北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.048

基于随机森林分类的微博机器用户识别研究

刘勘[1,†]　袁蕴英[1]　刘萍[2]

1. 中南财经政法大学信息与安全工程学院，武汉　430074; 2. 武汉大学信息管理学院，武汉　430072;
† E-mail: liukan@znufe.edu.cn

摘要　针对网络上机器用户大量散布谣言，发布虚假信息，误导网民舆论，严重影响网络环境的问题，以微博中的机器用户为研究对象，结合其自动化程度高、伪装能力强、信息发布有针对性的特点，从行为模式、微博内容、用户关系和发布平台 4 个维度分析机器用户的特征指标，利用信息熵、内容重复率等 8 个指标构建微博用户的特征向量，通过随机森林算法设计微博中机器用户的识别模型。最后，在真实的新浪微博数据集上进行验证，结果表明本模型识别机器用户的准确度达到 96.7%，可以有效地区分微博中的机器用户和普通用户。

关键词　机器用户；微博；随机森林
中图分类号　TP391

A Weibo Bot-users Indentification Model Based on Random Forest

LIU Kan[1,†], YUAN Yunying[1], LIU Ping[2]

1. School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430074; 2. School of Information Management, Wuhan University, Wuhan 430072; † E-mail: liukan@znufe.edu.cn

Abstract　Bot-users spread rumors or fake information widely, misleading the public opinion, seriously affecting the normal network environment. Weibo Bot-users are the main focus. Considering their high-level automation, strong disguise power and targeted ability to release, a four-dimensional characteristic index of information entropy, content repetition rate etc is proposed to construct a feature vector and a identification model based on Random Forest algorithm is designed to recognize the bot-users. Finally, the Sina Weibo set are used to verify the efficiency and effectiveness of the model, with the accuracy of 96.7%. The result shows that the model will undoubtedly be good at distinguishing the Bot-users from the ordinary ones.

Key words　Bot-users; Weibo; RandomForest

2007 年以来，微博依靠其简单便捷的操作方式、及时高效的互动模式以及低门槛的平台设计等优势在全球得到迅猛发展，已经成为最受欢迎的网络社交平台之一[1−2]。但是，也有一些不法分子借助微博平台发布虚假信息，大量散布谣言，误导网民舆论，严重地破坏了网络环境。网上招募大量廉价的水军，曾是这些虚假信息制造者常用的手段。随着互联网技术的升级，他们不满足于这种低效的方式，转而使用比如微博批量发布助手、微博自动广播器等自动化软件来操控账户，大大提高了虚假信息和谣言发布和传播的效率，更加强烈地冲击着网络的正常秩序。

所谓机器用户，指那些被第三方软件平台操控，可以自动地进行发布、转发、评论等行为，帮助操纵者达到扩散虚假信息、宣传垃圾广告等目的的账户。机器用户由于受自动化软件控制，可以通过设置时间间隔，筛选特定性别与区域的账户，调用特定内容库等功能，提高推送的精度和广度。另外，

国家自然科学基金(71203164)和国家社会科学基金(14BXW033)资助
收稿日期: 2014−07−27; 修回日期: 2014−10−14; 网络出版时间: 2014-11-28 15:20

由于软件设计的灵活性, 机器用户使用的语言不再是简单词汇的重复, 而是带有较强的主观性、原创性与独特性。参与转发的内容也不再局限于某一方面, 而是趋于多样化。这些特点使机器用户的仿真度较高, 极易被人误认为是普通用户, 增加对其信息的认可度。因此, 机器用户的识别是一个紧迫而困难的工作, 有效地识别微博中的机器用户, 对减少网络谣言的传播、净化网络环境有着积极的意义。

1 相关研究

早期的研究多是针对垃圾用户开展的。垃圾用户也是发布和传播虚假信息和谣言的用户, 但多为人工产生, 通过网络水军来散播信息和制造舆论。前几年网络水军较为活跃, 因此, 不少研究专注于对网络水军或垃圾用户的识别。Yardi 等[3]通过追踪一个 Twitter 话题的发展历程, 利用 URLs 信息、用户名字的规律性等识别出 Twitter 中的垃圾用户。Stringhini 等[4]通过建立 Twitter 用户的行为分析模型, 成功地将垃圾账户与正常账户区分开。Thomas 等[5]详细分析了垃圾账户的行为规律、生命周期和网页特征。Zhang 等[6]分析了 Twitter 上利用 URL 的推广活动, 建立了基于 URL 的垃圾用户识别算法。Lee 等[7]利用 Twitter 用户的行为特点、关注与被关注网等信息区分垃圾用户。Yang 等[8]设计了基于图像、邻居等的垃圾用户识别算法。赵斌等[9]考虑了微博消息序列中的文本相关性和时间相关性, 设计出基于重用检测模型的垃圾用户检测算法, 该算法包括语句级检测和词项级检测两个方法。郭浩等[10]通过提取用户个人信息、用户间关系等多种类别特征, 训练了一个垃圾互粉用户和正常用户的分类系统, 该系统识别垃圾互粉用户的有效性达到 80%以上。丁兆云等[11]针对微博中关注网络的有向特性, 给出有向网络中局部三角形数量统计算法 SirTriangleC, 提出基于统计特征与双向投票的算法 AttriBiVote。Shen 等[12]将微博中的垃圾信息分为新闻和广告两大类, 然后主要利用微博中的文本信息进行识别。

机器用户区别于一般水军操作的垃圾用户, 其主要特点是自动化的行为模式以及良好的伪装能力, 所以, 原有的垃圾用户识别算法无法准确识别出机器用户。机器用户出现的时间并不长, 对它的研究还不多, 只有一些学者针对 Twitter 展开了一些初步研究。如 Zhang 等[13]构建了一个基于每条 Twitter 发布时间的检测机器用户方法, 并用此模型得到 Twitter 中大约有 16%的活跃账户具有较高自动化行为。Chu 等[14]从用户行为、Twitter 内容和账户属性 3 个方面建立分类系统, 将 Twitter 中的用户分成机器用户、人类用户和半机器用户。Amleshwaram 等[15]利用设计特征描述构造了机器用户识别模型 CATS, 并证明该模型在数据很少的时候也能取得较好的效果。Wang[16]提取 3 个基于图模型的 Twitter 账户特征和 3 个基于 Twitter 内容的属性并设计算法, 识别出 Twitter 中的机器账户。

