



基于传播时空特性的社交网络灰帽用户检测

何欢¹, 朱焱¹, 李春平²

(1. 西南交通大学 信息科学与技术学院, 成都 611756; 2. 清华大学 软件学院, 北京 100091)

摘要: 社交网络灰帽用户极易隐藏且类型多样, 导致现有检测算法适用性较差。提出一种基于传播时空特性的社交网络检测算法。构建用户生成内容传播网络度量白帽和灰帽用户在传播空间上的不同特性, 融合时空传播特性并调节权重比例以提高分类性能。实验结果表明, 该算法能有效检测不同类型灰帽用户, 与用户特征分析、社交网络链接分析、多视图融合等主流灰帽用户检测算法相比, 其在 CAVERLEE、CRESCI-15、CRESCI-17 等多个数据集上的准确率及 AUC 值最高分别提升 26.08% 和 30.54%。

关键词: 社交网络; 灰帽用户; 网络传播; 特征融合; 用户检测

开放科学(资源服务)标志码(OSID):



中文引用格式: 何欢, 朱焱, 李春平. 基于传播时空特性的社交网络灰帽用户检测[J]. 计算机工程, 2021, 47(12): 192-199.

英文引用格式: HE H, ZHU Y, LI C P. Grey hat user detection in social network based on spatiotemporal characteristics of diffusion[J]. Computer Engineering, 2021, 47(12): 192-199.

Grey Hat User Detection in Social Network Based on Spatiotemporal Characteristics of Diffusion

HE Huan¹, ZHU Yan¹, LI Chunping²

(1. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China;

2. School of Software, Tsinghua University, Beijing 100091, China)

[Abstract] Grey hat users in social networks are diverse and good at disguising, which reduces the generability of detection algorithms. To address the problem, this paper proposes a social network detection mechanism based on the spatiotemporal characteristics of diffusion, Diffusion Spatio-Temporal Characteristics (DSTC). By using the characteristics of the diffusion time sequence of the content generated by white hat/grey hat users, a User Generated Content (UGC) diffusion network is constructed to measure the different characteristics of the white/grey hat users in the diffusion space. Based on the spatio-temporal characteristics of diffusion, a mechanism is designed for social network user detection, which integrates the spatio-temporal characteristics of diffusion and adjusts the weight ratio to improve classification performance. Experimental results show that compared with the mainstream grey hat user detection algorithms, which are based on user characteristic analysis, social network link analysis or multi-view fusion, the proposed algorithm exhibits a significant improvement in accuracy and AUC value, which is increased by up to 26.08% and 30.54% respectively on data sets like CAVERLEE, CRESCI-15, CRESCI-17, et al, indicating that DSTC can effectively detect different types of potential grey hat users.

[Key words] social network; grey hat user; network diffusion; feature fusion; user detection

DOI: 10.19678/j.issn.1000-3428.0059636

0 概述

Twitter、Facebook、YouTube、新浪微博等在线社交网络(Online Social Network, OSN)的扩散模式为“去中心化”, 该模式能使用户生成内容(User

Generated Content, UGC)在用户间建立的“关注-被关注”社交网络上广泛传播, 并呈现出传播速度快、覆盖范围广、社会影响力大等特点^[1]。但由于其自带的开放性、普适性、低成本、便捷性等优势, 容易成为攻击目标。

基金项目: 四川省科技计划项目(2019YFSY0032)。

作者简介: 何欢(1996—), 女, 硕士研究生, 主研方向为社交网络、数据挖掘; 朱焱(通信作者), 教授、博士、博士生导师; 李春平, 副教授、博士。

收稿日期: 2020-10-01 修回日期: 2020-12-02 E-mail: 1309825508@qq.com

常见灰帽用户(非正常用户)有僵尸粉、营销号、垃圾用户等,与白帽用户(正常用户)通过OSN实时分享生活、交友聊天、获取信息等不同,灰帽用户利用OSN平台不断扩大自身影响力以提高可信度,而后进行推广营销、引导舆论导向、散步谣言、盗取泄露他人信息、散布非法链接、钓鱼攻击等不友好甚至非法活动,严重威胁平台安全性及性能。因此,检测OSN中的灰帽用户至关重要,有利于OSN管理、广告、新闻媒体与读者等之间的交互优化。

为检测OSN中的灰帽用户,ERŞAHIN等^[2]通过分析用户名、个人资料、背景图片、朋友和关注者数量、推文内容、用户描述、推文的数量等用户属性信息进行分类检测。根据UGC的静态属性信息,RAYMOND等^[3]基于自然语言处理的文本分类,通过分析评论文本与正常用户评论的差异发现网络用户发布的虚假评论。ZHANG等^[4]使用基于链接相似性的方法关联用户活动,并采用基于机器学习的方法对可能的用户活动进行检测。以上方法简单有效,但需要UGC中的垃圾信息(如广告、非法字段等)含有明显关键字或是恶意链接,因此灰帽用户容易通过修改相关信息躲避检测。此外,上述方法只能针对特定数据而无法应对新的威胁,因此不具有普适性。

