意用户. 文献[83]提出一种基于小波的模型，来检测 OSNs 中信息传播主体. 该模型根据用户文本内容得到频谱图，并从离散小波变换和基于词法的系数衰减加权方案中创建特征向量，最后使用随机森林算法将用户分为正常人、合法机器人和恶意机器人. Fazil 等人[84]首先根据 OSNs 中 Twitter 用户与社交机器人的交互行为特征，将 Twitter 用户分为活跃、被动和不活跃用户. 其次利用 Twitter 中活跃、被动和不活跃用户的性别、年龄、位置等静态特征和用户交互的人、交互内容、交互主题等动态特征，运用朴素贝叶斯、减少误差修剪决策树和随机森林 3 种机器学习方法对社交自然人进行分类. Shi 等人[85]通过利用社会情境分析理论和“点击流”序列，选取转移概率和行为时间间隔作为特征，运用半监督 K-means 聚类算法[86]来检测 CyVOD[87]社交平台中的恶意社交机器人.

随着大数据、云计算时代的到来，深度学习引发了新一代的人工智能全球化浪潮. 在社交机器人检测过程中，研究人员通过提取社交自然人和社交机器人的传播行为特征和内容特征，运用深度学习模型来区分这两类主体. 深度学习模型由深层的神经网络构成[94]. 循环神经网络(Recurrent Neural Network, RNN)是一种前馈神经网络，在对句子或时间序列等可变化长度的序列信息建模方面具有优势[95]. 由于社交自然人的转发行为可以看作是一个时间序列，在两个连续的转发信息之间没有固定的时间间隔，并且每一条信息源的转发序列可以有不同长度，因此研究人员通常选取 RNN 作为分类模型. RNN 形式化定义如下:输入序列为 $(x_1,x_2,\cdots,x_T)$，模型更新的隐藏状态为 $(h_1,h_2,\cdots,h_T)$，输出向量为 $(o_1,o_2,\cdots,o_T)$，其中 T 为输入的长度. 从 1 到 T 按如下的方程进行迭代：

$$h_t=\tanh(\boldsymbol{U}x_t+\boldsymbol{W}h_{t-1}+\boldsymbol{b}) \tag{7}$$

$$o_t=\boldsymbol{V}h_t+\boldsymbol{c} \tag{8}$$

其中 $\boldsymbol{U}$, $\boldsymbol{W}$ 和 $\boldsymbol{V}$ 分别是输入层到隐藏层、隐藏层到隐藏层、隐藏层到输出层的权重矩阵，$\boldsymbol{b}$ 和 $\boldsymbol{c}$ 为偏置向量，$tanh(\ )$为双曲正切非线性函数. 在隐藏层中门控循环单元(Gated Recurrent Unit, GRU)按如下的方程：

$$z_t=\sigma(x_t\boldsymbol{U}_z+h_{t-1}\boldsymbol{W}_z) \tag{9}$$

$$r_t=\sigma(x_t\boldsymbol{U}_r+h_{t-1}\boldsymbol{W}_r) \tag{10}$$

$$\tilde{h}_t=\tanh(x_t\boldsymbol{U}_h+(h_{t-1}\cdot r_t)\boldsymbol{W}_h) \tag{11}$$

$$h_t=(1-z_t)\cdot h_{t-1}+z_t\cdot\tilde{h}_t \tag{12}$$

其中重置门 r_t 决定如何将新的输入与以前的内存进行组合，更新门 z_t 定义了将以前的内存级联到当前时间步长的大小，$\tilde{h}_t$ 表示隐藏状态 h_t 的候选激活状态[96].

Cai 等研究人员[88]提出一种行为增强的深度模型(BeDM)，首先将用户内容视为时间文本数据以提取潜在的时间模式，其次提出由卷积神经网络(CNN)[97]和长短期记忆网络（LSTM）模块[98]组成的深度学习框架，将用户内容信息和行为信息进行融合来检测社交机器人. 实验结果表明该方法准确率达到 88.41%、召回率为 86.26%、F1 为 83.32%. 此外，Ping 等人[89]提出了一个基于深度学习算法（DeBD）的社交机器人检测模型，该模型由内容特征提取层、推文元数据时间特征提取层和内容时间特征融合层构成. 在内容特征提取层，运用 CNN 来提取用户之间发布内容的关系. 在推文元数据时间特征提取层，运用 LSTM 来提取发布元数据的潜在时间特征. 在内容特征融合层，将时间特征与内容特征融合，实现对社交机器人的检测. 文献[90]通过选取用户内容和元数据作为检测特征，提出了一种基于上下文长短期记忆（LSTM）架构的深度神经网络进行社交机器人检测. 实验结果表明，该架构将社交机器人和社交自然人分离时可以实现高分类精度（AUC>96%）. 图 6 给出基于深度学习方法的社交机器人检测，研究人员常用的检测特征和检测模块.

在 2016 年美国总统大选政治事件中，社交机器人通过社交媒体平台传播低可信度来源新闻受到研究人员的广泛关注. 国防科技大学邵成成和美国印第安纳大学 Filippo Mencze 等研究人员[27]在世界顶级学术期刊《自然・通讯》上发表相关研究成果，

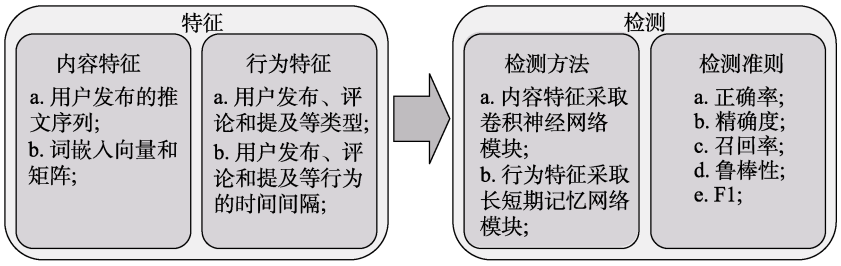

图 6 基于深度学习模型的社交机器人检测框架

该文献分析了在 2016 年美国总统大选期间（2016 年 5 月中旬至 2017 年 3 月）Twitter 上 1400 万条推文和转发的 40 万篇文章，发现社交机器人经常会在低可信度来源（经常发表各类错误信息的网站，这些网站需由信誉较好的第三方新闻和事实核查组织确认）的文章发表后和疯传前进行大量传播. 社交机器人还会通过回复和提到功能，将目标指向那些粉丝众多的有影响力的用户. 这种策略之所以取得成功，是因为人类比较容易受到机器人操控的影响，进而转发一些社交机器人发布的内容. 并且还通过分析发现，如果在研究期间封禁一小部分（约 10%）最像社交机器人的账号，几乎能消除低可信度内容链接的传播. 文献[99]也针对 2016 年美国总统大选，研究 Twitter 上社交机器人虚假新闻传播的事实. 发现社交机器人在虚假新闻传播的早前尤为活跃，并且更倾向于针对有影响力的用户，从而使得虚假新闻广泛地被分享. 因此，人们也更容易受到社交机器人发布虚假新闻的影响. 此外，Gilani 等学者[100]通过在 Twitter 上设置一个社交机器人帐户，并从 Web 服务器上对社交用户点击日志数据集进行分析，结果表明尽管社交机器人的数量比较少，但它们对社交平台上的内容流行度和活动产生了巨大的影响.

