doi:10.11959/j.issn.1000-436x.2015156

基于交互行为的在线社会网络水军检测方法

陈侃，陈亮，朱培栋，熊岳山
(国防科学技术大学 计算机学院，湖南 长沙 410073)

摘 要：网络水军对广告、谣言、木马和恶意链接进行传播，不仅干扰用户对在线社会网络的正常访问，还可能引发网络安全、社会稳定等方面的问题。针对网络水军信息传播的特点，提出基于交互行为的信息传播模型。模型根据不同传播主体间的交互定义特征来量化传播行为，使用决策树方法对水军传播的信息进行检测。通过新浪微博的真实数据分析传播模型并验证检测方法，结果表明检测方法能够对微博中水军信息进行有效检测。
关键词：网络水军；检测；交互行为；信息传播；在线社会网络
中图分类号：TP393 **文献标识码**：A

Interaction based on method for spam detection in online social networks

CHEN Kan, CHEN Liang, ZHU Pei-dong, XIONG Yue-shan
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: In online social networks, advertisements, rumors and malicious links are propagated by spammers arbitrarily. They not only disturb users'usualaccess, but also bring about network security threats and social panics. In an attempt to deal with the spam problems, an information diffusion model was proposed to capture the features of spam propagation. Propagation behaviors are quantitatively analyzed to detect spam messages with a decision tree-based method. The effectiveness of proposed detection model is evaluated with real data from the micro-bloggingnetwork of Sina. The experimental results show that proposed model can effectively detect spams in Sina micro-bloggingnetwork.
Key words: spam; detection; interaction; information diffusion; online social network

1 引言

网络水军出于政治或经济等目的对在线社会网络中的信息进行推广，使目标信息在极短的时间内大范围扩散，同时利用数量优势影响用户对其真实性的判断。根据内容和功能的不同，常见目标信息包括广告、木马和恶意链接、谣言等。广告水军以病毒营销的方式发布目标产品的不实描述，诱导用户对产品真实质量产生误判。病毒、木马和钓鱼网站被隐藏在正常内容中，或以中奖等方式吸引用户点击，通过超链接重定向到恶意程序所在的页面感染用户。谣言传播目的在于散布谣言并说服他人，不仅能够引导社会舆论，还可能引发大范围社会恐慌，甚至对国家安全和社会稳定造成威胁[1]。近年来爆发了多起网络造谣事件，例如“抢盐风波”[2]、“地震谣言”[3]等，对人民生活和社会治安造成严重困扰和威胁。

网络水军已成为工业界和学术界面临的重要课题，多种网络水军检测方法也被提出，如基于文本的方法[4]、基于黑名单的方法[5]和基于用户特征[6]的方法等。其中基于文本的方法适用于具有明显关键字的水军信息，如广告等；基于黑名单的方法适用于检测包含恶意链接的水军信息；基于用户行为模式的方法适用于检测具有明显水军特征的水军用户。这些检测方法局限性在于都只能检测单一种类的水军，在海量信息的条件下为保证低漏检率需要综合使用，从而增加检测的复杂性和时空耗费。因此设计一个通用性的检测方法具有重要意义。

收稿日期：2014-07-21；**修回日期**：2015-01-07
基金项目：国家自然科学基金资助项目（61170285, 61379103）
Foundation Item: The National Natural Science Foundation of China (61170285, 61379103)

本文提出了一种基于传播交互的水军检测方法。在线社会网络中，用户交互是引起信息传播的根本途径。水军虽然种类多样，但在交互行为上具有共同特性，而且与正常用户的交互行为表现出明显差异，因此从传播交互角度出发进行检测更具有通用性。

2　相关研究工作

近年来，随着在线社会网络的流行，网络水军越来越多地以在线社会网络作为水军活动的主要平台，知名网站如 Facebook、Twitter 和 Myspace 等都已经成为了水军活动的重要场所[5~7]。其他诸如论坛[8]、视频共享网站[9]、博客[10]等在内的在线网络也都已成为网络水军发动水军攻击的平台[11]。

水军检测可分为人员检测和信息检测，二者检测对象不同。人员检测针对水军成员，信息检测针对水军传播的信息。检测的一般观点是抽取特征，并利用特征分离水军成员或水军信息。