机器用户的特点使得信息能以更快的速度, 在更大的范围, 更有目的性地扩散, 使普通用户难以辨别信息的真实性, 给社会带来的危害远远大于垃圾用户。本文针对于国内微博的特点, 研究微博在内容、行为、关系、平台等多个角度所蕴含的有效维度特征, 提取各个维度的特征值并建立相应的分类模型来实现微博中机器用户的识别。

2 基本思路

国外的 Twitter 和国内的微博虽然同属于微博/博客网站, 但两者在形式、分享内容、用户行为等方面存在一定区别。比如惠普实验室公布的研究报告[17]说明: 国内微博的内容多样, 可以包含图片、视频和链接, 而 Twitter 上发布的内容只包含文字和链接; 国内微博中的热门话题大多是娱乐类的, Twitter 的热门话题则主要与新闻相关; 国内微博的转发频次远远高于 Twitter。

为了更好地了解机器用户的传播特点以及控制这些机器用户的自动化软件的推广原理, 我们特意申请了微博账户, 提供给某自动化软件公司作为机器用户。在两天时间内, 该账号自动发布了 11 条微博, 转发 23 条微博, 并关注 144 个微博账号。同时, 我们还从该公司获得部分机器用户的账号, 从这些账号提取的数据将被用于本文后面的实验中。另外, 我们联系了该自动化软件的销售人员, 亲身体验了该软件的批量发布功能。在指定的时间点, 我们要求的 10 条微博成功在指定类型的机器账户上发布。通过深入观察和分析, 发现不同的微博用户在行为模式、微博内容、用户关系和发布平台方面存在区别, 因此本文对这 4 个维度的特征进行深入分析。主要思路可以分为两部分: 首先, 提取每

位微博用户的 8 个特征值, 分别为描述行为特征的条件信息熵、表示内容特征的内容重复度、代表关系特征的账户关注度、互粉比例、@比例、评论比例、发私信率以及表达平台特征的平台个数; 然后, 利用微博用户和他们的特征值向量, 构建基于随机森林的分类算法, 通过有监督的训练得到用于区分机器用户和普通用户的识别模型。整个过程的总体框架如图 1 所示。

3 主要特征

3.1 行为特征

微博用户在微博上的主要行为是发布信息。机器用户发布微博的时间间隔是根据客户的需求进行设置的, 所以机器用户发微博的时间可能存在一定的规律。Shannon[18]提出的信息熵可以用来衡量离散随机事件的出现概率。Porta 等[19]和 Rosipal[20]都曾把信息熵运用到过程的复杂程度检测中。

本文将微博用户发微博的时间间隔 X_i 作为离散的随机变量, 每位用户的所有时间间隔构成一个随机序列 $X=\{X_i\}$, X_i 表示第 i 条微博和第 $i-1$ 条微博之间的时间间隔, 其熵定义为自信息量的数学期望, 记为

$$H(X_1,...,X_m)=E[I(x)]=-\sum_{X_1,...,X_m}P(x_1,...,x_m)\log P(x_1,...,x_m), \tag{1}$$

其中 $P(x_1, \ldots, x_m)$是 $P(X_1=x_1, \ldots, x_m)$的联合概率。当已知该序列的前 $m-1$ 项时, 其条件信息熵可以表示为

$$\mathrm{CE}(X_m/X_{m-1})=H(X_m/X_1,...,X_{m-1})=H(X_1,...,X_m)-H(X_1,...,X_{m-1})。 \tag{2}$$

用户发微博的时间间隔构成的序列都是有限序列, 而信息熵衡量的是一个无穷随机过程, 无法直接用来计算有限的序列。参考 Gianvecchio 等[21]的观点, 引入修正的条件信息熵来解决序列有限性所带来的问题。修正的条件信息熵的公式如下:

$$\mathrm{CCE}_m=\mathrm{CCE}(X_m/X_{m-1})=CE(X_m/X_{m-1})+\mathrm{perc}(X_m)gEN(X_1), \tag{3}$$

其中 perc(X_m)是在长度为 m 的序列里面只出现过一次的序列所占的比例, EN(X_1)是当 $m=1$ 时的信息熵。当序列长度取[2, m]中的不同值时, 分别计算出相应的修正条件信息熵的值 CCE_2, CCE_3, ..., CCE_m。如果该用户是机器用户, 那它的行为会有一定的规律性, 在已知前 $m-1$ 项时间间隔后, 推测出第 m 项的值所需的信息量较小, 因而其修正条件信息熵的值会较小。与之相反, 普通用户的行为随机化程度较高, 已知前 $m-1$ 项的信息对于推测第 m 项的时间间隔并没有太多帮助, 需要较大的信息量, 修正的条件信息熵值也会较大。考虑到本研究的目的是识别出微博中的机器用户, 如果计算得到的 CCE_2, CCE_3, ..., CCE_m 中的最小值都较大, 则该用户存在较规律的发微博行为的概率就较低, 为机器用户的概率也较低。因此微博用户的最终修正条件信息熵为