4 OSNs 虚假信息控制关键问题描述

在社交网络平台中，按照虚假信息传播的时间戳将传播控制分为传播前、传播中和传播后控制. 通过调研发现，当前大部分社交平台中虚假信息传播只能进行传播中和传播后控制，从而采取“封号”、“禁言”、“删帖”等一系列被动落后的方式来控制虚假信息传播. OSNs 访问控制是保护社交网络信息安全的重要组成部分，也是保护信息能被合法用户访问和防止信息泄露的一项关键技术. 文献[101]提出了 OSNs 访问控制体系结构，如图 7 所示. 该体

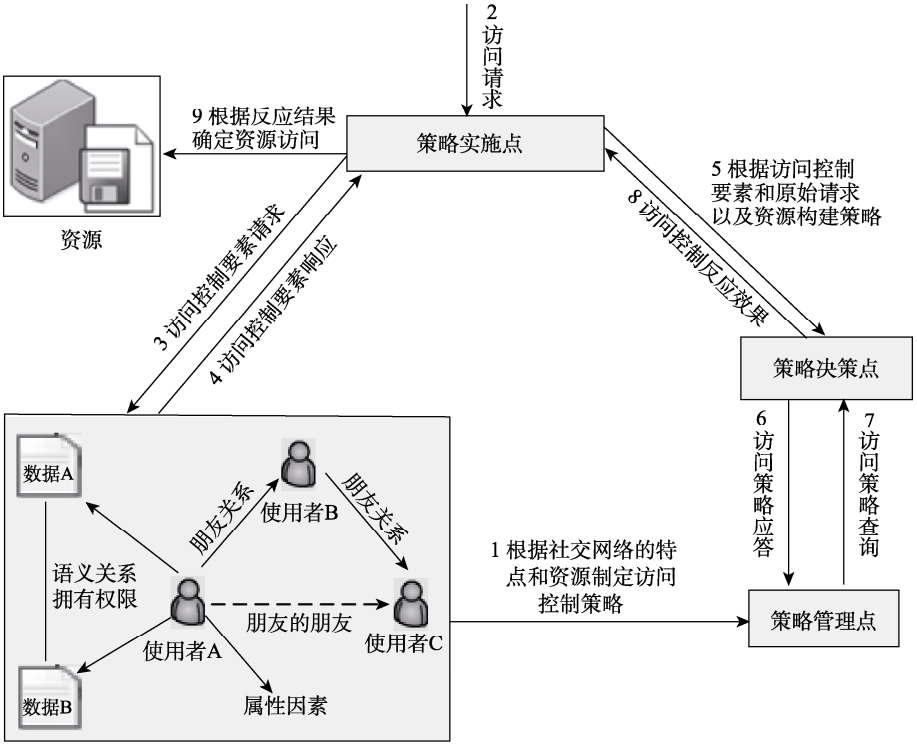

图 7 OSNs 访问控制体系结构[101]

系结构可适用于社交网络虚假信息传播控制应用场景，从而有效地解决虚假信息传播中的访问控制问题. 在人工智能视角下，如何将新一代人工智能技术与传统社交网络访问控制和使用控制技术相结合，实现虚假信息传播前的有效控制，是当前社交网络空间安全亟待解决的开放问题. 下面将从OSNs虚假信息传播节点控制方法、访问控制方法和使用控制方法三个维度，分别对当前的相关研究文献进行综述.

4.1　虚假信息传播的节点控制

虚假信息传播控制方法的研究，通常利用时间因素、拓扑结构关系和语义情感分析等. 主要分为两大类，（1）基于时间戳的节点控制方法，该方法直接采用封号等方式来控制传播虚假信息的账号，它的弊端在于没有考虑节点之间的相互影响，从而不能快速、有效地控制一些具有影响力的社交账号；（2）基于影响力的节点控制模型是通过运用PageRank等排序算法，识别出具有一些高度影响力的节点，该方法的弊端没有考虑虚假信息语义之间的拓扑关系，忽视了节点之间的相互影响程度. 针对上述两类方法的不足之处，文献[102]提出一种基于溯源的虚假信息传播控制策略，来及时删除大量优质源头节点和具有高影响力的节点，从而取得很好的虚假信息传播控制效果.

Wang等人[103]通过分析OSN中舆情传播的网络拓扑特征、用户网络地位的不对称性、社会强化效应、用户感知价值和信息的时效性等因素，提出一种基于用户相对权重的舆情传播控制模型，并通过仿真实验对结果进行分析得出网络舆情的传播与初始传播源的网络地位具有密切的关系. 文献[104]基于搜索引擎中的PageRank算法，提出了一种虚假信息传播控制方法 Fidic，该方法与随机传播控制算法、基于出度的传播控制算法和基于入度的传播控制算法相比具有很高的预测精度. He等人[61]提出一种基于异构网络的流行病传播动力学模型，该模型通过运用实时优化策略和脉冲式传播真实信息同时连续阻断谣言的策略，来阻止移动 OSNs 中谣言的传播.

4.2　虚假信息传播的使用控制

访问控制模型主要分为基于角色的访问控制模型、基于属性的访问控制模型和基于关系的访问控制模型. 这些模型分别将角色、属性和关系作为主要元素来控制对信息的访问. 在 OSNs 信息分享过程中，基于角色的访问控制通常利用多重关系、关系强度、方向关系、用户到用户的关系和用户到资源的关系等来控制信息的传播. 基于关系的访问控制根据社交用户之间的各种关系进行授权访问，来实现社交用户对资源的传播控制，提高了信息共享的安全性. 表 4 对 OSNs 中具有代表性的访问控制模型，进行对比分析.

表 4　OSNs 访问控制模型对比分析

模型或思想	代表性文献	特点
基于关系的访问控制模型	[105-107]	文献[105]和文献[106]分别基于用户间关系以及用户、用户间关系和公共信息，提出 OSN 访问控制模型，文献[107]提出基于属性访问控制的关系访问控制模型
基于属性的访问控制模型	[108]	提出用户访问控制规则、数据流控制规则和安全策略冲突消解的方案
基于群组的访问控制模型	[109-110]	将基于属性的访问控制与信息流策略相结合，实现对群组内和群组间分享信息时的控制
面向网络空间的访问控制模型	[111]	通过对网络空间中主体和客体进行概括，提出基于场景的访问控制模型
基于加密的访问控制模型	[112]	提出一个隐藏在群体中（Hide In The Crowd，HITC）系统,该系统给 OSN 平台中用户发布的每个数据客体分配解密特权，加强对共享数据的细粒度访问控制

国外Pang等人[106]针对OSNs的访问控制问题，从现有的访问控制方案中归纳确定了 OSNs 访问控制新需求，进而从用户可调节资源访问的角度，关注于社交媒体的公共信息安全，提出了一个包含用户、用户间关系和公共信息的 OSNs 新模型，并采用混合逻辑（Hybrid logic）形式化描述了主要访问控制策略. Bui 等人[107]首先将基于关系的访问控制定义为基于属性访问控制的面向对象扩展，其中关系是引用其他对象的字段表示，路径表达式用于跟踪对象之间的关系链. 其次，提出了两种从访问控制列表和以对象模型表示的属性数据中挖掘基于关系的访问控制策略算法，分别是启发式引导的贪婪算法和基于语法的进化算法. 文献[108]通过运用多媒体社交网络中的用户属性、环境属性和资源属性等，建立基于属性的访问控制模型，并对模型进行形式化描述，提出了一些访问控制规则、数据流控制规则和安全策略冲突消解的方案，最后将该模型应用到 CyVOD 社交平台上，实现对平台资源安全

可控访问.

为了防止用户在群组内或群组间共享信息时出现隐私泄漏，造成信息被恶意攻击者获取，Hu等人[109]在以群组为中心的安全信息共享（G-SIS）模型的基础上，提出了一个正式的基于群组访问控制（oGBAC）框架，该框架通过将群组运用到OSN，并对群组和群组之间的信息流施加一些限制，确保在OSN中与朋友共享信息时，用户的操作不会导致隐私泄露. 文献[111]提出一种面向网络空间的访问控制模型（Cyberspace-oriented access control model），其典型使用场景是用户通过网络利用移动设备访问具有时间和空间特性的敏感客体. 通过对网络空间中主体和客体进行概括，提出基于场景的访问控制模型. Ma 等人[113]指出普适社交网络支持在线、即时的社交活动和通信，移动社交用户通常从中得到一些有价值信息，但是也面临着一些恶意内容的分享. 基于此，提出一种基于信任管理的控制器系统PSNController，并在大量的恶意内容入侵和攻击场景下，进一步评估该控制器系统性能.