Irani 通过用户注册信息对水军成员进行检测[6]，这种方法使检测可以在用户注册时进行，但准确性较低，水军用户也可以随时更改信息逃避检测。Benevenut 使用 SVM 分类器对 Twitter 中网络水军进行检测[12]，使用的特征包括信息中包含链接的比例、用户账号使用时间、关注者的关注比例等。Wang 利用 Twitter 中 25 847 个用户信息对网络水军进行检测[13]，检测特征包括关注与被关注度、转发数量、双向交互数量以及链接比例等。

信息检测主要是根据信息内容分析水军特征，例如信息中链接特征以及基于自然语言处理的文本分类[14]。Zhang 使用基于链接相似性的方法关联水军活动[15]，并采用基于机器学习的方法对可能的水军活动进行检测。Blacklist 方法利用知名的 blacklist 站点来检测包含恶意链接的水军信息。Gao 使用此方法对 Facebook 留言墙中包含恶意链接的信息进行分析[16]。Grier 研究了 Twitter 传播信息中的恶意链接[5]，结果表明 Twitter 上 8%的链接都被重定向到恶意网站。他的工作还证明 blacklist 无法解决新的威胁，当一个恶意链接被标注为 blacklist 之前已经有超过 90%的用户被感染。文本内容也是水军检测的重要特征。Raymond 通过分析评论文本与正常用户评论的差异来发现网络水军发布的虚假评论[4]。Chen 利用回复、积极性及语义特征对新闻网络上的网络水军信息进行检测，可以提供 95%的检测准确率[17]。

当前网络水军检测的难点一方面在于检测的准确性有待提高，另一方面在于水军种类多样，账号多变，而检测方法大都只面向于单一种类的水军，无法提供通用的检测方案。为了保证检测的准确性需要同时使用多种检测机制，造成系统复杂性的提升和计算量的增加。

3　基于交互行为的信息传播模型

雇用网络水军的目的在于信息传播，雇主将产品、言论或观点在在线社会网络中推广，一方面需要增加信息传播广度，使其对更多用户可见；另一方面需要增加信息可信度，从而能够更好地影响用户，这些都是通过用户交互来实现的。

用户交互是信息传播的基本方式和根本动力。根据平台不同，交互类型也有不同，例如关注、转发、评论、点赞、收藏等。其中关注、转发和评论是在线社会网络中通用的交互方式。

关注：A 关注 B 之后，B 新发布的信息会实时推送给 A。

转发：A 转发 B 的信息，该信息从 B 的页面复制到 A 的页面，引起信息传播。

评论：A 评论 B 的信息，评论内容仍在 B 的页面显示，不会引起信息传播，但会对信息可信性和说服力造成影响。

虽然水军种类多样，而且水军账号不断变化，但从信息传播的角度来看，无论水军信息还是正常信息都有其固有的传播模式。这些模式体现在用户之间的交互上，从这 3 种交互行为入手对网络水军和正常用户在信息传播中的行为差异进行分析，就能为水军信息检测提供通用性的检测方案。

用 $F(u)$、$R(u)$、$C(u)$分别表示用户 u 的关注、转发和评论集合。其中，$F(u)$是由其他用户组成的无序集合；R 和 C 中的元素为类似<*user*, *time*>的二元组，*user* 代表发布信息的用户，*time* 为信息发布时间，集合按照 *time* 排序。

由于转发和评论在行为上都表现为信息的再发布，行为表现和特征描述都具有相似性。为避免重复将转发和评论通称为传播，传播集合用 $D(u)$表示。根据交互主体的不同，将传播特征分为关注者−传播者、发布者−传播者、传播者−传播者 3 种类型。其关系如图 1 所示。

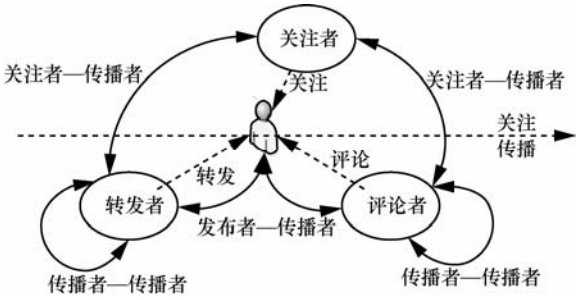


图 1 基于交互行为的信息传播模型

3.1 关注者–传播者特征定义

信息传播的前提是信息可见，在线社会网络中用户 A 发布的信息对用户 B 可见的方式主要有以下几种。

1) B 关注 A，B 就可以实时获得 A 的更新。由于在线社会网络中推送机制的广泛使用，新的信息发布后会立即推送给关注者。

2) B 关注 C，C 转发 A 的信息。那么 B 就可以通过 C 间接访问到 A 发布的信息。B 和 A 之间可能存在多跳。

3) B 直接获取 A 发布信息的链接，通过链接访问。

通过观察发现一般用户主要通过前 2 种方式访问信息，而网络水军则主要通过第 3 种方式访问目标信息。这是因为水军与雇主之间通常不存在直接的关注关系，只能通过雇主给出的链接进行信息传播。这使网络水军与正常用户在关注–传播关系上产生明显差异。

