基于复合分类模型的社交网络恶意用户识别方法

谈 磊[1,2,3,4] 连一峰[1,3] 陈 恺[1]

[1](中国科学院软件研究所信息安全国家重点实验室 北京 100190)

[2](中国科学院研究生院信息安全国家重点实验室 北京 100049)

[3](信息安全共性技术国家工程研究中心 北京 100190)

[4](信息网络安全公安部重点实验室(公安部第三研究所) 上海 201204)

摘 要 社交网络近年发展迅速,微博类社交网络的用户数目及规模急剧增大的同时也带来了诸多安全问题,为了保护用户的隐私和个人、集体的利益,需要针对这些恶意行为进行识别并对恶意用户进行处理。提出一种采用复合分类模型对用户进行分类的方法,并开发了一个对微博类社交网络用户进行分类的系统。通过研究用户的属性和行为特点,比较属性间的相关性,从两方面兼顾了分类的准确性和效率。

关键词 新浪微博 社交网络 自动分类 特征选择 恶意用户

中图分类号 TP309 **文献标识码** A

DOI: 10.3969/j.issn.1000-386x.2012.12.001

MALICIOUS USERS IDENTIFICATION IN SOCIAL NETWORK BASED ON COMPOSITE CLASSIFICATION MODEL

Tan Lei[1,2,3,4] Lian Yifeng[1,3] Chen Kai[1]

[1](*State Key Laboratory of Information Security, Insititute of Software, Chinese Academy of Science, Beijing 100190, China*)

[2](*State Key Laboratory of Information Security, Graduate University of Chinese Academy of Sciences, Beijing 100049, China*)

[3](*National Engineering Research Center of Information Security, Beijing 100190, China*)

[4](*Key Laboratory of Information Network Security of Public Security (the Third Research Institute of Public Security), Shanghai 201204, China*)

Abstract While having sharp increase in users and network size as in social network of microblogging, the rapid development of social network in recent years also brings lots of security problems. To protect user privacy, personal and collective interest against violations of these security issues, it is necessary to identify malicious behaviours and deal with malicious users. This paper presents a new method for classifying social network users on composite classification model and develops a system to classify users in social network of microblogging. The system analyses many features of the properties and behaviours of users and compares the correlation between the properties, and is able to take the account of both accuracy and efficiency.

Keywords Sina microblogging Social network Automatic classification Feature selection Malicious users

0 引 言

社交网络使得互联网从研究部门、学校、政府、商业应用平台扩展成一个人类社会交流的工具,越来越多的用户开始使用社交网络进行交流。2007 年 5 月饭否上线,是中国第一家引入美国微博概念的网站。新浪微博自 2009 年 8 月 14 日开始内测,8 月 28 日开始对外公测,11 月 2 日用户达 100 万;截至 2010 年 10 月底,新浪微博用户数已达 5000 万,新浪微博用户平均每天发布超过 2500 万条微博内容。2010 年 9 月 9 日,《中国微博元年市场白皮书》的数据显示,新浪微博月覆盖人数约为 4400 万,新浪微博每天产生的微博数超过 300 万,平均每秒有近 40 条微博产生。作为中国用户数最多的微博产品,公众名人用户众社交网络在进一步吸引了用户的同时也带来了一些严重的安全问题。2011 年 6 月,新浪微博出现大范围“中毒”事件,这次攻击像蠕虫一样大规模传播,对社交网络的可用性造成了影响,但没有对用户的账户和隐私信息造成侵害。它暴露出使用短 url 的一些缺点,用户很有可能在不知情的情况下点击短 url 而访问恶意页面,以此可能可以触发另一些更加严重的攻击。社交网络通常会要求用户使用真实资料注册,并在网站上提供了包括身份资料、学校资料、单位资料、账号、联系方式(手机、QQ、MSN 等)在内的大量真实资料及照片。由于类似的潜在安全隐患这些资料被他人所获,将有可能对用户本人造成种种危害,小则受到广告骚扰,大则遭遇诈骗、人身攻击、身份盗用等,甚至遭受网络钓鱼的危害[1]。

收稿日期:2012-03-23。国家自然科学基金项目(61100226);国家高技术研究发展计划项目(2011AA01A023);北京市自然科学基金项目(4122085);公安部三所开放基金课题(C10606)。谈磊,硕士生,主研领域:信息安全测试评估,社交网络安全。连一峰,副研究员。陈恺,博士。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

