

社交网络异常用户检测技术研究进展

曲强, 于洪涛, 黄瑞阳

(国家数字交换系统工程技术研究中心, 河南 郑州 450002)

摘要: 在社交网络中, 异常用户检测问题是网络安全研究的关键问题之一, 异常用户通过创建多个马甲进行虚假评论, 网络欺凌或网络攻击等行为严重威胁正常用户的信息安全和社交网络的信用体系, 因此大量研究人员对该问题进行了深入研究。回顾了近年来该问题的研究成果, 并总结出一个整体架构。数据收集层介绍数据获取方式与相关数据集; 特征表示层阐述属性特征、内容特征、网络特征、活动特征与辅助特征; 算法选择层介绍监督算法、无监督算法与图算法; 结果评估层阐述数据标注方式与方法评估指标。最后, 展望了该领域未来的研究方向。

关键词: 社交网络; 异常用户检测; 网络安全

中图分类号: T309

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2018025

Research progress of abnormal user detection technology in social network

QU Qiang, YU Hongtao, HUANG Ruiyang

National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China

Abstract: In social networks, the problem of anomalous users detection is one of the key problems in network security research. The anomalous users conduct false comments, cyberbullying or cyberattacks by creating multiple vests, which seriously threaten the information security of normal users and the credit system of social networks, so a large number of researchers conducted in-depth study of the issue. The research results of the issue in recent years were reviewed and an overall structure was summarized. The data collection layer introduces the data acquisition methods and related data sets, and the feature presentation layer expounds attribute features, content features, network features, activity features and auxiliary features. The algorithm selection layer introduces supervised algorithms, unsupervised algorithms and graph algorithms. The result evaluation layer elaborates the method of data annotation method and index. Finally, the future research direction in this field was looked forward.

Key words: social network, anomalous users detection, cyber security

1 引言

随着移动互联网技术发展的广泛应用, 在线社交网络因其具有便捷、灵活和内涵丰富等

特性而快速成为人们网络生活的重要组成部分。然而, 社交网络蕴含的巨大用户隐私信息及其广阔的商业价值, 也往往成为不法分子图谋不轨的目标, 其中, 异常用户是不法分子攻

收稿日期: 2018-01-05; 修回日期: 2018-02-27

通信作者: 曲强, 1498018641@qq.com

基金项目: 国家自然科学基金创新群体资助项目 (No.61521003)

Foundation Item: The Innovation Research Group of the National Natural Science Foundation of China (No.61521003)

击社交网络常见的手段之一。例如，商家为了商业利益，歪曲商品价值导向^[1~3]；罪犯利用多个马甲，欺骗网络用户，盗取信息甚至网上欺凌^[4~6]。据统计，异常用户的类型由于社交网络类型不同也呈现出多种形态，如表 1 所示，它们在不同的社交平台上广泛存在。例如，截至 2012 年 6 月，Facebook 存在大约 8.7% (8.3×10^7 个) 的虚假用户，而 Twitter 面临同样的问题，大约 5% 的用户是虚假用户，而有的专家则认为这一比例有可能高达 10%^[7]。

表 1 异常用户种类

异常用户种类	主要存在平台
Spammer (垃圾信息制造者)	Twitter, 微博平台, CNN, 亚马逊, Yelp
Hoaxes (文化攻击者)	维基百科
Bot (机器人)	Twitter
Sockpuppet (马甲)	维基百科, Twitter, Facebook, CNN
Troll (钓鱼者)	Slashdot zoo, 维基百科
Fake user (虚假用户)	Google, Twitter, Amazon
Vandal (恶意用户)	Twitter, 维基百科, TCP

社交网络中异常用户具有广泛存在性与严重危害性，相关学者对社交网络异常用户检测技术研究进行了总结。Kumar 等^[8]总结了社交网络中异常用户检测与异常行为检测的现状和相关技术；Jiang 等^[9]重点剖析了基于特征、基于空间与基于密度等恶意用户检测算法及其应用；Cui 等^[10]针对传统垃圾邮件、虚假评论、垃圾信息以及链接工厂 4 个方面，提取了网络特征、内容特征与行为特征，阐释了聚类算法、分类算法与图算法的应用。Beutel 等^[11]针对虚假信息的检测问题，研究基于图的算法，将其分成子图分析与挖掘算法，标签传递算法与隐因子分解算法。上述研究工作仅从特征或算法角度对现有工作进行总结，不够全面。

本文总结了社交网络异常用户检测技术的实现流程，如图 1 所示。数据收集层介绍数据获取方式与相关数据集；特征表示层阐述属性特征、内容特征、网络特征、活动特征与辅助特征；算法选择层介绍监督算法、无监督算法与图算法；结果评估层阐述数据标注方式与方法评估指标。

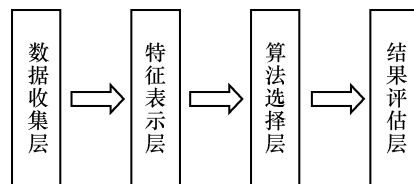


图 1 社交网络异常用户架构

2 数据收集层

2.1 数据获取方式

数据获取是开展社交网络异常用户检测技术研究的基本前提，根据任务的不同，数据获取方式也呈现多样性，数据获取方式如表 2 所示。不同的数据获取方式都存在一定的弊端。例如，合成数据集与真实数据集有一定的差别，实验结果不能令人信服；爬取数据集很难突破隐私保护机制；API 获取数据集是一个局部样本数据集，需要验证符合总体数据分布与否；合作数据集的公开获取范围相对有限；公开数据集的各类和数量目前仍十分有限，无法适应广泛的异常用户研究需求。

表 2 数据获取方式

数据获取方式	相关工作
合成数据集	文献[12~16]
爬取数据集	文献[17~20]
API 获取数据集	文献[21~25]
合作数据集	文献[26, 27]
公开数据集	文献[13, 28~36]