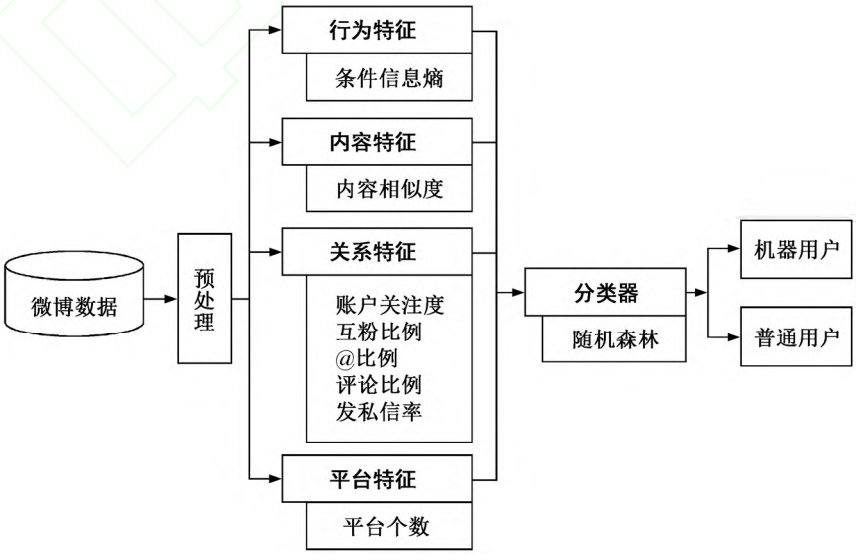

图 1　机器用户识别模型的基本框架

Fig. 1　Framework for bot-users indentification model

$$CCE_u = MIN\{CCE_2, CCE_3, ..., CCE_m\}。 \quad (4)$$

3.2 内容特征

通过观察机器用户的微博可以发现，现在他们已经不像普通垃圾用户那样在单个账户内大量发布同样的内容，但为了提高工作效率，仍不可避免地在多个不同的账户里发布内容完全相同的微博，其中包括一些个性化程度很高的原创微博，比如“今天吃了太多易长胖的东西了”，“今天吃了饭还得了礼物，蓝色果然是我的幸运色，好开森”等等。这些内容大大提高了机器用户的伪装能力，让普通用户难以区分与自己交流的到底是机器还是人。针对机器用户的这个特点，本文设计内容重复度指标来描述用户发布的微博与其他用户微博的相似程度。对于某用户发布的第 i 条微博，计算该微博与其他用户微博的重复度，公式如下：

$$Similarity_i = \frac{IdenticalWeibo_i}{TotalWeibo_i}, \quad (5)$$

其中 $TotalWeibo_i$ 是参与内容重复度匹配的微博总数，一般取某一固定值；$IdenticalWeibo_i$ 是参与匹配的微博中由其他用户发布的并与检测微博内容完全相同的微博数量。这里采用内容完全相同的匹配作为判断标准，而不是将微博分词后取特征值匹配，原因为：一方面一般只有机器用户才会大量发布与其他用户完全相同的微博，另一方面由于舆论热点的存在，特征值相同的微博会很多，特征值匹配会使普通用户的内容重复度值偏高，不利于区分机器用户与普通用户。

将该用户所有微博的重复度值取平均得到其最终的内容重复度，公式如下：

$$Similarity_u = \frac{1}{n}\sum_{i=1}^{n} Similarity_i, \quad (6)$$

其中 n 是该用户发布的微博总数。

机器用户的微博都是由自动化软件批量发布的，不同的机器用户总会存在发布内容完全相同的微博的情况，所以该内容重复度指标 $Similarity_u$ 的值会较大；普通用户发微博是一种原创的个性化的行为，发布与其他用户内容完全一致的微博的概率很低，该内容重复度指标的值也会较小。

3.3 关系特征

微博是基于用户关系构建的个性化信息传播、共享和获取平台。用户在使用微博的过程中，往往会在微博网络空间形成种种关系[22]，这些关系代表信息的传播方向。机器用户和普通用户使用微博的目的不同，形成的关系网络也存在差异。

3.3.1 关注关系

关注关系是微博中最主要的一种关系。将某用户接触的账户分为“朋友”和“粉丝”两类，其中“朋友”指该用户主动关注的微博账户，“粉丝”指关注该用户的微博账户。在微博中，该用户只能在首页上看到自己“朋友”发布的微博，而其“粉丝”可以看到该用户发布的微博。

根据郭浩等[23]的研究可知，积极关注别人，保持较高的发文数量，就可以吸引更多的粉丝，获得更高的关注度，使社会化网络媒体营销更加有效。机器用户会随机关注更多的人，寄希望于这些用户关注自己，从而看到并扩散自己发布的信息，使得其“朋友”数会远远高于“粉丝”数。普通用户在现实生活中的关系会或多或少地映射到其微博关系中，因而其“朋友”数和“粉丝”数相差不大。需要指出的是，微博中存在一些名人或者意见领袖，其言论受到很多人的关注，所以其“粉丝”数会远远大于“朋友”数。根据以上分析，定义账户关注度指标，计算公式如下：

$$Reputation_u = \frac{Followers_u}{Friends_n + Followers_u}, \quad (7)$$

其中 $Followers_u$ 是该用户的“粉丝”数，$Friends_u$ 是该用户的“朋友”数。$Reputation_u$ 值越大，说明该用户受到的关注度越高。因此机器用户的账户关注度较低，而普通用户(包括名人或者意见领袖)的账户关注度较高。

账户关注度仅仅衡量了单方面的关注关系，无法反映用户之间相互关注的情况，所以定义互粉比例来进一步分析用户之间的关注关系，计算公式如下：

$$MuturalRatio_u = \frac{Mutural\_F_u}{Friends_n + Followers_u - Mutural\_F_u}, \quad (8)$$

其中 $Mutural\_F_u$ 是该用户的互粉数。如果该用户是普通用户，基于互动的需要，更有可能存在相互关注的朋友，其互粉比例值较大；对于机器用户，与它相互关注的用户会比较少，其互粉比例值会比较小。

3.3.2 @关系

用户在发布微博时，为了提醒与该微博相关的用户及时查看该微博内容，往往会使用微博提供的@方法。被@的用户登录微博时就能够看到该微博

的提醒信息, 提高了沟通的效率。之前的研究认为机器用户会通过大量@一些不相关的人, 吸引这些用户的注意, 以达到快速扩散信息的目的[15]。为了衡量微博中@的情况, 定义@比例指标, 计算公式如下:

$$\text{MentionRatio}_u = \frac{\text{Mention}_u}{\text{TotalWeibo}_u}, \quad (9)$$

其中 Mention_u 是该用户存在@情况的微博数, TotalWeibo_u 是其发布的所有微博数。

3.3.3 评论关系

微博用户经常会在其他用户发布的微博下面进行评论, 发表自己对于该用户的状态或某事件的观点, 同时其他用户也会对该用户的微博进行评论, 这样就形成一种相互的评论关系。这里用评论比例来衡量这种关系, 计算公式如下:

$$\text{CommentRatio}_u = \frac{\text{Comment}_u}{\text{Commented}_u}, \quad (10)$$

其中 Comment_u 是用户 u 发出的评论数量, Commented_u 是被评论的数量。微博中评论的方式主要有两种, 一种是直接在原微博下进行评论, 另一种是在转发微博时对原微博进行评论。目前自动化的软件基本都带有自动评论的功能, 且一般来自于第二种。

3.3.4 私信关系

微博为了满足用户在较为公开的平台上也能进行较为私密的交流, 引入了私信功能。普通用户偶尔会采用这种方式进行直接的沟通。目前一些自动化软件也提供自动发送私信的功能, 所以机器用户的私信行为很可能会与普通用户不同。定义私信率指标, 计算公式如下:

$$\text{Message}_u = \frac{\text{SendM}_u}{\text{SendM}_u + \text{ReceiveM}_u}, \quad (11)$$