使用控制模型作为下一代访问控制模型的基础，具有决策连续性和属性可变性的显著特性. 使用控制系统是由主体及其主体属性、客体及其客体属性、权限、授权、职责和条件6种成分组成，其中授权、职责和条件是使用控制决策的组成部分. 图8给出OSNs虚假信息传播的使用控制模型结构图.

目前，国内外已经对此开始了广泛而深入的研究工作. 文献[114]在 Lamport 提出的行为时态逻辑（Temporal logic of actions, TLA）扩展形式的基础上，提出了使用控制的一个形式化模型和逻辑规范. 此模型的构建模块包括基于主体、客体和系统属性的一组系统状态序列、基于主体和客体属性的授权谓词、用于更新属性的使用控制操作和使用过程的访问状态和基于系统属性的职责动作和条件谓词. 使用控制策略定义为满足系统状态变化的一组时态逻辑公式. Wu 等人[115]研究了工业系统中社会网络无线传感器跨域细粒度数据使用控制机制，包括跨域细粒度访问控制和用于传感数据高效分析的模糊聚类. 此外，针对数据的使用提出了动态服务组合. 文献[116]提出了一种基于 Web 的社交网络表达性使用控制模型 $SoNeUCON_{ABC}$，该模型扩展了包括关系管理的 $UCON_{ABC}$，指定了相关的实体和元素以及访问控制策略语言. 此外，通过使用正则表达式对策略构造进行了详细描述，同时也对访问控制执行单元进行描述.

5 OSNs 虚假信息数据采集和描述

5.1 虚假信息数据集采集和标注

虚假信息数据的采集和标注是研究虚假信息检测、传播和控制过程中一项重要环节，数据采集和标注的质量直接影响后期研究人员验证研究方法效果的精确度.

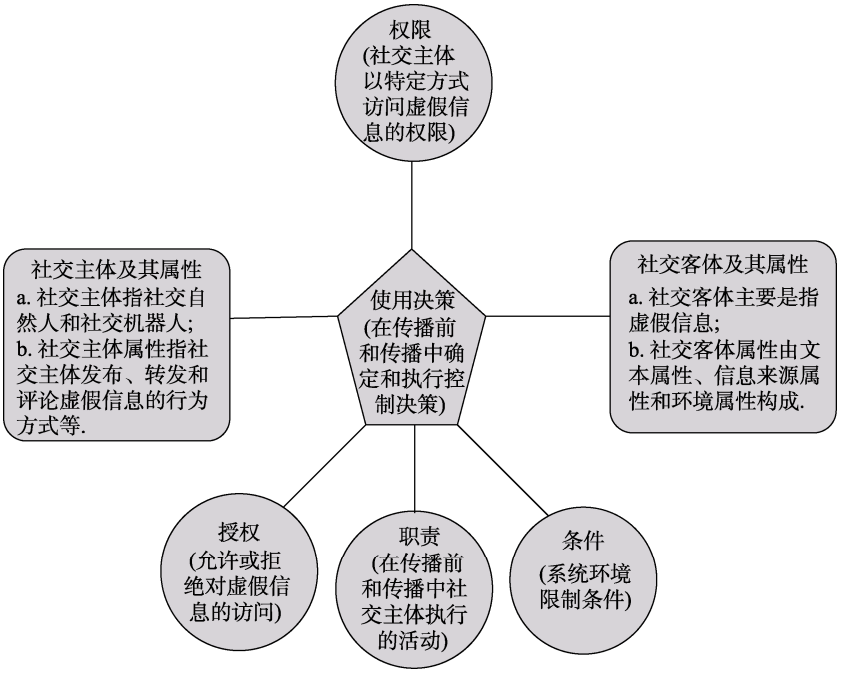

图 8 虚假信息传播使用控制模型

虚假信息的数据采集主要来源于新闻机构主页、搜索引擎和社交媒体网站等. 数据采集内容包括文本内容信息、用户基本信息以及用户发布、转发评论等社交上下文信息. 数据采集的数量和比例主要包括虚假信息话题的数量、每一个话题对应的虚假信息数量（即虚假信息的微博数量）以及虚假信息和真实信息之间的比例等. 对于单文本虚假信息检测和传播的研究通常选取 1 至 2 个话题，比如关于自然灾害、恐怖主义等方面虚假信息检测的研究. 基于多文本的虚假信息检测和传播研究，研究人员通常选取几十或上百种不同的虚假信息话题，针对每一种话题收集一定数量的虚假信息. 在虚假信息和真实信息的比例上，以文献[31][57][68][117]为代表的研究人员通常采取 1:1 的分布比例. 但是在真实的 OSNs 平台中，真实信息的数量通常大于虚假信息的数量，对此文献[118]从 Facebook 中选取官方认证的 333547 条真实信息和 51535 条虚假信息，大致比例为 6:1. 针对虚假信息数据采集的具体方法，通常研究人员首先确定感兴趣的虚假信息话题，然后使用与话题相关的关键字进行筛选收集. 其中通过调用 Twitter、Facebook、新浪微博和人人网等社交平台的 API 接口是一种常用的方法. 由以上可知，虚假信息数据的收集涵盖了内容特征和社交上下文特征等的多维度信息.

在虚假信息数据集的标注过程中，数据清洗操作是数据标注流程中的首要环节. 针对虚假信息数据采集过程中出现的噪声数据、缺失数据和重复数据等问题，执行数据清洗操作，从而获得高质量数据. 对清洗后的数据主要运用专家记者、事实核查网站、自动化检测器和众包等方式进行数据集标注. 通过调研发现，事实核查是标注虚假信息和真实信息的一项主流技术，主要分为面向专家的事实核查模型（Expert-oriented fact checking models）、面向众包的事实核查模型（Crowdsourcing-oriented fact checking models）、面向计算的事实核查模型[29]（Computational-oriented fact checking models）. 其中面向专家的事实核查通常采用 Snopes 和 FactCheck.org 等网站，依赖于专家的认知来评估信息的真实性，其弊端在于需要消耗一定的时间和财力. 面向众包的事实核查模型是通过利用群体的智慧来标注虚假内容，比如 Fiskkit 平台. 面向计算的事实核查模型是基于算法、知识图谱（knowledge graphs）和开放的网络等来评估.

5.2　虚假信息公开数据集

随着社交平台的安全性和隐私性不断提高，社交用户个人隐私数据获取受到严格的保护. 通过调研发现，当前用于研究虚假信息检测、传播和控制的公开数据集比较少，主要来自于 Twitter、Facebook 等国际知名媒体. 我们通过对现有的虚假信息公开数据集进行整理、分析（表 5），给出相应的数据描述，为后期研究人员获取数据集提供参考.