2.2 公开数据集

公开数据集是主要的数据集获取方式之一，也是社交网络异常用户检测相关算法评测对比的重要依据，因此本文介绍一些相关的公开数据集。Prakash 等^[13]公开包括电话网络、引文网络等网络的数据集；Akoglu 等^[33]公开包含邮件网络、政治网络、网关协议网络与引文网络等网络的数据集；卡耐基梅隆大学的 Shin 等^[37]、斯坦福大学的 Kumar 等^[38]、Rayana 等^[1]公开相关的数据集。目前，在公开数据集中，比较典型的几种数据集如下。

1) Twitter 数据集：Lee 等^[25]公开 TwitterH 数据集，该数据集含有 4 453 380 条 Twitter，223 115 条

回复 Twitter, 38 400 个用户, 19 200 个异常用户, 8 739 105 条用户关系; Ratkiewicz 等^[39]公开 2010 年 9 月到 10 月有关政治的 Twitter 数据集; Giatsoglou 等^[40]公开包含 298 个用户, 130 422 条 Twitter 的数据集, 用来探测虚假评论。

2) 评论数据集: Horne 等^[35]公开关于 2016 年总统竞选的虚假新闻与虚假评论数据集。

3) 邮件数据集: Xu 等^[29]在 TREC 会议公开 2005 年、2006 年、2007 年 TREC 数据集与 2008CEAS 数据集, 分别含有 92 189、37 822、64 620、75 419、137 705 个实例。

3 特征表示层

针对社交网络异常用户带来的威胁, 学术以及工业界对于异常用户的行为进行研究与分析, 提取出异常用户的特征, 目前这些特征可以总结为 5 类: 属性特征、内容特征、网络特征、活动特征、辅助特征。接下来, 3.1~3.6 节分别介绍这种 5 特征以及 5 种特征的对比。

3.1 属性特征

属性特征是指社交网络中注册用户的基本属性信息, 由于用户隐私保护, 部分属性信息难以获取, 而某些属性信息(如用户的注册名、性别以及信誉度评分等)则可以获取。属性特征主要分为注册名基本属性特征与信誉度特征。

1) 基本属性特征。一些研究人员利用注册名基本属性特征并结合其他特征检测异常用户。Zafarani 等^[7]利用注册名的复杂度、多样性、性别等特征检测 Twitter 网络中的恶意用户。Subrahmanian 等^[41]利用注册名、头像、关注数目、粉丝数目、位置等特征检测 Twitter 网络中的机器人水军。Kumar 等^[38]利用注册年龄等特征检测维基百科中的恶意用户。Yamak 等^[17]利用用户主页空间目录等特征检测维基百科中的马甲用户。研究人员^[18,25]利用用户全面的属性信息区探测社交网络中的虚假消息制造者。

2) 信誉度特征。这种特征只有在某些社交网络上可以获取。Adler 等^[21]利用信誉度等特征检测维基百科中的文化攻击者。Perez 等^[42]利用信誉度、粉丝数目、关注数目等特征检测 Twitter 网络中异常用户。Adler 等^[43]利用信誉度、国家荣誉度等

特征检测维基百科中的文化攻击者。

3.2 内容特征

内容特征是指从社交网络中用户发布的内容信息中提取的特征。内容特征又可以称为文本特征, 主要分为: 语义特征、语法特征、某些社交网络的特色特征。

1) 语义特征指发布文本内容表达的信息、话题或情感等。Subrahmanian 等^[41]利用 LDA 模型解析语义特征, 获取发布内容的文本检测 Twitter 网络中的机器人水军。Xu 等^[29]利用文本语义的 n-gram 模型特征检测邮件网络的垃圾邮件。Lim^[2]、Cheng^[44]、Kumar^[45]等利用文本相似度、自动可读性指标、词分布 LIWC 等特征检测 CNN、IGN、Breitbart 社区网络中的异常用户。Adler 等^[21]利用信息的发布与回复质量等特征检测维基百科中的文化破坏者。Dickerson^[46]、Mukherjee^[47]等利用语义特征以及个性特征等特征检测 Twitter 社交网络中的机器人。

2) 语法特征指蕴含在社交网络发布内容中的篇章或语句的结构、词性统计、题材格式以及标点使用情况等。Kumar 等^[38]利用文本长度、文本组成比例等特征检测维基百科中的异常用户。Solorio 等^[48]利用格式、语法、题材等特征探测维基百科的马甲用户。Adler^[43]、Gani^[49]等利用文本特征、语法格式特征探测 Twitter 社交网络中的马甲。

3) 特色特征指在某些社交网络中存在的特色符号特征, 如在微博中经常出现“#”“@”“http://...”“About me”等特殊符号。目前, 从事这方面的研究人员很少。例如, Gao 等^[50]利用文本与 URL 的特征检测 Facebook 网络中的垃圾信息制造者; Ratkiewicz^[31,39]、Perez^[42]、Benevenuto^[51]等利用标签、提及、URL、文本特征检测 Twitter 网络中的政治滥用者和异常用户。Lee 等^[25]利用“关于我”部分的文本特征检测 Twitter 社交网络中的垃圾信息制造者。

3.3 网络特征

网络特征指社交网络用户彼此之间的联系所构成的关系网络中提取的微观到宏观的网络结构特征。目前, 网络特征分为两类: 通联关系和结构指标。

1) 通联关系指将网络结构的邻接矩阵或者连接矩阵,通过分解矩阵进行降维,识别异常用户或者异常用户群组。Parakash^[13]、Jiang^[15]、Hu^[23,24,34]、Li^[27]、Wu^[28]、Shah^[52]、Kunegis^[53]等将通联关系转换成矩阵形式,进一步分解矩阵评分或者聚类,检测 Twitter 网络中的垃圾信息制造者与通信网络的异常用户。Beutel 等^[26]将通联关系转换成矩阵形式,最优化随机游走问题解,识别 Facebook 中的链路工厂问题。Shin 等^[37]将网络结构的邻接矩阵分解成多维列信息,进行密度块的挖掘,识别维基百科中的 TCP 恶意攻击者。