其中 SendM_u 是用户 u 发出的私信数量, ReceiveM_u 是该用户收到的私信数量。

3.4 平台特征

微博支持多种平台发布信息, 常见的发布平台可以分为五类。第一类是微博自身开发的服务平台, 这类平台供特定的用户发布微博, 比如微博weibo.com、专业版微博、媒体版微博、微博桌面等; 第二类是第三方微博管理应用工具, 这类工具由第三方公司开发, 服务于微博用户, 比如皮皮时光机、月光宝盒等; 第三类是移动设备, 不同的手机或者平板上有不同的微博客户端, 通过这些客户端发布的微博会显示相应的客户端版本, 比如iPhone 客户端、iPad 客户端、Android 客户端、小米手机 3S 等; 第四类是浏览器, 浏览器一般都会为微博提供扩展插件, 方便用户发布微博, 比如搜狗浏览器、360 浏览器等; 第五类是相关网站, 一些网站提供一键发布微博的功能, 帮助用户快速分享信息, 比如新浪博客、南方周末、译言网等。

机器用户由自动化软件控制, 微博的发布平台可能比较单一, 而普通用户则可能随机地选择平台操作微博账号, 本文用平台个数指标 $NumOfPlatform_u$ 来衡量某用户使用的不同平台数量。

4 随机森林分类

从前面提出的行为、内容、关系和平台 4 个特征维度, 共可以得到 8 个特征, 分别是修正条件信息熵(CCE_u)、内容重复度(Similarity_u)、账户关注度(Reputation_u)、互粉率(Mutural_u)、@比例(MentionRatio_u)、评论比例(CommentRatio_u)、发私信率(Message_u)和平台个数(NumOfPlatform_u)。识别某用户是否为机器用户的算法就是根据该用户的这 8 个特征值, 即

$$u\{\text{CCE}_u, \text{Similarity}_u, \text{Reputation}_u, \text{Mutural}_u, \text{MentionRatio}_u, \text{CommentRatio}_u, \text{Message}_u, \text{NumOfPlatform}_u\} \xrightarrow{\text{classify}} \{\text{Bot,Ordinary}\}。$$

本文选用随机森林[24]作为分类阶段的分类器。主要是基于以下原因。

1) 微博用户具有不同的背景, 来自不同的行业, 使用微博的习惯也各不相同, 因而从这些账户里提取的特征值个体差异性会较大, 单个特征值对区分机器用户和普通用户的贡献不会十分显著。随机森林算法对每个指标的要求不是很高, 每个指标都只需要包含少量区分信息即可。

2) 每位微博用户的行为、发布的内容、关系和使用平台之间很可能会存在一些潜在的相关性, 但对这些相关性又很难进行准确衡量, 所以那些对特征之间相关性十分敏感的算法会不适合。随机森林对于特征之间的多重共线性不敏感, 也不需要进行特征选择, 运用到本次分类研究中比较合适。

3) 随机森林算法可以评估所有输入特征的重要性, 可以找出那些显著区分机器用户和普通用户的特征, 为进一步研究机器用户的行为模式打下基

础。

4) 随机森林算法对离散点相对不敏感，鲁棒性较好。由于微博用户的多样性，难免会存在一些噪音数据，随机森林可以很好地避免这些数据对最终模型的影响。

随机森林综合运用了Boostrap方法[25]结合随机子空间的方式提取特征的思想，是由一组随机生成决策树组成的分类器。当一位未知类别的微博用户的8个特征值输入到训练好的随机森林后，随机森林里面的每棵决策树都独立地对该用户进行判断，根据多数表决原则决定该用户是否为机器用户。

每棵决策树的训练过程与一般的决策树有两点不同。首先，如果一共有 N 个样本，那么每棵决策树的训练集是从这些样本中采用有放回的抽样方法随机选取 N 个组成[26]。显然，每个样本被选中的概率是 $1/N$，未被选中的概率是 $1-1/N$，那么某样本没有出现在训练集中的概率就是$(1-1/N)^N$。当 N 趋于无穷大时，这一概率趋近于 $e^{-1}=0.368$，所以留在训练集中的样本大概占原来总样本的 63.2%。其次，决策树的最佳划分节点并不是从所有 M 个特征中(本文中 $M = 8$)直接选择得到，而是需要先从 M 个特征中随机选出 m 个($m<<M$)作为候选特征集，再利用 GINI 指数从候选特征集中选出最佳划分节点。

在随机森林中，每棵决策树都是完全分裂的，但并不需要进行剪枝。这是由于每棵树的输入样本都是随机选择形成的，并不是全部样本，因此不会出现过度拟合。由不同训练集训练而成的所有决策树组合在一起就形成最终的随机森林。

5 实验及结果分析

5.1 数据获取

5.1.1 原始数据集

本文实验所用的数据来自新浪微博。借鉴赵斌等[9]“深度优先”的方法，利用新浪的 API 接口获取一些已知用户的基本信息和发布的微博，然后获取他们的粉丝列表，将粉丝列表中的用户加入种子用户中，继续得到他们的基本信息和发布的微博。反复迭代，最终共收集到 1500 名用户的微博数据，其中存在少数用户在个别指标上有缺失的情况，实验中将该项指标的值取为 0。然后，用网络爬虫爬取我们在试用某自动化软件时获取的 120 个机器用户的微博数据。经过整理，获取的用户基本信息包括如表 1 所示的属性，微博信息包括如表 2 所示的属性。

为了客观地分析用户的行为特点，保证不同用户的行为特征值能够进行横向比较，本文将用于分析的微博发布时间限定在一个月内，具体时间为 2014 年 3 月 21 日 00:00 到 2014 年 4 月 20 日 24:00。获取的每位用户都只保留在这段时间内的微博信息。月微博发布量较小的用户对网上舆论的影响不大，研究价值较低，所以本文只挑选月微博发布量大于 30 条的用户，这样的用户共有 782 位。

5.1.2 划分机器用户和普通用户

通过以上方法得到的 782 个微博用户中也可能有机器用户。本文基于图灵测试[27]的思想，在新浪微博中搜索这些用户的 UID，找到他们的微博主页，由三位经常使用新浪微博的学生查看该用户的基本信息以及至少两页的微博内容，人工判断该用户是否为机器用户。遵循的统一判断标准为：如果该用户的微博主页有用心装扮过，个人资料里面的信息较为完整，相册中包含一些个人的照片，微博中包含一些主观原创的内容，并且经搜索后，这些内容没有与其他用户所发的微博存在完全一样的情况，同时该用户还在微博的评论中与其他用户有互动，