如果能够提前发现可疑用户或者恶意用户，让社交网络运营商采取措施排除恶意节点标记可疑用户，提示用户拒绝掉可疑用户的好友申请要求，拒绝接收他们发送的私信消息等，能够在一定程度上保护用户的安全和隐私。目前针对用户分类采用的模型多为单一的判别分析模型，如朴素贝叶斯模型[2,3]、k 邻近模型等；这些模型本身存在一定的缺陷性，由于朴素贝叶斯模型基于一个各属性相互独立的假设，在现实生活中一个用户的各项属性之间往往可能存在一定的联系，因此对分类精度会产生影响，而 K 邻近模型虽然精度较好，但是开销非常大。本文提出了一种对这两个模型进行优化的复合模型，在保证分类精度的情况下减少了计算开销，我们研究了两类用户在属性和行为上的特点，并比较了该算法在准确性和计算开销上的优势。本文实现了一个分类系统使用该方法对实际社交网络进行实验，系统包括数据分析处理部分和决策部分，它对用户的众多特点属性进行分析来决定该用户是否是正常用户，我们的实验证明了该分类系统的有效。

1 相关工作

国外对微博类社交网络的研究工作大多是针对 twitter 的研究。研究人员调查了人们为何使用社交网络，例如可以找到一些兴趣和活动的共同爱好者，或者在工作上进行一些非正式的交流；Honeycutt，Herring 和 boyd 等人对 twitter 的一些交流实践都进行了研究；Huberman 等人的研究则在人际关系方面发现 twitter 用户只和一小部分他们的社交关系在社交网络中有交互，而这些结果中并没有讨论社交网络中存在的恶意用户以及恶意行为。

在 twitter 恶意用户检测方面：文献[4]用蜜罐账户的方式采集社交网络用户信息，并从中发掘特征开发了一个检测恶意用户的工具；文献[5]研究了超过 50 万 twitter 用户，并归纳总结了多个属性将 twitter 用户分为 human、bot、cyborg 这 3 类，该分类采用了基于朴素贝叶斯模型的线性判别分析，该方法的基础之一属性间的独立性假设对真实社交网络不一定成立，因此在对真实社交网络进行检测时会影响分类的准确性。而同时文献[6]研究了超过 7 万个 twitter 用户并在对他们推文的方式进行分类的基础上，通过链接结构对用户进行分类，他们在工作的同时还研究了 twitter 的网络属性和地域描述；文献[7]研究了超过 10 万个 twitter 用户，并以 follower-to-following 的比例为依据对用户进行了分类，他们的工作同样发现 twitter 用户在近期的超过线性的增长，但是他们的分类工作更加简单，只基于单个属性的判定的准确性有待提升。

在用户行为研究方面：文献[8]对 twitter 用户行为的研究发现，一旦攻击者获取了用户的信任，用户即使在现实生活对其一无所知的情况下也会点击攻击者发送的消息，这反映了社交网络中恶意节点的巨大危害；文献[9]则发现，如果攻击者窃取到了用户的具体信息，那么针对该用户的钓鱼攻击就有更好的成功率。他们的工作基本是对攻击行为和模式的研究总结，却没有给出具体的应对方法。

本文提出了一种对微博类社交网络用户进行分类的方法，并实现了一个对潜在恶意用户进行检测的工具，第一个对新浪微博的潜在恶意用户进行了检测，收集了大量新浪微博用户基本信息以进行分类；在对用户分类时当前多使用基于朴素贝叶斯模型的线性判别分析，考虑到属性之间的不独立关系，本文在符合相关性的样本子集中使用 k 邻近模型有更高的准确率；而在不符合相关性的样本子集中使用朴素贝叶斯分类算法较 k 邻近模型准确率相差不大而能提高效率；相比单项指标的判别，本文根据用户特征属性提出的多项指标综合判定提高了分类的准确性；实验首先确定存在相关性的用户参数，然后按照是否符合该相关性对样本首先分类，再分别使用不同的分类模型进行分类；在评估过程中对不同情况的两种分类模型的效率和准确率均进行了比较；评估结果证明这样的方法是可行的。

2 用户数据收集分析

在这一章，我们首先说明了用户数据的收集情况并详细描述了我们对有助于进行自动分类的用户属性的观察，然后建立了一个训练集作为实验的基础，再讨论了用于对属性参数进行分析进而对用户进行分类的模型。

