doi:10.11959/j.issn.1000-436x.2017275

恶意社交机器人检测技术研究

刘蓉[1]，陈波[1]，于泠[1,2]，刘亚尚[1]，陈思远[1]
（1. 南京师范大学计算机科学与技术学院, 江苏 南京 210023；2. 江苏省大规模复杂系统数值模拟重点实验室，江苏 南京 210023）

摘 要：攻击者利用恶意社交机器人窃取用户隐私、传播虚假消息、影响社会舆论，严重威胁了个人信息安全、社会公共安全，乃至国家安全。攻击者还在不断引入新技术实施反检测。恶意社交机器人检测成为在线社交网络安全研究的一个重点和难点。首先回顾了当前社交机器人的开发与应用现状，接着对恶意社交机器人检测问题进行了形式化定义，并分析了检测恶意社交机器人所面临的主要挑战。针对检测特征的选取问题，厘清了从静态用户特征、动态传播特征，以及关系演化特征的研究发展思路。针对检测方法问题，从基于特征、机器学习、图论以及众包 4 个类别总结了已有检测方案的研究思路，并剖析了几类方法在检测准确率、计算代价等方面的局限性。最后，提出了一种基于并行优化机器学习方法的恶意社交机器人检测框架。

关键词：社交机器人；在线社交网络；特征工程；机器学习；图论；众包；并行化
中图分类号：TP391 **文献标识码**：A

Overview of detection techniques for malicious social bots

LIU Rong[1], CHEN Bo[1], YU Ling[1,2], LIU Ya-shang[1], CHEN Si-yuan[1]
(1. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China;
2. Jiangsu Provincial Key Laboratory for Numerical of Large Scale Complex System, Nanjing 210023, China)

Abstract: The attackers use social bots to steal people's privacy, propagate fraud messages and influent public opinions, which has brought a great threat for personal privacy security, social public security and even the security of the nation. The attackers are also introducing new techniques to carry out anti-detection. The detection of malicious social bots has become one of the most important problems in the research of online social network security and it is also a difficult problem. Firstly, development and application of social bots was reviewed and then a formulation description for the problem of detecting malicious social bots was made. Besides, main challenges in the detection of malicious social bots were analyzed. As for how to choose features for the detection, the development of choosing features that from static user features to dynamic propagation features and to relationship and evolution features were classified. As for choosing which method, approaches from the previous research based on features, machine learning, graph and crowd sourcing were summarized. Also, the limitation of these methods in detection accuracy, computation cost and so on was dissected. At last, a framework based on parallelizing machine learning methods to detect malicious social bots was proposed.

Key words: social bots, online social network, feature engineering,machine learning, graph, crowdsourcing, parallelism

1 引言

虚拟机器人随着互联网技术的普及与发展而逐渐受到人们的青睐，如帮助人类从繁杂的数据获取任务中解脱出来的搜索引擎爬虫机器人。社交机器人是目前活跃于社交网络领域的一种虚拟机器人，如作为商业化数字营销工具的涨粉机器人。社交机器人是一种自动化程序[1]，能够控制社交账号，

收稿日期：2017-07-30
通信作者：陈波，bchen@njnu.edu.cn
基金项目：赛尔网络下一代互联网技术创新基金资助项目（No.NGII20160509）；江苏省高等教育教学改革重点课题基金资助项目（No. 2015JSJG034）
Foundation Items: CERNET Innovation Project (No.NGII20160509), Key Subject of Higher Education Teaching Reform of Jiangsu Province (No.2015JSJG034)

运用人工智能等相关技术模仿人类行为在社交网络中活动。

互联网安全公司 Imperva Incapsula 公布的《2016 年机器人流量报告》指出，2016 年人类在线流量降为 48.2%，而机器人流量则达到了 51.8%[2]。这一报告同时还指出，互联网中恶意机器人的流量已经达到 28.9%，占总流量的较大比重。社交网络中也活跃着大量社交机器人。据报道，Twitter 上存在 2 300 万社交机器人，约占总用户的 85%。其中，一类社交机器人以服务人类、提高人类生活质量为目的，而另一类恶意社交机器人不仅通过伪装和身份窃取诱使受害者暴露敏感信息，实现社会工程攻击[3]，还成为了干扰政治选举、操纵社会舆论、左右金融决策的工具。因此，恶意社交机器人也被称为 Sybil（女巫）账户[4]。对于恶意社交机器人的检测和识别技术研究已经成为保障公民隐私信息安全、维护社交网络秩序和社会稳定的一项重要工作。

2 社交机器人的应用与开发现状

2.1 社交机器人的应用现状

社交机器人从诞生之初是以服务人类、提高人类生活质量为目的的。例如，全球最大新闻机构之一的美联社从 2014 年 7 月开始使用 Automated Insights 公司开发的“语言专家”批量生产财经新闻。在 2016 年里约奥运会期间，国内的《今日头条》网站也开始利用其实验室研发的机器人“张小明”编写新闻稿专门报道冷门赛事，解决了人类记者无法关注到每一场比赛的问题。这些报道的内容虽然是机器人程序利用自然语言处理、机器学习和视觉图像处理等技术产生的，但阅读量仍然非常可观。

利用社交机器人能够快速传播信息的特点，人们研发了 SF QuakeBot，它能够在社交网络上及时发布与地震相关的信息，减少地震给人类造成的破坏[5]。

一些商业平台上，社交机器人不但能够充当客服的角色，为消费者提供各种产品信息，还能够增加与消费者的互动，引导消费者参与到品牌的营销活动中。例如，2016 年 10 月 eBay 推出的 ShopBot 就已经支持用户用自然语言或者上传图片的方式与机器人对话，帮助用户搜索最实惠的商品。2017 年 F8 开发者大会上，Facebook 宣布了一系列对聊天机器人 chatbot 支持的新功能特性，以广泛实现用户与机器人对话等。

社交机器人还能够在网络社区中增强公众对公共事务的参与意识，促进公众之间的合作，帮助人们更好地处理事务。百度公司在 2015 年百度世界大会上推出了虚拟社交机器人“度秘”，通过自然语言对话的方式架起真实世界里商家与用户之间的桥梁。与“度秘”相比，微软公司研发的“微软小冰”在社交能力上更胜一筹。依靠微软在大数据、机器学习、人工智能和自然语言分析等方面的技术积累，微软小冰的思维方式与人类更加接近，并且具备实时决策能力，能够与人类进行长程会话[6]。

但是，技术是一把双刃剑。Imperva Incapsula 公司的报告提醒人们，从网络流量上来看，恶意机器人流量已经超过了普通机器人[2]。恶意社交机器人对个人信息安全、社会安全，乃至国家安全都有着极大的影响，已经成为网络空间安全的毒瘤。

恶意社交机器人会伪装成独立的实体，创建一些虚假的账户，实施窃取用户隐私[7]、发送垃圾邮件、传播恶意链接、发动 DDoS 攻击等活动，给无辜用户造成伤害[8]。

由于社交机器人可以表现得像人类用户一样在社交媒体上评论和转发其他用户的状态，干扰政治辩论，或被政治家当作操纵舆论的工具[9]，因此，恶意社交机器人的加入将会影响人们对某一事件的判断，促使网民的观点变得更加极端和难以控制，甚至将政治动员从线上发展到线下。例如，有的恶意社交机器人能够从在线社交网络中提取用户观点预测股票市场，干扰在线交易活动[10]。2010 年的美国中期选举中，大量社交机器人在社交平台上散布成千上万的虚假消息，支持某位候选人，并污蔑其竞争对手[11]。而在 2016 年的美国大选中，很多人相信是社交机器人参与并干扰大选，使特朗普能够反败为胜成功当选了总统[12]。

这些不符合人类社会道德标准的行为，已然严重影响了人类社会的正常生活秩序。

此外，社交网络上充斥的各种水军和僵尸粉实际上也与恶意社交机器人相关。网络水军是指出于政治或经济等目的对在线社交网络中的信息进行推广，使目标在短时间内大范围扩散的网络用户群体[13]。其中网络水军有一部分是由人类用户充当的，他们在主流论坛中大量发帖炒作话题[14]；还有一部分则是软件机器人，它们受控于攻击者并在互