表 1 用户基本信息

Table 1 Typical attributions for users

属性	说明
Id	用户 UID
Followers	粉丝数
Friends	关注数
MuturalF	互粉数
Comment	评论数
Commented	被评论数
SendM	发私信数
ReceiveM	收私信数

表 2 微博信息

Table 2 Typical attributions for Weibo

属性	说明
MID	微博 MID
Time	创建时间
Content	微博内容
Platform	发布平台

则认为该用户为普通用户的概率较高。如果该用户发布的原创微博经搜索后存在与其他用户完全一致的情况, 在转发微博时使用的转发语缺乏个人特质, 存在@非好友的情况, 发布的时间存在一定的规律, 并且在评论中完全没有任何互动, 则认为该用户是机器用户的概率较高。最终挑选出 330 名最具有普通用户特征的微博, 与 120 名机器微博用户共同作为实验数据。

5.1.3 实验过程

本文实验先从 120 名机器用户随机取出其中75%的数据, 即 90 名机器用户, 再从 330 名普通用户中随机取出 90 名用户, 最后将抽取的两个数据集合并在一起, 完成随机森林模型的训练。模型的评估阶段分成两部分, 一方面将剩余的 30 名机器用户与 30 名普通用户混合在一起构成平衡数据集, 去掉类别标签后输入到模型中进行实验; 另一方面将剩余的 30 名机器用户与更多的普通用户混合在一起, 构成不平衡数据集, 去掉类别标签后输入训练好的随机森林中, 模拟评估真实场景下该模型的有效性。

5.1.4 信息熵分析

针对行为特征中的修正条件信息熵, 将每位用户相邻两条微博的发布时间相减, 得到发微博时间间隔序列, 如表 3 所示。

将机器用户和普通用户的发微博时间间隔序列输入后, 利用式(4)得到每位用户的修正条件。图 2 显示机器用户和普通用户各自修正条件信息熵的累积分布函数。

由图 2 可知, 机器用户的修正条件信息熵明显比普通用户的修正条件信息熵小, 说明机器用户的发微博行为存在较强的规律性, 而普通用户的发微博行为比较随机, 验证了前面对用户发微博行为的分析结果。

5.1.5 内容重复度分析

针对内容特征中的内容重复度, 首先对用户的

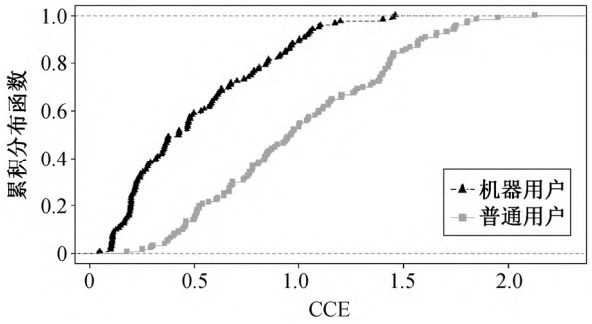

图 2 CCE 的累积分布函数

Fig. 2 Cumulative distribution function of CCE

微博内容进行预处理: 1) 对于转发的微博, 去除被转发的微博内容, 即“//@”后面的内容, 只保留原创的内容; 2) 去除系统自动添加的“转发微博”4 个字; 3) 去除微博中的地址、表情、@及@后的账户名称; 4) 去除内容为空或者是空格的微博。

然后, 将用户的每条微博内容为关键句, 通过新浪微博的搜索页面进行重复微博的爬取, 爬取前 3 页搜索结果(约 60 条, 实验中发现 60 条结果已经足以反映机器用户和普通用户的内容重复度)。

最后, 使用正则表达式匹配的方式解析出网页上原创的微博内容(假设共有 n 条), 与用于搜索的微博进行比较, 统计出完全相同的微博条数(假设有 m 条)。由于搜索到的页面中可能存在用于搜索的那条微博, 为了排除该微博对内容重复度的影响, 将搜索得到的总微博数和完全匹配的微博数同时减去 1, 即以 $n-1$ 为内容重复度计算式(5)中的 $\mathrm{TotalWeibo}_i$, 以 $m-1$ 为内容重复度计算式(5)中的 $\mathrm{IdenticalWeibo}_i$。

通过计算得到所有机器用户和普通用户的内容重复度, 其累积分布函数如图 3 所示。

由图 3 可知, 大多数的机器用户的内容重复度值都很高, 基本都在 30%以上, 有些甚至达到 90%以上, 而普通用户的内容重复度都很低, 集中在30%以下。这说明机器用户中的确存在批量发布完

表 3 用户的发微博时间间隔(部分)

Table 3 Weibo interval time table

用户 uid	发微博时间间隔序列(以“天”为单位)
1021738871	0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, …
1001333992	0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …
1006252657	2, 3, 0, 0, 2, 0, 0, 0, 1, 0, 2, 1, 0, 3, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, …
1000650441	2, 0, 1, 0, 2, 1, 0, 0, 1, 0, 1, 0, 2, 0, 1, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, …

