基于行为的新浪微博恶意用户识别

林成峰[1]，陈　凯[2]，周　异[2]，周　曲[2]

（1. 上海交通大学信息安全工程学院，上海　200240；
2. 上海交通大学电子信息与电气工程学院，上海　200240）

摘要：以新浪微博为研究对象，提出一种基于行为特征检测微博恶意用户的方法。利用蜜罐账户、爬虫程序、淘宝购买等多种方法收集恶意用户样本。根据行为模式将恶意用户样本进行分类，得到 3 种恶意用户类型：过度广告、重复转发和过度关注。然后提取用户行为特征，通过数据统计分析以及与正常用户的比较，找出恶意用户的行为特点。最后，利用机器学习工具构造自动分类器用于自动鉴别恶意用户，并且在对分类器进行测试之后证实了该方法的可行性和准确性。

关键词：计算机工程；新浪微博；恶意行为；恶意用户检测；机器学习
中图分类号：TP393　　**文献标识码**：A　　**文章编号**：1674-2850(2014)04-0322-10

Behavior-based identification of spammers in Sina weibo

LIN Chengfeng[1], CHEN Kai[2], ZHOU Yi[2], ZHOU Qu[2]

(1. *School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai* 200240, *China*;
2. *School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai* 200240, *China*)

Abstract: This paper proposed a method to identify spammers in Sina weibo based on their behavior features. We collected spammer samples using different methods such as using honeypot, using crawler and buying from online merchants. We divided spammers into three categories according to their behavior pattern: aggressive advertising, repeated duplicate reposting and aggressive following. We analyzed behavior and compared them with the legitimate users to extract features that could tell spammers from legitimate users. We built identification system with the help of machine learning toolkits. The evaluation result showed the effectiveness and accuracy of the identification system.

Key words: computer engineering; Sina weibo; spamming behavior; spammer detection; machine learning

0　引言

长期以来，互联网上的恶意行为给网络用户带来各种麻烦，威胁着用户的隐私和财产安全，破坏网络环境。恶意行为起初主要以电子邮件为载体：垃圾邮件、邮件炸弹等恶意行为常令邮件服务商和用户头疼。如今，随着社交网络的兴起，各种恶意行为已经蔓延到这一新平台。以 Facebook 为代表的社交网络饱受其苦[1]，这些恶意行为给运营商带来了巨大损失。

最近，微博作为一种新的交流、信息发布以及市场营销平台正在飞速发展。微博的代表 Twitter 拥有超过 5 亿用户，每天产生 3.4 亿的推文。许多重大社会事件都有 Twitter 参与的身影：自然灾害、总统选举、犯罪追踪等。重大社会事件在微博上开始传播的时间往往要远早于主流媒体的报道时间。这种实时、快速的信息发掘和传播方式使得微博成为了各种恶意行为滋生的温床。而短链接技术在微

基金项目：高等学校博士学科点专项科研基金（20120073120030）；信息网络安全公安部重点实验室开放课题（C12607）
作者简介：林成峰（1989—），男，硕士研究生，主要研究方向：社交网络数据挖掘
通信联系人：陈凯，副教授，主要研究方向：数据挖掘、社交网络分析. E-mail：kchen@sjtu.edu.cn

博中的广泛应用使得通过链接传播的有害信息很难被检测出来。GRIER 等[2]曾对 Twitter 中的 URL 链接做过一个统计，约有 8%的短链接最终指向包含钓鱼信息、恶意软件或病毒的页面，这些链接的点击率大约在 0.13%左右。

在国内，以新浪微博为代表的各大微博平台，也同样遭受恶意行为的威胁。这里以新浪微博为主要研究平台，介绍一种基于行为检测的微博恶意用户检测机制：通过抓取恶意用户样本、分析样本行为、提取行为特征、构造机器学习自动分类器等一系列步骤，设计、实现了一种通过检测行为检测微博恶意用户的系统，最后的测试也表明该系统的有效性。

1 相关研究

传统的反恶意行为机制通常从以下 2 个角度展开：检测恶意信息和检测恶意用户。恶意信息的检测往往基于内容的分析和统计特征，如 HUANG 等[3]利用数理统计方法分析恶意信息的特征；YIN 等[4]则通过研究内容和上下文特征检测恶意信息。对于社交网络来说，由于信息产生的速度和量都非常巨大，因此针对恶意信息的检测效率较低。相反地，针对恶意用户的检测则会有较好的效果。如，IRANI 等[5]分析了 190 万 MySpace 用户的档案，利用机器学习技术，设计并实现了一套恶意用户检测系统。WEBB 等[6]利用蜜罐手段，在 MySpace 上成功捕获 1 570 个恶意用户。LEE 等[7]在 Facebook 上采用了类似的手段，并且取得了不错的效果。WANG 等[8]则提出了一种更为通用的跨平台恶意用户检测机制。总的来说，在社交网络领域，特别是微博领域，针对恶意用户的检测手段更为常见且有效。