联网中制造、传播虚假意见和垃圾信息[15]。网络水军不但干扰正常的网络流量，影响用户体验，还会散布不实信息，威胁公共秩序。

僵尸粉则是指由特定软件生成的恶意账号[16]。除了为特定账号增加粉丝数营造虚假的繁荣景象，僵尸粉还会传播各种营销信息，严重威胁社交平台的公信力。

这些充当网络水军的软件机器人、由软件产生的僵尸粉和恶意社交机器人一样都会破坏正常的网络生态环境，但它们与恶意社交机器人仍然存在一些差异。

首先，从产生机制上来看，与传统的网络水军和僵尸粉相比，恶意社交机器人是一种全自动的软件程序，能够实现自学习，具备一定反检测能力，对人类行为的模仿更加彻底；其次，从活动意图上来看，恶意社交机器人不再局限于仅完成推送垃圾信息给用户[17]或攻击其他用户的任务，它们还能搜集用户的账户资料和相册等，通过售卖、勒索等方式对用户造成更加严重的伤害；第三，从运行过程来看，网络水军和僵尸粉内部并无过多的交流，大多是以孤立点的形式存在。而在社交机器人网络中，机器人之间经常相互转发状态和提及对方[18]，容易形成社交关系图。

社交网络拥有庞大的用户群体，越来越多的人开始使用各种社交平台，将自己的日常生活与社交平台紧密结合在一起。为了保障用户利益，过滤不良信息，维护社会稳定，保障社会公共安全，研究恶意社交虚拟机器人的检测技术是非常必要的。

2.2　社交机器人的开发技术

了解社交机器人的开发与应用现状，是研究检测技术的基础。社交机器人首先需要在社交网站上拥有账号，才能模仿人类在社交网络中活动。

通常要想拥有一个社交账号，除了劫持正常用户的账号外，社交机器人还可以创建属于自己的账号，这里必须解决 3 个问题。

第一，提供有效的邮箱。由于目前账号创建时各网站都需要用户提供可用的邮箱进行注册，在点击邮箱中的激活链接确认用户身份后，用户才能真正成为该账号的所有者。获取有效的邮箱并不困难，如 10MinuteEmail 等网站可以提供免费的临时邮箱地址，并且不需要用户注册；而 MailRu 等网站则不会限制每次浏览器会话时所创建的邮箱数量。若申请者盗用了他人的邮箱创建账号，但又需要获得激活链接，那么，可以编写简单的脚本下载激活邮件，然后点击激活链接发送 HTTP 请求，即可完成激活工作并获得社交账号。

第二，创建用户简介。社交网络中通常需要用户完善自己的个人简介，很多好友关系都是建立在这种基础之上[19]。因此，社交机器人需要填写足够引人注意的个人简介，才能达到吸引粉丝、成为意见领袖的目的。其中，最重要的就是用户头像的设置。有报道表明，用户在社交媒体上的影响力与使用的头像有关，拥有一个好看的头像可以提高浏览量与关注度。许多人会将个人照片上传到 HotOrNot 之类的网站上以获得关注度，这些网站还会根据性别和年龄对照片进行分类。社交机器人可以很方便地从网站上爬取有吸引力的照片作为自己的账户头像，或将这些照片添加到个人相册中以提高关注度。

第三，解决验证码问题。全自动区分计算机和人类的图灵测试（CAPTCHA, completely automated public turing test to tell computers and humans apart）是一种广泛应用的反机器人技术。该技术认为可以设计一些人类能轻易通过但计算机却难以通过的任务来区分人类和机器人。但随着图像处理技术的进步，机器人有着越来越强的识别能力，使得 CAPTCHA 的区分越来越困难。另外，攻击者将验证码识别任务外包给廉价劳动力则更加简单高效，识别的准确率也更高[20]。

为了能与人类用户交流，社交机器人必须模仿人类行为，具备与人类对话的能力。机器学习（machine learning）是一门专门研究计算机如何模拟或实现人类的学习行为，从而重新组织已有的知识结构使之不断改善自身性能的学科[21]。一般用机器学习的方法解决问题时，需要经过数据预处理、特征提取、特征选择和识别预测等过程。通过选取合适的特征，计算机能够逐渐提高自身的学习能力并改善解决问题的结果。而神经网络技术则试图模仿大脑的神经元之间传递、处理信息的方式，不断训练、评估、纠错以提高预测准确率。例如，DeepDrumpf 是 Hayes 创建的一个运行在 Twitter 上的机器人，它是一个基于深度学习神经网络的人工智能系统。DeepDrumpf 以特朗普在 Twitter 上的言语作为训练语料，通过不断学习，最终实现在社交网络上模仿特朗普发布信息。

目前许多社交网络平台，包括国内的新浪微博、阿里巴巴等平台都推出了反机器人机制，攻击者在设计社交机器人还需要考虑如何避开这种检测。文献[8]中介绍了有些社交机器人在实施恶意行为时往往会将活动分布在多个进程上，让每个进程都承担一部分任务。而所谓的反机器人机制一般都是针对单进程行为的检测，因此，这种利用多进程执行任务的机器人能够轻易逃过检测。

从最初简单完成抓取数据的任务，到干扰社会舆论、窃取用户隐私，再到现在越来越灵活地躲避反僵尸检测机制，社交机器人表现得越来越智能化，其学习能力不断提高，这对检测恶意社交机器人也提出了越来越高的要求。

3 恶意社交机器人的检测问题

自从意识到社交机器人会带来负面影响，各界学者已经尝试了许多方法去分析和识别社交机器人。一些社交平台上正逐渐意识到恶意社交机器人的危害，并提出了一些应对措施，效果却不尽人意。此外，恶意社交机器人的检测与社交机器人的检测不同，具有比较明显的差异。恶意社交机器人通常会表现出与正常机器人不同的行为特征，如它们会向用户邮箱中投放大量广告、发送指向恶意网站的链接，或者在社交平台上疯狂刷屏、窃取用户个人信息，并且还会隐瞒自己的地理位置等。这些特征对检测工作具有明确的导向性，可以用来识别其中的恶意社交机器人。

3.1 问题定义

目前社交网络中的用户主要由人类用户、恶意社交机器人和正常机器人构成。但与恶意社交机器人相比，人类用户和正常机器人从事恶意行为的可能性较小，且其行为特征与恶意机器人相比具有比较明显的区别。因此可以将人类用户与正常机器人统一定义为正常用户，在线社交网络中恶意社交机器人的检测就变成一个二元分类问题，其形式化定义如下。

定义 1 $U=\{u_1,u_2,\cdots,u_{|U|}\}$ 是待检测社交网络平台上的账号集合，类别集合 $C=\{C_{\mathrm{M}},C_{\mathrm{N}}\}$，$C_{\mathrm{M}}$ 代表恶意社交机器人（malicious socialbot）集合，C_{N} 代表正常用户（normal account）集合。社交网络中的恶意社交机器人检测即发现账号 u_i 是否属于恶意社交机器人集合 C_{M}，决策函数为

$$\varphi(u_i,c_j):U\times C\to\{0,1\}(1\leqslant i\leqslant|U|,j\in\{M,N\})$$

其中，$\varphi(u_i,c_j)$ 是一个二分类函数，即

$$\varphi\left(u_i,c_j\right)=\begin{cases}1, & u_i\in C_M\\ 0, & u_i\in C_N\end{cases}$$

3.2 问题描述

恶意社交机器人检测需要解决 2 个重要问题。

1) 提取区分度明显、确保检测准确率的识别特征。

特征的选取主要从 2 个方面展开。一是利用相关领域的专业知识，根据对某些属性的统计结果从中选取具有代表性和区分度的属性。这种方法提取出来的特征还需要通过实验进一步验证和完善，以达到更好分类的目的；第二种方法则是利用机器学习的方式产生。例如，Filter 方法就是对每一维的特征“打分”，给每一维特征赋予权重，用权重表示相应特征的重要性。然后按照权重对特征进行排序，从中选择更加合适的特征完成分类过程。在选取多个特征后，为了便于计算和实验，还需要继续进行属性约简，降低特征的维度。有些特征并不会明显提高分类的精确度，因此，可以从属性集中选取一个最小的属性集，这个属性集能完全确定分类，并且约简前后的属性集对论域的分类结果相同。由于社交机器人处于不断地变化和进化中[22]，人工智能技术和反检测机制往往会使一些刚提出的检测特征很快失效，对识别恶意社交机器人没有帮助，因此，需要仔细分析恶意社交机器人的活动，提出更具有区分度、演化性更强的特征。