2.1 数据收集

我们使用了新浪微博的公共 API 来收集我们需要的用户数据，在 2011 年 6 月 - 8 月的时间里我们共收集了 400,000 条推文及其用户的数据。要完成一个自动分类鉴别的系统，我们还需要一个包含已知正常用户和可疑或恶意用户的训练集。我们从获取的数据中随机的选取了一些不同的样本，然后通过人工检查他们的用户日志和主页的形式对这些样本进行判别。训练集包含 1500 个正常用户的样本和 1500 个恶意用户的样本。和正常用户相区别，大部分恶意用户存在一些共同的特征，如推文大部分含有 URL、followers 数和 friends 和 favourites 数都非常少、大部分最近才注册、推文基本以 Web 发布等。

2.2 参数定义和数据分析

(1) 参数定义

对于一条具体的推文 x 定义其 8 个参数{U,F1,F2,F3,T,S,f,M}

U(x) ∈{0,1}：推文内容中是否含有 URL

F1(x),F2(x),F3(x) ∈ Z：发布推文的用户的 followers，friends，favourites 数

T(x) ∈{2009,2010,2011}：发布推文的用户的注册时间

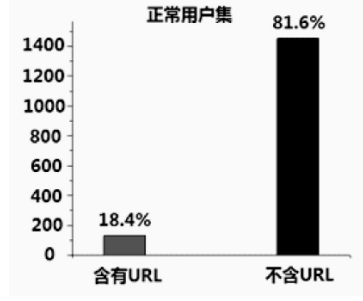
S(x) ∈ Z：发布推文的用户的总发布状态数

f(x) = F1(x)/F2(x)：用户的 F/F 率，用户的 followers 个数/friends 个数比例

M(x) ∈{0,1}：该用户是否属于恶意用户

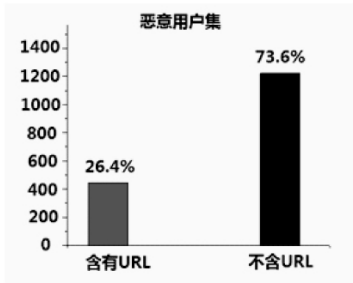
(2) 推文 URL 分析 U(x)

恶意用户要通过推文传递恶意信息的话很大程度上需要依赖推文中附带的指向恶意链接的 URL，但是正常用户也有发布链接分享页面的需求，我们检查了测试集得到了以下结果图图 1(a) 和图 1(b)。



(a) 正常用户

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

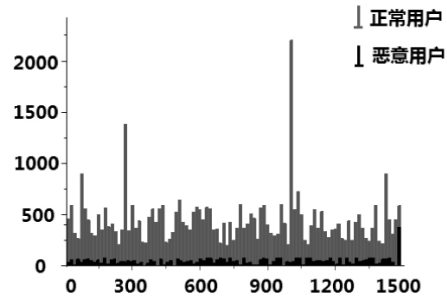


(b) 恶意用户

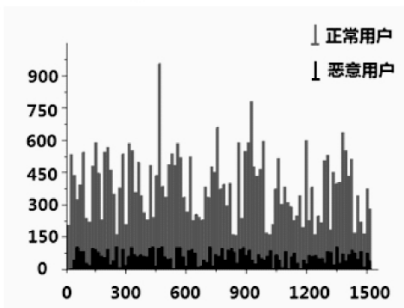
图 1　推文 URL 率比较图

(3) followers, friends, favourite 数分析 F1(x), F2(x), F3(x)

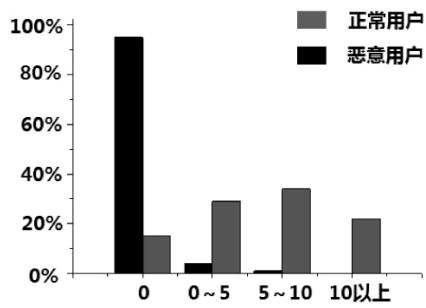
这 3 个属性基本反映了用户的交际和喜好状况,从图 2 反映的 2 类用户状况来看,恶意用户的交际范围和喜好比正常用户要少的多,很多恶意用户基本没有发布推文之外的交际操作。



(a) followers个数比较图



(b) friends个数比较图



(c) favourites个数比较图

图　2

(4) 注册时间分析 T(x)

虽然在新浪微博中不存在长时间不使用微博就会被注销的情况,但是如图 3 所示我们的研究发现发布恶意信息的用户大多是新近注册的用户而非更早注册的用户,试图发布恶意信息的攻击者通常会注册许多账户进行活动,而且不会在集中的一个账户中发布太多恶意信息,更倾向于频繁更换账户发布恶意信息。