针对上述问题,有研究人员从“用户-关注-用户”社会关系网络入手提出有限攻击边缘假设,该假设认为白帽用户很少与灰帽交朋友,即白帽用户与灰帽用户之间的友谊链接数量有限。基于该假设,研究人员提出大量检测算法^[5-7]。然而,有研究人员发现灰帽用户能产生更多的攻击边缘^[8-10],即有限攻击边缘假设在现实世界的OSN中不成立。这导致基于该假设基础提出的监测方法存在缺陷,检测精度有待提高^[11-12]。因此,研究人员尝试通过分析用户关注、转发、回复、提及、共享话题等更具可靠性的用户交互行为的方法进行检测。ZHANG等^[13]开发了社交活动网络(Social Activity Network, SAN),通过2层超图统一用户的关注和行为,充分利用用户的行为模式以描述灰帽用户活动到达其受众的方式,并揭示主导信息传播功能的因素。CRESCI等^[14]受生物学遗传信息DNA的启发,通过对垃圾收集器的集体行为进行深入分析,提高了灰帽用户检测的有效性。理论上,与用户关注谁相比,用户在选择与谁互动上更具选择权和可信度。但实际上,该类方法仍只适用于检测具有明显异常行为的灰帽用户。

与单一视图检测方法局限于检测特定种类灰帽用户不同,多视图融合模型能在海量信息中综合使用各类特征或算法,从而保证了低漏检率。MATEENETAL等^[15]提出一种基于用户、内容和图这3类特征的混合检测技术,通过整合特征区分用户,获得更高的效率和精确度。与MATEENETAL类似,LI等^[16]和LIU等^[17]分别针对融合多视图特征提

出了检测机制。LI提出一种半监督混合模型,基于用户、用户社交信任网络、UGC和用户评论转发结构这4类特征检测用户,通过阶梯网络融合过滤各类特征区分用户,并获得更高的效率和精确度。结果表明,融合多类特征的混合模型检测精度更高,其针对不同种类灰帽的检测效果更具有鲁棒性和稳定性。然而,混合方法需要考虑多种视图,检测复杂且时空耗费巨大,且当出现新的种类时仍需重新考量评估参数,不具有普适性。

用户交互是OSN中信息传播的根本途径,灰帽用户虽然种类多样、善于伪装并极易衍生出新种类,但因其最终目的均是通过OSN散布信息扩大自身影响力,故在交互行为上具有共同特性。此外,因为灰帽用户与正常用户的交互行为有明显差异,所以从传播交互角度出发进行检测将更简单有效且通用性更高。本文提出一种基于时空传播的灰帽用户检测机制,从用户UGC传播交互角度出发,在时序、空间2个维度挖掘正常用户与灰帽用户的本质区别。同时在静态属性、社交网络基础上,进一步利用传播网络信息寻找潜在灰帽用户,使灰帽用户识别算法更具普适性。

1 社交网络灰帽用户检测机制

现阶段社交网络灰帽用户检测机制因检测对象极易隐藏且类型多样,目前存在2个问题:1)单一且普适性低,只能针对某一特定数据;2)适配性低,当灰帽用户出现新种类时,需重新评估并改变检测模型。然而,灰帽用户虽然种类多样且善于伪装,但因最终目的均是扩大自身影响力,故在交互行为上具有共同特性,即在其UGC或参与他人UGC传播过程中与白帽用户相比有明显差异。具体来讲,灰帽用户可通过伪装诸如性别、年龄、爱好等属性使自身与白帽用户差异性减小,也能通过发布正常UGC使之不包含垃圾关键字躲避平台检测。但研究数据表明,所有灰帽用户的目的是为了扩大自身在整个社交网络中的比重,以便达成自己营销、宣传、发布广告等最终目的,因此可以从用户UGC传播角度考虑。一方面,社交网络用户影响力主要取决于用户UGC的传播能力;另一方面,灰帽用户经常活跃在其他用户UGC传播链中以便达到宣传目的。此外,用户UGC在发布后越短时间内(时序特性)影响的用户人群(空间特性)越多,传播能力就越强,所以传播特性可以从传播时序和传播空间两方面体现。

本文提出一种基于传播时空特性(Diffusion Spatio Temporal Characteristics, DSTC)的社交网络灰帽用户检测机制,融合传播时序和传播空间2类特性进行最终检测,其具体过程如图1所示。