2) 在现有研究基础上设计准确率更高、综合代价更小的检测算法。

虽然可以借用对于检测网络水军和僵尸粉的已有研究成果，但这些方法仍然存在检测准确率不高、算法通用性差等问题。文献[13]中提出，现有的检测方法大多针对某一类网络水军，在面对海量数据时需要将这些方法结合使用，但也会因此增加检测的复杂性和时空代价。此外，各检测方案的准确率差异较大，即使使用相同的检测特征和实验数据，若基于不同的算法构建分类器也会影响检测结果。所以上述问题也是检测恶意社交机器人所必须要解决的。文献[18]中提出，如果检测方案是针对同时攻击某一组账户的社交机器人，在检测独立攻击不同组账号的社交机器人的时候，这种检测方案很可能失效。所以，必须考虑到社交机器人的运行

情况，结合各种机器学习算法的特点，提出一种检测效果更好、综合代价更低的检测方案。

4　恶意社交机器人的检测方法研究

目前，围绕上述两大问题的研究现状及存在的局限性分析如下。

4.1　检测特征的选取研究内容

检测特征的选取研究主要从特征选取的重要性分析、选取的原则、选取的特征以及特征的有效性验证等方面展开。

4.1.1　检测特征选取的重要性

恶意社交机器人的检测过程中最重要的就是选取恰当的识别特征。有价值的识别特征在确定识别范围和提高识别准确率等方面能够起到关键性作用。

张宇翔等[23]对此进行了研究，他们以新浪微博的数据为基础，将不同分类器使用同一特征组和同一分类器使用不同特征组的分类结果进行对比分析，得出特征组的选择比分类器的改进更加重要的结论，并认为要从用户行为、内容信息和社会关系等多方面定义检测特征。

Fazil 等[24]在研究恶意社交机器人攻击对象的时候也对检测特征进行了研究，并通过特征选择和排序算法确定检测的主导特征，这对分析社交机器人的攻击目标具有重要作用。

4.1.2　检测特征选取的原则

事实上，在选取怎样的相关属性作为检测特征时，本研究小组也针对微博中僵尸粉的演化特征进行了实证研究，并提出了一些特征选取的原则。本文在进一步归纳总结后认为选取检测特征时至少需要遵循以下 3 条原则。

1) 根据统计指标选取区分度大的特征，尤其是恶意社交机器人与普通用户差异较大的特征属性，必要时可以先验证特征有效性。

2) 确保特征之间的相关性最小，即尽量保证特征之间不相关，避免评价维度上的重合。

3) 选取的特征尽可能全面，除了应包含用户的个人基本信息等静态特征以外，还应包括用户的行为传播、内容和情感等动态特征，以及网络关系特征，特别是一些演化性较强、恶意社交机器人难以学习的特征。通常选取的特征包含范围越广，对用户的分析就越精确。

另外，为了便于实验和检测，所提出的各种检测特征必须能够量化。

4.1.3　检测特征的选取

针对已有恶意社交机器人的研究，本文分析了相关文献中提出的检测特征，将检测特征分为“三类五种”，即用户静态特征、动态传播特征和关系演化特征 3 类，涉及用户特征、内容特征、情感特征、传播特征和演化特征 5 种。具体内容如图 1 所示。

在早期研究中，学者们倾向于选择用户的自然属性，如用户的地理位置、粉丝数等作为检测特征。但随着社交机器人自学习能力的增强，它们可以通过修改定位、关注更多账号等方式伪造成合法用户，若只采用用户属性作为检测特征，检测准确率将较低；之后研究者们注意到消息内容、情感等动态传播特征，如原创微博的数量和微博中蕴含的情绪等。由于恶意社交机器人大多转发他人状态，原创微博较少，这类特征对提高检测准确率有一定积极作用。随着检测技术的进一步发展，研究者们开始提出一些关系演化特征，如社交结构中的聚类系数和节点核数等。

随着恶意社交机器人的演化升级，其在特征复杂性和伪装程度等方面都发生了巨大变化，其行为逐渐趋同正常用户，导致检测复杂度增加。但是，恶意社交机器人往往会相互配合活动，通常是由一个机器人发布状态，其他机器人有组织地评论、转发。这种合作式的活动能够使社交机器人在社交网络上形成较为明显的图结构，并且区别于普通用户的网络图结构。正常用户的状态在得到其他用户的转发或评论后，博主会与其他用户交流，而恶意社交机器人则不会。因此，从社交网络图结构来看，恶意社交机器人网络图结构中很可能会出现大多数节点的出度为 0 或入度为 0 的现象，从而造成恶意社交机器人的社交图结构中聚类系数和节点核数与正常用户有明显差异，节点间的连线较为稀疏。

此外，恶意社交机器人大都存在个人简介填写不全、内容发布混乱并包含大量链接、隐瞒地理位置及关注用户多为机器人等问题，与正常用户存在较大差异。在选取检测特征的时候，可以从这些方面着手。

4.1.4　检测特征的有效性验证

为了更好地对比恶意社交机器人与正常用户在特征上的差异，确保检测特征的可靠性与高效性，在进行分类检测前还必须验证所选取特征的有

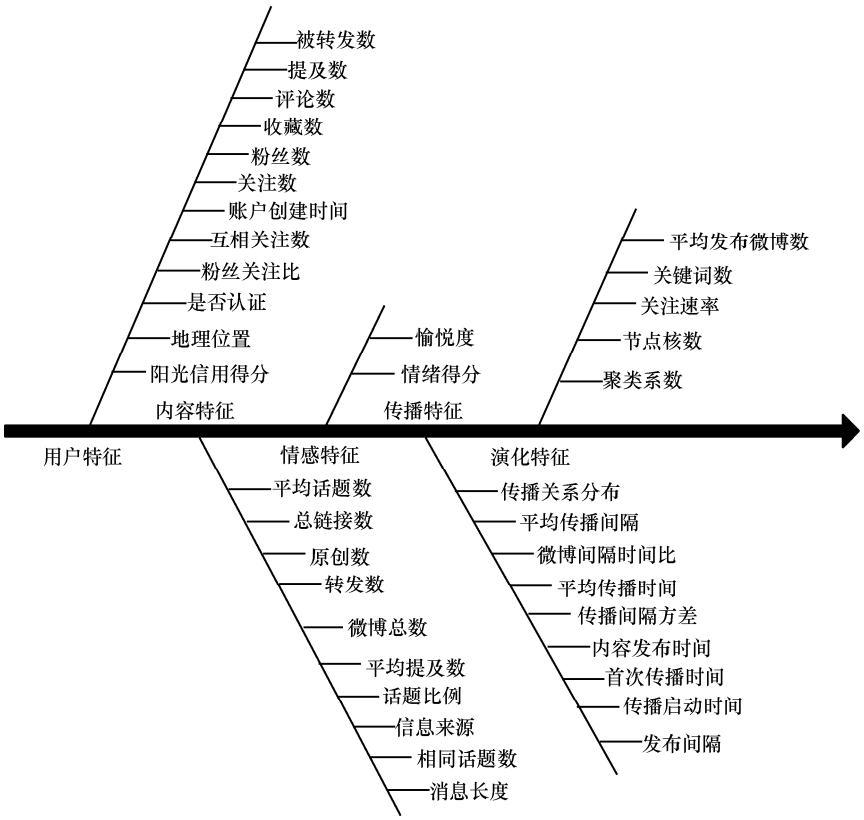

图 1　检测特征研究现状分析

效性。本研究小组曾使用 R 语言绘制每个特征的累积分布函数图[25]，分析某一特征在各区间内的数据分布，从而得到该特征对检测效果的贡献。还可以使用前向搜索方法，从 0 开始每次增量地从 n 个特征中选一个加入特征集，交叉验证并计算错误概率。如此循环 n 次，选择错误率最低的一组作为最终的检测特征。