定义传播关系分布用来衡量传播者与关注者之间的关系，用 *DR* 表示传播关系分布，*p* 为一条信息，*u* 为信息发布者，*DR* 的计算式为

$$DR(p)=\frac{|F(u)\cap(\cup D(p)user)|}{|D(p)|}, p\in p(u)$$

其中，*P*(*u*)为用户 *u* 发布的所有信息，|*D*(*p*)|表示集合 *D*(*p*)的元素数量。*DR* 用来衡量传播者同时也是关注者的比例，正常用户主要通过关注关系获取信息，而网络水军主要通过链接方式获取信息，因此造成 *DR* 值的差异。

3.2 发布者–传播者特征定义

发布者与传播者之间进行直接交互，从交互时间的角度定义了平均传播时间（ADT）、首次传播时间（FDT）和传播启动时间（DST）3 个特征。

1) 平均传播时间

传播时间为信息从发布到最末一次传播的总时间，平均传播时间用来描述每一条转发/评论的平均持续时间。用 *ADT* 表示平均传播时间，计算式为

$$ADT(p)=\frac{\max(d_i\cdot time)-p\cdot time}{N},$$
$$d_i,\in D(p),1\leqslant i\leqslant N$$

其中，*N*=|*D*(*p*)|，由于 *D*(*p*)是按照时间排序的，因此 $\max(d_i\cdot time)=d_N\cdot time$ 。

网络水军通过完成雇主发布的传播任务获取报酬，而报酬是有限的，如果任务完成数量超出奖励限额就不会获得报酬。因此网络水军期望在任务期限内尽可能早地完成任务，而且任务完成数量一旦达到限额就不会再对信息进行传播。正常信息的传播仅受限于用户的使用习惯，传播时间与传播范围都没有具体的界限。

2) 首次传播时间

首次传播时间用来描述从信息发布到获得第一条转发/评论所等待的时间，用 *FDT* 表示首次传播时间，计算式为

$$FDT(p)=d_1\cdot time-p\cdot time$$

其中，d_1 为 *D*(*p*)中第一个元素。

由于消息实时推送机制的广泛使用以及移动终端应用的大力推广，很多在线社会网络都具有“类实时”特性，用户之间能够以近似实时的方式进行交互，信息也能够在第一时间被关注者传播。而网络水军访问目标信息的方式通常不是通过对被关注者的推送，而是通过给定的链接，因此难以体现出实时特性。同时水军活动任务的发布、接受、和实施都需要耗费一定的时间，使水军信息的首次传播时间比正常信息更长。

3) 传播启动时间

传播启动时间用来描述一条信息变“可信”所需要的时间。当一条信息的转发和评论量达到一定程度时，能够吸引更多用户关注并影响用户对信息内容的判断。用 *DST* 表示传播启动时间，计算式为

$$DST(p)=d_m\cdot time-p\cdot time,(d_i,\in D(p))$$

其中，*m* 为可信参数，用来描述一条信息产生影响力所需要的转发/评论的数量。本文中定义 *m*=1 000。即认为一条信息的转发/评论量超过1 000就能对用户判断产生影响。

3.3 传播者–传播者特征定义

传播者与传播者之间并没有或很少直接交互，只是在与发布者交互时产生时序关系。对该时序关

系进行分析可以更好地理解传播者参与的积极性和行为规律。从传播者−传播者角度定义了平均传播间隔（ADI）和传播间隔方差（VDI）2 个特征。

1) 平均传播间隔

传播时间间隔为每两条相邻信息之间的时间间隔，平均传播间隔为所有传播时间间隔的均值。其计算式为

$$ADI(p)=\sum_{i=1}^{N-1}\frac{d_{i+1}\cdot time-d_i\cdot time}{N-1},d_i,\in D(p),N=|D(p)|$$

由于水军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都很小。而正常用户发布的信息出于个人使用习惯的差异，时间间隔相对更大。