在国外，虽然微博反恶意用户的课题得到了广泛研究，但是其研究对象基本上都是以 Twitter 为代表的英语微博平台，其用户以国外用户为主。对于中文微博的研究则比较少见。由于文化和用户群体的差异，中文微博平台中的恶意用户和 Twitter 上的恶意用户可能存在较大不同。因此，专门针对中文微博平台的研究也十分必要。研究以新浪微博为对象，提出一种基于用户行为的恶意用户检测系统。通过检测恶意行为判断用户是否是恶意用户。研究的主要成果有如下 3 点：

1）为获取用于分析的恶意用户样本，研究采取了多种方法：主动式蜜罐账户引诱、爬虫程序抓取、直接在网上购买微博推广服务来获得恶意用户样本。

2）通过分析捕获的样本，总结归纳出新浪微博平台中常见的 3 种恶意行为：过度广告、重复转发和过度关注。通过数据统计，提取出了这 3 种行为的一些特征。通过与正常用户比较，找出了具有区分度的那些特征。

3）利用上一步骤找到的特征，为每一种行为建立机器学习分类器。将这些分类器整合成为一个自动检测系统，并且对其进行测试。测试结果表明，分类器可以较高的准确度识别未知用户是否为恶意用户。

2 恶意用户样本收集

2.1 使用蜜罐账户

STRINGHINI 等[9]关于 Twitter 和 Facebook 的研究成果表明，这 2 个社交网络平台中的恶意用户会主动关注其他用户来建立关系。相对地，新浪微博中的恶意用户也很有可能拥有这种特性。于是建立了 25 个主动式的蜜罐账户，利用程序让他们自动运行，模拟人类用户的一些行为，例如发布微博、转发微博、关注用户等。笔者进行了长达 8 个月的实验，最终捕获了 517 个用户[10]。

浏览这些用户的个人主页和微博正文，通过内容判断他们是不是恶意用户。将拥有以下行为的人定义为恶意用户：1）发布包含广告和钓鱼链接的微博；2）发布指向恶意网页和带毒网页的 URL 链接；3）反复发布包含相似或相同内容的微博；4）所有其他可能会给其他用户带来骚扰的行为。

在 517 个用户中，发现了 114 个恶意用户。其中，只有 9%的用户和蜜罐账户互相关注，这说明大部分恶意用户都是主动地关注蜜罐账户（蜜罐账户不会主动地关注自己的新粉丝）。这也直接验证了之

前的猜想：新浪微博中的恶意用户会积极主动地关注其他用户来建立社会关系。

此外，蜜罐的活跃程度也影响着吸引恶意用户的效率。表 1 显示的是吸引到最多恶意用户的 5 个蜜罐的具体数据。对每个蜜罐都设置了不同的运行参数，他们的活跃程度不同，发帖内容也有很大差异。由表 1 可以看出：绝大多数恶意用户都是由最活跃的那几个蜜罐吸引得到，而活跃度较低的蜜罐账户能吸引到的恶意用户数量则相对非常少；有 84％的恶意用户是由前 3 个蜜罐吸引得到，而剩余的所有蜜罐无论是在微博数量、关注数量还是吸引到的恶意用户的数量上都无法和前 3 名相比。

表 1　蜜罐账户捕获情况统计

Tab. 1　Results of honeypot capture

微博数量/条	粉丝数量/个	关注数量/个	捕获恶意用户/个	非好友粉丝比例/%
1 595	192	853	45	50.00
1 052	95	644	37	41.05
1 919	76	597	12	32.89
1 072	50	355	3	22.00
295	6	51	2	50.00

2.2　使用爬虫程序

由于蜜罐收集恶意用户的数量无法满足研究需求，因此需要采取另外的方法收集更多恶意用户样本。在新浪微博中，恶意用户会经常出现在一些热门微博或者明星微博的转发列表中。他们通过转发这些微博参与到热门话题或者某个明星的粉丝群体中，试图建立更多的社会关系，以获取目标用户。因此，设计并实现了一个微博爬虫程序，用这个爬虫程序监控微博名人的微博，将积极参与热门微博的用户抓取出来，希望能在其中找到恶意用户。