2) 结构指标指复杂网络的结构特征指标,如度、聚类系数、中心性、模度等特征,区分异常用户与正常用户。Akoglu 等^[33]利用节点数量、边数量、权重和、特征值、好友数量^[25]探测多种社交网络的异常用户。Hosseinmardi^[20]、Kumar^[38]等利用聚类系数^[45,54]的特征识别维基百科中的虚假信息制造者。Fire^[16]、Ratkiewicz^[39]、Subrahmanian^[41]、Gani^[49]等利用度、聚类系数、社区存在数、平均度特征以及偏离度特征检测谷歌等网络的虚假用户。Ratkiewicz^[31]、Kumar^[55]等设计 8 种中心度度量方式 Freaks、FMF、Prestige、M-PR、SSR、NR、SEC、M-HITS、BAD 根据假设进行聚类,识别 Slashdot Zoo 中的钓鱼者。

除了上面这些基本结构指标,研究人员设计一些独特的结构指标识别异常用户。Cao 等^[4]设计出一种信任程度指标,识别 Facebook 中的虚假用户。Jiang 等^[30]设计同步性度量和常态度量指标,进行聚类,识别 Twitter 网络中的虚假信息制造者。Ye 等^[12]设计自相似性与节点多样性指标,识别合成数据集中的垃圾信息制造者。Hooi 等^[14]设计一种怀疑分数与密度度量指标,识别 Twitter 网络中伪装的虚假信息制造者。Dickerson 等^[46]设计 NCR、NAR、NDR 等邻居排序指标识别 Twitter 网络中的机器人。

3.4 活动特征

活动特征指社交网络中用户的行为模式。目前使用的活动特征一般可以分为用户自身活动特征、邻居反馈活动特征、监管者监管活动特征。

1) 用户自身活动特征指社交网络中用户自己的行为模式,如发布信息的数量、频次、规律

等。Shachaf 等^[56]分析了维基百科中钓鱼者的活动,如钓鱼者在百科领域是活跃的,一般孤立行动,经常做出重复性损害性极大的行为等。Lee^[25]、Subrahmanian^[41]、Kumar^[45]、Dickerson^[46]、Benevenuto^[51]、Xu^[54]等利用发布消息频次、时间间隔等特征识别 Twitter 中垃圾信息制造者或通信网络中异常用户。Wu 等^[18]利用发布信息数量矩阵特征检测新浪微博中的异常用户。

2) 邻居反馈活动特征指对某个用户发布信息后,该用户的邻居节点对其发布信息这一行为的反馈行为模式,如回复信息数量、时间间隔等。Cheng 等^[5,57]分析 CNN 网络,指出如果用户周围存在钓鱼者,则该用户很可能成为钓鱼者;同时分析 Twitter 网络,指出消极回复增加产生恶意用户的可能性。Tsikerdakis 等^[32]利用回复消息数量在不同空间(主页、讨论区、文章)的分布构建 GR 指标,达到识别维基百科中马甲的目的。Mukherjee^[47]、Jindal^[58]等利用回复消息数量^[54](或占比^[20,50])、有益消息占比等特征识别亚马逊网络中的虚假评论制造者。Yamak 等^[17]利用回复消息频次以及回复消息后产生的删除增加比例探测维基百科中的马甲。

3) 监管者监管活动特征指对用户发布信息这一行为,社交网络监管者做出相应的判断,实行奖惩行为的模式。Cheng 等^[44]利用监管者删除、拦截发布信息的比例探测谷歌等网络的异常用户。

3.5 辅助特征

社交网络异常用户检测除了利用上述 4 类主要的特征之外,还可以利用一些辅助的特征。这类辅助特征指可以协助主要用户特征识别异常用户的特征,一般不能独立存在,必须依赖主要特征存在。本节主要介绍 2 种辅助特征,分别是时间特征、评价特征。

1) 时间特征指活动发生的时间信息,该特征主要从时间的角度观察内容特征、网络特征与活动特征的变化。Xu 等^[54]利用每周发生活动的时间平均值、方差等特征识别垃圾信息。Adler 等^[21,43]利用回复的时间差^[32]、活动发生的小时数^[40]等特征识别维基百科中的恶意用户。Yamak 等^[17]利用用户注册与第一次发布信息

的时间差探测维基百科中的马甲。Subrahmanian 等^[41]利用发布信息的最长持续时间以及时间信息熵等特征识别 Twitter 中的机器人。Gupta^[22]、Gao^[50]等利用发生活动的时间集中性进行聚类,识别 Facebook 的异常用户群组。Beutel 等^[26]将邻接矩阵转换成邻接时间矩阵,利用该矩阵进行聚类,识别 Facebook 中的异常用户群组。Wu 等^[18]构建事件发生持续时间矩阵^[59]等矩阵识别新浪微博中的流言制造者。

2) 评价特征指在某些网络中,用户可以评价他人或者物品,而形成评价分数的信息。Hooi 等^[60]利用评分矩阵和随时间变化评分差矩阵,进行相关的概率模拟,来识别 Flikpart 的虚假评论制造者。Rayana 等^[1]利用用户的评分平均值^[12]、评分方差、评分偏离度^[61]、评分分布熵以及评分时间差等特征 Yelp 的虚假评论制造者。