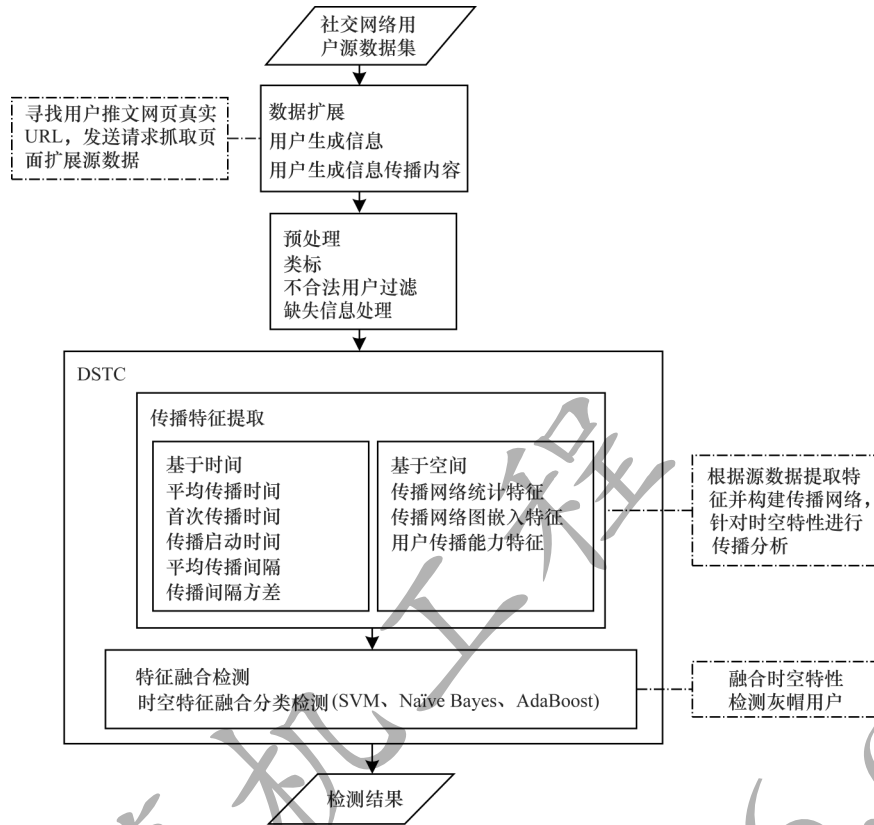


图1 时空特性传播过程

Fig.1 Process of diffusion spatio temporal characteristics

由图1可知,对社交网络用户源数据集进行扩展并预处理,可得到用户UGC及UGC传播过程源数据。基于DSTC对预处理后社交网络用户UGC传播数据进行的检测具体可分为2部分:1)对预处理后的源数据进行时空特征提取工作并得到两类传播特征,包括时序特征和空间特征,时序特征即传播过程在时序上的特性,空间特征即UGC形成的传播网络图所体现的特性;2)融合传播时空两类特征,并分别采用判别式模型代表(SVM)、生成式模型代表(Naive Bayes)、集成学习代表(AdaBoost)这3类分类算法检测灰帽用户,得到最终检测结果。

2 传播特征提取

为更好定义传播特征提取过程,现给出相关重要符号定义:用 $U_{UGC}(u)$ 表示用户 u 的用户生成内容, $u \in U, U \subseteq V$ 。其中 U 表示评论过用户 u 该条UGC的所有用户, V 表示整个网络中的所有用户。假设用户 u 的一条UGC被发布后收到 $n-1$ 条UGC评论,设三元组集合 $U_{UGC}(u) = \{ \langle u_i, t_i, u_j \rangle \}$, $u_i, u_j \in U$, $i > 1, j < m$ 表示用户 u_i 在源UGC发布 t_i 时间后评论用户 u_j ,三元组顺序按时间 t 升序排序。集合 $U_{UGC}(u)$ 中时间 t_i 直接反映 $U_{UGC}(u)$ 的传播时间特性,数量 n 及三元组 $\langle u_i, t_i, u_j \rangle$ 则间接表明UGC的传播空间特

性。根据已定义好的符号,提取传播时序和空间特征。

2.1 传播时序特征的提取

白帽用户发布的UGC能达到的传播范围与自身在社交网络中重要程度、UGC内容包含的模式、UGC文本情感倾向等诸多因素有关。因此,白帽用户 $U_{UGC}(u)$ 中体现的传播时间与传播范围没有具体的界限,随机性较强。而灰帽用户一般在在特定时间有目的地发布UGC,过了特定时间段不再传播,传播时间上相似性更强。综上所述,鉴别灰帽用户可以从 $U_{UGC}(u)$ 的传播时间角度考虑。

平均传播 A_{ADT} 代表 $U_{UGC}(u)$ 传播开始至结束收到每个用户评论所用的时间间隔。灰帽用户 A_{ADT} 较白帽而言更加稳定,数值相差小。平均传播时间的计算公式如式(1)所示:

$$A_{ADT} = \frac{t_n - t_1}{n} \quad (1)$$

首次传播时间 F_{FDT} 代表 $U_{UGC}(u)$ 从传播开始至收到第1个用户评论的时间间隔。灰帽用户评论其他用户UGC的通道较单一,通常是经过给定的链接直接进入,且灰帽用户UGC一般只会收到灰帽用户评论。所以,白帽用户发布UGC后,关注该白帽的其他用户在接收推送后对其进行评论互动具有实时特性,灰帽则