利用新浪提供的应用程序界面（application program interface，API），完成了用于监控用户最新微博更新情况和抓取微博转发列表的爬虫程序。对于输入的指定用户，该程序会自动抓取该用户最新的微博转发列表，将参与转发用户的转发记录存入数据库，之后通过统计和排序，挑选出其中最活跃、最积极的那部分用户。这里选择新浪微博影响力排行榜前 100 的用户作为监控对象，这些用户所发布的微博大多数会成为成千上万人转发的目标。他们通常是在现实生活中拥有极大影响力的名人，常常拥有几十、甚至上百万的微博粉丝。利用爬虫程序过滤得到了一批最活跃用户的名单，之后通过浏览他们的微博和主页，过滤得到了 879 个恶意用户。

在这些恶意用户中发现了一种特别的恶意用户。他们对于同一条微博会重复转发多次，此外，他们所转发微博的来源往往只有特定的一个或几个人，这些人都是一些著名的歌手或电影演员。一个典型的例子就是著名演员吴奇隆，他在新浪微博中拥有数百万粉丝。879 个恶意用户中有 100 多人都几乎转发了吴奇隆的每一条微博，他们对一条微博会转发多次，每次附带的评论内容非常类似甚至几乎相同。与这些恶意用户不同的是，正常的用户即使转发了吴奇隆的微博，也很少有重复转发的现象。通过统计这些用户之间的社会关系，发现他们大多数都互相关注，形成一个集团，甚至可以看出一些人为操纵的迹象。

2.3　购买微博推广服务

随着新浪微博的流行，越来越多的人尝试使用微博进行各种营销活动。一种低成本的“微博粉丝服务”由此产生，在淘宝等电子购物平台上，有许多商家出售这种服务：利用自己控制的大量账号，通过关注顾客的账号提高顾客账号的人气（也有转发、评论、点赞等功能），帮助用户完成微博的推广传播。这种最“简单粗暴”的手段往往带有恶意行为色彩，因为这些账号在关注“顾客”的同时也会胡乱关注其他用户，这时常会对他人造成骚扰。笔者通过直接在淘宝上购买这种“粉丝”，获得恶意用户样本。从几家不同的商铺，购买了大约 8 000 个“粉丝”，并且随机选取其中的一部分作为恶意用户样本。

3 恶意用户特征分析

在利用第 2 节中提到的几种方法收集了足够的恶意用户样本之后，通过人工浏览用户主页和微博内容的方式，根据恶意用户的主要恶意行为对用户进行了分类。在所收集的恶意用户中，发现了 3 种主要恶意行为：过度广告、重复转发和过度关注。3 种恶意行为的用户数量如表 2 所示，其中还包括 3 198 个正常用户样本。

表 2 用于特征分析的样本集合

Tab. 2 Spammer samples for feature analysis

主要恶意行为	样本数量/个
过度广告	716
重复转发	710
过度关注	1 000
正常用户	3 198

过度广告指的是频繁发布或转发包含广告信息的微博。这些广告信息大多数与某些产品相关，如保健品、化妆品和服装首饰等，有时也会与一些提供收费服务的网站相关。

重复转发是指用户反复多次转发同样的微博。拥有这种行为的恶意用户通常会以极高的频率转发一个用户的微博，对于一条微博会在短时间内多次转发。

过度关注是指用户关注他人的频率和数量远超过常人。拥有这种行为的恶意用户会主动地关注大量用户，甚至达到关注数量的上限，他们还会频繁地变换关注的对象。与积极的关注行为相对，他们自己微博相关的活动非常少，有的用户完全是一个不活动的账号，不发布微博也不转发微博。

在完成了恶意用户样本收集以后，针对 3 种不同行为的用户进行了数据统计和分析。统计各种不同的行为特征，将恶意用户的特征数据与正常用户的特征数据进行比较，找出两者差别较大的特征，这些特征可以看成恶意行为所带来的具有辨识度的“行为特征”。下文将针对 3 种恶意行为进行特征分析，将特征分为社会关系相关特征、微博发布行为特征和微博内容相关特征。社会关系相关特征主要包括用户的关注、粉丝、好友相关的特征；微博发布行为特征主要指发布微博的频率等；微博内容相关特征指与微博的具体内容相关，但是与微博的文本含义无关的一些统计特征，例如链接数量、图片数量等。

下面将依次分析 3 种恶意行为不同种类的特征，为方便比较和分析，将 3 种恶意用户共有的社会关系特征和微博发布行为特征放到一起进行比较。

3.1 社会关系特征与微博发布特征分析

为找出恶意用户和正常用户在社会活动和微博发布行为上的差异，对每一个样本用户收集以下信息：关注数量、粉丝数量、好友数量、微博数量和微博账户年龄。

利用收集的数据，画出不同分类用户的关注数、粉丝数以及好友数量的累积分布函数（cumulative distribution function，CDF）曲线，如图 1 所示。