4.2　检测方案的研究内容

针对在线社交网络中恶意社交机器人带来的各种危害，各界学者提出了一些检测方案。虽然和社交机器人相比，网络水军与僵尸粉的行为表现、日常活动等有所不同，但它们都属于在线社交网络中的异常用户，其社交行为都偏离了正常社交方式，因此，在检测恶意社交机器人时可以借鉴网络水军和僵尸粉检测的思想。根据这些方案采用的核心思想不同，本文将它们分为 4 类，分别是：基于特征的方案、基于机器学习的方案、基于图论的方案和基于众包的方案。下面分别对这些方案的检测思想及其局限性作一些介绍。

4.2.1　基于特征的检测方案

利用恶意社交机器人与正常用户的个人简介、内容发布以及传播行为等方面的不同，基于特征和基于机器学习的方案都是将检测恶意社交机器人看作是一个分类问题。

为了实现扩大粉丝数量、增强个人影响力的目的，恶意社交机器人向其他用户发送大量好友请求，并在短时间内发布大量内容迅速刷屏。由于社交机器人本质上仍然是一段自动运行的程序，其语言组织与识别能力、情感表达和社交时间等与正常用户存在较大差异。这些差异恰恰可以作为判断该用户是否是恶意机器人的依据，并且已经在实践中得到了应用。

文献[4]中提出的 BotOrNot 最早于 2014 年发布，是 Twitter 公开的第一个检测社交机器人的接口。该系统从网络、用户、交友、时间、内容和情感等 6 类特征入手，提取 1 000 多种属性特征进行分析从而判断待检测用户属于恶意社交机器人还是正常用户。系统先后采用随机森林模型（random forest）、AdaBoost、线性回归模型（logistic regression）和决策树模型（decision tree）作为分类器分别进行预测，经过十折交叉验证后发现随机森林模型的分类效果最好，准确率达到 95%，因此，将该

模型作为最终的检测模型。系统运行时，先采集这 2 类账号的样本数据，包括账号名称、最近发布的状态等历史社交信息进行样本训练。然后利用 Twitter 的搜索接口，收集待检测账号最近的 200 个帖子和最近被提及的 100 个帖子，分析上述 1 000 多种特征所包含的信息，判断该账号属于恶意机器人的可能性。

与传统的检测方案不同，BotOrNot 系统的特点体现在以下几个方面。

1) 考虑用户发布内容中蕴含的情绪和态度，包括愉悦度、PAD 情绪分数等，分析账号所有者在发布内容时的情绪状况。

2) 提取内容特征时利用自然语言处理技术提取语言特征，尤其是词性标注方面。例如，推文中出现的名词、动词以及形容词的频率等。

3) 除了简单地提供分类结果，BotOrNot 还提供一系列交互式的可视化工具，在显示分类结果的同时展示特征分析结果。

与 BotOrNot 类似，文献[28]中介绍的 Botometer 也是通过收集用户账号的各种信息，包括用户元数据、好友数据、时间特征以及情感分析结果后，最终给出该用户与机器人的相似程度得分。

然而考虑到社交机器人对人类行为的模仿程度越来越高，对于那些更加复杂并且兼有人类和机器人行为特征的半机器人，现有检测系统还无法准确检测到这些半机器人的存在。例如，社交机器人会从互联网上下载相关信息填充自己的个人简介，或者模仿正常用户对他人的状态加以评论，只凭借账号内容等信息判断待检测账号属于恶意机器人的可能性还不足以应对机器人的智能化发展和更新换代。

但是，近年来机器学习和众包模式的蓬勃发展提供了检测半机器人的新思路：利用一些机器人无法学习的特征并结合机器学习或许能够提高检测准确率，或采用众包模式将识别任务分发给多个标注人员完成。

4.2.2 基于机器学习的检测方案

机器学习算法是基于数据，先从数据中挖掘或分析出数据规律或模型，然后利用规律或模型对未知数据进行分析或预测的算法。它的计算过程更多地依赖于数学模型、统计和概率相关知识。除此之外，机器学习算法更多地利用数据或历史经验进行自我改善，可以分为有监督、无监督和半监督的学习方式[33]。

由于无监督分类事先不需要对分类器进行先验知识的学习，而是直接使用样本数据进行分类，对分类后的类别特征不能确定，并且需要进行大量的分析才可能获得较好的分类结果，因此，在恶意社交机器人的检测过程未见用这种方法。有监督分类则需要提供训练样本不断计算，从样本中学习选择特征参数，建立判别函数后才可以对被识别的样本进行分类。这种方法能够形成符合特征的分类模型，有效利用先验数据信息，还减少了人为因素的干预，在实际问题解决中得到了广泛应用。朴素贝叶斯算法、*K* 近邻算法、C4.5 决策树算法和随机森林算法等都已被用于识别社交网络中的网络水军和僵尸粉。

1) 基于贝叶斯模型的检测方案

朴素贝叶斯预测是一种以动态模型为研究对象的预测方法，它不但需要分析模型信息和数据信息，还充分运用了已知的先验信息。基于朴素贝叶斯算法的检测方案的主要思想是：假设各特征属性之间相互独立，根据某对象的先验概率利用贝叶斯公式计算其后验概率，选择具有最大后验概率的类作为该对象所属的类。这种检测方法简单易懂，在实际操作中也较为简便，得到了许多学者的青睐。

张艳梅等[21]提出了一种基于贝叶斯模型的检测网络水军的算法，该算法使用了 6 个属性特征：粉丝关注比、平均发布微博数、互相关注数、综合质量评价、收藏数和阳光信用数。由于网络水军一般是带有特殊目的地去关注用户，其粉丝数与关注数的比值与正常用户相比，差距较大；其次，正常用户在社交网站上注册账号后，其在线时间往往与个人的空闲时间有关，而水军为了达到保持活跃状态和吸引用户的目的，会在短时间内大量刷屏，具体的表现形式就是其平均发布微博数明显上升。自 2015 年起，新浪微博为每个账号添加了阳光信用指标。这一属性由官方发布，具有一定的可信性，符合作为检测特征的条件。之后，分别建立水军与非水军的属性阈值矩阵和概率矩阵，通过人工标注的方式获取水军和非水军的数量，并计算填写上述四个矩阵的内容。

一个未分类的数据 x 表示为 $x=\{a_1,\cdots,a_m\}$，每个 a_i 分别表示 x 对应于属性 i 的值，类别集 $B=\{y_1,y_2\}$，y_1 代表非水军，y_2 代表水军，则 $P(y_1|x)$、$P(y_2|x)$中值较大的 y_i 即为分类结果。

根据贝叶斯定理得

$$P(y_i \mid x)=\frac{P(x \mid y_i)P(y_i)}{P(x)} \tag{1}$$

其中， $P(x)$ 为常数。由于测试数据集 $P(y_i)$确定，只需计算 $P(x|y_i)$，$P(x|y_i)$计算公式为

$$P(x \mid y_i)=P(a_i \mid y_i)P(a_2 \mid y_i)\cdots P(a_m \mid y_i) \tag{2}$$

若判定为非水军，则不计数，否则在水军计数变量上加 1，最终得到判定为水军的用户个数。

实验结果表明，这种基于朴素贝叶斯算法的检测方案的水军识别率较高，特别是加入阈值优化算法之后能够明显提高非水军的识别率，具有一定的应用前景。

但是，朴素贝叶斯模型本身就存在一些缺陷。贝叶斯模型是建立在各特征属性相互独立的基础之上[28]，这在真实的社交网络中难以实现。因为各检测特征之间通常存在一定的相关性，不可能完全独立。此外，从分类器性能角度来看，文献[24]中针对同一数据分别使用了朴素贝叶斯、剪枝决策树和随机森林 3 种算法构建分类器，最终结果表明朴素贝叶斯算法的分类性能并不是最优，分类效果有待提高。