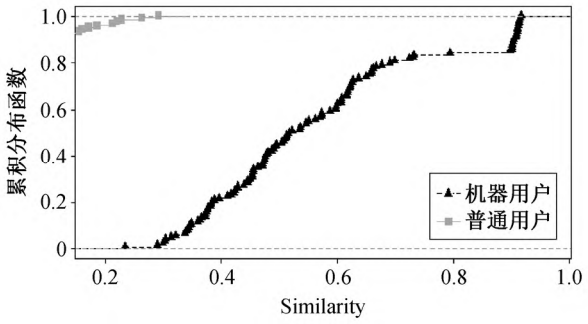

图 3　内容重复度的累积分布函数
Fig. 3　Cumulative distribution function of Similarity

全相同的微博的情况，该指标能够较好地区分机器用户和普通用户。

但是按照自动化软件的工作原理，机器用户的内容重复度应该是非常接近 100%，而实际结果显示，机器用户的内容重复度分布在 20%~100%之间，主要原因如下。

1) 部分自动化软件会将发布的微博内容进行微调，而本文设计的内容重复度特征值是匹配完全相同的微博，所以计算内容重复度时没有考虑这部分微博的个数。

2) 新浪微博会自动将用于搜索的句子拆开成关键词进行检索，所以爬取到的微博可能会存在一些关键词相同，但是内容不同的微博。

同时，普通用户的内容重复度也不全为 0，而是存在一些与其他用户完全相同的微博内容，主要原因如下。

1) 普通用户发布的微博内容可能是一些已经被广泛使用的俗语、成语、网络用语等，比如“子不教父之过”，所以不同的微博用户很可能会发布完全相同的微博。

2) 新浪微博中存在一些转发就送奖励的活动，普通用户会因为这些奖励而发布完全相同的内容。

3) 某位名人的公共言论，名著上的某段话，甚至某段广告词，可能被大多数普通用户所认同，这些普通用户会在自己的微博上发布这些完全相同的内容。

4) 注册网站或者首次使用某款软件后，该网站或软件会询问用户是否分享到微博平台，其中一些用户会选择分享，这些分享的内容也基本上是完全相同的。

5.1.6　账户关注度和互粉比例分析

针对关系特征中的账户关注度，将获取的用户基本信息中关注数和粉丝数代入式(7)和(8)中，分别计算得到每位用户的账户关注度和互粉比例，其累积分布函数分别如图 4 和 5 所示。

由图 4 和 5 可知，机器用户的账户关注度和互粉比例都明显比普通用户对应的值小。进一步的统计发现，75%机器用户的关注度在 0.5 以下，表明大部分机器用户的“粉丝数”远远低于“朋友数”。并且，机器用户最高的互粉比例都低于 50%，普通用户的互粉比例则可以接近 80%，说明普通用户间的互动活动是比较普遍的，而机器用户以获取粉丝为目的，随机关注他人，互动的情况很少。

5.1.7　@比例分析

统计每位用户发布的总微博数和存在@情况的微博数，带入式(9)，得到用户的@比例，对应的累积分布函数如图 6 所示。

与此前的研究结果[15]不同，本文的实验结果表明，大部分机器用户的@比例低于普通用户，表明他们在发布微博时并不会大量@其他用户。这主要是由于大量@他人会使得被@的用户使用微博时的体验很差，新浪微博会对扰民账号进行封号处理。现在的机器用户团队为了保证账号的长期有效性，已经不再采取这种方式来扩散信息。普通用户为了

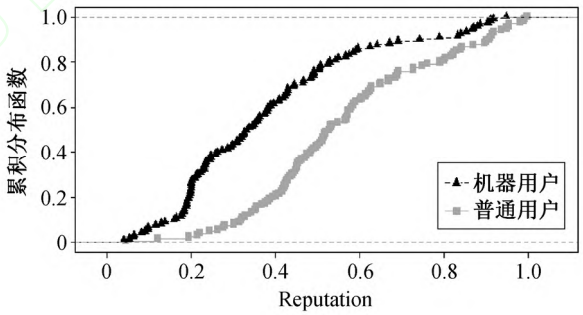

图 4　账户关注的累积分布函数
Fig. 4　Cumulative distribution function of Reputation

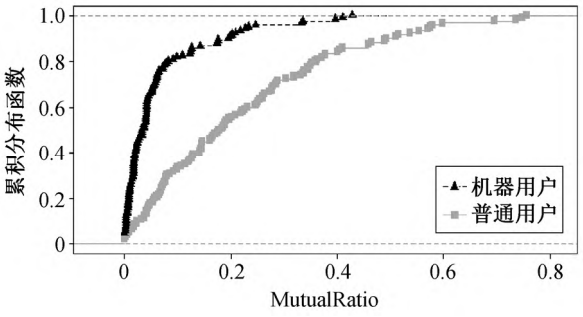

图 5　互粉比例的累积分布函数
Fig. 5　Cumulative distribution function of MutualRatio

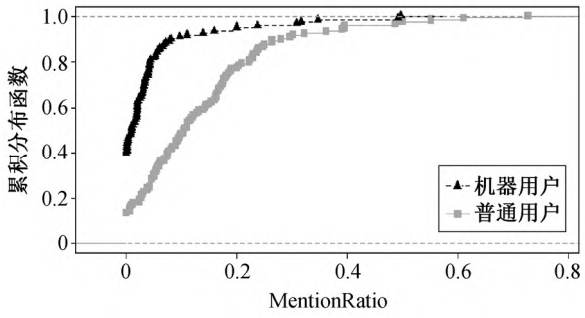

图 6　@比例的累积分布函数
Fig. 6　Cumulative distribution function of MentionRatio

互动的需求, 会在微博中进行一些@行为, 机器用户极少甚至根本不@其他用户的做法与普通用户不一样, 因而@比例仍然可以作为区分机器用户和普通用户的重要指标。

5.1.8　评论比例分析

根据获取的评论数量和被评论数量, 计算出每位用户的评论比例, 机器用户和普通用户的累计分布函数如图 7 所示。

由图 7 可知, 普通用户的评论比例较低, 机器用户则较高。机器用户可以通过大量转发评论某目标微博, 营造该微博十分热门的假象; 另一方面, 转发评论普通用户的微博也可以吸引普通用户的注意, 从而增加粉丝数量, 进一步增强自己信息扩散的能力, 因而机器用户的 Comment_u 数值较大。然而, 机器用户自身发出的微博质量可能较低, 得到的评论数并不多, Commented_u 的数值较小, 两者的比值较大。普通用户发布评论的行为一般不会像机器用户那么频繁, Comment_u 数值偏小。然而, 普通用户原创的微博质量往往较高, 会收到粉丝的评论或者由于成为机器用户的目标对象而被大量评论, Commented_u 的数值较大, 因而两者比值与机器用户相比会偏小。

5.1.9　私信数分析

实验中, 将用户发出的私信数和收到的私信数带入式(11), 计算得到用户的发私信率, 其累积分布函数如图 8 所示。

由图 8 可知, 机器用户和普通用户的发私信率差别非常大。机器用户的私信行为非常集中, 可以明显地分成三部分: 极少发私信、收发私信数量基本持平和大量发私信三种, 而普通用户的行为则比较分散, 收发私信行为比较随机。