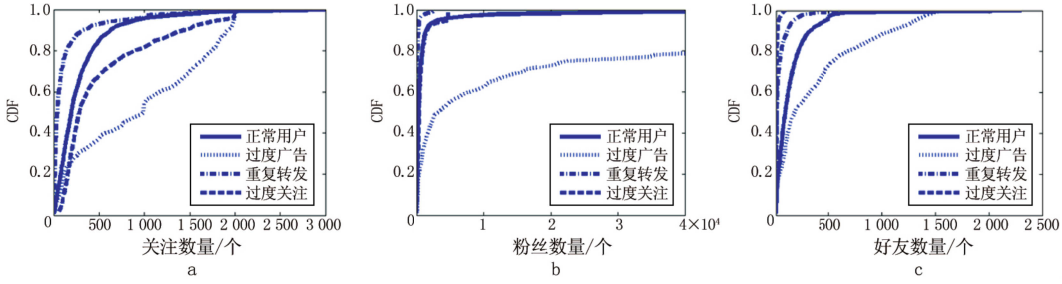

图 1 4 种用户社会关系特征 CDF 曲线比较

Fig. 1 CDF curves of four social features

a—关注数量；b—粉丝数量；c—好友数量

a-Following number; b-Follower number; c-Friend number

可以看出，以过度广告行为为主的恶意用户会关注更多的人，拥有更多的粉丝和好友。50%的过度广告用户关注了超过 1 000 人，而其他 3 类用户中有 80%未达到这个数字。超过 20%的过度广告用户拥有超过 40 000 粉丝，比其他类别的恶意用户和正常用户都高出不少。这意味着过度广告用户会更

加积极主动地建立社会关系，他们通过主动关注他人获取反关注以扩大粉丝网络，使得他们发布的广告能被更多的人看到。

相对而言，以重复转发行为为主的恶意用户则显得不那么活跃。他们专注于转发特定用户的微博，因此不需要关注太多用户也不需要太多的粉丝。所以，他们的好友数也会比较少。

而过度关注用户和前面 2 种恶意用户都不同，他们关注了大量的用户，但是拥有的粉丝和好友却非常少。图 2 为用户的关注-粉丝比例以及好友-粉丝比例的 CDF 曲线，过度关注用户的关注-粉丝比例整体偏高，而好友-粉丝比例则较低。原因可能在于，他们除了关注他人以外，很少发微博，因此也无法吸引其他用户关注他们。

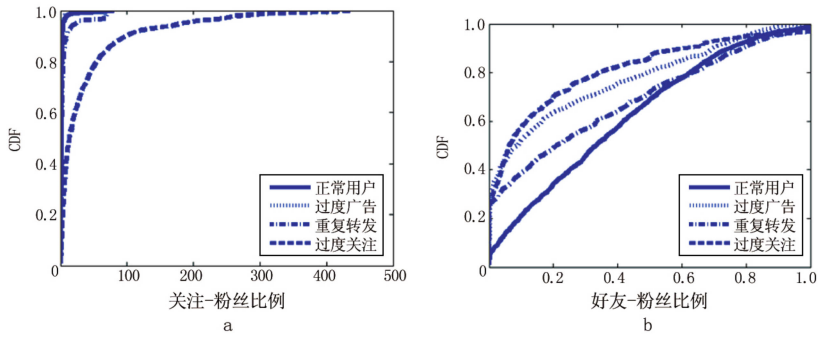

图 2　关注-粉丝和好友-粉丝比例 CDF 曲线比较

Fig. 2　CDF curves of following-follower ratio and friend-follower ratio

a—关注-粉丝比例；b—好友-粉丝比例

a-Following-follower ratio; b-Friend-follower ratio

关注用户发布微博的行为，根据用户的微博总数和账户存活时间，统计用户每日发布微博的平均数量。发现过度广告用户发布微博的频率远高于正常用户，而过度关注用户发布微博的频率则远低于正常用户。重复转发用户的微博发布频率和正常用户在同一水平。

3.2　微博内容相关特征分析

3.2.1　过度广告

一条广告微博中通常包含 URL 链接用以指向广告页面，也会有商品图片来引起用户点击的兴趣。猜测过度广告用户发布的微博更多的会是这种类型，因此选取以下特征，通过分析和比较验证猜想：用户每条微博中平均的 URL 数量、平均每天发布的 URL 数量、包含图片的微博占所有微博的比例以及@符号的数量。

图 3a 和图 3b 分别为每条微博 URL 数量和每日 URL 数量的 CDF 曲线。传播 URL 可能是最常见的广告方法，因此可以看到，无论是每条微博中包含的 URL 平均数还是每天用户发布的 URL 数量，