表 5　虚假信息公开数据集

数据集名称	相关文献	平台	信息数量	数据链接
BS Detector	[29]	BS detector	–	https://github.com/bs-detector/bs-detector
FakeNewsNet	[119]	Twitter	201921 篇文章	https://github.com/KaiDMML/FakeNewsNet
BuzzFeedNews	[120]	Facebook	1627 篇文章	https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data
BuzzFace	[121]	Facebook	2263 篇文章 160 万条评论	https://github.com/gsantia/BuzzFace
FacebookHoax	[122]	Facebook	15500 条帖子	https://github.com/gabll/some-like-it-hoax
LIAR	[123]	PolitiFact	12836 条简短陈述	https://www.cs.ucsb.edu/ william/software.html
CREDBANK	[124]	Twitter	6000 万条推文	http://compsocial.github.io/CREDBANK-data/

在表 5 中，FakeNewsNet 数据集由 PolitiFact 数据集和 GossipCop 数据集组成，其中 PolitiFact 数据集由 624 篇真实文章和 432 篇虚假文章构成，GossipCop 数据集由 16817 篇真实文章和 6048 篇虚假文章构成. BuzzFeedNews 数据集包含 9 家新闻通讯社在美国大选前一周（2016 年 9 月 19 日至 23 日、9 月 26 日和 27 日），通过 Facebook 发布的全部新闻. 这些新闻由 1627 篇文章、826 篇主流文章、356 篇“左翼”文章和 545 篇“右翼”文章组成. 每篇文章都由 5 名 BuzzFeed 记者逐条核实. 其缺点是每条新闻只包含标题和文本，缺少社交上下文信息. BuzzFace 数据集是在 BuzzFeed 数据集的基础上进一步扩展得到的，由 2263 篇新闻文章和 160 万条新闻内容的评论组成. FacebookHoax 数据集由科学新

闻相关的帖子和使用 Facebook Graph API 收集的虚假信息页面组成，包含了 32 页（14 页虚假文章和 18 页正常文章）15500 篇文章. LIAR 数据集来源于 PolitiFact 事实核查网站，一共收集了 12836 条人工标记的简短陈述，这些陈述被标记为错误、几乎都不正确、一半正确、大部分正确和正确等六种类别，其缺点是只包含大多数简短的语句，而不能涵盖完整的新闻文章. BS Detector 数据集是从 244 个网站上利用 Chrome 的 BS Detector 扩展工具识别出的虚假新闻数据，BS detector 输出为信息真实性的标签. CREDBANK 数据集是由 6000 万条推文组成的大规模众包数据集，该数据集从 2015 年 10 月开始收集，利用 Amazon Mechanical Turk 的 30 个标注者来评估每一条信息的可信度.

6 基于社会情境安全分析的跨平台传播和控制

随着在线社交平台数量的不断增加，社交用户的规模也在不断扩大，虚假信息在社交平台上的大量传播，给社交用户带来了严重的影响. 在过去的研究中，研究人员主要针对在同一个社交平台上虚假信息检测和传播展开研究，并取得了一些研究成果. 这些研究成果在制止虚假信息的传播上，主要采取“封号”、“禁言”、“删除”等补救措施，来阻止虚假信息的进一步扩散. 随着一种以人为中心新的计算范式—情境分析（Situation analytics）的出现[125-127]，为跨平台虚假信息传播和控制带来了全新的思路. 当前，如何将情境分析方法与新一代人工智能技术进行巧妙结合，实现对虚假信息“传播前”的控制，将成为研究者未来关注的焦点.

我们在情境分析的基础上，结合社交网络特征，进一步地提出了适用于 OSNs 应用场景的社会情境分析（Social situation analytics）方法[128]. 社会情境分析方法能有效地挖掘社交用户虚假信息传播过程中频繁的行为序列模式，但对于虚假信息传播和控制研究不能完全适用. 因此我们在社会情境分析方法的基础上，又进一步提出了可适用于虚假信息传播和控制研究的社会情境安全和层次化分析框架，如图 9 所示. 该框架自下而上，分别从社交实体层（内容安全）、社交环境层（环境安全）、社交行为层（行为安全）、社交意图层（意图安全）和社交目标层（服务安全）入手逐层展开，最终构建了一套涵盖五层和六要素的社会情境安全分析体系框架. 每一层及其对应的基本要素概述如下：第一层：社交实体层（包括社交自然人、半社交机器人和社交机器人三类社交主体，真实和虚假的社交信息等社交客体，OSNs 服务平台等）；第二层：社交环境层（时间和空间、群组和角色）；第三层：社交行为层（动作 Action）；第四层：社交意图层（意愿 Desire、情绪 Emotion）；第五层：社交目标层（受众目标 Target、趋势 Trend）. 最后重点介绍虚假信息在跨平台传播及使用控制方面，面临的挑战和进一步值得研究和探讨的问题.

（1）半社交机器人检测技术研究

由于社交主体在跨平台虚假信息传播中扮演着重要的角色，为了更加高效地研究社交主体虚假信息的传播模式，首先需要对半社交机器人检测方法进行研究. 当前社交机器人检测已取得了初步性的研究成果，随着新一代人工智能技术的不断发展，OSNs 平台中的半社交机器人具有社交自然人和社交机器人共同拥有的特征. 因此，只有选取能精准有效地区分社交自然人、半社交机器人和社交机器人传播虚假信息的特征和算法，才能够实时准确检测出 OSNs 平台中存在的半社交机器人账户. 在社会情境分析技术中，利用社会情境要素，即社交主体所处的环境、身份、动作、意愿和目标，来及时捕获半社交机器人传播虚假信息的特征，从而得出半社交机器人在传播虚假信息的过程中，表现出与社交自然人和社交机器人明显不同的传播虚假信息模式. 因此，基于社会情境分析理论，来区分社交自然人、半社交机器人和社交机器人是研究虚假信息传播和控制的首要问题.

（2）虚假信息交叉传播意图检测研究

在社交网络平台中，数据的隐私和安全受到了严格的保护. 在大数据驱动的人工智能技术下，为了对跨平台的虚假信息交叉传播开展研究，需要大量精准的数据作为支撑. 因此，首先通过运用联邦学习技术[129]，实现不同平台之间的用户传播虚假信息的数据融合，解决跨平台的数据孤岛问题. 其次通过利用社会情境分析技术，建立一套可计算的 OSNs 用户虚假信息传播模型理论与方法，来准确地预测社交用户的传播行为和意图，掌握虚假信息传播的整体趋势，目前该方向尚未得到很好的研究和突破性进展. 因此，通过将联邦学习和社会情境分析理论相结合，对跨平台的社交用户虚假信息交叉传播研究，将作为研究者今后关注的新问题.