2) 传播间隔方差

传播间隔方差为所有的传播间隔之间的方差，用来描述一条信息的所有转发或评论的时间间隔的差异程度，计算方法为

$$VDI(p)=\sum_{i=1}^{N-1}\frac{[d_{i+1}\cdot time-d_i\cdot time]^2}{N-1}-ADI(p)^2$$
$$d_i,\in D(p),N=|D(p)|$$

水军行为的突发性不仅表现在时间间隔短，而且间隔分布也处于一个相对较小的范围内。而普通用户的转发和评论受访问习惯的影响表现出更大的差异性。

4　基于决策树的水军检测方法

将网络水军检测问题看作二分类问题，设 P 为在线社会网络中所有信息集合，$P=\{Ps\cup Pn\}$，其中，Ps 为网络水军推广的信息集合，Pn 为正常信息集合。设 p 为一条信息,使用特征向量表示为 $P=\langle DR,ADT,FDT,DST,ADI,VDI\rangle$。目标函数为 $\phi(p):p\to\{0,1\}(p\in P)$，其中，$\phi(p)$ 为二分类函数，$\phi(p)=\begin{cases}1,p\in Ps\\0,p\in Ps\end{cases}$。网络水军检测即发现信息 p 是否属于集合 Ps。

针对二分类问题当前已经有多种方案，例如决策树、SVM、Bayes、神经网络方法等。分类流程包括训练和分类 2 部分，训练过程通过特征选取和分类训练构造分类器，分类过程使用分类器对新的样本实现分类。本文选取决策树 C5 算法作为分类检测算法。C5 算法采用 Boosting 方式提高模型准确率，更适合在线社会网络这类数据量较大的场景。

决策树的根节点为数据样本集，分支节点对应着对单一属性的测试，该测试将数据空间分割为多个子集。每条分支对应该属性的不同属性值，而叶节点是带有分类标记的样本集分割。决策树需要使用训练集构建，然后实现对新样本的分类检测。

首先定义相关概念如下。

信息熵：在样本集 S 中，依据目标属性（是否为水军信息）将 S 分为 NS 和 SS 这 2 个子集，则 S 的信息熵计算为

$$E(S)=-\frac{1}{|S|}(|NS|)\mathrm{lb}\frac{|NS|}{|S|}+|SS|\mathrm{lb}\frac{|ss|}{|s|}$$

条件熵：设属性 D 根据取值将样本集 S 划分为 m 个子区间，使 $S=S_1\cup S_1\cup\cdots\cup S_m$，则属性 D 的条件熵计算为

$$E(S|D)=-\sum_{i=1}^{m}\frac{|S_i|}{|S|}E(S_i)$$

信息增益：属性 D 的信息增益 $Gain(D_i)$计算为

$$Gain(D)=E(S)-E(S|D)$$

信息增益比率：属性 D 的信息增益比率计算为

$$GainRatio(D)=\frac{Gain(D)}{SplitI(D)}$$

其中，

$$SplitI(D)=-\frac{1}{|S|}(\sum_{i=1}^{m}|S_i|\mathrm{lb}\frac{|s_i|}{|s|})$$

借助各属性的信息增益比率构建检测决策树。设训练数据集 $S=D_1D_2D_3D_4D_5D_6$ 为 6 维向量空间，其中，$D_i(1\leqslant i\leqslant 6)$分别对应模型中定义的 6 种特征。决策树构建算法如下。

算法 1　基于传播特征的决策树构建算法

输入　训练数据集 S

输出　决策树 DT

1) 初始化，设 $t=S$ 为 DT 的根节点。

2) 计算当前样本节点 t 的信息熵，以及 t 中每个特征属性 D_i 的信息增益比率 $GainRatio(D_i)$。

3) 令 $D_k=\max\{GainRatio(D_i)\}$，根据 D_k 的取值将 t 划分为 m 个子集，每个子集为 t 的一个分支，对应一个新的决策树节点。

4) 依次设每个新的决策树节点为当前样本节点，重复步骤 2)~4)，直到所有新样本节点中的样本满足：①都属于同一目标类；②所有属性都处理完毕；③样本的剩余属性取值完全相同。并将这样的节点标记为叶节点。

5) 用所有叶节点中占多数的目标分类属性值来标记该叶节点，决策树构建完成，返回 DT。

构造成功之后，就可以使用决策树对新的样本值进行目标属性的分类检测。从决策树的根节点开始，测试比较这个节点对应的属性值，然后选择正确分支向叶节点移动，重复比较和分支过程，直到到达叶节点，叶节点的类别属性即为最终的分类检测结果。