没有。因此,灰帽用户的 F_{FDT} 一般要比白帽用户更长。首次传播时间的计算公式如式(2)所示:

$$F_{\text{FDT}} = t_2 - t_1 \quad (2)$$

传播启动时间的计算公式如式(3)所示:

$$S_{\text{SDT}} = t_m - t_1 \quad (3)$$

其中: m 为传播启动的阈值,即当 $U_{\text{UGC}}(u)$ 中 $n > m$ 时(UGC至少收到 m 条评论),认为该条UGC达到传播认定条件。本文设 $m = 100$ (OSN中UGC评论数量中位数),即当转发量达到100后UGC被认为是启动传播,可以对整个OSN存在影响。 S_{SDT} 越小,影响范围越大。过滤用户发布的不重要UGC,只考虑传播范围较大能对OSN产生影响的UGC。此外, m 所花费的时间大小表明UGC的受欢迎程度,能侧面体现用户 u 在社交网络中的重要性。灰帽用户由于经常发送重复相似垃圾UGC,不被大多数用户认可,被关注的可能性小,影响力一般较小。

平均传播间隔如式(4)所示:

$$A_{\text{ADI}} = \frac{\sum_{i=2}^n (t_i - t_{i-1})^2}{n-1} \quad (4)$$

传播间隔方差如式(5)所示:

$$V_{\text{VDI}} = \frac{\sum_{i=2}^n (t_i - t_{i-1})^2}{n-1} - A_{\text{ADI}}^2 \quad (5)$$

其中:平均传播间隔 A_{ADI} 和传播间隔方差 V_{VDI} 分别代表元组 $\langle u_i, t_i, u_j \rangle$ 之间的时间间隔及分布情况。因灰帽用户行为多集中在短时间之内进行,呈现出突发特性,因此每2条相邻信息之间的时间间隔相对白帽用户要小。另外,突发性不仅表现在时间间隔短,时间间隔分布与白帽相比也会处于一个相对较小的范围内。而白帽用户的转发评论等受访问随机推送影响表现出更大的差异性。

2.2 传播空间特征的提取

以用户 u_i 为节点, $U_{\text{UGC}}(u) = \{\langle u_i, t_i, u_j \rangle\}$ 中元组 $\langle u_i, t_i, u_j \rangle$ 代表的“用户-评论-用户”为边构造传播网络,可反映在UGC动态传播空间中用户的节点信息特性。

基于图结构的检测方法通常比其他检测方法效率高,因为灰帽用户虽然能伪造信息躲避检测,但是其行为模式却不能轻易改变。本文从传播空间上提取以下几类特征。

2.2.1 传播网络结构统计特征

直接由图结构统计计算获取,诸如 PageRank、clustering、betweenness 等常见图节点结构信息。

2.2.2 传播网络图嵌入特征

图嵌入技术能对网络中的用户节点进行低维向量表示,且该低维特征向量能较好地保留原有网络

的拓扑结构。Node2vec 模型^[18]认为网络结构上相似节点具有相似的嵌入表示,属于同一社区的节点在低维空间的距离更相近。本文采用 Node2vec 模型对传播网络图进行图嵌入特征提取,得到用户节点特征向量。

2.2.3 用户传播能力特征

用户传播特征由以下指标表征:

1)一阶自我中心网络环路路径数量,用以评估用户传播的量级程度。用户一阶传播网络如图2所示。

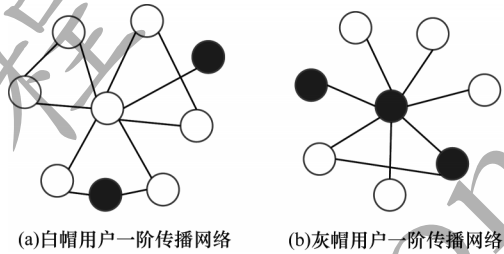


图2 用户一阶传播网络

Fig.2 First order diffusion network of user

在图2中,白色为白帽用户,黑色为灰帽用户。可以看到,图2(a)中有3条环路,图2(b)只有1条环路,证明2类用户在自我一阶传播网络之间确实存在差异。

一阶自我中心网络环路路径数量是指由当前用户出发经过一阶邻居后最终能回到用户并构成回路路的数量。构成环路的用户集实际是社交网络中的一个社区,代表了因同一兴趣形成的社交圈。白帽用户的自我中心网络通常呈现局部分散而整体聚集的状态,这是因为白帽用户兴趣多元交叉,扩散开后又形成多个社区;而灰帽用户由于不关心其他用户,其社交圈也并非由兴趣促使形成,因此其传播网络呈现整体、局部都分散的状态。所以,灰帽用户自我中心网络所形成的回路路径数量一般比白帽用户少。

2)用户传播指数,用以评估用户在网络中的传播能力的指标。借鉴微博传播指数 BCI (Blog Communication Index),通过微博的活跃度和传播度反映用户传播能力和传播效果,利用用户传播指数 $U_{\text{UCI}}(u)$ 评估用户 u 的传播能力,其计算公式如式(6)所示:

$$U_{\text{UCI}}(u) = 0.2 \times W_1 + 0.8 \times W_2 \quad (6)$$

其中: W_1 代表用户活跃度; W_2 代表用户在传播网络中能达到的传播度;计算公式如式(7)和式(8)所示:

$$W_1 = 0.3 \times \ln(X_1 + 1) + 0.7 \times \ln(X_2 + 1) \quad (7)$$

$$W_2 = 0.2 \times \ln(X_3 + 1) + 0.2 \times \ln(X_4 + 1) + 0.25 \times \ln(X_5 + 1) + 0.25 \times \ln(X_6 + 1) + 0.1 \times \ln(X_7 + 1) \quad (8)$$