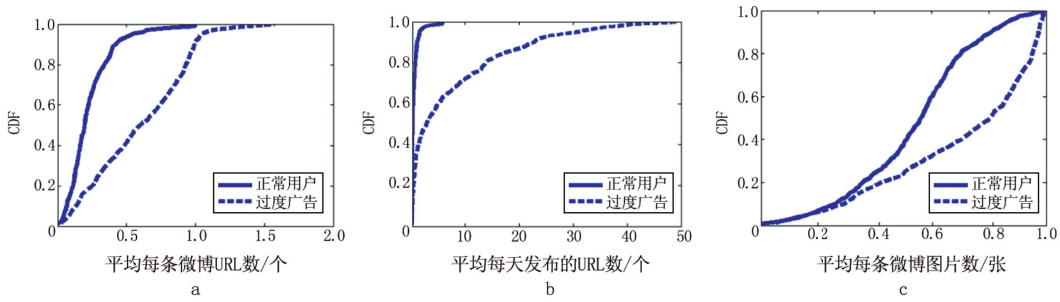

图 3　微博内容特征 CDF 曲线（URL 和图片）

Fig. 3　CDF curves of content features (URL and images)

a—平均每条微博 URL 数；b—平均每天发布的 URL 数；c—平均每条微博图片数

a-Average URLs in each microblog; b-Average URLs posted per day; c-Average images in each microblog

过度广告用户都远超正常用户。此外，由图 3c 也可以看出，过度广告用户发布的图片也更多，这两点都直接证实了之前的猜想。

考虑@符号的数量，统计发现，过度广告恶意用户的微博中@数量整体上要少于正常用户。曾经猜测这些恶意用户会使用@符号将广告信息推送给其他用户，但事实上却并非如此。一般用户使用@符号有 2 种情况：转发别人的微博，以及使用@符号将微博推送给指定用户。过度广告恶意用户更倾向于发布原创的广告微博而不是转发他人。此外，他们也尽量避免使用@符号，因为这会增加他们被其他用户举报的风险。

3.2.2 重复转发

重复转发用户最大的特点就是他们会对同一条微博重复转发多次，而正常用户很少有这种行为。提取并统计出以下行为特征用以比较。

1）重复转发的微博比例：所有转发微博中，来源重复部分的比例。例如，用户转发了微博 A 2 次，微博 B 3 次，那么属于重复转发的次数分别为 1 次和 2 次，比例就是 (1+2)/5.

2）单条微博平均转发次数：用户平均对每一条不同的源微博进行转发次数的均值。

3）单条微博最高转发次数：用户对于一条微博的最高转发次数，上面的例子中为 3.

4）不同微博来源的数量：用户所转发微博的来源数量。例如 B 转发了 A 的微博，C 又转发了 B 的这一条微博，则对 C 而言，微博的作者是 A 而来源却是 B.

5）用户专注度：用户专注度用于表示用户专注于转发特定来源微博的程度，计算来自不同来源的微博占所有转发微博的比例，并且取最高者作为专注度。

重复转发比例的 CDF 曲线如图 4a 所示，大约有 80%的恶意用户拥有 60%以上的重复转发微博，而 50%正常用户的重复转发比例小于 60%. 图 4b 与图 4c 分别为平均转发次数和最高转发次数的 CDF 曲线。可以很明显看出，无论是平均值还是峰值，重复转发恶意用户都超过正常用户。这说明重复转发恶意用户习惯于对一条微博进行多次转发，而正常用户则很少这样。有一些正常用户的重复转发峰值可能会很高，但是他们的重复转发平均值很低，对他们而言这些行为只是偶然现象，而对于重复转发恶意用户来说则是普遍现象。

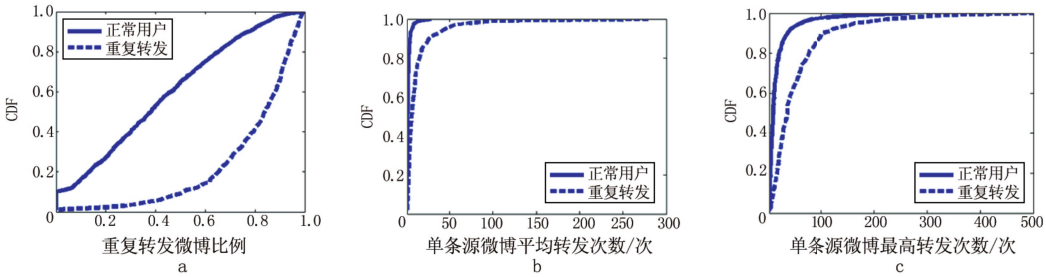

图 4 与转发行为相关的特征 CDF 曲线比较

Fig. 4 CDF curves of features related with reposting

a—重复转发微博比例；b—单条源微博平均转发次数；c—单条源微博最高转发次数

a-Duplicate reposting ratio; b-Average reposting for single source microblog;
c-Max reposting for single source microblog