因此，有人提出了基于复合分类模型的识别方法。

2) 基于复合分类模型的检测算法

考虑到朴素贝叶斯模型存在的固有缺陷会对分类算法的精度产生一定的影响[29]，而 K 近邻算法虽然精度高，但需要计算样本各训练点之间的距离，造成算法的开销很大[30]。谈磊等[31]将朴素贝叶斯模型与 K 近邻模型结合起来，提出了一种基于复合分类模型的算法来检测社交网络中的恶意用户。出于提高检测准确率的目的，该方案认为对于具体该使用哪种算法完成检测，必须确定一种算法选择策略来选择相应的检测算法。以文献[31]中的方案为例，该方案提出选择相关性最强的两个属性的比值作为检测算法的选择标准。先利用式(3) 计算特征属性 X 和 Y 之间的相关性。假设相关性最强的 2 个属性分别为 F_1 和 F_2，则将 $\left(\frac{F_1}{F_2}\right)$ 作为决定应该选择哪种分类算法的标准。

$$R(X,Y)=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2} \tag{3}$$

以 $m=\frac{F_1}{F_2}$ 作为阈值，若样本 $\boldsymbol{X}$ 的计算结果大于 m，则选择 K 近邻算法，根据式(5)计算样本 $\boldsymbol{X}$ 与训练集中各数据点之间的距离，并选择距离最小的数据点 $\boldsymbol{X_i}$ 所属的类别作为 $\boldsymbol{X}$ 的类别。

$$D(\boldsymbol{X}_i,\boldsymbol{X})=\sqrt{U(\boldsymbol{X}_i,\boldsymbol{X})+\sum_{j=1}^{3}\left(\frac{F_j(\boldsymbol{X}_i)+F_j(\boldsymbol{X})}{F_j(\boldsymbol{X}_i)-F_j(\boldsymbol{X})}\right)^2-\left(\frac{S(\boldsymbol{X}_i)+S(\boldsymbol{X})}{S(\boldsymbol{X}_i)-S(\boldsymbol{X})}\right)^2} \tag{4}$$

$$U(\boldsymbol{X}_i,\boldsymbol{X})=(U(\boldsymbol{X}_i)-U(\boldsymbol{X}))^2 \tag{5}$$

若计算结果小于阈值，则选择贝叶斯算法。与 1)中介绍的方法相似，先确定相应的阈值矩阵，然后根据给定待检测账号 $\boldsymbol{X}$ 的相关属性，计算每种类别的后验概率，选择后验概率值最大的类别作为 $\boldsymbol{X}$ 所属的类别即可。

对新浪微博上的数据进行实验的结果表明，这种方案的检测准确性与提供的数据相近。但是对于一些相关性不高的属性而言，上述方案中关于 K 近邻算法的部分亟待完善，是否需要在这部分中计算加权距离仍然值得研究。另外，如果要将恶意用户中的恶意机器人划分出来，还需要进一步考虑特征的选取问题。

3) 基于 C5 决策树的检测方案

虽然早有学者提出了多种检测网络水军的算法，包括基于黑名单的算法[32]、基于用户特征的算法[33]以及基于文本的方法[34]等，但是陈侃等认为这些检测算法的范围局限在单一类别的水军，难以应对当今大数据环境下水军种类多样的现状。这些方法在实际使用时反而会增加检测的复杂性，通用性也比较差。由于决策树具有决策过程非常直观、简单易懂的特点，是一种有监督的分类方式，产生的分类规则非常直观易理解[35]，他们提出了一种基于 C5 决策树的检测算法。与其他方案重点关注账号的个人特征如粉丝数、信息完善程度等不同的是，该方案中信息的传播和用户交互行为得到了更多的关注。

设 p 为一条信息，并用相应的特征向量表示为一个多元组。实验数据为已知所属类型的用户及其发布的信息 S（其中，正常用户的信息为 NS，网络水军的信息为 SS）和需要通过检测判别身份的用户数据 $\boldsymbol{X}$。首先，计算所有信息 S 的信息熵

$$E(S)=-\frac{1}{|S|}(|NS|)\mathrm{lb}\frac{|NS|}{|S|}+\frac{1}{|S|}(|SS|)\mathrm{lb}\frac{|SS|}{|S|} \quad (6)$$

对于信息 p 的每个属性 D，根据取值将样本集 S 划分为 m 个子区间，并计算 D 的条件熵，计算式为

$$E(S|D)=-\sum_{i=1}^{m}\frac{|S_i|}{|S|}E(S_i) \quad (7)$$

则相应的属性 D 的信息增益计算式为

$$Gain(D)=E(S)-E(S|D) \quad (8)$$

属性 D 的信息增益比率计算式为

$$GainRatio(D)=\frac{Gain(D)}{SplitI(D)} \quad (9)$$

$$SplitI(D)=-\frac{1}{|S|}\left(\sum_{i=1}^{m}|S_i|\mathrm{lb}\frac{|S_i|}{|S|}\right) \quad (10)$$

根据每条信息 p 的 6 种特征，计算相应属性的信息增益比率，并生成最终的 C5 决策树。决策树构造成功后，即可对需要检测的数据 $\boldsymbol{X}$ 进行判别。

对于同样的数据执行 SVM 算法和 RBF 算法后发现，基于 C5 决策树算法的分类器具有更高的准确率和召回率，各种性能评价指标的综合得分也证明了这种检测方案的有效性和准确性。但决策树对连续性的字段比较难预测，尤其是对存在时间顺序的数据还需要很多预处理工作，这就使基于决策树的检测方案具有很大的限制性，无法投入广泛使用。

基于机器学习的恶意社交机器人检测主要包括训练学习和检测 2 个阶段。虽然目前机器学习在分类问题中已经得到广泛使用，但存在 2 个对分类结果影响非常大的因素：语料库和算法模型。若使用有监督的学习方法，则需要使用大量的训练样本，而社交网络中用户数据类型多、数据量大，采用人工标注的方式将产生很大的代价。因此可以考虑结合无监督学习提出新的检测方案。

无监督学习作为一种直接对数据建模的学习方式，能够用于标记样本，不但可以降低付出的人工代价，还能减少人工干预提高样本标记的准确性。此外，各种算法模型都存在一些缺陷，如朴素贝叶斯算法必须基于各特征相互独立的假设，决策树算法不利于预测连续性字段等，这些都将对分类结果产生一定的影响。

4.2.3 基于图论的检测算法

基于图论的方案则重点关注恶意社交机器人与正常用户各自所形成的社交网络图。根据图的不同结构和连接方式，将检测恶意社交机器人的问题转化为图中异常子图的检测问题，并利用图挖掘等相关算法来识别恶意社交机器人。

随着人工智能技术的发展，机器向人类学习的能力不断增强，社交机器人发布的内容越来越趋向于人类的交流习惯。大量使用重复形容词吸引用户注意的方式已被社交机器人所抛弃，因此，基于内容分析识别恶意社交机器人的方法效果有限，难以适应时代的发展[36]。

考虑到社交网络中存在一个重要特点，即账号之间存在联系[37]。恶意社交机器人在从事社交活动时的显著特征之一就是通过发布大量带有特定标签的状态为相关话题宣传造势，从而吸引更多的用户关注。往往是一个机器人发布状态后其他机器人转发或评论，由此形成的社交图结构具有明显的中心点，接近于星型模型。这是因为社交机器人的粉丝大多互不相识，彼此之间互动较少。但正常用户的关注者与粉丝大多来源于现实好友，且会花时间与他人交流以维护好友关系。正常用户发布状态后，其好友圈中的对象互相评论，互动操作比较频繁，因此，正常用户的社交图结构明显区别于机器人的图结构，节点之间的连线分布更为均匀。此外，由于社交机器人是出于一定目的接近人类用户，会在短期内大量关注其他用户，从而造成其关注数与粉丝数的比例严重失调，因此，恶意社交机器人的关注关系与正常用户有较大差别。

基于图论的检测方案就是利用这些区别来检测恶意社交机器人，不但在提取检测特征时重点关注用户的关系特征，在确定检测方法时也运用了图论思想。这种方案实际上是以恶意社交机器人与正常用户在社交网络中形成的网络图具有不同结构为基础的，所以基于图论的检测方案的关键在于构建用户的社交网络图，然后利用图挖掘等相关算法分析社交网络图，从而判断该用户是否属于恶意社交机器人。