3.6 不同特征的对比

不同特征具有不同的应用场景,如表 3 所示,其中列出了不同特征的不同特点。

4 算法选择层

目前,研究人员主要采用机器学习算法解决社交网络异常用户检测问题,比较主流的方法分为 3 种:监督算法、无监督算法、图算法,如表

4 所示。其中,图算法属于无监督算法,但由于它使用的数据是图数据这种特殊的形式并且图算法是一种应用广泛的异常用户检测算法,因此无监督算法与图算法分别陈述。4.1~4.4 节分别介绍监督算法、无监督算法、图算法以及不同算法的对比。

4.1 监督算法

当获取的数据含有标签时,研究人员基于分类的思想,设计出数值形式特征,将用户分成异常用户与正常用户,达到检测异常用户目的。监督算法可以分为单一分类算法与集成分类算法。

1) 单一分类算法指仅利用一种分类算法检测异常用户,常用的有逻辑回归、支持向量机、决策树等。Cheng^[5]、Zafarani 等^[7]采用逻辑回归检测 Twitter 中的恶意用户与 CNN 中钓鱼者,发现名字语言模式特征是区别恶意用户与正常用户最显著的特征;Jindal 等^[58]采用逻辑回归检测 Amazon 中的虚假评论,根据产品与评论的分类情况,强调检测 4 个危害性较大的虚假评论区域,发现回复活动特征对该问题最为有效,而 Lim^[2]发现总体评分偏离度对检测虚假评论没有任何作用。Solorio 等^[48]利用支持向量机检测维基百科的马甲,通过实验发现了对该问题贡献度最大的 30 种特征;Benevenuto 等^[51]利用支持向量机检测

表 3 检测特征对比

检测特征	特点	关键	特征评估
属性特征	采用人为设计方法,容易被攻击者绕过,算法设计简单,效率低,准确率相对较低,数据量级小,具有严格隐私保护	突破隐私保护	不常用
内容特征	采用自然语言处理方式,容易被攻击者绕过,算法设计复杂,效率低,准确率相对较低,数据量级大,具有轻微隐私保护	设计复杂算法,合理语言模式	常用
网络特征	采用复杂网络处理方式,不易被攻击者绕过,算法设计简单,效率低,准确率相对较低,数据量级大,不具有隐私保护	掌握全局结构	主流
活动特征	采用行为模式分析处理方式,不易被攻击者绕过,算法设计简单,效率高,准确率高,数据量级大,具有轻微隐私保护	选取区分度最大的活动信息	主流
辅助特征	采用时间序列模型分析,不易被攻击者绕过,算法设计复杂,效率高,准确率高,数据量级小,具有轻微隐私保护	有效利用时间维度信息	热门

表 4 检测算法分类

算法	思想	算法关键问题
监督算法	单一分类算法 集成分类算法	分类 选取具有最大区分度的关键特征
无监督算法	从上到下的分解挖掘 从下到上的聚类挖掘	聚类 选取合理的相似性指标
图算法	谱分解 随机游走	图异常点检测 处理大规模稀疏的图数据

Twitter 中的垃圾信息制造者, 通过向支持向量机中引入参数 J 给出垃圾信息制造者占比的先验分布, 调节 J 来平衡准确率或者召回率的提升, 并且通过 gain 与卡方检验选出了几种比较具有区分性的特征。Adler 等^[21]采用决策树检测维基百科中的文化攻击者, 并且牺牲了一部分准确度要求达到更高的召回率, 获得 2010 年 PAN 比赛的冠军。Gupta 等^[22]采用决策树与贝叶斯网络算法检测在 Twitter 有关 Sandy 飓风的虚假消息中, 发现决策树的效果较好且文本特征的贡献度比较大。

2) 集成分类算法指将多种单一分类算法集成, 以获取更高的准确率的集成分类算法, 如随机森林、Adaboost 等。Cheng 等^[44]利用随机森林算法检测 CNN 等讨论社区中的异常用户, 发现采样样本时间越长, 方法预测准确度越差, 侧面证实了用户行为的改变容易导致被判定为异常用户。Yamak 等^[17]采用了包含随机森林与贝叶斯网络在内的 6 种分类算法, 检测维基百科中的马甲, 通过实验发现最好的特征依次为回复频次、增加字节与平均贡献度。Ratkiewicz 等^[31]利用 Adaboost 方法检测 Twitter 中的政治水军, 并且利用卡方检验给出贡献度最大的 10 种特征, 分析了发现的政治水军的特点。Tsikerdakis 等^[32]采用支持向量机、随机森林与 Adaboost 方法, 检测社交网络中的欺骗, 发现 Adaboost 方法效果更好而且这种新定义指标的效果更好。Xu 等^[29]将包括贝叶斯、NSNB、Winnow 等算法集成为线性联合算法, 通过优化权重得到每个算法对该问题的有效性。

4.2 无监督算法

当样本数据不含标签或含有很少标签时, 研究者基于聚类的思想, 提出利用无监督学习算法解决异常用户检测问题。无监督算法分为从上到下的分解挖掘和从下至上的聚类挖掘。

1) 从上到下的分解挖掘算法。在没有样本数据的标签时, 研究人员通过分解社交网络图检测异常用户群组。Bu 等^[62]基于话题相似性构建 SVN 网络, 基于文本特征相似性删除一部分边形成 SPN 网络, 基于模度的相似性进行聚类挖掘异常用户的群组社区, 并且给出了方法准确率的形式。TIA 算法^[55]根据不同的中心度值界限初始化正常用户和恶意用户, 然后根据不同的攻击模式采取

不同的分解图操作, 不断更新恶意用户与正常用户群组, 达到预测 Slashdot 网络中恶意用户目的。D-CUBE 方法^[37]通过删除基数或者密度最大的属性值维度来分解关系张量, 直到最后剩下一个异常群组, 不断迭代得到多个异常用户群组, 该方法采用分布式算法形式, 适用于大规模图数据形式。ND-SYNC 方法^[40]直接基于 RTFRAUD 方法对构建的特征空间进行社区发现, 利用内在与外在同步性的偏离度进行群组异常性检测, 发现 Twitter 中虚假用户。