（3）虚假信息传播目标识别研究

在虚假信息传播意图检测的基础上，准确识别虚假信息的传播目标，有利于及时控制虚假信息的扩散. 通过调研文献发现，当前关于虚假信息传播

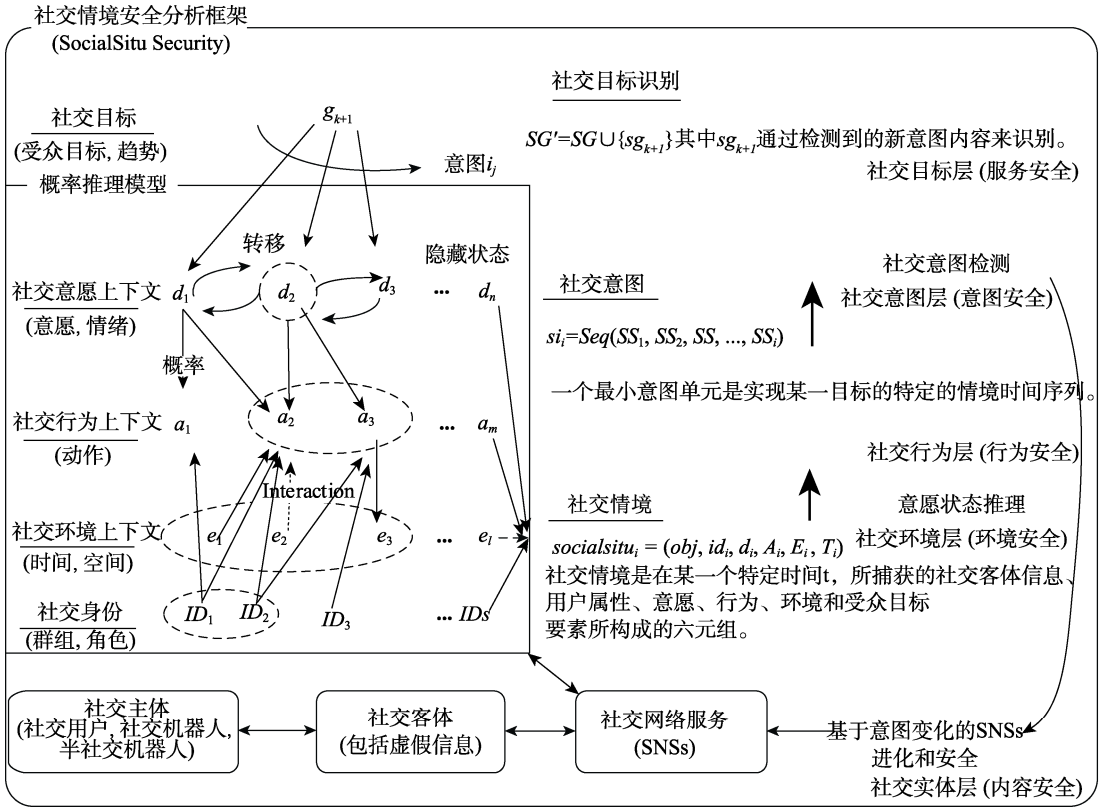

图 9　社会情境安全分析框架

的目标识别相关研究鲜有报道. 伴随着社会情境意图的动态变化，社交用户会随时改变社交目标，从而造成传播行为的改变. 社交用户在社交平台上传播虚假信息的目标，具有动态性、随机性和易变性等特点. 在未来的研究中，如何利用社会情境分析技术，采用合适的机器学习和深度学习算法，在 OSNs 虚假信息传播行为和意图检测的基础上，建立一套可计算的社交用户传播目标和趋势识别理论和方法，是未来又一个值得深入研究的问题.

（4）虚假信息传播的使用控制研究

社交网络服务正呈现出跨平台、跨社群、多时空等重要特征，传统的访问控制模型已不再适用于新兴的 OSNs 应用实际需求. 在虚假信息传播意图检测和目标识别的基础上，进一步融合社会情境分析技术，研究跨平台、跨社群的细粒度时空使用控制机制和方法，对虚假信息传播实施传播前和传播中及时有效地使用控制，提高主动式虚假信息传播控制能力，是未来的又一个研究趋势，同时也为建立开放、共享、安全、可控的 OSNs 空间环境提供了安全技术保障.

7　结束语

OSNs 虚假信息检测、传播和控制已经成为社交网络安全领域的热点问题，受到国内外研究人员的广泛关注，已开展了相关研究，并取得一定进展. 从当前国内外研究趋势来看，首先 OSNs 虚假信息的检测主要从信息内容和社交上下文辅助信息两个方面展开研究，其次针对社交用户虚假信息传播问题，分别从基于 SIR 等传染病传播模型研究、基于社交网络统计性质的虚假信息传播研究和跨媒介虚假信息传播研究等 3 个方面展开论述，针对社交机器人虚假信息传播研究问题，从社交机器人检测方法和传播模式等方面进行分析总结，然后针对社交网络虚假信息传播控制研究，分别从虚假信息传播节点控制和使用控制方面展开论述，分析了用于虚假信息传播研究的使用控制模型，同时也对虚假信息数据采集和标注以及公开数据集进行总结. 最后提出了社会情境安全和分析框架，进一步讨论了 OSNs 虚假信息传播和控制研究面临的一些挑战和未来可能的研究发展趋势.

未来，随着大数据、人工智能和新兴前沿计算技术的深入发展，政府和学术界、工业界人员面对在线社交网络普遍存在且不断产生的虚假信息，一方面需要将检测、传播和控制三者更加综合地考虑和加以实施，另一方面社会情境安全将为构建虚拟化社交网络空间安全，实现网络信息内容生态治理，

提供理论基础、技术支撑和应用场景.

致 谢 衷心感谢各位评审专家和编辑部的老师们对本文提出的宝贵的修改意见和建议!

参 考 文 献

[1] Zhang Zhi-Yong, Zhao Chang-Wei, Wang Jian. Social media networks security theory and technology. Beijing: Science and Technology Press, 2016 (in Chinese)
(张志勇, 赵长伟, 王剑. 社交媒体网络安全理论与技术. 北京: 科学出版社, 2016)

[2] Zhang Lei, Cui Yong, Liu Jing, et al. Application of machine learning in cyberspace security research. Chinese Journal of Computers, 2018, 41 (9): 1943-1975 (in Chinese)
(张蕾, 崔勇, 刘静, 江勇, 吴建平. 机器学习在网络空间安全研究中的应用. 计算机学报, 2018, 41(9): 1943-1975)

[3] Zhang Z Y, Gupta B B. Social media security and trustworthiness: overview and new direction. Future Generation Computer Systems, 2018, 86: 914-925

[4] Garg S, Kaur K, Kumar N. Hybrid deep learning-based anomaly detection scheme for suspicious flow detection in SDN: a social multimedia perspective. IEEE Transactions on Multimedia, 2019, 21(3): 566-578

[5] Zhang Z, Wang K. A trust model for multimedia social networks. Social Network Analysis and Mining, 2013, 3(4): 969-979

[6] Bondielli A, Marcelloni F. A survey on fake news and rumour detection techniques. Information Sciences, 2019, 497: 38-55

[7] Kumar K, Geethakumari G. Detecting misinformation in online social networks using cognitive psychology. Human-centric Computing and Information Sciences, 2014, 4(1): 14-26

[8] Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science, 2018, 359(6380): 1146-1151

[9] Chen Yan-Fang, Li Zhi-Yu, Liang Xun, et al. Review on rumor detection of online social networks. Chinese Journal of Computers, 2018, 41 (7): 1648-1677 (in Chinese)
(陈燕方, 李志宇, 梁循, 齐金山. 在线社会网络谣言检测综述. 计算机学报, 2018, 41 (7): 1648-1677)

[10] Lazer D M J, Baum M A, Benkler Y, et al. The science of fake news. Science, 2018, 359(6380): 1094-1096

[11] Wu L, Morstatter F, Hu X, et al. Mining misinformation in social media. Florida, USA: CRC Press, 2016

[12] Wang P, Angarita R, Renna I. Is this the era of misinformation yet: combining social bots and fake news to deceive the masses //Proceedings of the Companion Proceedings of the Web Conference. Lyon, France, 2018: 1557-1561

[13] Khaldarova I, Pantti M. Fake news: The narrative battle over the ukrainian conflict. Journalism Practice, 2016, 10(7): 891-901

[14] Li De-Yi, Yu Jian. Introduction to artificial intelligence. Beijing: China Science and Technology Press, 2018 (in Chinese)
(李德毅, 于剑. 人工智能导论. 北京: 中国科学技术出版社, 2018)

[15] Ferrara E, Varol O, Davis C, et al. The rise of social bots. Communications of the ACM, 2016, 59 (7): 96-104

[16] Lokot T, Diakopoulos N. News bots: automating news and information dissemination on twitter. Digital Journalism, 2016, 4(6): 682-699.