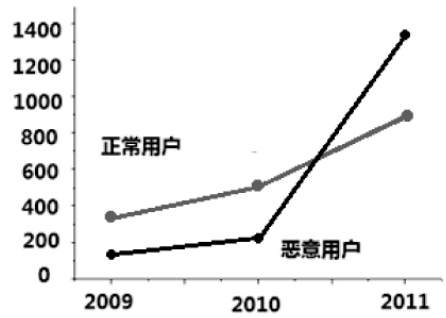


图 3　注册时间分布比较图

(5) statuses 数分析 S(x)

用户可以随时发布不超过 140 字的消息,数据会被记录在状态数中,已发布的状态数这个属性一定程度反映了用户的活跃程度,如图 4 所示,一个更新频繁的用户通常会经常发布推文分享消息和心情等等,因此反映出的状态数会比恶意用户高。

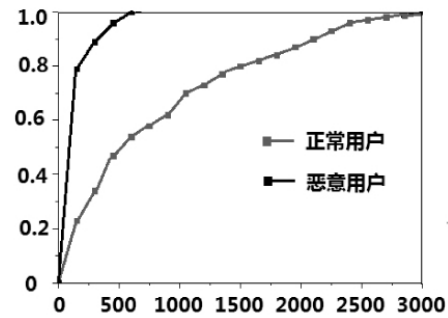


图 4　状态数 CDF 比较图

(6) F/F 率分析 F(x)

F/F 率指的是用户的 followers 个数/friends 个数,这个比例反映的是用户 follow 他人和被他人 follow 的数量比例,很多恶意用户 follow 了他人但是没有多少用户去 follow 该恶意用户,因此图 5 反映出来的恶意用户的 F/F 比例会较正常用户高出许多。

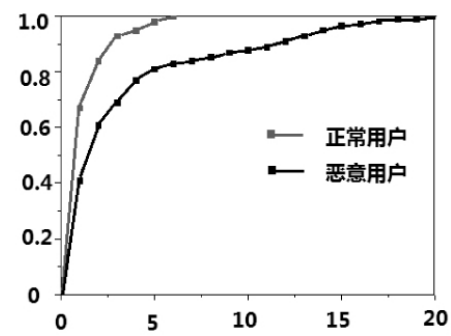


图 5　F/F 率 CDF 比较图

(7) 发文方式分析

用户可以通过网页、WAP 页面和手机短信、彩信发布消息或上传图片,此外还可通过 API 用第三方软件或插件发布信息。大部分用户采取的是网页访问新浪微博的方式,在正常用户中也很有一部分使用手机发送微博,而在恶意用户中使用 Web 方式的攻击者占绝大多数(如表 1 所示)。

表 1　发文方式比较

正常用户			恶意用户	
网页	手机	其他	网页	其他
65.3%	27.2%	14.5%	89.2%	10.8%

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

3 恶意用户识别

3.1 相关性计算

这里首先我们对测试集的用户账户的以下几个参数进行相关性计算: followers 数 F1(x),friends 数 F2(x),favourites 数 F3(x),statuses 数 S(x)。每 2 个参数的相关系数记为 R(X,Y),经典概率论相关系数计算公式如下:

$$R_{X,Y}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2} \tag{1}$$

由于数据本身的原因我们选取的是测试集中 1500 名正常用户的 4 项属性进行的计算,经过计算我们得到了这四个参数的相关性。

表 2 中可以看出最具有相关性(最接近 1)的一对属性是 F1(x)和 F2(x),即 followers 数和 friends 数,则我们用单个样本的 F/F 率 F(x)来判断是否符合总体的趋势,进而选择不同分类算法。

表 2 相关系数计算

参数	F1(x)	F2(x)	F3(x)	S(x)
F1(x)	1			
F2(x)	0.8982	1		
F3(x)	-0.2081	0.1296	1	
S(x)	0.5826	0.4963	0.0281	1

3.2 分类方法选择

目前针对社交网络用户的分类算法有朴素贝叶斯、KNN、决策树、向量空间模型、神经网络等,根据被测属性是否符合相关性选择朴素贝叶斯和 KNN 这两种不同的算法进行分类。朴素贝叶斯分类算法中的类条件独立性假设在许多实际问题中并不一定成立,这一点将会引起分类的误差;而 KNN 算法的不足之处是计算量较大,因为对每一个待分类的样本都要计算它到全体已知样本的距离,才能求得它的 K 个最近邻点。从相关性的角度考虑,我们首先把样本分为两类,如果该样本的具体属性符合属性间的相关性,那么我们使用 KNN 分类算法来对这个子集进行分类,不符合的子集则使用朴素贝叶斯算法进行分类。