程晓涛等[38]据此提出了基于社交关系图的检测方案。他们认为，社交机器人的行为特征会随着技术的进步而发生变化，但其在社交网络结构上的关系不会轻易发生变化。如果将社交网络中的任一用户 u 视为一个节点，那么用户 u 的好友关系网络可以看作一个有向图 $G=\{V, E\}$。其中，V 表示用户 u 的所有关注者集合，E 表示这些关注者之间的关注关系。将节点 i 的度表示为 k_i，E_i 是节点 i 的

k_i 个邻居节点之间实际存在的邻居对的个数。通过提取聚类系数分析用户 u 的好友中彼此存在好友关系的概率，衡量 u 的好友圈的紧密程度。C_i 的计算式为

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{11}$$

C_i 越大，表示好友关系越多，好友圈越紧密。双向关注比也是一个有效的检测特征。由于社交机器人通常是随机选择要关注的对象，无法强求其他用户关注自己。而正常用户则是根据自己的兴趣爱好和现实生活有选择地关注他人，这就造成了社交机器人与正常用户的双向关注比存在较大差异。

此外，用户的节点核数也可以作为检测特征之一。一个图的 k 核是指反复去掉网络图中度小于或等于 k 的节点后所剩余的子图。若一个节点存在 k 核，而在 k+1 核中被移除，则称这个节点的核数为 k。通过计算用户的节点核数，分析用户与其邻居节点之间的连通关系。社交机器人由于朋友关系不够紧密，其核数一般较低。文献[38]中还针对是否使用图结构特征进行了对比实验，结果表明社交网络图结构特征能够明显提高检测结果。

鉴于社交机器人与正常用户的社交网络结构具有较大差异，可以将恶意社交机器人的检测问题转化为异常子图的识别问题。除了分析用户的好友关系之外，恶意社交机器人与正常用户的状态发布和传播也存在不同的图结构。但是考虑到现实中社交网络的复杂性，在应用基于图的检测方案时需要对方案进行严格的验证，特别是图特征的选取。

4.2.4 基于众包的检测方案

以时下热门的众包策略为基础，基于众包的方案将检测任务分发给多个标注人员，通过投票决策的方式判断该用户是否为恶意社交机器人。

众包是一种分布式的问题解决和生产模式，即公司或机构把内部的任务以公开的方式外包给非特定的大众网络的行为[39]。谭婷婷等[40]认为，众包是一种网络化社会生产，把需要员工完成的工作任务分发给普通大众，使每个人不至于负担过重；其次，众包具有一定的弹性，方便招聘新的人员；最重要的是，众包可以解决一些自动化技术无法解决的问题。例如，自 2006 年以来，“众包”便开始大量出现在网络和媒体上。目前广泛应用的百度百科词条编辑、将未知号码加入黑名单等，都是互联网引导普通用户作为志愿者参与到特定项目的平台建设中的典型案例。

而在商业化应用中，出于品牌推广和产品炒作的考虑，各公司也会采用众包策略。例如，雪铁龙公司曾在 Facebook 上推出一个应用程序，允许用户为该公司的一款新车选择他们所喜欢的设计方案。这种做法既达到了产品和企业宣传的目的，提高了用户的参与度，还迎合了用户的需求。Google 旗下的 Waze 地图根据用户提交的报告，能够向司机提供到达目的地最快捷的路线，甚至发现一些不为人所知的道路。该应用遵循了“我为人人，人人为我”的理念，并且能够及时更新路况。这种做法也确实使许多人受益。

亚马逊土耳其机器人（AMT, Amazon mechanical Turk）是一种 Web 服务应用程序接口，开发者认为在许多任务的执行过程中，人类的执行结果比机器人更有效、更可靠，如识别照片或视频中的对象、研究数据细节等。AMT 是一种提供基于众包思想的网络服务的平台，请求者发布任务并设定工作报酬。每个任务都称为一个 HIT（human intelligence task），工作者申请并完成任务后将获得相应的报酬。目前该服务已经被许多研究者用于采集数据、完成一些机器不方便执行的任务。

借助 AMT 平台，Wang 等[41]提出了基于众包的异常账号检测算法。他们认为，人类在察觉交流中的细微差别和分辨照片真伪的能力远远超过了社交机器人，能够非常容易地判断用户是否为社交机器人。文献[9]中就曾指出，人类具有一定的辨识能力，对于一些明显可疑的异常账户，真正的人类用户并不会转发这些账户发布的状态。Wang 等设计了一个可扩展的检测系统来识别社交机器人，并在 AMT 上招募 Turker 参与检测。该系统分为 2 层：第一层为过滤层，系统根据一些特征判断用户是否可疑，若可疑则收集该账户的相关信息；第二层是众包层，将第一层中获取的可疑账户信息分发给标注者进行标注。实验数据来源于 Facebook 和人人网。实验中，标注者通过网页查看可疑账户的资料，包括用户的基本信息、主页和相册等，根据以上信息判断账户的身份。作为对比，实验中邀请语言学专家和研究生组成专家组作为第一组标注者，另一组则是由社会科学专业的本科生组成。最终结果表明专家组的识别准确率最高，而 Turker 组的准确率最低。

这一结果同时也表明，虽然基于众包的检测算

法准确率较高，但会随着工作时间的增加而降低。此外，基于众包的检测算法也暴露出一些缺陷。第一，由于众包用户的匿名性和组织方式的开放与自由，如何审查众包工作者的资质和能力需要进一步讨论。很可能会出现一些仅为获得报酬而不认真完成任务的作弊标注者[42]，这将干扰检测结果的准确性；第二，这种检测算法在大规模数据上使用成本太高。如果为了节约成本减少雇佣的人数，又将导致检测准确率下降；第三，将用户的个人信息暴露给外部工作人员会产生新的隐私保护问题[43]；第四，若为了提高检测准确率引入投票系统，将可能出现众包系统被恶意攻击的情况，即多数标注者故意扰乱决策结果。但是正如《人民日报》中关于众包的报道一样，“众包模式，大势所趋”。目前总结的这些问题，将来也许会出现更妥善的解决方案。

4.3　研究小结

事实上，上述这些检测方案之间并没有明确的界限，如基于机器学习的方案也使用了检测特征基于图论的检测方案中很可能也包含机器学习的相关算法。不管提出何种检测方案，最重要的是方案中使用的特征和算法能够提高检测的准确率，有助于更好地区分正常用户与恶意机器人。

本研究小组今后工作的思路如下。

1) 更加注重关系类特征和演化类特征的选取

从图 1 的特征选取演化过程中可以看出，早期对用户个人信息的静态特征研究已经比较充分，然而，恶意社交机器人能够利用互联网上海量的图片和文字等信息伪装成合法用户并绕过基于用户信息的检测方案。考虑到恶意社交机器人为了扩大影响力，往往会在初期大量发布带标签的状态，并且频繁转发，从而给话题炒出一定热度。它们通常是相互配合活动，即由一个社交机器人发布状态，其他机器人直接转发，相互之间没有交流，发布状态者与转发者之间缺少互动。因而从社交结构图上来看，恶意社交机器人的社交结构与正常用户存在较大差别，图中孤立点的个数和节点的出入度等明显异于正常用户，各节点之间的连线也较为稀疏。因此，今后在选取检测特征时，除了要遵循本文提出的 3 个基本原则外，在特征类别方面倾向于选取关系类特征，如节点的度数等。

此外，虽然恶意社交机器人具备一定的自学习能力，但可以设定活动频率、节点核数等演化类特征加以判断。活动频率是指单位时间内用户发布、转发的状态总数，节点核数则用于衡量节点之间连通关系的强弱程度。由于恶意社交机器人本质上是一段自动运行的程序，且活动目的之一就是煽动舆论、吸引他人注意。然而正常用户是有选择地进行社交活动，一般要符合自己的喜好和意愿，其发布、转发状态的操作有一定的间隔。时间是一种无法改变的自然属性，因此，恶意社交机器人很难模拟正常用户的社交活动频率。此外，正常用户经常会互相评论和分享信息，但社交机器人则缺少互动。这种互动行为的缺乏将导致各节点之间联系不够紧密，易遭到破坏。