[17] Savage S, Monroy-Hernandez A, Hollerer T. Botivist: calling volunteers to action using online bots//Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing. California, USA, 2016: 813-822

[18] Global D. Bot traffic report 2019. California, USA: Imperva Incapsula, 2019

[19] Davis C A, Varol O, Ferrara E, et al. BotOrNot: a system to evaluate social bots//Proceedings of the 25th International Conference Companion on World Wide Web. Montreal, Canada, 2016: 273-274

[20] Dewangan M, Kaushal R. SocialBot: behavioral analysis and detection. Singapore: Springer, 2016

[21] Woolley S C. Automating power: social bot interference in global politics. First Monday, 2016, 21(4): 12

[22] Bessi A, Ferrara E. Social bots distort the 2016 US presidential election online discussion. First Monday, 2016, 21: 11-17

[23] Kollanyi B, Howard P N. Junk news and bots during the German federal presidency election: what were German voters sharing over twitter?. Computational Propaganda Project Working Paper Series, 2017: 1-5

[24] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?// Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 591-600

[25] Ferrara E. Disinformation and social bot operations in the run up to the 2017 french presidential election. First Monday, 2017, 22 (8): 1-2

[26] Glenski M, Weninger T, Volkova S. Propagation from deceptive news sources who shares, how much, how evenly, and how quickly?. IEEE Transactions on Computational Social Systems, 2018, 5 (4): 1071-1082

[27] Shao C C, Ciampaglia G L, Varol O, et al. The spread of low-credibility content by social bots. Nature Communications, 2018, DOI: 10.1038/s41467-018-06930-7

[28] Varol O, Ferrara E, Davis C A, et al. Online human-bot interactions: detection, estimation, and characterization. arXiv preprint arXiv:1703.03107, 2017

[29] Shu K, Sliva A, Wang S, et al. Fake news detection on social media: a data mining perspective. ACM Sigkdd Explorations Newsletter, 2017, 19(1): 22-36

[30] Tacchini E, Ballarin G, Della Vedova M L, et al. Some like it hoax: automated fake news detection in social networks. arXiv preprint arXiv:1704.07506, 2017

[31] Ruchansky N, Seo S, Liu Y. CSI: A hybrid deep model for fake news detection//Proceedings of the 2017 ACM Conference on Information and Knowledge Management. Singapore, 2017, 1: 142-150

[32] ReisJCS, CorreiaA, MuraiF, etal. Supervised learning for fake news detection. IEEE Intelligent Systems, 2019, 34:76-81

[33] Conroy N J, Rubin V L, Chen Y. Automatic deception detection: methods for finding fake news//Proceedings of the 78th Annual Association for Information Science and Technology. Washington, USA, 2015: 1-4

[34] Baeth M J, Aktas M S. Detecting misinformation in social networks using provenance data//Proceedings of the 13th International Conference on Semantics, Knowledge and Grids.

Beijing, China, 2017: 85-89

[35] Zhou X, Zafarani R, Shu K, et al. Fake news: fundamental theories, detection strategies and challenges//Proceedings of the 12th ACM International Conference on Web Search and Data Mining. Victoria, Australia, 2019: 32-39

[36] 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

[37] Rath B, Gao W, Ma J, et al. From retweet to believability: utilizing trust to identify rumor spreaders on twitter//Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Sydney, Australia, 2017: 179-186

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

[39] Liu Ya-Hui, Jin Xiao-Long, Shen Hua-Wei, et al. A survey on rumor identification over social media. Chinese Journal of Computers, 2018, 41 (7): 1536-1558 (in Chinese)
(刘雅辉, 靳小龙, 沈华伟, 鲍鹏, 程学旗. 社交媒体中的谣言识别研究综述. 计算机学报, 2018, 41(7): 1536-1558)

[40] Kakol M, Nielek R, Wierzbicki A. Understanding and predicting web content credibility using the content credibility corpus. Information Processing and Management, 2017, 53(5): 1043-1061

[41] Saez-Trumper D. Fake tweet buster: a web tool to identify users promoting fake news on Twitter//Proceedings of the 25th ACM Conference Hypertext Social Media. Santiago, Chile, 2014:316-317

[42] Castillo C, Mendoza M, Poblete B. Information credibility on twitter //Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 665-674

[43] Canini K R, Suh B, Pirolli P. Finding relevant sources in Twitter based on content and social structure//Proceedings of the NIPS MLSN Workshop. California, USA, 2010: 1-7

[44] Sikdar S, Kang B, Donovan J, et al. Cutting through the noise: Defining ground truth in information credibility on Twitter. Human, 2(3): 151-167, 2013

[45] Canini K R, Suh B, Pirolli P L. Finding credible information sources in social networks based on content and social structure //Proceedings of the 3th International Conference on Privacy, Security, Risk and Trust. Boston, USA, 2011: 21-28

[46] Shu K, Wang S, Liu H. Beyond news contents: the role of social context for fake news detection//Proceedings of the 12th ACM International Conference on Web Search and Data Mining. Cambridge, England, 2017: 90-95

[47] Saif M Mohammad, Parinaz S, Svetlana K. Stance and sentiment in tweets. ACM Transactions on Internet Technology, 2017, 17(3): 26-37

[48] Jin Z, Cao J, Jiang Y. News credibility evaluation on microblog with a hierarchical propagation model//Proceedings of the 14th IEEE International Conference on Data Mining. ShenZhen, China, 2014: 230-239

[49] Jin Z, Cao J, Zhang Y, et al. News verification by exploiting conflicting social viewpoints in microblogs//Proceedings of the 13th AAAI Conference on Artificial Intelligence. California, USA, 2016: 33-42

[50] Manzoor S I, Singla J, Nikita. Fake news detection using machine learning approaches: A systematic review//Proceedings of the 3th International Conference on Trends in Electronics and Informatics. Tirunelveli, India, 2019: 230-234

[51] Gilda S. Evaluating machine learning algorithms for fake news detection//Proceedings of the 15th Student Conference on Research and Development. Putrajaya, Malaysia, 2017: 110-115

[52] Granik M, Mesyura V. Fake news detection using naive bayes classifier//Proceedings of the 1st Ukraine Conference on Electrical and Computer Engineering. Kyiv, Ukraine, 2017: 900-903

[53] Benamira A, Devillers B, Lesot E, et al. Semi-supervised learning and graph neural networks for fake news detection//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York, USA, 2019: 568-569

[54] Alrubaian M, Al-Qurishi M, Hassan M M, et al. A credibility analysis system for assessing information on twitter. IEEE Transactions on Dependable and Secure Computing, 2018, 15 (4): 661-674

[55] Alrubaian M, Al-Qurishi M, Al-Rakhami M, et al. A multistage credibility analysis model for microblogs//Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris, France, 2015: 1434-1440

[56] Abbasi M, Liu H. Measuring user credibility in social media social computing. New York, USA: Springer, 2013: 441-448

[57] Liu Y, Wu Y F. Early Detection of fake news on social media through propagation path classification with recurrent and convolutional networks//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Hong Kong, China, 2018: 354-361

[58] Campan A, Cuzzocrea A, Truta T M. Fighting fake news spread in online social networks: actual trends and future research directions //Proceedings of the 2017 IEEE International Conference on Big Data. Boston, USA, 2017: 4453-4457

[59] Wen S, Haghighi M S, Chen C, et al. A sword with two edges: propagation studies on both positive and negative information in online social networks. IEEE Transactions on Computers, 2015, 64(3): 640-653

[60] Cao Jiu-Xin, Wu Jiang-Lin, Shi Wei, Liu Bo, et al. Sina microblog information diffusion analysis and prediction. Chinese Journal of Computers, 2014, 37(4): 779-790 (in Chinese)
(曹玖新, 吴江林, 石伟, 刘波, 郑啸, 罗军舟. 新浪微博网信息传播分析与预测. 计算机学报, 2014, 37(4): 779-790)

[61] He Z B, Cai Z P, Yu J G, et al. Cost-efficient strategies for restraining rumor spreading in mobile social networks. IEEE Transactions on Vehicular Technology, 2017, 66(3): 2789-2800

[62] Zanette D H. Dynamics of rumor propagation on small-world networks. Physical Review E Statistical Nonlinear and Soft Matter Physics, 2002, 65(1):041908-041916