5.1.10　微博发布平台分析

针对微博用户的平台特征, 实验中对用户发布每条微博所使用的平台进行了统计, 得到每位用户使用平台个数, 如图 9 所示。由图 9 可知, 机器用户使用的平台比较单一, 约 80%的机器用户使用过的平台数量在 5 个以下。普通用户发布的平台则比较多样化, 有些用户甚至会使用 20 个以上的不同发布平台。这主要是因为机器用户由自动化软件操控, 一般只能选择少量的特定的方式发布微博, 而普通用户则会根据操作方式的便捷程度选择最方便的方式发布微博, 尤其是现在很多网站都会提供微博的一键发布功能, 使得普通用户能够通过多个不同的网站发布信息, 使用的平台个数大大增加。

5.1.11　随机森林分类

本文实验利用 R 语言中的 RandomForest 包[28]实现随机森林算法。该包实现了 Breiman 和 Cutler 的随机森林算法, 并可以提供各种特征的重要性分析。

首先使用 R 中的 Sample 函数从机器用户和普通用户中分别随机取出 75%的数据(共 180 位用户)作为训练集。然后使用 RandomForest 包中的 TuneRF( )函数确定最佳的特征候选集个数 m=3。

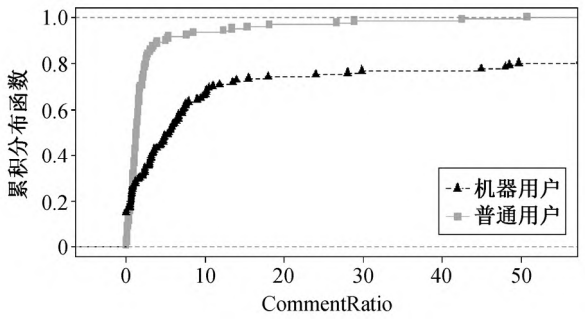

图 7　评论比例的累积分布函数
Fig. 7　Cumulative distribution function of CommentRatio

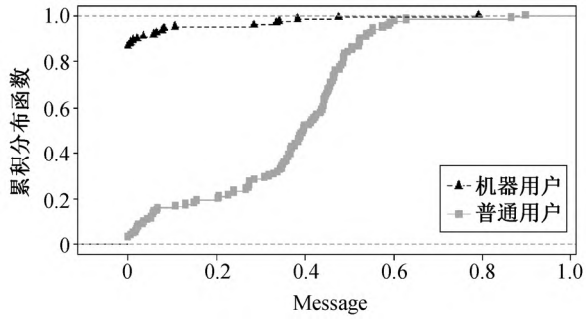

图 8　发私信率的累积分布函数
Fig. 8　Cumulative distribution function of Messages

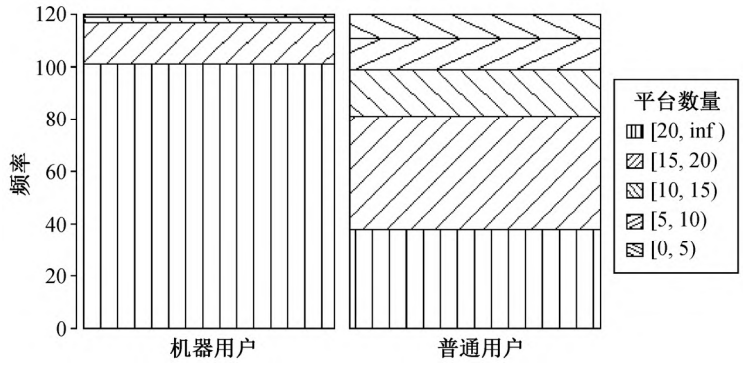

图 9　平台数量条形图
Fig. 9　Bar graph of platform number

使用 RandomForest 方法建立起含有 500 棵决策树的随机森林模型。根据前面的描述可知, 随机森林算法在建立每棵决策树时都会预留约 36.8%的数据, 这些数据称为袋外数据。由于这些数据没有参与决策树的训练过程, 可以作为测试集用于检测模型的有效性。经证明, 使用该袋外数据进行的错误估计是无偏的[24]。本实验模型的袋外数据错误估计率是 1.11%, 得到的混淆矩阵如表 4 所示。由表 4 可知, 本模型准确地识别出 98.9%的机器用户,只将一个普通用户误识别成机器用户, 说明本模型区分机器用户和普通用户的能力较强, 能够较为准确地识别出微博中的机器用户。

为了评估模型中某个特征的重要性, 将袋外数据中该特征的值进行随机交换, 形成新的测试集, 然后输入到随机森林中, 得到新的预测结果。通过交换前与交换后预测准确率的下降百分比来判断该特征的重要性。实验中对 8 个特征都进行了分析, 得到各特征的重要性如图 10 所示。

由图 10 可知, 各特征按照重要性由高到低依此为: 内容重复度(Similarity), 发私信率(Message), 修正的条件信息熵(CCE), 平台个数(NumOfPlatform), 互粉比例(MutualRatio), @比例(MentionRatio), 评论比例(CommentRatio)和账户关注度(Reputation)。其中内容重复度最为重要, 接近 55%, 说明在不同的微博账户中发布相同内容的情况十分普遍, 是识别机器用户的一个重要特征。此外, 关系、行为、平台特征也发挥了重要作用。虽然这些单个特征的重要性不是很明显, 但是模型整体的识别能力较强, 这也符合随机森林的思想, 即多个弱分类器组合后可以形成强分类器的功能。

表 4　随机森林的混淆矩阵
Table 4　Confusion matrix of RandomForest

识别结果	机器用户	普通用户	分类错误
机器用户	89	1	0.01111111
普通用户	1	89	0.01111111

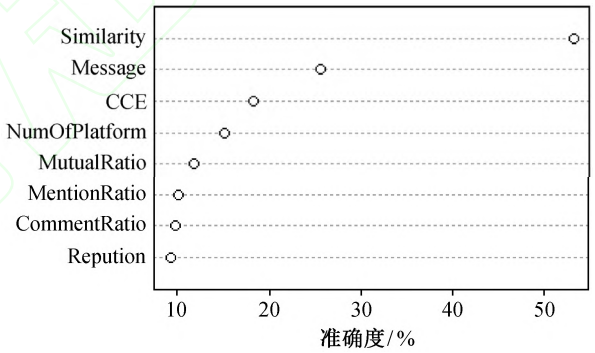

图 10　各特征的重要性
Fig. 10　Mean decrease accuracy of each feature

接下来的实验中, 将余下的没有参与模型训练的 60 位用户(机器用户 30 名, 普通用户 30 名)数据输入到建立好的随机森林中, 同时考虑到实际微博中的机器用户数量远远小于普通用户数量, 又将未参与训练的 30 名机器用户分别与 120 名、240 名普通用户混合在一起, 构成不平衡数据集进行测试。结果显示, 机器用户的准确率分别为 96.7%, 93.5%, 93.5%, 召回率始终为 96.7%, 普通用户的识别结果也具有较高的准确率和召回率, 分类器表现出较强的识别能力。测试结果如表 5 所示。