2) 利用机器学习，特别是深度学习在自然语言处理和图像识别方面的优势，设计新的检测方法

深度学习作为机器学习的一个重要分支已成为当下学术界的研究热点，其强大的学习能力和在特征提取方面的良好表现使深度学习在计算机视觉、语音识别和自然语言处理等领域得到广泛使用[44]，在识别恶意社交机器人方面也有一定积极作用。

除了运用机器学习方面的算法外，还可以结合图论和众包思想，充分发挥各算法的优越性。

3) 融合已有检测方法，取长补短

图 2 是本文提出的一种恶意社交机器人检测框架，该方法结合了无监督学习和有监督学习的优点。首先以要检测的社交平台作为数据来源，通过爬虫获取用户数据。在数据预处理阶段，完成数据清洗工作。为了提高样本标记的可信度和客观性，本文采用无监督聚类分析的方法标记样本，摆脱人工标记费时费力和主观性强的问题，提高样本标记的效率和客观性，减少标记误差对检测结果造成的干扰；最后根据选取的特征，利用分类效果最优的有监督 SVM 算法对待检测数据进行分类，识别其中的恶意社交机器人账号。

4) 检测算法的并行优化

社交网络中用户数据类型多、数据量大的特点造成传统基于单机实验环境下的检测方案准确率不够高、时空代价大，本文选择从并行化角度解决这一问题。例如，在聚类分析的过程中对数据和算法作并行化，即在分布式计算框架 Hadoop 平台上，利用 MapReduce 编程模式对海量数据实现并行化处理，采用基于聚类的有放回随机抽样的数据划分方法，在实现数据并行的同时有效地保证划分到各个子节点上的训练集的类分布覆盖原始数据集的类分布，从而减少运算的时间和存储空间，提高运

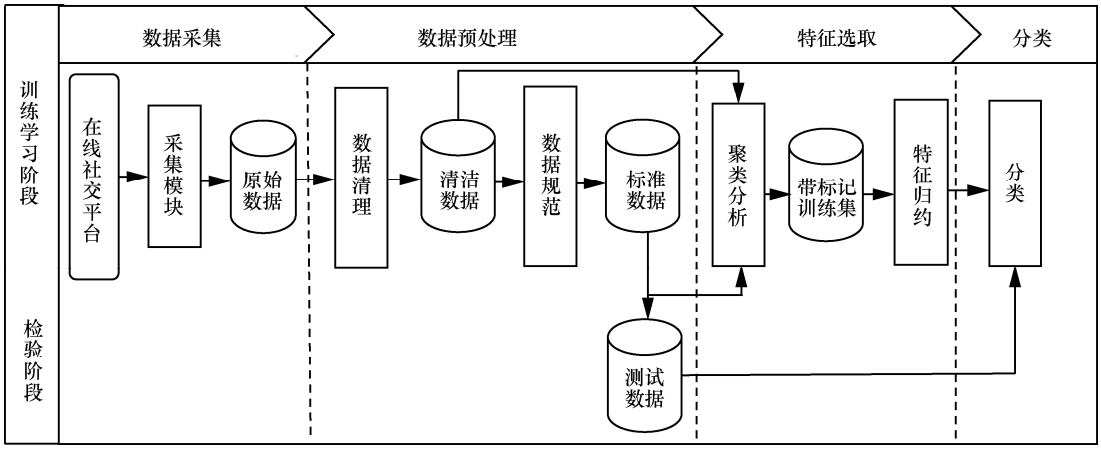

图 2　恶意社交机器人检测框架

算精度；在识别恶意社交机器人时，对 SVM 算法作并行化实现，并利用遗传算法对改进后的并行 SVM 算法的核函数参数及惩罚因子进行更进一步的优化，以实现检测模型的优化，提高训练精度，确保检测结果的准确性。

5　结束语

恶意社交机器人窃取用户隐私、传播虚假消息、影响社会舆论，给个人信息安全、社会公共安全，乃至国家安全造成了严重威胁，对恶意社交机器人的检测研究已经成为在线社交网络安全研究的一个重要问题。

本文在分析社交机器人开发与应用现状的基础上，对目前恶意社交机器人检测特征和检测方法进行了梳理，理清了研究思路，总结出今后在特征选取的过程中更加注重关系演化类特征，提出了一种从并行化角度设计的恶意社交机器人检测方案，并对检测算法的实现作一定的优化，以降低计算代价、提高检测准确率。

应当看到的是，攻击者还在不断引入新技术实施反检测。因此，与恶意社交机器人的对抗将是一个长期的工作。

参考文献：

[1] BOSHMAF Y, MUSLUKHOV I, BEZNOSOV K, et al. Key challenges in defending against malicious social- bots[C]//5th USENIX Conference on Large-scale Exploits and Emergent Threats. Berkeley, CA, USA, 2012: 12-15.

[2] IGAL Z. Bot traffic report 2016[R]. California: Imperva Incapsula, 2017.

[3] DEWANGAN M, KAUSHAL R. SocialBot: behavioral analysis and detection[M]//Singapore: Springer, 2016: 450-46.

[4] DAVIS C A, VAROL O, FERRARA E, et al. Botornot: a system to evaluate social bots[C]//25th International Conference Companion on World Wide Web. Montreal, Quebec, Canada, 2016: 273-274.

[5] CAROLINA ALVES de L S, NICHOLAS B. Is that social bot behaving unethically?[J]. Communications of the ACM, 2017, 60(9): 29-31.

[6] 杜鸣皓. “社交机器人”入侵[J]. 中国品牌, 2017 (2): 36-41.
DU M H. “Social bot” invades[J]. China Brand, 2017 (2): 36-41.

[7] BRITO F, PETIZ I, SALVADOR P, et al. Detecting social-network bots based on multiscale behavioral analysis[C]//The Seventh International Conference on Emerging Security Information, Systems and Technologies. Barcelona, Spain, 2013: 81-85.

[8] JI Y, HE Y, JIANG X, et al. Combating the evasion mechanisms of social bots[J]. Computers & Security, 2016, 58(C): 230-249.

[9] STIEGLITZ S, BRACHTEN F, BERTHELé D, et al. Do social bots (still) act different to humans? – comparing metrics of social bots with those of humans[C]//International Conference on Social Computing and Social Media. Vancouver, BC, Canada, 2017:379-395.

[10] BOSHMAF Y, MUSLUKHOV I, BEZNOSOV K, et al. Design and analysis of a social botnet[J]. Computer Networks, 2013, 57(2): 556-578.

[11] 李娜, 刘洋, 宋明黎. 社交机器人的兴起[J]. 中国计算机学会通讯, 2016, 12(8): 78-86.
LI N, LIU Y, SONG M L. The rise of social bots [J]. Communications of the CCF, 2016, 12(8):78-86.

[12] BESSI A, FERRARA E. Social bots distort the 2016 U.S. Presidential election online discussion[J]. First Monday, 2016, 21(11).

[13] 陈侃, 陈亮, 朱培栋, 等. 基于交互行为的在线社会网络水军检测方法[J]. 通信学报, 2015, 36(7): 120- 128.
CHEN K, CHEN L, ZHU P D, et al. A method of online water army based on interactive behavior[J]. Journal on Communications, 2015, 36(7): 120-128.

[14] 吕晨. 基于用户行为的网络论坛水军检测研究与实现[D]. 成都: 西南交通大学, 2017.
LV C. Research and realization of water army detection based on user behavior[D]. Chengdu: Southwest Jiaotong University, 2017.

[15] 韩忠明, 杨珂, 谭旭升. 利用加权用户关系图的谱分析探测大规模电子商务水军团体[J]. 计算机学报, 2017(4): 939-954.
HAN Z M, YANG K, TAN X S. Using spectrum of wei-ghted graph of users to analyze and detect large-scale e-commerce water army[J]. Chinese Journal of Computers, 2017(4): 939-954.

[16] 陶永才, 王晓慧, 石磊, 等. 基于用户粉丝聚类现象的微博僵尸用户检测[J]. 小型微型计算机系统, 2015, 36(5): 1007-1011.
TAO Y C, WANG X H, SHI L, et al. Detection of zombies on microblog based on the phenomenon of user fanclustering[J]. Journal of Chinese Mini-Micro Computer Systems, 2015, 36(5): 1007-1011.