考虑转发微博来源的有关特征，正常用户会关注并且转发许多不同用户的微博，而重复转发恶意用户却只会转发特定几个甚至一个用户的微博。图 5a 为用户微博来源数量的 CDF 曲线，恶意用户更加倾向于转发某些特定用户的微博，这些用户可能是他们“推广服务”的客户，而正常用户转发微博来源的选择面就广得多。除此之外，恶意用户的用户关注度也比正常用户要高得多，如图 5b 所示。重复转发的恶意用户可能关注了大量的人，但是他们转发的大多数微博却来自于一个人。在实验样本中，发现有大量的恶意用户所“服务”的对象是同一个用户，这说明这些恶意用户可能是受到人为操纵，被用于某些微博营销推广活动中去了。

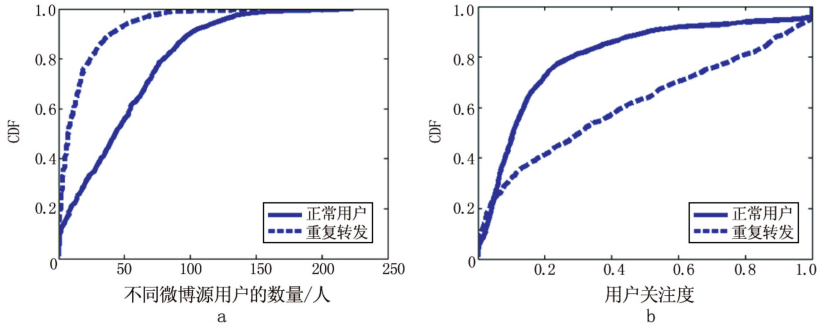

图 5　与转发来源相关的特征 CDF 曲线比较

Fig. 5　CDF curves of features related with reposting sources

a—不同微博源用户的数量；b—用户关注度

a-Number of source users; b-User focusing metric

3.2.3　过度关注

拥有过度关注行为的恶意用户会关注大量的人，但是却缺少其他的微博活动，包括发布微博以及通过微博评论等方式与他人进行互动。这里选取以下特征用以分析和比较：微博被转发的比例、微博被评论的比例、在所有被评论微博下面评论数量的平均值。

如图 6a 和图 6b 所示，正常用户的微博会收到更多的转发和评论。有 40%的过度关注恶意用户的微博从未被转发；60%的微博没有评论。即使是那些有评论的微博，他们也很少回复这些评论。图 6c 展示了用户所有被评论微博下方的评论数量的平均值，这一数值计算的是所有被评论微博下方的评论数量的平均值，如果一个用户的微博从未被评论则该值为零，否则该值至少为 1. 由图 6c 可以看出，只有 20%的恶意用户的平均评论数超过了 1，这说明绝大部分恶意用户不会回复微博下方的评论。而正常用户则不同，他们微博下的评论数量更多，他们也会利用评论功能与其他人进行交流。

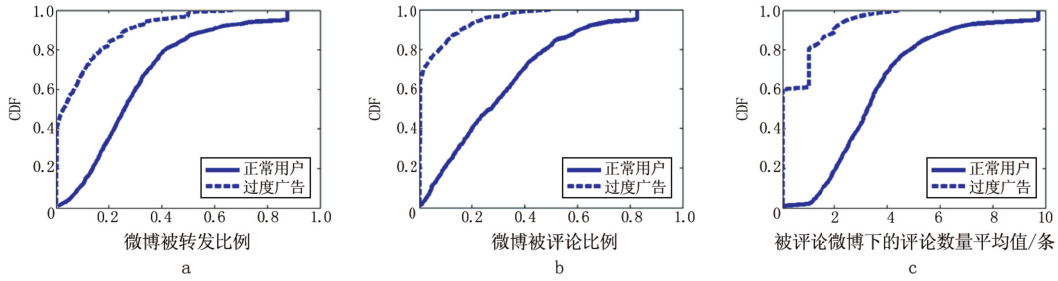

图 6　微博互动相关特征 CDF 曲线

Fig. 6　CDF curves of features related with microblog communication

a—微博被转发比例；b—微博被评论比例；c—被评论微博下的评论数量平均值

a-Ratio of microblogs being reposted; b-Ratio of microblogs being commented; c-Average comments of microblogs being commented