其中: X_1 为 UGC (总数 UGC 代表用户传播能力); X_2 为原创 UGC 数; X_3 为转发 UGC 数; X_4 为评论 UGC 数; X_5 为原创 UGC 转发数; X_6 为原创 UGC 评论数; X_7 为原创 UGC 点赞数。式(8)中每个 X 特征代表一类评价指标,对每个 X 特征进行 $X = \ln(X + 1)$ 的标准化处理后分配权重。

3) 用户传播信任度,用以评估传播用户在网络中传播信任的能力。通过用户传播网络中其一阶邻居用户给予的信任度可大致判断其种类。通常来说,白帽用户更倾向与白帽交互,故传播网络中节点的一阶出度邻居为白帽的越多,该用户为白帽的可能性就越大,即他人给予的信任度越高。反之,当入度节点的灰帽节点越多,代表越信任灰帽用户,自身为灰帽用户的可能性越大,他人给予的不信任度越高。

借鉴 PageRank 算法的思想,定义节点 u 的信任度 $t_{\text{trust}}(u)$ 与不信任度 $d_{\text{distrust}}(u)$ 的计算公式分别如式(9)和式(10)所示:

$$t_{\text{trust}}(u) = \alpha \sum_{p: u \rightarrow p} \frac{t_{\text{trust}}(p)}{i_{\text{indegree}}(p)} + (1 - \alpha)s(u) \quad (9)$$

$$d_{\text{distrust}}(u) = \alpha' \sum_{q: q \rightarrow u} \frac{d_{\text{distrust}}(q)}{o_{\text{outdegree}}(q)} + (1 - \alpha')s'(u) \quad (10)$$

其中: p 代表用户 u 的出度节点,即用户 u 评论用户 p 的 UGC; $t_{\text{trust}}(u)$ 代表节点 p 拥有的信任值; $i_{\text{indegree}}(p)$ 为 p 的所有入边数量,代表 u 信任 p ; 两者相除代表 p 分配给 u 的信任值,求和得到 u 从自身一阶邻居所得到的信任值; $s(u)$ 代表用户 u 的初始信任值,通过参数 α 调节自身信任值与从一阶邻居获取分配的信任值,更新信任用户 u 为白帽的信任值。不信任值计算原理与信任值一样,不同的是不信任值从用户 u 的出度节点 q 获取,且 q 的不信任值分配通过 q 的出度数量 $o_{\text{outdegree}}(q)$ 计算。

4) 用户传播率,用以评估用户传播占整个 OSN 的比重。传播率是指信息接受人群占传播对象的百分比,即 UGC 自身网络节点数与整个研究对象网络的比率。

$$D_{\text{diffusion rate}}(u) = \frac{U_{\text{UGC}}^{\text{number}}(u)}{A_{\text{All}}} \quad (11)$$

其中: A_{All} 为所有 UGC 传播网络中的节点数; $U_{\text{UGC}}^{\text{number}}(u)$ 为用户 u 的 UGC 参与传播的用户数量。

2.3 传播特征融合

将传播时序和空间两类特征结合后更能反映用户特性,故借鉴早期先融合多层特征再训练预测的思想,选择并行策略将时序、空间两类特征向量组合成复向量。对于输入的时序特征 x 和空间特征 y ,通过超参数 β 调节权重得到社交网络用户特征向量 $z = \beta \times x + (1 - \beta) \times y$ 。最终选取判别式模型代表 SVM、生成式模型代表 Naive Bayes 及集成学习分类算法代

表 AdaBoost 检测社交网络灰帽用户,并对检测结果进行比较分析。

3 实验结果与分析

3.1 数据集

为分析验证 DSTC 的适用性和有效性,本文实验共用了 4 个数据集,各数据集统计信息如表 1 所示。

表 1 DSTC 数据集数据分布

Table 1 Distribution of DSTC dataset

数据集	灰帽用户数量	白帽用户数量	UGC	Diffusion
CRESCI-17 ^[14]	10 894	3 474	√	√
CAVERLEE ^[19]	22 179	19 276	√	×
YANG ^[20]	42 446	8 092	×	×
CRESCI-15 ^[21]	1 715	1 440	√	√

在表 1 中,UGC 和 diffusion 分别表示数据集中是否包含用户发布的 UGC 及对应传播信息,√代表包含,×代表不包含。当源数据不包含 UGC 或 UGC 传播信息时,通过网络爬虫对社交网络源数据进行数据扩展,根据源数据中的用户信息匹配查找并确定用户,爬取用户最新的信息和最近 50 条 UGC 及其传播过程,保证源数据最新且数量足够用来分析 UGC 及 UGC 传播过程信息。如果出现用户已注销或源 UGC 已删除等错误,则忽略该用户或该 UGC。

Caverlee 数据集由 RYUMINL 等^[19]提供,包含从 2009 年 12 月 30 日至 2010 年 8 月 2 日在 Twitter 上收集的社交蜜罐数据集。该数据集包含用户基本属性信息,用户粉丝数随时间的变化及这段时间内用户发布的推文。