具体步骤如下:

输入:训练集 $S(X1,X2,\cdots,Xn)$ 和其中一个样本 X

输出:样本所属的类别 $M(X)$

Step1 计算平均 F/F 率

$$F(S)=\frac{1}{n}\sum_{i=1}^{n}F(Xi) \tag{2}$$

和样本的 $F(X)$。

Step2 若 $0.5F(S)<F(X)<1.5F(S)$,使用 KNN 算法对 X 进行分类

$Xi(U(Xi),F1(Xi),F2(Xi),F3(Xi),S(Xi))$ 和 $X(U(X),F1(X),F2(X),F3(X),S(X))$

参数的距离:

$$D(Xi,X)=\sqrt{U(Xi,X)+\sum_{1}^{3}\left(\frac{Fj(Xi)+Fj(X)}{Fj(Xi)-Fj(X)}\right)^2-\left(\frac{S(Xi)+S(X)}{S(Xi)-S(X)}\right)^2} \tag{3}$$

$$U(Xi,X)=(U(Xi)-U(X))^2 \tag{4}$$

对 $i=1,2,\cdots,n$ 计算 $D(Xi,X)$,找到

$\mathrm{MIN}_{0<i<n+1}D(Xi,X)$ 的 $X\min,M(X)=M(X\min)$ (5)

Step3 否则使用朴素贝叶斯算法对 X 进行分类

先计算训练集中的恶意用户的比例 $P(M(x)=1)$ 和 $P(M(x)=0)$,对每个参数 C(包括 $U,F1,F2,F3,S$)先行划分两类用户的范围 $C1$ 和 $C2$,计算每个参数的 4 个条件概率:

$P(X\in C1\mid M(X)=1)$,$P(X\in C1\mid M(X)=0)$

$P(X\in C2\mid M(X)=1)$,$P(X\in C2\mid M(X)=0)$

计算后验概率:

$$P0=P(M(X)=0)P(X\mid M(X)=0)=P(M(X)=0)\times\prod_{k\text{为每个参数所属}}P(X\in Ck\mid M(X))=0) \tag{6}$$

$$P1=P(M(X)=1)P(X\mid M(X)=1)=P(M(X)=1)\times\prod_{k\text{为每个参数所属}}P(X\in Ck\mid M(X))=1) \tag{7}$$

若 $P0>P1$ 则 $M(X)=0$ 反之 $M(X)=1$。

Step4 返回类别算法结束

我们从训练集实验结果的准确性和时间消耗两方面验证这样的分类方法,从样本分类开始向全部 KNN 和全部 Bayes 开始分成几个部分测试准确性和时间如表 3 所示。

表 3 不同比例分类效率准确性对比表

参数	分组 1	分组 2	分组 3	分组 4	分组 5
准确性 p(%)	96.51	95.82	95.66	95.50	95.48
时间 t(ms)	306	291	280	272	259
参数	分组 6	分组 7	分组 8	分组 9	
准确性 p(%)	93.86	92.34	91.93	90.26	
时间 t(ms)	213	167	102	55	

其中第 1 号为全部使用 KNN 分类器,第 10 号为全部使用 Bayes 分类器,第 5 号为按照相关性分类分别使用两类分类器,中间则均匀调整使用两类分类器的样本数。我们对这一组参数使用最小二乘法进行线性拟合,得到的结果为 $p=88.9751+0.0239t$。

我们希望通过使用复合的分类模型来提升至少 10% 的程序运行时间并达到至少 95% 的准确率,将这两个限制条件带入之后我们得到 p 和 t 的取值范围在 $95.00<p<95.55$ 和 $252<t<275$。那么按照样本的具体属性是否符合属性间的相关性来分类的结果(第 5 组)是符合我们的期望的,加快了 15.3% 的程序运行时间并达到了 95.48% 的准确率。

由此可见在不符合相关性的样本子集中采用朴素贝叶斯算法提高效率,在符合相关性的样本子集中采用 KNN 算法提高精度是可行的;Chu 等人单一的使用基于朴素贝叶斯模型的线性判别分析将 twitter 用户分为 3 类,实验训练集 3000 样本的分类评估的准确性分别为 94.90%、82.80%、93.70% [5]。