5 实验和分析

5.1 数据准备

从新浪微博中抓取真实数据分析传播特征。水军活动以很多方式存在，如广告水军、意见水军、木马病毒水军等。尽管内容和功能各有不同，但都以同样的方式被组织和传播。其中广告水军更常见也更容易区分，因此使用广告水军作为原型来分析其传播特征。

首先通过人工方式对新浪微博中的水军广告进行标注，然后提取这些广告信息中的关键字。利用新浪微博提供的搜索引擎使用这些关键字进行搜索，并保存搜索结果。

一般地，很多用户在看到广告时会选择忽略，极少参与转发或评论。在搜索结果中，80%的广告微博的转发和评论次数少于 10 次，大多数为 0 次。此外约 10%的微博具有很高的转发和评论量，认为它们较大概率来自于网络水军。过滤掉少于 100 次评论和转发的微博，最后得到 1 424 条水军数据集。

为了与水军数据进行对比，还搜集了正常用户的微博数据。采用手动方式挑选一些较小概率雇用网络水军的用户，选取方式是：①熟悉的用户，如朋友或老师；②教育或科学界的知名人士。选择教育或科学界人士是因为认为相比其他行业，这些用户更小概率会雇用网络水军。抓取了这些用户在 4 月 1 日到 4 月 14 日之间的所有微博。同样过滤掉少于 100 次评论和转发的微博，最后得到 1 687 条正常数据集。

5.2 传播特征统计分析

使用抓取到的数据集对水军用户和正常用户的传播特征进行分析，各项特征的累积分布如图 2 所示。

图 2 给出了传播关系分布（DR）特征的累积分布，可看出水军信息的特征值远小于正常信息的特征值。在转发特征图中，80%的水军信息的 *DR* 值小于 0.2，说明 80%的水军信息中，由关注者给出的转发不到总量的 20%。与之形成对比的是约 80%的正常信息的 *DR* 值大于 0.2。这一对比在评论特征图中更加明显，80%的水军信息的 *DR* 值小于 0.1，说明 80%的水军信息中，仅有不到 10%的转发和评论来自于关注者。这一分布证明了正常信息的转发和评论主要来源于关注者，而水军信息的转发和评论主要来源于陌生人。

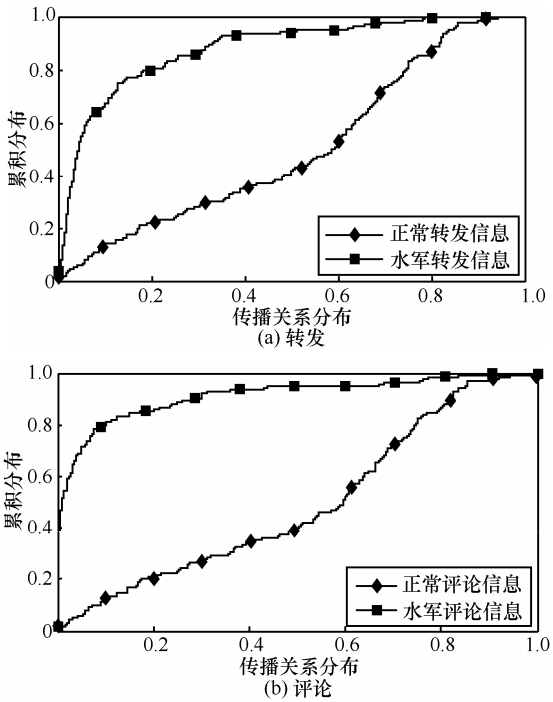
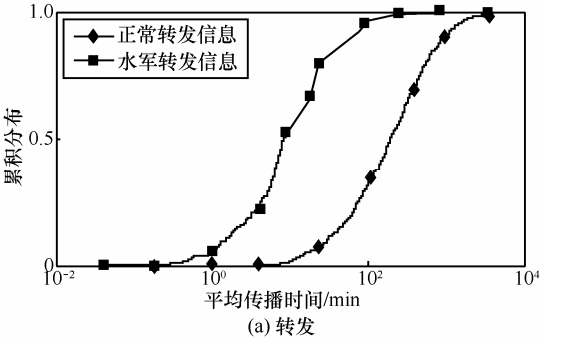