根据 2018 年美国中期选举期间收集的政治推文,美国印第安纳大学复杂网络与系统研究中心的 YANG 等^[20]筛选收集了相关用户及数据,并手动确定了一些真正参与了有关选举和在线讨论的真实人类用户及发现的机器人帐户。在选举后,大多数机器人程序帐户都已被 Twitter 暂停,证实了作者标注标签的正确性。

CRESCI-17^[14]和 CRESCI-15^[21]均由 CRESCI 团队提供。CRESCI-15 包含手动标注的真实和虚假 Twitter 帐户。CRESCI-17 数据集的僵尸用户包含更细粒度的分类:传统的垃圾用户、社交垃圾用户和假粉丝。传统的垃圾用户监听程序是简单的漫游器,会反复发布相同的内容;社交垃圾用户模仿普通用户的个人资料和行为,可以躲避某些检测方法;假粉丝是某用户为了扩大影响力而购买的用户。本文将 3 类不同类标的灰帽用户统一为灰帽用户(不区分灰帽类型,类标一致)。

3.2 实验设计及结果分析

3.2.1 传播特征有效性验证

为了验证所提传播时空特征是否有效,另提取传统方法所用的用户属性特征和UGC文本特征。用户属性特征包括粉丝数量、关注数量、UGC总数、F-F比率、性别、年龄、是否为认证用户等特征;UGC文本特征包括最近一周发布UGC的数量、包含超链接的UGC占UGC总数的比率、评论他人的UGC占UGC总数的比率、转发他人的UGC占UGC总数的比率、@他人的UGC占UGC总数的比率、参与话题的UGC占UGC总数的比率、UGC之

间的相似性等特征。针对3类特征分别采用SVM、Naïve Bayes、Adaboost分类算法进行检测,实验结果如表2所示。评价指标采用准确率(Accuracy)、F1-score和AUC(Area Under Curve)。其中F1-score代表precision(正确预测的正样本数占有预测为正样本的数量的比值)和recall(正确预测的正样本数占真实正样本总数的比值)的调和平均,F1-score越高说明试验方法越有效;AUC代表ROC曲线(以假正率(FP_rate)和真正率(TP_rate)为轴的曲线)的面积,AUC越高,分类性能越好。

表2 不同分类器在不同数据集下特征分类性能对比
Table 2 Comparison of feature classification performance of different classifiers on different datasets

数据集	分级机	用户特征			文本特征			扩散特性		
		准确率	F1-score	AUC	准确率	F1-score	AUC	准确率	F1-score	AUC
CRESCI-17 ^[14]	SVM	0.801 2	0.798 8	0.800 9	0.798 9	0.799 9	0.782 3	0.900 2	0.901 9	0.901 3
	NB	0.823 4	0.819 9	0.810 5	0.809 3	0.810 4	0.807 8	0.910 9	0.909 2	0.909 8
	AdaBoost	0.857 8	0.851 1	0.826 7	0.821 7	0.817 9	0.817 3	0.929 0	0.930 5	0.929 9
CAVERLEE ^[19]	SVM	0.937 0	0.942 4	0.939 6	0.909 1	0.900 8	0.907 0	0.944 4	0.937 6	0.933 5
	NB	0.959 7	0.953 2	0.949 5	0.890 9	0.889 5	0.892 8	0.996 2	0.995 9	0.995 6
	AdaBoost	0.998 3	0.999 1	0.999 0	0.931 8	0.931 5	0.934 2	0.998 9	0.998 8	0.998 8
YANG ^[20]	SVM	0.845 7	0.891 7	0.832 2	0.803 8	0.832 9	0.812 3	0.912 4	0.920 9	0.913 0
	NB	0.867 3	0.889 2	0.872 7	0.863 2	0.865 9	0.859 9	0.901 2	0.913 4	0.898 4
	AdaBoost	0.932 4	0.923 2	0.917 8	0.891 7	0.900 1	0.892 2	0.931 2	0.933 3	0.934 2
CRESCI-15 ^[21]	SVM	0.813 4	0.816 7	0.817 8	0.801 9	0.813 5	0.806 7	0.912 4	0.910 3	0.911 6
	NB	0.847 8	0.850 0	0.811 2	0.837 0	0.824 9	0.830 6	0.927 8	0.923 9	0.929 7
	AdaBoost	0.876 3	0.866 8	0.859 9	0.892 0	0.882 6	0.880 1	0.942 3	0.940 8	0.942 2

表2中加粗数据表示不同分类方法针对同一分类器下在同一数据上分类指标最优的数据。由表2可知,本文提出的DSTC方法所提取的传播时空特征在各个数据集上的分类效果均优于传统方法所用的用户属性和UGC文本特征,证明了DSTC所提传播时空特征的有效性。以研究应用最广且分类效果差别不大的Caverlee数据集为例,选用集成学习AdaBoost方法时,通过对比用户特征和文本特征,发现传播特征在AUC值上也能分别提高0.000 8和0.064 6。用户特征性能优于文本特征是因为相比用户特征单一选项更改性不强,灰帽用户更容易通过发布正常UGC文本来隐藏自身,而传播特征直接反映用户行为特性,可以更好地揭示用户之间的差异,故分类效果更好。