2) 从下至上的聚类挖掘算法。在拥有一部分数据样本标签时, 研究人员利用相似性和已知标签的样本进行图结构的聚类, 解决异常用户群组检测问题。CopyCatch 方法^[26]主要构建了社交网络的时间矩阵, 以最大化 TNBC 核心中的异常用户数量为目标进行聚类, 检测 Facebook 中的异常用户攻击组, 并且给出稳定性与收敛性证明。Gao 等^[50]根据文本的 MD5 相似性以及 URL 指向目标的相同性聚类构建相似的群组, 再通过虚假用户的分布覆盖以及时间爆发判定每个群组是否为虚假用户群组。Prakash 等^[13]发现了一种称为 EigenSpoke 的新模式, 并且利用该模式的分数将样本聚类, 直到模度不再增加, 以此检测社交网络用户群组。Subrahmanian 等^[41]初始化少部分 Twitter 中的机器人, 利用文本信息相似性进行聚类, 直到找到满足数量的机器人终止。CatchSync 方法^[30]重新定义了同步与常态, 在此基础上绘制了 SN 图, 基于距离进行聚类, 识别网络中的欺骗者。

4.3 图算法

随着以图数据形式存在的网络拓扑特征和活动特征成为检测异常用户的主要依据特征, 图算法成为检测异常用户的流行方法。在基于图的算法中, 比较主流的方法有 2 种: 基于谱分解的算法、基于随机游走的算法。

1) 基于谱分解的算法指对通过处理得到的特征矩阵进行谱变换或分解, 得到不同的群组。在利用基于谱的分解算法解决异常用户检测问题方面, 学者做了大量工作。Wu 等^[28]构建了分层树结构, 将内容矩阵与分层树网络结构的稀疏表示融合, 利用谱分解模拟方式迭代出最佳权重,

进一步识别欺骗者。Hu 等^[23,34]将内容矩阵和构造的随机游走网络矩阵结合，利用谱分解方式求解最佳权重，识别 Twitter 中垃圾信息制造者。后来 Hu 等^[24]又将情感信息矩阵融入内容矩阵和邻接矩阵中，优化目标函数求得最佳权重。

当然，其他的研究人员也做出了一些贡献。Li 等^[27]根据一定阈值的度与选取的种子创建子图，基于子图利用谱分解形成更小的子空间，检测子空间内用户是否为 YouTube 的虚假用户。面对影响维度的增加，尤其是时间因素，FEMA 算法^[63]根据一定的正则化限制条件，对不同时间的三维张量进行分解得到映射矩阵和核心张量，进一步得到属于异常用户概率。Shah 等^[52]利用 SVD 矩阵分解方法重构网络节点的度，通过设定度的界限检测隐藏性比较好的攻击。

2) 基于随机游走的算法指通过计算节点的转移路径，给出未知节点与已知节点的关系或者转移概率，检测未知节点是否为异常节点。目前，该算法主要应用于社交网络马甲账户的检测。SybilGuard^[64]方法利用修改的随机游走算法计算节点的转移路径，通过节点路径与正常节点的路径是否有交集判定是否为正常用户，但是每条攻击路径依旧最多容忍 $O(\sqrt{n} \log n)$ 个马甲节点。为降低误报率与容忍马甲节点数，SybilLimit^[65]通过判定节点路径最后一条边与正常节点路径最后一条边是否有交集判定是否为正常用户，将每条攻击路径的容忍马甲数降低为 $O(\log n)$ ，但是上述 2 种方法每次迭代只能判定一个节点效率较低。SybilDefender^[66]从正常节点出发进行 f 次随机游走，得到 $f+1$ 个判定

节点，通过 R 次随机游走得到节点访问次数的平均值以及偏离值标准，之后通过对一个节点进行同样的操作、利用同样标准判定是否为马甲节点，然后利用已经发现的马甲节点和同样的原理快速进行多个马甲账号的检测。SybilRank 方法^[4]利用随机游走算法将已知的多个正常节点的可信值分配给其他节点，根据标准化后的可信值与度值结果进行排序，将可信值较小的节点判定为可疑节点。SybilBelief 方法^[67]利用马尔可夫随机场检测马甲账号，首先给网络中未知的节点都分配一个随机值标识是正常节点还是马甲节点，然后通过马尔可夫随机场计算每个节点的后验概率，即节点是正常节点的概率。

4.4 不同算法的对比

各类检测算法各有优劣，有各自的应用场景，如表 5 所示，其中列出了不同检测算法的优点与缺点。

5 方法评估层

选取合适的特征表示与算法选择之后，研究人员需要对方法的效果进行一定的评估，需要获取数据标注以及方法评估指标。5.1~5.2 节介绍数据标注方式以及方法评估指标。

5.1 数据标注方式

如何标注数据标签是评估异常用户检测方法的前提，目前数据标注方式如表 6 所示。不同的数据标注方式都有一定的特点。例如，依据其他特征虽然简单易行，但是标注结果不能令人信服；黑名单虽然准确率很高，但是不易从网站管

表 5

检测算法对比

检测算法	优点	缺点
监督算法	1) 准确率高 2) 检测速度快，效率高 3) 算法设计成熟，部署技术成熟 4) 实时性好	1) 需要含标签数据 2) 需要提前训练 3) 需要选取区分度特征 4) 选取特征易被攻击者绕过，未知模式检测效果差
无监督算法	1) 仅仅需要不含标签数据 2) 不需要提前训练 3) 有效检测未知模式	1) 准确率低 2) 算法设计复杂，效率低 3) 实时性差
图算法	1) 只需要图数据 2) 不需要提前训练 3) 有效检测未知模式	1) 准确率低，实时性差 2) 理论假设条件复杂，现实不成立 3) 算法设计复杂，效率低 4) 社交网络存在不同的差异