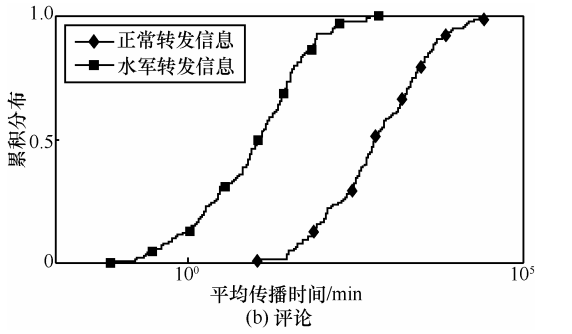
图 2 基于转发和评论的 *DR* 累积分布

图 3 给出了平均传播时间的累积分布，从图中可以看出，80%的水军转发信息平均持续时间少于 20 min，而 90%的正常用户的平均持续时间都大于 20 min。此外，80%的水军评论信息平均持续时间少于 30 min，相同时间下正常评论信息只有不到 5%。

图 4 给出了首次传播时间的累积分布，可以看出水军和正常信息在 *FDT* 上分布差异性明显。约 90%以上的正常信息都可以在 10 min 之内获取到第一条转发和评论。而在相同时间之内，水军信息中只有 10%能够获取到第一条转发，18%能获取到第一条评论。在 1 min 内，约 45%的正常信息可以获得第一条转发和评论，而水军信息中只有 2%可以获得第一条转发，7%获得第一条评论。

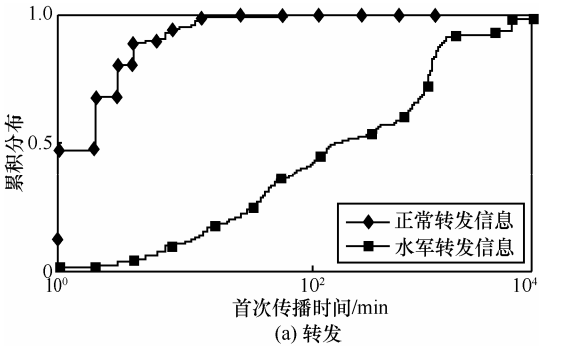


(a) 转发

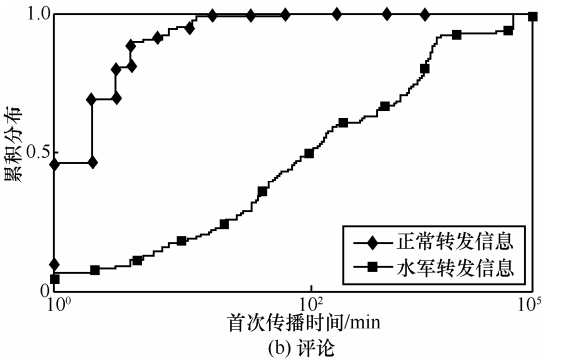


(b) 评论

图 3　基于转发和评论的 *ADT* 累积分布



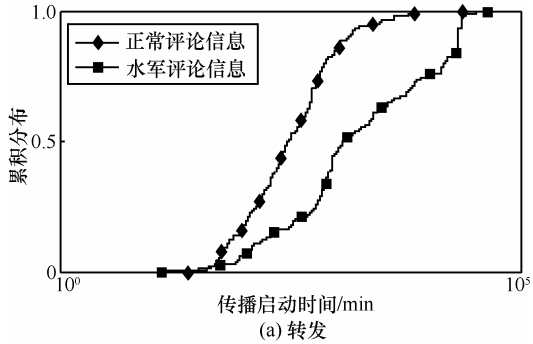
(a) 转发



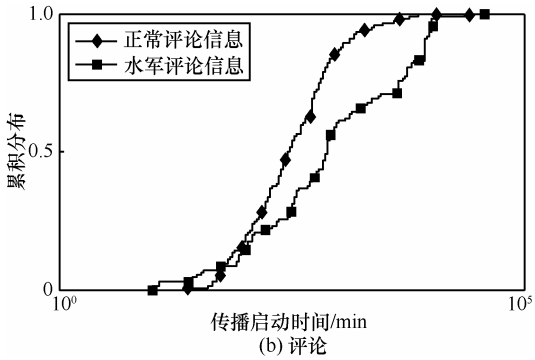
(b) 评论

图 4　基于转发和评论的 *FDT* 累积分布

图 5 给出了传播启动时间的累积分布。从图中可看出正常信息的启动时间一般小于水军信息。60%的正常信息的转发启动时间小于 200 min，该时间之内只有 20%的水军信息获得应有的转发。*DST* 特征的差异性不如其他特征明显，评论特征更为相近。