从表 5 还可以看出, 识别结果并未受到数据规模的影响。这表明模型所设计的 8 个指标可以较好

表 5　预测结果
Table 5　Prediction results for test data1

识别结果	机器用户 (30 名)	普通用户 (30 名)	普通用户 (120 名)	普通用户 (240 名)
机器用户	29	1	2	2
普通用户	1	29	118	238

地描述机器用户与普通用户的特点, 随机森林算法能够根据这些特征做出较为准确的判断。只要指标显示出相应的特征, 即会被准确地识别, 而不受数据分布、数据集不平衡等因素的影响, 这也提升了模型在实际应用中的抗干扰能力。

6　结语

本文针对国内微博中机器用户的特点, 从行为模式、微博内容、用户关系、发布平台 4 个维度提取出用户的信息熵、内容重复度、关注度、互粉比例、@比例、评论比例、平台个数和私信率等 8 个特征指标, 并利用随机森林算法实现了对机器用户的识别。真实数据的实验表明, 本文采用的指标和分类模型能有效地识别微博中的机器用户。本文的工作对避免虚假、有害信息的扩散, 营造积极健康的网络环境有重要意义。由于网络信息传播模式变化较快, 机器用户采用的方法也会不断变化, 因此, 进一步的研究需要关注机器用户的最新特点或变化趋势, 及时调整或构建新的识别模型。比如, 目前不同的机器用户往往发布相同内容的信息, 这也是识别机器用户的重要特征。但是, 如果机器用户发布文字不同但语义相同的信息, 则需要进一步分析微博的语义特征。另外, 机器用户除出现在微博中外, 在论坛、贴吧、新闻评论、商品评论等其他网络平台上也较活跃, 针对这些平台中的机器用户也需要构建相应的识别模型。

参考文献

[1] 中国互联网网络信息中心. 第 33 次中国互联网发展状况调查统计报告[R/OL]. (2014-03-05)[2014-07-01]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201403/t20140305_46240.htm

[2] 王莹莉, 张敏. 国内微博研究现状综述. 图书馆学研究, 2012, 33(12): 2-8

[3] Yardi S, Romero D, Schoenebeck G. Detecting spam in a twitter network. First Monday, 2009, 15(1): 1-13

[4] Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks // Proceedings of the 26th Annual Computer Security Applications Conference. New York: ACM, 2010: 1-9

[5] Thomas K, Grier C, Song D, et al. Suspended accounts in retrospect: an analysis of twitter spam // Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement. New York: ACM, 2011: 243-258

[6] Zhang X, Zhu S, Liang W. Detecting spam and promoting campaigns in the twitter social network // Proceedings of the 2012 IEEE 12th International Conference on Data Mining. Brussels: IEEE Computer Society, 2012: 1194-1199

[7] Lee K, Eoff B D, Caverlee J. Seven months with the devils: a long-term study of content polluters on Twitter // AAAI Conference on Weblogs and Social Media (ICWSM). Barcelona: 2011: 185-192

[8] Yang C, Harkreader R C, Gu G. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers // Recent advances in intrusion detection. Berlin: Springer, 2011: 318-337

[9] 赵斌, 吉根林, 曲维光, 等. 基于重用检测的微博垃圾用户过滤算法. 南京大学学报: 自然科学版, 2013, 49(4): 456-464

[10] 郭浩, 陆余良, 王宇, 等. 多特征微博垃圾互粉检测方法. 中国科技论文, 2012, 7(7): 548-551

[11] 丁兆云, 周斌, 贾焰, 等. 微博中基于统计特征与双向投票的垃圾用户发现. 计算机研究与发展, 2013, 50(11): 2336-2348

[12] Shen Yang, Li Shuchen, Ye Xiaoxiao, et al. Content mining and network analysis of microblog spam. Journal of Convergence Information Technology, 2010, 5(1): 135-140

[13] Zhang C M, Paxson V. Detecting and analyzing automated activity on twitter // Passive and active measurement. Berlin: Springer, 2011: 102-111

[14] Chu Z, Gianvecchio S, Wang H, et al. Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing, 2012, 9(6): 811-824

[15] Amleshwaram A A, Reddy N, Yadav S, et al. CATS: characterizing automation of Twitter spammers // Fifth IEEE International Conference on Communication Systems and Networks (COMSNETS). Bangalore: 2013: 1-10

[16] Wang A H. Detecting spam bots in online social networking sites: a machine learning approach // Data and Applications Security and Privacy XXIV. Berlin: Springer, 2010: 335−342

[17] Yu L, Asur S, Huberman B A. What trends in Chinese social media // Proceeding of the 5th ACM Workshop on Social Network Mining and Analysis. SanDiego: 2011: 978−988

[18] Shannon C E. A mathematical theory of communication. Bell System Technical Journal, 1948, 27(3): 379–423

[19] Porta A, Baselli G, Liberati D, et al. Measuring regularity by means of a corrected conditional entropy in sympathetic outflow. Biological Cybernetics, 1998, 78(1): 71−78

[20] Rosipal R. Kernel-based regression and objective nonlinear measures to assess brain functioning[D]. Scotland: University of Paisley, 2001

[21] Gianvecchio S, Wang H. Detecting covert timing channels: an entropy-based approach // Proceedings of the 14th ACM Conference on Computer and Communications Security. Alexandria, VA: 2007: 307−316

[22] 傅颖斌, 陈羽中. 基于链路预测的微博用户关系分析. 计算机科学, 2014, 41(2): 201−206

[23] 郭浩, 陆余良, 王宇, 等. 多特征微博垃圾互粉检测方法. 中国科技论文, 2012, 7(7): 548−551

[24] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5−32

[25] Efron B, Tibshirani R J. An introduction to the bootstrap. Boca Raton: CRC Press, 1994

[26] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences, 2003, 43(6): 1947−1958

[27] Turing A M. Computing machinery and intelligence. Mind, 1950, 59: 433−460

[28] Lia A w, Wiener M. RandomForest: breiman and cutler's random forests for classification and regression [R/OL]. (2014−07−17) [2014−07−20]. http://cran.r-project.org/web/packages/randomForest/index.html