理者获取；人工标注准确率比较高，但是花费时间与经济成本比较大；软件评测准确度高而且成本少，但是只适用于固定问题检测，具有局限性；社交网络自身功能也面临准确性不高的问题。因此需要根据研究的环境，选择合适的标注方式。

表6 数据标注方式

数据标注方式	相关工作
依据其他特征	文献[58,61,62]
黑白名单	文献[31,50]
人工标注	文献[5,16,43,44,46,59]
软件评测	文献[2]
社交网络自身功能	文献[44]

5.2 方法评估指标

一般比较常用的方法评估指标有准确率、召回率、精确率、 F 指数、马修斯相关系数与 ROC (AUC) 曲线指标。其中，ROC 曲线是将基于混淆矩阵的真正率和假正率值作为坐标轴画出的曲线，AUC 表示 ROC 曲线下的面积。这些指标都基于混淆矩阵，各种指标的定义形式如表 7 所示。

表7 方法评估指标

方法评估指标	形式定义
精确率	$A = \frac{TP + TN}{TP + TN + FP + FN}$
准确率	$P = \frac{TP}{TP + FP}$
召回率	$R = \frac{TP}{TP + FN}$
F 指数	$F = \frac{2PR}{P + R}$
MCC	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FN)}}$

另外，还有许多使用比较少的方法评估指标。比如，Jiang 等^[63]采用 MAE、RMSE、hit ratio 等指标评测 FEMA 算法对于微博的异常用户检测效果；Rayana^[1]、Lim 等^[2]针对异常用户怀疑分数排序问题，采用 NDCG 指标整体衡量检测出的虚假信息制造者排序情况。

除了需要对方法进行评估之外，有时需要对特征的重要性进行排序，除了仍用上述方法评价

指标进行比较，还可以使用检验方法对特征重要性进行排序，如置信度^[48]、卡方检验^[31]、升力曲线^[58]等。

6 结束语

随着社交网络影响力逐渐增强，越来越多的恶意用户将攻击目标集中于社交网络，其中，异常用户对于社交网络的危害严重威胁了社交网络的信息安全甚至用户的财产生命安全。为此，针对社交网络异常用户检测问题，本文提出了一个整体的架构，分别从数据收集层、特征表示层、算法选择层和结果评估层对该问题的处理架构进行了阐述，并且对不同数据来源、不同特征、不同算法以及不同数据标注方式、评估标准进行了详细的总结与对比。但是，社交网络异常用户检测是一个不断发展的过程，针对现有的问题与挑战，研究人员提出了一些解决方案。但是，这些方案依旧存在不足，以下几点将是未来的研究方向。

1) 设计多特征多维度的检测方法。首先，本文系统总结了异常用户检测的许多特征维度，但是这些方法往往利用几种部分特征，如何融合多种维度特征是一个比较关键的问题。其次，时间维度特征成为研究热点，如何利用时间特征动态检测社交网络异常用户备受关注。

2) 有效对抗经过伪装的攻击方法。针对社交网络异常用户检测问题，各种各样的方法层出不穷。但是异常用户为避免被检测出，开始选择隐蔽攻击形式，模仿正常用户。例如，异常用户模仿正常用户的写作方式避免被用内容特征的方法检测出，异常用户创建大量指向正常用户的链接避免被用网络结构特征的方法检测出。由于异常用户变得越来越狡猾，伪装性越来越强，因此需要设计出一些抵抗针对特定伪装^[11]问题的方法。

3) 检测未知的攻击模式。通过分析异常用户的攻击方式，笔者发现社交网络总会面对未知的攻击。所有研究工作都是基于已知攻击模式的特征，对于未知攻击模式难以提取有效特征。最近，研究人员提出将网络结构特征映射到傅里叶空间中形成新的特征表示，能否利用新空间的特征表示有效区分异常用户与正常用户值得深思。

4) 检测异构网络下的异常账户。绝大多数研究都是基于二维网络进行异常用户检测研究, 而且绝大多数网络只含有一种关系。因此, 异构网络^[1]中异常用户检测对社交网络异常用户检测是一个值得思考的问题。

5) 设计高效轻量级的算法。目前尽管这个领域的研究方法层出不穷, 但是这些方法仅用于小规模数据甚至需要大量的假设条件, 与实际中社交网络真实情况不相适应。未来, 如何针对大规模数据, 设计出高效实时监测算法是一个值得关注的重点。

参考文献:

- [1] RAYANA S, AKOGLU L. Collective opinion spam detection: Bridging review networks and metadata[C]//The 21st ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. 2015: 985-994.
- [2] LIM E P, NGUYEN V A, JINDAL N, et al. Detecting product review spammers using rating behaviors[C]//The 19th ACM International Conference on Information and Knowledge Management. 2010: 939-948.
- [3] MALBON J. Taking fake online consumer reviews seriously[J]. Journal of Consumer Policy, 2013, 36(2): 139-157.
- [4] CAO Q, SIRIVIANOS M, YANG X, et al. Aiding the detection of fake accounts in large scale social online services[C]//The 9th Usenix Conference on Networked Systems Design and Implementation. 2012: 15-15.
- [5] CHENG J, BERNSTEIN M, DANESCU-NICULESCU-MIZIL C, et al. Anyone can become a troll: causes of trolling behavior in online discussions[C]//ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017). 2017.
- [6] HINDUJA S, PATCHIN J W. Bullying, cyberbullying, and suicide[J]. Archives of Suicide Research, 2010, 14(3): 206-221.
- [7] ZAFARANI R, LIU H. 10 bits of surprise: Detecting malicious users with minimum information[C]//The 24th ACM International on Conference on Information and Knowledge Management. 2015: 423-431.
- [8] KUMAR S, CHENG J, LESKOVEC J. Antisocial behavior on the Web: characterization and detection[C]//The 26th International Conference on World Wide Web Companion. 2017: 947-950.
- [9] JIANG M, KUMAR S, SUBRAHMANYAN V S, et al. Data-driven approaches towards malicious behavior modeling[C]//The 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). 2017, 19: 42.
- [10] JIANG M, CUI P, FALOUTSOS C. Suspicious behavior detection: current trends and future directions[J]. IEEE Intelligent Systems, 2016, 31(1): 31-39.
- [11] BEUTEL A, AKOGLU L, FALOUTSOS C. Graph-based user behavior modeling: from prediction to fraud detection[C]//The 21st ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. 2015: 2309-2310.
- [12] YE J, AKOGLU L. Discovering opinion spammer groups by network footprints[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2015: 267-282.
- [13] PRAKASH B A, SRIDHARAN A, SESHADRI M, et al. Eigenspokes: surprising patterns and scalable community chipping in large graphs[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2010: 435-448.
- [14] HOOI B, SONG H A, BEUTEL A, et al. Fraudar: bounding graph fraud in the face of camouflage[C]//The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 895-904.
- [15] JIANG M, CUI P, BEUTEL A, et al. Inferring strange behavior from connectivity pattern in social networks[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2014: 126-138.
- [16] FIRE M, KATZ G, ELOVICI Y. Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies[J]. Human Journal, 2012, 1(1): 26-39.
- [17] YAMAK Z, SAUNIER J, VERCOUTER L. Detection of multiple identity manipulation in collaborative projects[C]//The 25th International Conference Companion on World Wide Web. 2016: 955-960.
- [18] WU S, LIU Q, LIU Y, et al. Information credibility evaluation on social media[C]//The 30th AAAI Conference on Artificial Intelligence. 2016: 4403-4404.
- [19] FRIGGERI A, ADAMIC L A, ECKLES D, et al. Rumor cascades[J]. Dalton Transactions, 2014, 43(16):6108-19.
- [20] HOSSEINMARDI H, GHASEMIANLANGROODI A, HAN R, et al. Towards understanding cyberbullying behavior in a semi-anonymous social network[C]//2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2014: 244-252.
- [21] ADLER B, DE ALFARO L, PYE I. Detecting wikipedia vandalism using wikitrust[J]. Notebook Papers of CLEF, 2010, 1: 22-23.
- [22] GUPTA A, LAMBA H, KUMARAGURU P, et al. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy[C]//The 22nd International Conference on World Wide Web. 2013: 729-736.
- [23] HU X, TANG J, ZHANG Y, et al. Social spammer detection in microblogging[C]//The International Joint Conference on Artificial Intelligence. 2013: 2633-2639.
- [24] HU X, TANG J, GAO H, et al. Social spammer detection with sentiment information[C]//2014 IEEE International Conference on Data Mining (ICDM). 2014: 180-189.
- [25] LEE K, CAVERLEE J, WEBB S. Uncovering social spammers: social honeypots+ machine learning[C]//The 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010: 435-442.
- [26] BEUTEL A, XU W, GURUSWAMI V, et al. Copycatch: stopping