以差异最明显的CRESCI-17数据集为例,选用集成学习AdaBoost方法对比用户特征和文本特征,发现传播特征在AUC值上分别提高0.103 2和0.112 6。此外,虽然同样是传播特征且在不同数据集不同分类器中传播特征分类表现有差异,但整体分类性能表现良好。而用户属性、文本特征的分类性能虽然在某个数据集上优于DSTC传播特征,但在其他数据集的分类效果并不理想,证明传统方法并不适合所有数据集,其鲁棒性不高。本文DSTC方法提出

的传播特征适用性更高。

3.2.2 DSTC方法有效性验证

为验证本文DSTC检测方法的有效性,与其他同类检测方法进行对比,包括与传统检测方法和当前较为流行或新颖的灰帽用户检测算法进行对比,如CRESCI提出关于用户UGC传播相似性的社交指纹数字DNA检测方法(DDNA)、通过常用混合模型方法检测的SSDMV方法和最近提出的集成用户社交网络和活动图网络的SAN方法,AUC值对比如图3所示,实验对比结果如表3所示。

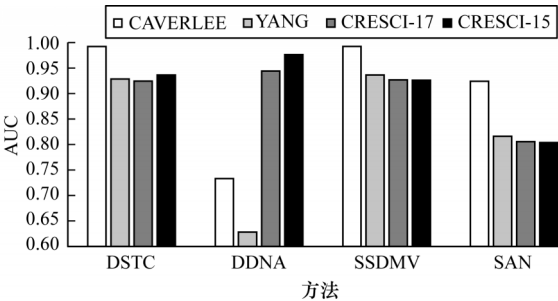


图3 不同方法在不同数据集下AUC的对比
Fig.3 Comparison of Area Under Curve of different methods on different datasets

表3 不同方法在不同数据集下分类性能对比

Table 3 Comparison of Classification performance of different methods on different datasets

数据集	DSTC			DDNA ^[14]			SSDMV ^[16]			SAN ^[13]		
	准确率	F1-score	AUC	准确率	F1-score	AUC	准确率	F1-score	AUC	准确率	F1-score	AUC
CRESCI-17 ^[14]	0.929 0	0.930 5	0.929 9	0.952 2	0.949 7	0.949 9	0.949 8	0.939 7	0.932 4	0.798 2	0.812 8	0.809 2
CAVERLEE ^[19]	0.998 9	0.998 8	0.998 8	0.738 1	0.729 3	0.735 6	0.998 3	0.999 0	0.999 1	0.923 3	0.919 7	0.929 8
YANG ^[20]	0.931 2	0.933 3	0.934 2	0.639 1	0.605 3	0.628 8	0.932 4	0.946 7	0.942 4	0.823 1	0.809 4	0.820 1
CRESCI-15 ^[21]	0.942 3	0.940 8	0.942 2	0.989 8	0.971 7	0.982 9	0.931 0	0.932 3	0.931 9	0.801 6	0.808 8	0.807 4

由图3和表3可以看出,DDNA方法在其他数据集上的效果并不理想,这是因为DDNA通过作者自定义设计的数字DNA转换方法将用户UGC转为DDNA序列,并通过计算序列之间的相似性学习两类用户之间的差异。DDNA方法虽简单高效,但因为设计主观性太强,普适性并不高,只在针对表现差异明显的CRESCI数据集时有较好表现。

SAN方法通过统一用户社交网络与UGC传播活动网络,并耦合3种基于随机游动的算法检测灰帽用户,该方法在各个数据集上表现良好。但因SAN所采取的一半监督信任传播策略本身存在实验效果稳定但精度不够的问题,虽然已解决普适性和适配性问题,但该方法在各个数据集上的表现也并非最优。

SSDMV方法效果与DSTC差异不大甚至在有些数据集上优于DSTC,能解决普适性和适配性问题,但SSDMV方法需提取用户、文本、社交网络关注图结构、用户回复图结构等4类特征后将各个视图特征通过阶梯网络设计过滤门组件融合训练,方法复杂且难于计算,时空耗费太高。

基于DSTC的用户检测性能在多个数据集上优于其他方法,例如准确率最高提升26.08%,AUC值最高提升30.54%。这是因为DSTC提取的基于传播时序和空间特性能更好地反映各类灰帽与白帽用户之间的差异,简化检测算法的同时增强了检测算法的鲁棒性和普适性。

综上所述,本文所提DSTC方法能有效检测社交网络灰帽用户,不仅解决了灰帽用户检测算法只能针对特定种类的问题,而且更加简单,检测精度和适用性更高。

4 结束语

本文针对社交网络灰帽用户检测算法适用性较差的问题,提出一种基于传播时空特性的检测算法。根据社交网络UGC传播中的时空特性定义提取相关特征,从UGC传播角度区分灰帽白帽之间的差异性,并融合传播时序和传播空间特征进行分类检测。实验结果表明,该算法在CAVERLEE、CRESCI-15、CRESCI-17等多个数据集上效果较好,在保证检测精度的前提下,简化了检测算法,