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

[64] Tan Zhen-Hua, Shi Ying-Cheng, Shi Nan-Xiang, et al. Rumor

propagation analysis model inspired by gravity theory for online social networks. Journal of Computer Research and Development, 2017, 54(11): 2586-2599 (in Chinese)
(谭振华, 时迎成, 石楠翔, 杨广明, 王兴伟. 基于引力学的在线社交网络空间谣言传播分析模型. 计算机研究与发展, 2017, 54(11): 2586-2599)

[65] Tambuscio M, Ruffo G, Flammini A, et al. Fact-checking effect on viral hoaxes: a model of misinformation spread in social networks //Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 977-982

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

[67] Volkova S, Shaffer K, Jang J Y, et al. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 647-653

[68] Pal A, Chua A. Propagation pattern as a telltale sign of fake news on social media//Proceedings of the 5th International Conference on Information Management. Cambridge, UK, 2019: 269-273

[69] Glenski M, Weninger T, Volkova S. How humans versus bots react to deceptive and trusted news sources: a case study of active users. arXiv preprint arXiv:1807.05327v1, 2018

[70] Glenski M, Weninger T, Volkova S. Identifying and understanding user reactions to deceptive and trusted social news sources. arXiv preprint arXiv:1805.12032v1, 2018

[71] Rao Yuan, Wu Lian-Wei, Zhang Jun-Yi. A survey of information propaganda mechanism under the cross-medium. Science China Information Sciences, 2017, 47(12): 27-49 (in Chinese)
(饶元, 吴连伟, 张君毅. 跨媒介舆情网络环境下信息传播机制研究与进展. 中国科学: 信息科学, 2017, 47(12): 27-49)

[72] Zhang J, Zhang R, Sun J, et al. TrueTop: A Sybil-resilient system for user influence measurement on twitter. IEEE/ACM Transactions on Networking, 2016, 24(5): 2834-2846

[73] Jia J, Wang B, Gong N Z. Random walk based fake account detection in online social networks//Proceedings of the 47th IEEE International Conference on Dependable Systems and Networks. Denver, USA, 2017: 273-284

[74] Mehrotra A, Sarreddy M, Singh S. Detection of fake Twitter followers using graph centrality measures//Proceedings of the 2nd International Conference on Contemporary Computing and Informatics. Noida, India, 2016: 499-504

[75] Dickerson J P, Kagan V, Subrahmanian V S. Using sentiment to detect bots on twitter: are humans more opinionated than bots? //Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Beijing, China, 2014: 620-627

[76] Alarifi A, lsaleh M, Al-Salman A. Twitter turing test: Identifying social machines. Information Sciences, 2016, 372: 332-346

[77] Gilani Z, Kochmar E, Crowcroft J. Classification of twitter accounts into automated agents and human users//Proceedings of the 9th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris, France, 2017: 22-26

[78] Chen Xia, Min Hua-Qing, Song Heng-Jie. Automatically identify users who cheat on crowdsourcing platform. Computer Engineering, 2016, 42(8): 139-145 (in Chinese)
(陈霞, 闵华清, 宋恒杰. 众包平台作弊用户自动识别, 计算机工程, 2016, 42(8): 139-145)

[79] Wang G, Zhang X Y, Tang S L, et al. Clickstream user behavior models. ACM Transactions on the Web, 2017, 11(4): 1-37

[80] Liu Rong, Chen Bo, Yu Ling, et al. Overview of detection techniques for malicious social bots. Journal on Communications, 2017, 38(Z2): 197-210 (in Chinese)
(刘蓉, 陈波, 于泠, 刘亚尚, 陈思远. 恶意社交机器人检测技术研究. 通信学报, 2017, 38(Z2): 197-210)

[81] Morstatter F, Wu L, Nazer T H, et al. A new approach to bot detection: striking the balance between precision and recall //Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Washington, USA, 2016: 533-540

[82] Costa A F, Yamaguchi Y, Traina A J M, et al. Modeling temporal activity to detect anomalous behavior in social media. ACM Transactions on Knowledge Discovery from Data, 2017, 11 (4): 1-23

[83] Jr S B, Campos G F C, Tavares G M, et al. Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets. ACM Transactions on Multimedia Computing Communications and Applications, 2018, 14(1s): 1-17

[84] Fazil M, Abulaish M. Identifying active, reactive, and inactive targets of socialbots in twitter//Proceedings of the International Conference on Web Intelligence. Leipzig, Germany, 2017: 573-580

[85] Shi P N, Zhang Z Y, Choo K R. Detecting malicious social bots based on clickstream sequences. IEEE Access, 2019, 7(1): 28855-28862

[86] Liu Jian-Wei, Liu Yuan, Luo Xiong-Lin, et al. Semi-Supervised learning methods. Chinese Journal of Computers, 2015, 38(8): 1592-1617 (in Chinese)
(刘建伟, 刘媛, 罗雄麟, 等. 半监督学习方法. 计算机学报, 2015, 38(8): 1592-1617)

[87] Zhang Z, Sun R, Zhao C, et al. CyVOD: a novel trinity multimedia social network scheme. Multimedia Tools and Applications, 2017, 76(18): 18513-18529

[88] Cai C Y, Li L J, Zeng D. Behavior enhanced deep bot detection in social media//Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics. Shenzhen, China, 2017: 128-130

[89] Ping H, Qin S J. A social bots detection model based on deep learning algorithm//Proceedings of the 2018 IEEE International Conference on Communication Technology. Chongqing, China, 2018: 1435-1439

[90] Sneha K, Emilio F. Deep neural networks for bot detection. Information Sciences, 2018, 467:312-322

[91] Cai C, Li L, Zeng D. Detecting social bots by jointly modeling deep behavior and content information//Proceedings of the 2017 ACM Conference on Information and Knowledge Management. Singapore, 2017: 122-127

[92] Clark E M, Williams J R, Jones C A, et al. Sifting robotic from organic text: a natural language approach for detecting automation on twitter. Journal of Computational Science, 2016, 16:1-7

[93] Walt E V D, Eloff J. Using machine learning to detect fake identities: bots vs humans. IEEE Access, 2018, 6(99):6540-6549

[94] Zhang Yu-Qing, Dong Ying, et al. Situation, trends and prospects of deep learning applied to cyberspace security. Journal of Computer Research and Development, 2018, 55(6):1117-1142 (in Chinese)
(张玉清, 董颖, 等. 深度学习应用于网络空间安全的现状、趋势与展望. 计算机研究与发展, 2018, 55(6):1117-1142)

[95] Lipton Z C, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv: 1506.00019, 2015

[96] Cho K, Van M, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches//Proceedings of the Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 103-111

[97] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014

[98] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[99] Shao C C, Ciampaglia G L, Varol O, et al. The spread of fake news by social bots. arXiv preprint arXiv: 1707.07592, 2017

[100] Gilani Z, Farahbakhsh R, Crowcroft J, et al. Do bots impact twitter activity?//Proceedings of the 26th International Conference on World Wide Web Companion. Perth, Australia, 2017: 781-782

[101] Chen Tian-Zhu, Guo Yun-Chuan, Niu Ben, Li Feng-Hua. Research progress of access control model and policy in online social networks. Chinese Journal of Network and Information Security, 2016, 2(8):1-9 (in Chinese)
(陈天柱, 郭云川, 牛犇, 李凤华. 面向社交网络的访问控制模型和策略研究进展. 网络与信息安全学报, 2016, 2(8): 1-9)

[102] Yang Jing, Zhou Xue-Yan, Lin Ze-Ming, et al. False information spread control method based on source tracing. Journal of Harbin Engineering University, 2016, 37(12): 1691-1697 (in Chinese)
(杨静, 周雪妍, 林泽鸿, 张健沛, 印桂生. 基于溯源的虚假信息传播控制方法. 哈尔滨工程大学学报, 2016, 37(12): 1691-1697.)