- group attacks by spotting lockstep behavior in social networks[C]//The 22nd International Conference on World Wide Web. 2013: 119-130.
- [27] LI Y, MARTINEZ O, CHEN X, et al. In a world that counts: clustering and detecting fake social engagement at scale[C]//The 25th International Conference on World Wide Web. 2016: 111-120.
- [28] WU L, HU X, MORSTATTER F, et al. Adaptive spammer detection with sparse group modeling[C]//The International AAAI Conference on Web and Social Media. 2017: 319-326.
- [29] XU C, SU B, CHENG Y, et al. An adaptive fusion algorithm for spam detection[J]. IEEE Intelligent Systems, 2014, 29(4): 2-8.
- [30] JIANG M, CUI P, BEUTEL A, et al. Catchsync: catching synchronized behavior in large directed graphs[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 941-950.
- [31] RATKIEWICZ J, CONOVER M, MEISS M R, et al. Detecting and tracking political abuse in social media[C]//The International Conference on Weblogs and Social Media(ICWSM). 2011: 297-304.
- [32] TSIKERDEKIS M, ZEADALLY S. Multiple account identity deception detection in social media using nonverbal behavior[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(8): 1311-1321.
- [33] AKOGLU L, MCGLOHON M, FALOUTSOS C. Oddball: spotting anomalies in weighted graphs[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2010: 410-421.
- [34] HU X, TANG J, LIU H. Online social spammer detection[C]//The 28th AAAI Conference on Artificial Intelligence. 2014:59-65.
- [35] HORNE B D, ADALI S. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news[C]//The 2nd International Workshop on News and Public Opinion. 2017.
- [36] KUMAR S, SPEZZANO F, SUBRAHMANIAN V S. Vews: a wikipedia vandal early warning system[C]//The 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 607-616.
- [37] SHIN K, HOOI B, KIM J, et al. D-cube: dense-block detection in terabyte-scale tensors[C]//The 10th ACM International Conference on Web Search and Data Mining. 2017: 681-689.
- [38] KUMAR S, WEST R, LESKOVEC J. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes[C]//The 25th International Conference on World Wide Web. 2016: 591-602.
- [39] RATKIEWICZ J, CONOVER M, MEISS M, et al. Truthy: mapping the spread of astroturf in microblog streams[C]//The 20th International Conference Companion on World Wide Web. 2011: 249-252.
- [40] GIATSOGLOU M, CHATZAKOU D, SHAH N, et al. Nd-sync: detecting synchronized fraud activities[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2015: 201-214.
- [41] SUBRAHMANIAN V S, AZARIA A, DURST S, et al. The DARPA Twitter bot challenge[J]. Computer, 2016, 49(6): 38-46.
- [42] PEREZ C, LEMERCIER M, BIRREGAH B, et al. Spot 1.0: scoring suspicious profiles on twitter[C]//The International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2011: 377-381.
- [43] ADLER B T, DE ALFARO L, MOLA-VELASCO S M, et al. Wikipedia vandalism detection: combining natural language, metadata, and reputation features[C]//The International Conference on Intelligent Text Processing and Computational Linguistics. 2011: 277-288.
- [44] CHENG J, DANESCU-NICULESCU-MIZIL C, LESKOVEC J. Antisocial behavior in online discussion communities [J]. Computer Science, 2015.
- [45] KUMAR S, CHENG J, LESKOVEC J, et al. An army of me: Sockpuppets in online discussion communities[C]//The 26th International Conference on World Wide Web. 2017: 857-866.
- [46] DICKERSON J P, KAGAN V, SUBRAHMANIAN V S. Using sentiment to detect bots on twitter: are humans more opinionated than bots?[C]//2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2014: 620-627.
- [47] MUKHERJEE A, VENKATARAMAN V, LIU B, et al. What yelp fake review filter might be doing?[C]//The International Conference on Web and Social Media(ICWSM). 2013.
- [48] SOLORIO T, HASAN R, MIZAN M. A case study of sockpuppet detection in wikipedia[C]//The Workshop on Language Analysis in Social Media. 2013: 59-68.
- [49] GANI K, HACID H, SKRABA R. Towards multiple identity detection in social networks[C]//The 21st International Conference on World Wide Web. 2012: 503-504.
- [50] GAO H, HU J, WILSON C, et al. Detecting and characterizing social spam campaigns[C]//The 10th ACM Sigcomm Conference on Internet Measurement. 2010: 35-47.
- [51] BENEVENUTO F, MAGNO G, RODRIGUES T, et al. Detecting spammers on twitter[C]//Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS). 2010: 12.
- [52] SHAH N, BEUTEL A, GALLAGHER B, et al. Spotting suspicious link behavior with fbbox: an adversarial perspective[C]//2014 IEEE International Conference on Data Mining (ICDM). 2014: 959-964.
- [53] KUNEGIS J, LOMMATZSCH A, BAUCKHAGE C. The slashdot zoo: mining a social network with negative edges[C]//The 18th International Conference on World Wide Web. 2009: 741-750.
- [54] XU Q, XIANG E W, YANG Q, et al. SMS spam detection using noncontent features[J]. IEEE Intelligent Systems, 2012, 27(6): 44-51.
- [55] KUMAR S, SPEZZANO F, SUBRAHMANIAN V S. Accurately detecting trolls in slashdot zoo via decluttering[C]//2014 IEEE ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2014: 188-195.
- [56] SHACHAF P, HARA N. Beyond vandalism: Wikipedia trolls[J]. Journal of Information Science, 2010, 36(3): 357-370.
- [57] CHENG J, DANESCU-NICULESCU-MIZIL C, LESKOVEC J. How community feedback shapes user behavior[C]//The International Conference on Weblogs and Social Media (ICWSM). 2014.

- [58] JINDAL N, LIU B. Opinion spam and analysis[C]//The 2008 International Conference on Web Search and Data Mining. 2008: 219-230.
- [59] MUKHERJEE A, LIU B, GLANCE N. Spotting fake reviewer groups in consumer reviews[C]//The 21st International Conference on World Wide Web. 2012: 191-200.
- [60] HOOI B, SHAH N, BEUTEL A, et al. Birdnest: bayesian inference for ratings-fraud detection[C]//2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. 2016: 495-503.
- [61] FEI G, MUKHERJEE A, LIU B, et al. Exploiting burstiness in reviews for review spammer detection[C]//The 7th International AAAI Conference on Weblogs and Social Media (ICWSM). 2013: 175-184.
- [62] BU Z, XIA Z, WANG J. A sock puppet detection algorithm on virtual spaces[J]. Knowledge-Based Systems, 2013, 37: 366-377.
- [63] JIANG M, CUI P, WANG F, et al. Fema: flexible evolutionary multi-faceted analysis for dynamic behavioral pattern discovery[C]//The 20th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. 2014: 1186-1195.
- [64] YU H, KAMINSKY M, GIBBONS P B, et al. Sybilguard: defending against sybil attacks via social networks[C]//ACM Sigcomm Computer Communication Review. 2006: 267-278.
- [65] YU H, GIBBONS P B, KAMINSKY M, et al. Sybillimit: a near-optimal social network defense against sybil attacks[C]//IEEE Symposium on Security and Privacy. 2008: 3-17.
- [66] WEI W, XU F, TAN C C, et al. Sybildefender: Defend against sybil attacks in large social networks[C]//IEEE INFOCOM. 2012: 1951-1959.
- [67] GONG N Z, FRANK M, MITTAL P. Sybilbelief: a semi-supervised

learning approach for structure-based sybil detection[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(6): 976-987.

[作者简介]



曲强(1994-), 男, 黑龙江齐齐哈尔人, 国家数字交换系统工程技术研究中心硕士生, 主要研究方向为网络空间安全、大数据分析处理、复杂网络异常用户检测。



于洪涛(1970-), 男, 辽宁丹东人, 博士, 国家数字交换系统工程技术研究中心研究员, 主要研究方向为网络大数据分析处理。



黄瑞阳(1986-), 男, 福建漳州人, 博士, 国家数字交换系统工程技术研究中心助理研究员, 主要研究方向为文本挖掘、图挖掘。