3.3　总结

综上所述，研究了不同恶意行为用户的特点，并且找出了与这些行为相关的一些统计特征。这些特征可以用来区分恶意用户与正常用户，这也是构建自动分类检测系统的基础。

过度广告行为的恶意用户热衷于社交行为。他们拥有更多的关注、粉丝、好友。他们会发布更多包含链接和图片的微博，却很少使用@符号。

重复转发行为的恶意用户粉丝与好友比正常用户少。他们经常重复地转发同一条微博，这些微博的来源也经常只有几个甚至一个用户。

过度关注行为的恶意用户会关注大量的人，但是粉丝和好友则极少。与他们积极的关注行为形成

鲜明对比的是他们完全不热衷于其他活动。他们极少发布或转发微博，也基本上不会回复自己微博下方的评论，这使得他们的微博很少被人转发或评论。

在收集到的样本中，有一些恶意用户呈现出了不止一种恶意行为的特征。例如，有的恶意用户不仅自己发布大量广告微博，同时也会疯狂转发某几个其他用户的广告微博，这种复合的行为模式给检测系统提出了要求。恶意行为的检测需要综合考虑不同的恶意行为，这需要将针对不同恶意行为的分类器整合起来，才能起到更好的检测效果。

4 恶意用户检测

基于上述分析和结论，设计并建立了一套基于恶意行为判定的恶意用户检测系统。该系统可以通过判断用户是否实行了恶意行为（上文所提到的3种典型行为）来判断一个用户是否是恶意用户。

以上文所提到的那些统计信息为特征，针对每一种恶意行为，训练了机器学习自动分类器。3种行为的分类器组合起来形成了检测系统的核心部分。该系统能够自动地收集统计用户的行为特征，通过分类器的判定，输出一个用户是否为恶意用户的结果。该系统可以找出那些有过恶意行为的恶意用户。

4.1 恶意行为分类器构建

利用著名的机器学习工具集Weka训练恶意行为分类器。对于每一种恶意行为都尝试了30多种不同的分类算法，在训练数据集上采取10次交叉验证法测试算法的性能和效率，选取其中表现最优秀的算法作为分类器的算法。

表3展示了作为训练集的样本信息，不同行为的分类器采用的是不同的训练样本集合，每个数据集中的特征均不同。

表3 训练用样本集合

Tab. 3 Dataset for training

训练集	正常用户	恶意用户
过度广告/条	698	716
重复转发/次	1 500	710
过度关注/个	1 000	1 000

过度广告行为：粉丝数、关注数、好友数、好友-粉丝比例、每天微博数、每天URL数、每条微博URL数、每条微博@符号数、含图片微博比例。

重复转发行为：粉丝数、好友数、重复转发的微博比例、单条源微博平均转发次数、单条源微博最高转发次数、不同来源用户数量、用户专注度。

过度关注行为：粉丝数、好友数、关注-粉丝比例、被转发的微博比例、被评论的微博比例、每条被评论微博下方评论的平均数量。

测试结果最优的3种算法如表4所示，最终选取了以下3种算法分别作为3种行为的分类器算法。过度广告——Random Committee；重复转发——AD Tree；过度关注——Random Forest.

表4 恶意行为分类器算法测试结果

Tab. 4 Evaluation results for classification algorithm

恶意行为	算法	准确率	召回率	综合评价指标
过度广告	Random Committee	0.937	0.936	0.936
	Random Forrest	0.934	0.934	0.934
	Decorate	0.924	0.923	0.923
重复转发	AD Tree	0.842	0.843	0.842
	Simple Logistic	0.842	0.842	0.842
	Smo	0.840	0.839	0.840
过度关注	Random Forrest	0.963	0.963	0.963
	Classification via Regression	0.961	0.961	0.961
	Decorate	0.958	0.958	0.958

4.2 检测系统测试

完成分类器和检测系统的建立之后，利用一些真实的用户数据对这个系统进行测试。邀请了一些志愿者帮助收集用于测试的样本，包括恶意用户和正常用户。最终的测试样本集如表5所示。将样本的数据输入检测系统，最后输出判定结果，并且将结果与之前人为标注的结果进行比较，如表 6 和表 7 所示。

表 5 测试用样本集合

Tab. 5 Dataset for system testing

用户标签	数量
正常用户/个	811
过度广告/条	336
重复转发/次	435
过度关注/个	812

表 6 系统测试判定结果

Tab. 6 Output of the identification system

人工标签	数量/个	判定为过度广告/条	判定为重复转发/次	判定为过度关注/个	任何恶意行为/个
正常用户	811	26	0	24	48
过度广告	336	295	5	67	314
重复转发	435	135	279	138	366
过度关注	812	75	18	580	619