提高了算法适用性。下一步将研究传播序列的上下文关系特性,同时结合特征融合算法实现更好的分类性能。

参考文献

- [1] 吴越,陈晓亮,蒋忠远. 微博信息流行度预测研究综述[J]. 西华大学学报(自然科学版),2017(1):1-6.
WU Y, CHEN X L, JIANG Z Y. Survey on predicting popularity of information in microblogs [J]. Journal of Xihua University (Natural Science Edition), 2017(1): 1-6. (in Chinese)
- [2] ERÇAĞIN B, AKTAŞ Ö, KILIÇ D, et al. Twitter fake account detection[C]//Proceedings of the 2nd International Conference on Computer Science and Engineering. Washington D. C., USA: IEEE Press, 2017: 388-392.
- [3] RAYMOND Y K, STEPHEN L, LIAO S Y. Text mining and probabilistic language modeling for online review spam detection [J]. ACM Transactions on Management Information Systems, 2011, 2(4): 25-30.
- [4] ZHANG X, ZHU S, LIANG W. Detecting spam and promoting campaigns in the twitter social network[C]//Proceedings of the 12th IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2012: 1194-1199.
- [5] NEIL Z G, MARIO F, PRATEEK M. Sybil belief: a semi-supervised learning approach for structure-based sybil detection[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(6): 976-987.
- [6] JIA J Y, WANG B H, GONG N Z Q. Random walk based fake account detection in online social networks [C]//Proceedings of the 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. Washington D. C., USA: IEEE Press, 2017: 273-284.
- [7] WANG B, ZHANG L, GONG N Z. Sybilscar: Sybil detection in online social networks via local rule based propagation[C]//Proceedings of 2017 IEEE Conference on Computer Communications. Washington D. C., USA: IEEE Press, 2017: 1-9.
- [8] YANG Z, WILSON C, WANG X, et al. Uncovering social network sybils in the wild [J]. ACM Transactions on Knowledge Discovery from Data, 2014, 8(1): 1-29.
- [9] SRIDHARAN V, SHANKAR V, GUPTA M. Twitter games: How successful spammers pick targets [C]//Proceedings of the 28th Annual Computer Security

- Applications Conference. New York, USA; ACM Press, 2012; 389-398.
- [10] BOSHMAF Y, MUSLUKHOV I, BEZNOSOV K, et al. The social bot network: when bots socialize for fame and money [C]//Proceedings of the 27th Annual Computer Security Applications Conference. New York, USA; ACM Press, 2011; 93-102.
- [11] KOLL D, SCHWARZMAIER M, LI J, et al. Thank you for being a friend: an attacker view on online-social-network-based sybil defenses [C]//Proceedings of the 37th IEEE International Conference on Distributed Computing Systems Workshops. Washington D. C., USA; IEEE Press, 2017; 157-162.
- [12] EFFENDY S, YAP R H. The strong link graph for enhancing sybil defenses [C]//Proceedings of the 37th IEEE International Conference on Distributed Computing Systems. Washington D. C., USA; IEEE Press, 2017; 944-954.
- [13] ZHANG X, XIE H, LUI J C. Sybil detection in social-activity networks: modeling, algorithms and evaluations [C]//Proceedings of the 26th IEEE International Conference on Network Protocols. Washington D. C., USA; IEEE Press, 2018; 44-54.
- [14] STEFANO C, ROBERTO D, MARINELLA P, et al. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling [J]. IEEE Transactions on Dependable and Secure Computing, 2018, 15(4): 561-576.
- [15] MATEEN M, IQBAL M A, ALEEM M, et al. A hybrid approach for spam detection for Twitter [C]//Proceedings of the 14th International Bhurban Conference on Applied Sciences and Technology. Washington D. C., USA; IEEE Press, 2017; 466-471.
- [16] CHAOZHUO L, SENZHANG W, LIFANG H, et al. SSDMV: semi-supervised deep social spammer detection by multi-view data fusion [C]//Proceedings of the 18th IEEE International Conference on Data Mining. Washington D. C., USA; IEEE Press, 2018; 247-256.
- [17] LIU Y, WU B, WANG B, et al. SDHM: a hybrid model for spammer detection in Weibo [C]//Proceedings of 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York, USA; ACM Press, 2014; 942-947.
- [18] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks [C]//Proceedings of 2016 ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. New York, USA; ACM Press, 2016; 855-864.
- [19] KYUMIN L, EOFF B D, CAVERLEE J. Seven months with the devils: a long-term study of content polluters on Twitter [EB/OL]. [2020-09-01]. https://www.researchgate.net/publication/221297999_Seven_Months_with_the_Devils_A_Long-Term_Study_of_Content_Polluters_on_Twitter.
- [20] YANG K C, VAROL O, HUI P M, et al. Scalable and generalizable social bot detection through data selection [EB/OL]. [2020-09-01]. <https://arxiv.org/abs/1911.09179>.
- [21] CRESCI S, PIETRO D R, PETROCCHI M, et al. Fame for sale: efficient detection of fake Twitter followers [J]. Decision Support Systems, 2015(80): 56-71.