4 评 估

4.1 准确性测试

我们按 Bayes 分类算法和 KNN 分类算法编写了分类器,在比较是否符合相关性之后分别对训练集进行了测试,测试的结果如下。对 3000 组数据以 10-折交叉验证计算的 10 折误检率 CDF 图如图 7 所示,10 折的大部分结果均显示使用复合模型误差能控制在 5% 左右。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

我们对测试集的以下两种情况进行了比较: 不符合相关性的样本子集和符合相关性的样本子集中分别使用者两种算法并比较分类结果的误差,结果如表 4 所示。

表 4　不同情况下分类算法对比

子集	不符合相关性的样本子集 10 折误检率				
KNN	2.17%	2.95%	3.71%	3.55%	2.72%
	4.29%	2.68%	3.27%	3.90%	3.49%
Bayes	6.35%	4.38%	4.49%	5.91%	4.17%
	4.11%	5.23%	4.70%	5.32%	5.46%
子集	符合相关性的样本子集 10 折误检率				
KNN	3.92%	3.78%	3.21%	2.17%	2.85%
	3.14%	2.69%	3.82%	4.29%	3.66%
Bayes	8.61%	10.21%	9.58%	8.35%	10.65%
	9.33%	11.05%	9.78%	9.19%	8.71%

从表 5 可知在不符合相关性的样本子集中采用朴素贝叶斯算法提高效率,在符合相关性的样本子集中采用 KNN 算法提高精度是可行的; 文献 [5] 单一的使用基于朴素贝叶斯模型的线性判别分析将 twitter 用户分为 3 类,实验训练集 3000 样本的分类评估的准确性分别为 94.90%、82.80%、93.70%。而我们在实验中得到的结果达到了 4.52% 的平均误检率,可见相对于单一朴素贝叶斯模型,复合模型在准确率上更为优秀,而相对单一 KNN 模型在效率上也得到了优化。我们使用这一模型对整个数据集进行了分类,在总数据集中,400000 用户中共包括32514 名恶意用户,占总人数的 8.13%。

表 5　实验结果比较

社交网络	分类方法	训练集大小	分类结果	平均误检率
twitter	复合模型	3000	32.3%	5.12%
twitter	Bayes	3000	37.5%	9.53%
Sina	复合模型	3000	8.13%	4.52%

在符合相关性的子集部分,由于属性独立性的失效采取 KNN 分类算法在准确性上能够提高; 在不符合相关性的子集部分采用朴素贝叶斯算法则兼顾了效率。但是对其他一些相关性不高的属性在 KNN 分类算法中是否需要对距离加权仍然值得研究,而且这样的方法在对恶意用户的检测上也很难再将恶意用户划分出 bot 或者 cyborg 等类别来,还需要对用户的行为和属性进行更精细的划分。

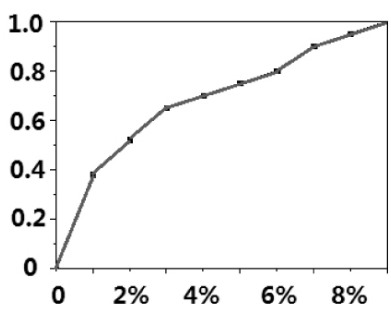


图 6　误检率 CDF 图

4.2　其他社交网络验证

我们使用同样的方法对 twitter 社交网络进行了验证: 我们共收集了 20,000 条推文及其用户的信息,按 3000 名用户的训练集分析了推文信息、发文方式、followers 个数、friends 个数、statuses 个数、爱好个数、创建时间等参数,采用复合模型进行分类的结果显示 32.3% 的用户属于恶意用户。我们检测到的恶意用户的结果与文献 [5] 的实验结果显示 twitter 中 bot 的数量 37.5% 的结果也相近,但分析来看,我们的结果准确率更高。

5　结　语

本文提出的这种对微博类社交网络用户进行分类的系统,将用户分为正常用户和可疑用户两类; 并研究了这两类用户在属性和行为上的特点,根据用户属性的相关性选择不同的分类算法进行分类以在保证高准确性的情况下兼顾效率。该系统包括数据分析处理部分和决策部分,它对用户的众多特点属性进行分析来决定该用户是否是正常用户,我们在新浪微博对真实用户数据的实验证明了该分类系统的有效性。

今后的研究工作包括进一步完善该系统,研究基于其他模型的分类算法,然后开发其在其他类似社交网络的应用,如 facebook、youtube 等,并从更加详细的区分恶意用户的角度考虑用户的行为模式和改进分类模型的方法,同时针对社交网络模型和用户隐私保护进行下一步的研究。

参　考　文　献

[1] Zhao Dejin, Mary Beth Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work [C] // Proceedings of the ACM 2009 International Conference on Supporting Group Work, Sanibel Island, FL, USA, 2009.