[17] CHU Z, GIANVECCHIO S, WANG H, et al. Detecting automation of twitter accounts: are you a human, bot, or cyborg?[J]. IEEE Transactions on Dependable & Secure Computing, 2012, 9(6):811-824.

[18] VAROL O, FERRARA E, DAVIS C A, et al. Online human-bot interactions: detection, estimation, and characterization[C]//International Conference on Web and Social Media (ICWSM). AAAI. Montreal, Canada, 2017.

[19] 俞轶楠. 微博用户个人特征、动机、行为和微博吸引力关系的研究[D]. 北京: 清华大学, 2012.
YU Y N. Research of relationship between micro-blog users' personal characteristics, motivation, behavior and attraction on microblog[D]. Beijing: Tsinghua University, 2012.

[20] MOTOYAMA M, LEVCHENKO K, KANICH C, et al. Re: CAPTCHAs-understanding CAPTCHA-solving services in an economic context[C]//USENIX Security Symposium, Washington, DC, USA, 2010: 435-462.

[21] RAMASUBRAMANIAN K, SINGH A. Machine learning using R[M]. Berkeley, CA: Apress, 2016: 2-3.

[22] FERRARA E, VAROL O, DAVIS C, et al. The rise of social bots[J]. Communications of the ACM, 2014, 59(7): 96-104.

[23] 张宇翔, 孙菀, 杨家海, 等. 新浪微博反垃圾中特征选择的重要性分析[J]. 通信学报, 2016, 37(8): 24-33.
ZHANG Y X, SUN W, YANG J H, et al. Analysis on the importance of feature selection in anti-spam in Sina Weibo[J]. Journal on Communications, 2016, 37(8): 24-33.

[24] FAZIL M, ABULAISH M. Identifying active, reactive, and inactive targets of socialbots in Twitter[C]// International Conference on Web Intelligence. ACM, 2017:573-580.

[25] 刘亚尚, 陈波, 朱汉, 等. 微博僵尸粉演化特征实证研究[J]. 情报探索, 2015(12): 1-9.
LIU Y S, CHEN B, ZHU H, et al. An empirical study on the evolutionary characteristics of zombie on microblog[J]. Information Research, 2015(12): 1-9.

[26] 刘凡平. 大数据时代的算法：机器学习、人工智能及其典型案例[M]. 北京: 电子工业出版社, 2017: 87.
LIU F P. Algorithms for big data age: machine learning learning, artificial intelligence, and typical cases[M].Beijing: Publishing House of Electronics Industry, 2017: 87.

[27] 张艳梅, 黄莹莹, 甘世杰, 等. 基于贝叶斯模型的微博网络水军识别算法研究[J]. 通信学报, 2017, 38(1): 44-53.
ZHANG Y M, HUANG Y Y, GAN S J, et al. Research on identification algorithm of internet water army on microblog based on Bayesian model[J]. Journal on Communications, 2017, 38(1): 44-53.

[28] 高岩. 朴素贝叶斯分类器的改进研究[D]. 广州: 华南理工大学, 2011.
GAO Y. Research on the improvement of naive Bayesian classifier[D]. Guangzhou: South China University of Technology, 2011.

[29] 唐姜贤. 拓展的朴素贝叶斯分类器的比较研究与优化集成[D]. 兰州: 兰州财经大学, 2015.
TANG J X. Comparative study and optimized integrati-on of extended naive bayesian classifier[D]. Lanzhou: Lanzhou University of Finance and Economics, 2015.

[30] 陆微微, 刘晶. 一种提高*K*-近邻算法效率的新算法[J]. 计算机工程与应用, 2008, 44(4): 163-165.
LU W W, LIU J. A new algorithm for improving the efficiency of *K*-nearest neighbor algorithm[J]. Computer Engineering and Applications, 2008, 44(4): 163-165.

[31] 谈磊, 连一峰, 陈恺. 基于复合分类模型的社交网络恶意用户识别方法[J]. 计算机应用与软件, 2012, 29(12):1-5.
TAN L, LIAN Y F, CHEN K. An identification method for malicious users in social network based on compound classification model[J]. Computer Applications and Software, 2012, 29(12):1-5.

[32] GRIER C, THOMAS K, PAXSON V, et al. @spam: the underground on 140 characters or less[C]//17th ACM Conference on Computer and Communications Security. Chicago, Illinois, USA, 2010:27-37.

[33] IRANI D, WEBB S, PU C. Study of static classification of Social spam profiles in mySpace[J]. Cancer Cytopathology, 2013, 121(10): 591-597.

[34] LAU R Y K, LIAO S Y, KWOK C W, et al. Text mining and probabilistic language modeling for online review spam detection[J]. ACM Transactions on Management Information Systems, 2012, 2(4):1-30.

[35] CHINCHORE A, XU G, JIANG F. Classifying sybil in MSNs using C4.5[C]// The 3rd International Conference on Behavioral, Economic, and Socio-Cultural Computing. Durham, NC, USA, 2016:145-150.

[36] 程晓涛. 微博网络水军识别技术研究[D]. 郑州: 中国人民解放军信息工程大学, 2015.
CHENG X T. Research on identification technology for Internet water army on microblog[D]. Zhengzhou: PLA Information Engineering University, 2015.

[37] 张玉清, 吕少卿, 范丹. 在线社交网络中异常账号检测方法研究[J]. 计算机学报, 2015, 38(10): 2011-2027.
ZHANG Y Q, LYU S Q, FAN D. Research on anomaly account detec-

tion method in online social network[J]. Chinese Journal of Computers, 2015, 38(10):2011-2027.

[38] 程晓涛, 刘彩霞, 刘树新. 基于关系图特征的微博水军发现方法[J]. 自动化学报, 2015, 41(9):1533-1541.
CHENG X T, LIU C X, LIU S X. Method for detecting water army on microblog based on the characteristics of the graph of relationship[J]. Acta Automatica Sinica, 2015, 41(9):1533-1541.

[39] 高梦超, 胡庆宝, 程耀东,等. 基于众包的社交网络数据采集模型设计与实现[J]. 计算机工程, 2015, 41(4):36-40.
GAO M C, HU Q B, CHENG Y D, et al. Design and implementation of data acquisition model in social network based on crowdsourcing[J]. Computer Engineeing, 2015, 41(4):36-40.

[40] 谭婷婷, 蔡淑琴, 胡慕海. 众包国外研究现状[J]. 武汉理工大学学报(信息与管理工程版), 2011, 33(2): 263-266.
TAN T T, CAI S Q, HU M H, Current situation of foreign studies on crowdsourcing[J]. Journal of Wuhan University of Technology (Information & Management Engineering), 2011, 33(2):263-266.

[41] WANG G, MOHANLAL M, WILSON C, et al. Social turing tests: Crowdsourcing sybil detection[J]. arXiv preprint arXiv:1205.3856, 2012.

[42] 陈霞, 闵华清, 宋恒杰. 众包平台作弊用户自动识别[J]. 计算机工程, 2016, 42(8):139-145.
CHEN X, MIN H Q, SONG H J. Automatically identify users who cheat on crowdsourcing platform[J]. Computer Engineering, 2016, 42(8): 139-145.

[43] ELOVICI Y, FIRE M, HERZBERG A, et al. Ethical considerations when employing fake identities in online social networks for research[J]. Science and Engineering Ethics, 2014, 20(4):1027-1043.

[44] DU X, CAI Y, WANG S, et al. Overview of deep learning[C]// Chinese Association of Automation (YA- C), Youth Academic Annual Conference. Wuhan, China, 2017: 159-164.

作者简介：

刘蓉（1994-），女，江苏泰州人，南京师范大学硕士生，主要研究方向为信息安全、社交网络等。

陈波（1972-），男，江苏南通人，南京师范大学教授、硕士生导师，主要研究方向为信息安全、社会计算等。

于泠（1971-），女，江苏金坛人，南京师范大学副教授，主要研究方向为信息安全、社会计算。

刘亚尚（1990-），女，河南郑州人，南京师范大学硕士生，主要研究方向为信息安全、社会计算。

陈思远（1993-），男，江苏淮安人，南京师范大学硕士生，主要研究方向为信息安全、Android 移动安全等。