表 7 恶意行为分类器的各评估指标

Tab. 7 Evaluation metrics of the classifiers

分类器	真阳性比例	假阳性比例	精确率	准确率	召回率	*F*1 综合指标
过度广告	0.877 9	0.032 1	0.941 5	0.919 0	0.878 0	0.898 0
重复转发	0.641 4	0	0.874 7	1	0.641 4	0.781 5
过度关注	0.714 3	0.029 6	0.842 3	0.960 3	0.714 3	0.819 2

就单一行为检测来说，过度广告的行为检测分类器在真阳性比例和准确率上都排在第一。重复转发行为分类器的真阳性比例虽然较低，但是却没有将任何正常用户误报为恶意用户。再从整体的检测效果来看，一些恶意用户样本虽然被标注为某种行为，但是却被另外一种行为的分类器检测出来。手动确认了这些样本情况，发现他们拥有多种恶意行为。例如一些被标注为“重复转发”行为的恶意用户其实也在不断地发布带有 URL 的广告微博。虽然单一的行为分类器也许无法对这种复杂的恶意用户进行很好的判定，但是综合了不同行为分类器的检测系统则可以较好地发现这种恶意用户。考虑整体的检测效果，系统一共标记了样本集中 82.06%的恶意用户，而只有 5.92%的正常用户被误报。

5 结论

以新浪微博为对象，研究了新浪微博中的恶意行为和恶意用户。通过蜜罐账户、爬虫程序和购买“微博推广服务”等多种手段获取了用于研究分析的恶意用户样本。在这些恶意用户中，发现了3 种主要的恶意行为：过度广告、重复转发和过度关注。提取统计出各种行为特征，通过将恶意用户与正常用户进行比较，找到了恶意用户的特征。

过度广告行为的恶意用户建立社会关系的时候更加积极且成功，他们会发布大量带有 URL 和图片的微博；重复转发行为的恶意用户粉丝和好友较少，他们更加倾向于转发某一特定用户的微博，对于每条微博通常会转发多次；过度关注行为的恶意用户会关注大量的人，自己拥有的粉丝和好友数量却极少，他们很少发布微博也很少被人转发和评论。

利用以上特征，设计并实现了一套自动化监测系统：以机器学习算法为核心，构建针对 3 种不同行为的分类器，将 3 个分类器整合成为一个系统。利用一些真实的用户样本对这一系统进行了测试并且发现系统整体表现令人十分满意。

在测试的过程中，发现有一些特殊的恶意用户的活动策略非常谨慎。例如，一些恶意用户仅在较短的一段时间内实施恶意行为，其他时间都表现得和正常用户一样。对于这种恶意用户，系统并不能够很好地进行识别。而且，此系统的识别是基于某些特征，如果恶意用户针对这些特征有意地进行改变和掩饰，很有可能使系统失效。因此，在未来的工作中，如何增加系统的灵活性和健壮性是一个非

常重要的课题。

[参考文献] (References)

[1] GAO H Y, HU J, WILSON C, et al. Detecting and characterizing social spam campaigns[C]//Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. New York, USA: ACM, 2010: 35-47.

[2] GRIER C, KURT T, PAXSON V, et al. @spam: the underground on 140 characters or less[C]//Proceedings of the 17th ACM Conference on Computer and Communications Security. New York, USA: ACM, 2010: 27-37.

[3] HUANG C R, JIANG Q C, ZHANG Y. Detecting comment spam through content analysis[J]. In Web-Age Information Management, 2010, 6185: 222-233.

[4] YIN D W, XUE Z Z, HONG L J, et al. Detection of harassment on Web 2.0[C]//In Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009. Madrid, Spain: CAW, 2009: 2-9.

[5] IRANI D, WEBB S, PU C. Study of static classification of social spam profiles in myspace[C]//Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010. Washington, DC: ICWSM, 2010: 82-89.

[6] WEBB S, CAVERLEE J, PU C. Social honeypots: making friends with a spammer near you[Z]. Mountain View, CA: CEAS, 2008.

[7] LEE K, CAVERLEE J, WEBB S. Uncovering social spammers: social honeypots + machine learning[C]//In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10). New York, USA: ACM, 2010: 435-442.

[8] WANG D, IRANI D, PU C. A social-spam detection framework[C]//In Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS'11). New York, USA: ACM, 2011: 46-54.

[9] STRINGHINI G, KRUEGEL C, VIGNA G. Detecting spammers on social networks[C]//In Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC'10). New York, USA: ACM, 2010: 1-9.

[10] ZHOU Y, CHEN K, SONG L, et al. Feature analysis of spammers in social networks with active honeypots: a case study of chinese microblogging networks[C]//2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Istanbul, Turkey: IEEE, 2012: 728-729.