[103] Wang Jia-Kun, Yu Hao, Wang Xin-Hua, et al. Dissemination and control model of public opinion in online social networks based on users' relative weight. Systems Engineering Theory and Practice, 2019, 39(6): 1565-1579(in Chinese)
(王家坤, 于灏, 王新华, 白丽. 基于用户相对权重的在线社交网络舆情传播控制模型. 系统工程理论与实践, 2019, 39(6): 1565-1579)

[104] Wang Yong-Gang, Cai Fei-Zhi, Eng Keong Lua, et al. A diffusion control method of fake information in social networks, Journal of Computer Research and Development, 2012, 49(S2): 131-137(in Chinese)
(王永刚, 蔡飞志, Lua E K. 一种社交网络虚假信息传播控制方法. 计算机研究与发展, 2012, 49(S2): 131-137)

[105] Cheng Y, Park J, Sandhu R. An access control model for online social networks using user-to-user relationship. IEEE Transactions on Dependable and Secure Computing, 2016, 13(4): 424-436

[106] Pang J, Zhang Y. A new access control scheme for facebook-style social networks. Computers and Security, 2015, 54 (44): 44-59

[107] Bui T, Stoller S D, Li J J. Greedy and evolutionary algorithms for mining relationship-based access control policies. Computer and Security, 2019, 80: 317-333

[108] Zhang Z, Han L, Li C, et al. A novel attribute-based access control model for multimedia social networks. Neural Network World, 2016, (6): 543-557

[109] Hu D, Hu C, Fan Y, et al. oGBAC-a group based access control framework for information sharing in online social networks. IEEE Transactions on Dependable and Secure Computing, DOI 10.1109/TDSC.2018.2875697, 2018

[110] Krishnan R, Sandhu R, Niu J, et al. A conceptual framework for group-centric secure information sharing//Proceedings of the 4th International Symposium on Information, Computer, and Communications Security. New York, UK, 2009, 384–387

[111] Li F H, Li Z F, Han W L, et al. Cyberspace-oriented access control: a cyberspace characteristics based model and its policies. IEEE Internet of Things Journal, DOI 10.1109/JIOT.2018. 2839065, 2018

[112] Abdulla A K, Bakiras S. HITC: Data privacy in online social networks with fine-grained access control//Proceedings of the 24th ACM Symposium on Access Control Models and Technologies. Canada, Toronto, 2019: 123-134

[113] Ma S, Yan Z. PSNController: An unwanted content control system in pervasive social networking based on trust management. ACM Transactions on Multimedia Computing, Communication and Application, 2015, 12(1s): 1-23

[114] Zhang X W, Parisi-Presicce F, Sandhu R, et al. Formal model and policy specification of usage control. ACM Transactions on Information and System Security, 2005, 8(4): 351-387

[115] Wu J, Dong M, Ota K, et al. Cross-domain fine-grained data usage control service for industrial wireless sensor networks. IEEE Access, 2017, 3: 2939-2949

[116] Gonzalez-Manzano L, González-Tablas A I, de Fuentes J M, et al. $SoNeUCON_{ABC}$, An expressive usage control model for web-based social networks. Computers and Security, 2014, 43: 159-187

[117] Rajdev M, Lee K. Fake and spam messages: detecting misinformation during natural disasters on social media //Proceedings of the 2015 IEEE International Conference on Web Intelligence and Intelligent Agent Technology. Singapore, 2015: 72-75

[118] Vicario M, Quattrociocchi W, Scala A, et al. Polarization and fake news: Early warning of potential misinformation targets. arXiv preprint arXiv:1802.01400v1, 2018

[119] Shu K, Mahudeswaran D, Wang S, et al. FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286, 2018

[120] Potthast M, Kiesel J, Reinartz K, et al. A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638, 2017

[121] Santia G C, Williams J R. Buzzface: A news veracity dataset with facebook user commentary and egos//Proceedings of the 12th International AAAI Conference on Web and Social Media, California, USA, 2018: 531–540.

[122] Tacchini E, Ballarin G, Della Vedova M L, et al. Some like it hoax: Automated fake news detection in social networks, arXiv preprint arXiv:1704.07506, 2017

[123] Wang W Y. Liar, liar pants on fire: a new benchmark dataset for fake news detection//Proceedings of the 55th Annual Meeting of

the Association for Computational Linguistics. Vancouver, Canada, 2017: 422–426
[124] Mitra T, Gilbert E. CREDBANK: a large-scale social media corpus with associated credibility annotations//Proceedings of the 9th International AAAI Conference on Web and Social Media. Oxford, UK, 2015: 258–267
[125] Chang C K, Jiang H Y, Hua M, et al. Situ: a situation-theoretic approach to context-aware service evolution. IEEE Transactions on Services Computing, 2009, 2 (3): 261-275
[126] Chang C K. Situation analytics: a foundation for a new software engineering paradigm. Computer, 2016, 49 (1): 24-33
[127] Chang C K. Situation analytics-at the dawn of a new software engineering paradigm. Science China Information Sciences, 2018, 61 (5): 050101:1-050101:14
[128] Zhang Z Y, Sun R R, Wang X X, et al. A situational analytic method for user behavior pattern in multimedia social networks. IEEE Transactions on Big Data, 2019, 5(4): 520-528
[129] Yang Q, Liu Y, Chen T, et al. Federated machine learning. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19

ZHANG Zhi-Yong, Ph. D., professor, Ph. D. supervisor, Henan Province Distinguished Professor. His research interests include cyberspace security and artificial intelligence, social big data analysis and mining, trusted computing and access control.

JING Jun-Chang, Ph. D. candidate. His research interests include social network security, machine learning and deep learning.

LI Fei, master, senior engineer. Her research interests include artificial intelligence and distributed database.

ZHAO Chang-Wei, Ph. D., lecturer. His research interests include artificial intelligence and network information security.

Background

With the rapid increase of Online Social Networks (OSNs) users, especially the emergence of social bots, the extensive spread of fake information has a severely negative effect on individuals and society. In order to control the spread of fake information in time, many researchers have begun to do research in this field. Therefore, many new findings have been published on the detection, dissemination and control methods of fake information. These research results are playing increasingly important roles in the security of OSNs, but there is rare literature report on research works on both social human and social bots potential and hidden propagation intentions detection and distribution usage control. The purpose of this paper is to make a survey on existIng relative theories and methods, and to outline some issues and challenges that can be addressed.

This paper reviews important research achievements made by computer scientists in fake information detection, dissemination and control from the perspective of artificial intelligence in recent years. Firstly, we present two aspects of the fake information detection methods on characterization and models. Secondly, we comprehensively and systematically analyze and compare the detection methods and propagation patterns and strategies of social human and social bots. We then summarize the control methods of fake information dissemination and provide a usage control model of fake information propagation. Moreover, we also discuss the method of fake information data collection and annotation, and some public datasets for detection, propagation and control. Finally, we propose an integrated SocialSitu (Social Situation) security analytics framework, and further discuss future research directions of cross-platform propagation and control of fake information based on social situation analysis theory and federal learning technology.

The paper is supported by National Natural Science Foundation of China Grant No. 61972133 and No. 61772174, as well as Project of Leading Talents in Science and Technology Innovation for Thousands of People Plan in Henan Province Grant No.204200510021, which are titled by “Social Situational Analytics Based Fake Information User Propagation Intention Detection and Usage Control”, “Research on Crowd Evaluation Method for Security and Trustworthiness of Social Media Platforms Based on Signaling Theory and Crowdsourcing” and “Research on SocialSitu Security Theory and Key Technologies”, respectively. The programs is oriented at the security and trust of online social networks, and propose a creative convergence between situational analytics (computing) and usage control (security) theories to explore the behavior-intention analysis computational theory and security control mechanism, with the aim at resolving the burning issue of the fake information propagation at user’s will. The programs are not only significantly fundamental study meanings of establishing the SocialSitu security theory for social computing and social intelligence, but also has better applicable visions for establishing virtual social network security.