[2] Yardi S, Romero D, Schoenebeck G, et al. Detecting spam in a twitter network [J]. First Monday, 2010, 15(1).

[3] Steven Gianvecchio, Xie Mengjun, Wu Zhenyu, et al. Measurement and classification of humans and bots in internet chat [C] //Proceedings of the 17th USENIX Security symposium, San Jose, CA, 2008.

[4] Gianluca Stringhini, Christopher Kruegel, Giovanni Vigna. Detecting Spammers on Social Networks [C] //ACSAC '10 Dec. 6-10, 2010, Austin, Texas USA.

[5] Chu Zi, Steven Gianvecchio, Wang Haining. Who is Tweeting on Twitter: Human, Bot, or Cyborg? [C] //ACSAC '10 Dec. 6-10, 2010, Austin, Texas USA.

[6] Akshay Java, Song Xiaodan, Tim Finin, et al. Why we twitter: understanding microblogging usage and communities [C] //Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, San Jose, CA, USA, 2007.

[7] Krishnamurthy B, Gill P, Aritt M. A few chirps about twitter [C] // USENIX Workshop on Online Social Networks, 2008.

[8] Bilge L, Strufe T, Balzarotti D, et al. All your contacts are belong to us: Automated identity theft attacks on social networks [C] //World Wide Web Conference, 2009.

[9] Jagatic T N, Johnson N A, Jakobsson M, et al. Social phishing. Comm [J]. ACM, 2007, 50(10): 94-100.

[10] Harris Interactive Public Relations Research. A study of social networks scams [R]. 2008.

[11] Baltazar J, Costoya J, Flores R. Koobface: The largest web 2.0 botnet explained [M]. 2009.

[12] Jeff Yan. Bot, cyborg and automated turing test [C] //Proceedings of the 14th International Workshop on Security Protocols, Cambridge, UK, March 2006.

(下转第 17 页)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

也是一个与 P 有着相同方法的合法的 Java 程序,且 P' 中所有方法与 P 中对应的所有方法路径等价。

假设给定对字段 f 的任意的访问语句 f = e 和引用语句 x = r(f) ,其中 e 为一个表达式,x 为一个变量,r(f) 为一个包含 f 的表达式,f 的类型为 type。自封装字段产生的访问函数 type GetF( ) 可以返回 f 的值,SetF(type t) 可以通过传参对 f 进行赋值。故重构后的访问语句变为 SetF(e) ,引用语句变为 x = r(GetF( )) 。设重构前的方法 m 对应于重构后的方法 m' 。现在考虑任意方法 m 的执行路径,要证明 m' 对 m 是路径等价的,可以分类讨论如下:

① 如果 m 中并没有对 f 的任何访问或引用语句,那么显然 m 和 m' 是路径等价的。

② 如果 m 中包含了对 f 的访问语句 f = e,则 m' 中使用 SetF(e) 代替该访问语句,由于访问语句是独立于其他读写操作的,故并不会改变数据流依赖,在控制流上即在路径上增加了对 SetF( ) 的调用,然后再在访问函数中调用 f = e 语句,即并不会改变执行路径和结果,从而 m' 对 m 是路径等价的。

③ 如果 m 中包含了对 f 的引用语句 x = r(f) ,则 m' 中使用 x = r(GetF( )) 代替该访问语句,同样由于引用语句是独立于其他读写操作的,故并不会改变数据流依赖,在控制流上即在路径上增加了对 GetF( ) 的调用,然后再在引用函数中调用 x = r(f) 语句,也不会改变执行路径和结果,从而 m' 对 m 是路径等价的。

④ 如果 m 中包含了同时对 f 引用和访问的语句,即 f = r(f) ,则 m' 中使用 SetF(r(GetF( ))) 代替,m 中对 f 的访问是数据流依赖于 r(f) 中的 f 的,而 m' 中根据函数嵌套调用的先后顺序也是 SetF( ) 数据流依赖于 GetF( ) 的,故也不会改变数据流依赖,在控制流上情况同②和③,从而 m' 仍然是对 m 路径等价的。