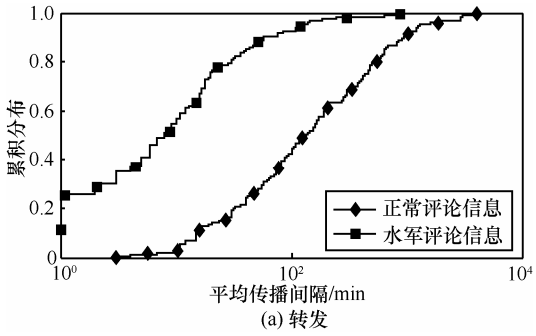
(a) 转发



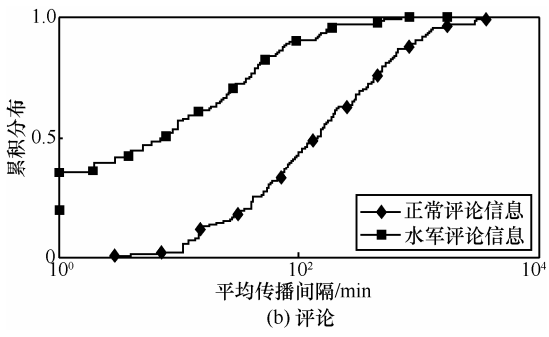
(b) 评论

图 5　基于转发和评论的 *DST* 累积分布

图 6 给出了平均传播间隔的累计分布。从数量上看，水军信息的平均传播间隔小于正常信息。60%的水军转发和评论间隔都小于 10 min，而在此范围内的正常信息不到 10%。此外 25%的水军转发和 35%的水军评论的平均传播间隔都在 1 min 之内，这证明了水军信息转发和评论时的突发特性。



(a) 转发



(b) 评论

图 6　基于转发和评论的 *ADI* 累积分布

图 7 给出了传播间隔方差的累计分布。水军信息传播间隔的方差更小，说明水军信息的传播间隔之间的差异性更小。原因是水军信息的突发特性使时间间隔都相对集中在一个小范围内，而正常信息受用户使用习惯的影响差异性更大。