反之,关于 m 对 m' 的路径等价可以用相似的讨论得出,这里省略之。至此便证明了自封装字段的正确性。

4　结　语

本文分析了基于分解的重构验证方法体系,从建立正确性标准、重构方法和目标语言的分解、验证和证明等方面给出了详细分析,并以具体的重构方法"搬移字段"为例进行验证,按照方法分解、方法规约、方法验证的步骤逐一对重构过程进行了分析,在完善重构验证这一研究方向作出了一定贡献。

本文的研究内容主要停留在理论分析和验证上,还未将验证好的重构加以实现,以生成可验证的自动化的重构工具。同时,由于重构方法种类繁多,仍然有很多方法需要得到分析和验证,因此对其他方法给出详细的验证过程仍然需要进一步的探索和研究。

参　考　文　献

[1] Max Schäfer, Torbjörn Ekman, Oege de Moor. Challenge proposal: verification of refactorings [C] //PLPV'09, January 20, 2009, Savannah, Georgia, USA.

[2] Torbjörn Ekman, Ran Ettinger, Max Schäfer, et al. Refactoring bugs [R/OL]. 2008. http://progtools.comlab.ox.ac.uk/refact oring/bugreports.

[3] Nik Sultana. Verification of refactorings in isabelle/hol [D]. University of Kent, 2008.

[4] H Christian Estler, Thomas Ruhroth, Heike Wehrheim. Modelchecking Correctness of Refactorings-Some Experiments [C]. Electronic Notes in Theoretical Computer Science, 2007, 187: 3-17.

[5] Gérard P Huet. The Zipper [J]. J. Funct. Program, 1997, 7(5): 549-554.

[6] Atsushi Igarashi, Benjamin C Pierce. On inner classes [J]. Information and Computation, 202, 177(1): 56-89.

[7] Sorin Lerner, Todd Millstein, Erika Rice, et al. Automated soundness proofs for dataflow analyses and transformations via local rules [J]. SIGPLAN Not, 2005, 40(1): 364-377.

[8] Xavier Leroy. Formal certification of a compiler back-end [C] //POPL, 2006: 42-54.

[9] Max Schäfer. Specification, Implementation and Verification of Refactorings [D]. Wolfson College, 2010.

[10] Max Schäfer, Torbjörn Ekman, Oege de Moor. Formalising and Verifying Reference Attribute Grammars in Coq [OL]. 2008. http://progtools.comlab.ox.ac.uk/projects/refactoring/formalising-rags.

[11] Alejandra Garrido, José Meseguer. Formal Specification and Verification of Java Refactorings [C] //SCAM, 2006.

[12] Max Schäfer, Oege de Moor. Specifying and Implementing Refactorings [C] //OOPSLA/SPLASH'10, October 17-21, 2010, Reno/Tahoe, Nevada, USA.

[13] Martin Fowler. 重构: 改善既有代码的设计 [M]. 熊节, 译. 人民邮电出版社, 2010.

(上接第 5 页)

[13] Il-Chul Moon Dongwoo Kim, Yohan Jo, Alice Oh. Analysis of twitter lists as a potential source for discovering latent characteristics of users [C] //To appear on CHI 2010 Workshop on Microblogging: What and How Can We Learn From It?, 2010.

[14] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, et al. Measurement and analysis of online social networks [C] //Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, CA, USA, 2007.

[15] Lewis D D. Naive (Bayes) at forty: the independence assumption in information retrieval [C] //The l0th European Conference on Machine Learning, New York: Springer, 1998: 4-15.

[16] Mladenic D, Grobelnik M. Feature Selection for Unbalanced Class Distridution and Naive Bates [C] //Prceedingos the Sixteenth International Conference on Machine Learning, 1999: 258-267.

[17] Xie Z, Hsu w, Liu Z, et al. SNNB: a selective neighborhood based naive bayes for lazy learning [C] //Proceedings of the Sixth Pacific Asia Conference on KDD, 2002: 104-114.

[18] Tan S. An efective refinement strategy for KNN text classifier [J]. Expert Systems with Applications, 2006, 30(2): 290-298.

[19] Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter [C] //Paper to be presented at the Hawaii International Conference on System Sciences. HICSS-43, 6 January 2010.

[20] Huberman B A, Romero D M, Wu F. Social networks that matter: Twitter under the microscope [C]. 2009, First Monday, 14.

[21] Honeycutt C, Herring S C. Beyond microblogging: Conversation and collaboration via Twitter [C] //Proceedings of the 42nd Hawaii International Conference on System Sciences. HICSS-42, 2008.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net