5.3 检测结果

将数据集分为训练集和测试集，比例为 7:3。按照算法 1 的描述对训练集进行分类训练，得到决策树如图 8 所示。利用该树可以直接进行水军检测。

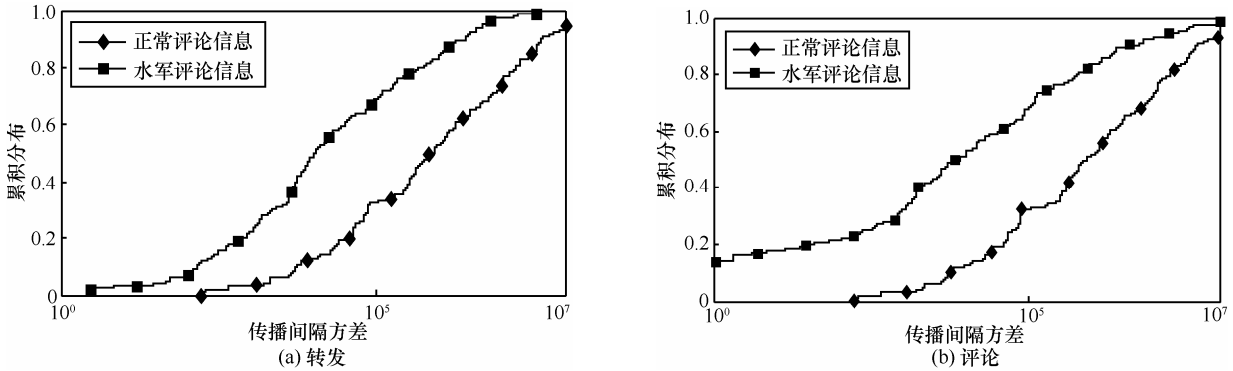


图 7 基于转发和评论的 *VDI* 累积分布

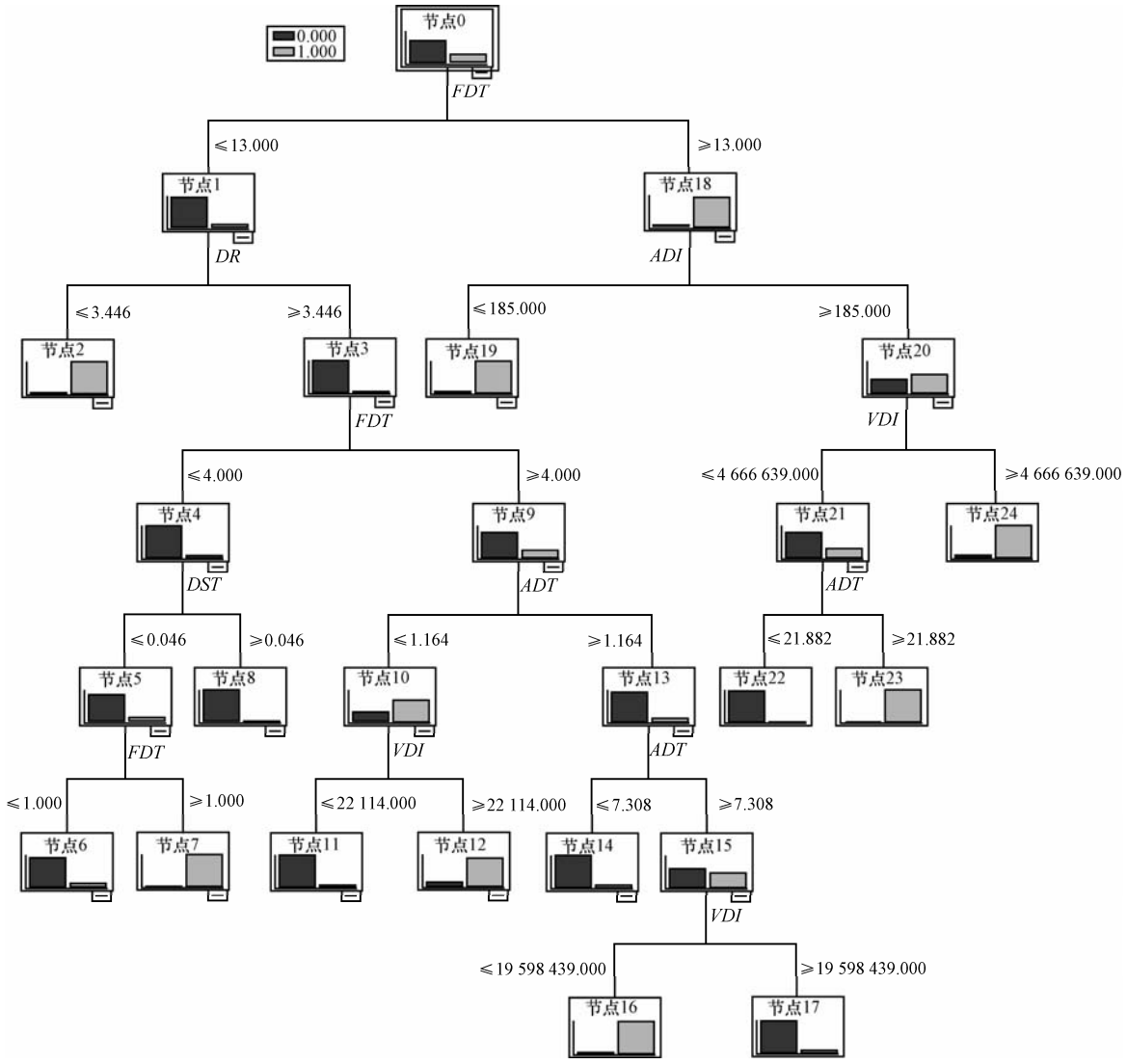


图 8 基于交互行为特征的水军检测决策树

使用测试集对决策树检测方法的有效性进行验证，并同时对比了SVM算法以及神经网络的RBF算法，验证结果如表 1 所示。

表 1　　检测方法结果对比

分类器	准确率	召回率	综合评价
决策树	0.999	0.956	0.977
SVM	0.985	0.670	0.798
RBF	0.969,	0.797	0.875

结果表明本文的决策树算法在基于传播模型的网络水军检测方面具有明显优势，准确率和召回率都高于其他 2 种方法。其中，SVM 算法可以提供较高的准确率，但召回率难以保证，漏检率较高。RBF 算法的召回率有所提升，但仍大幅度低于决策树算法。从综合评价来看，决策树算法性能最优，其次是 RBF 算法，SVM 算法由于召回率过低因此性能最差。

文献[18]使用概念图模型对新浪微博中的水军进行检测，基于平台和检测对象的相似性，使用它作为参考与本文的检测方法进行对比，结果如表 2 所示。可以看出，无论是在准确率、召回率还是综合评价上，本文的检测方法都表现出明显优势。这也证明了本文方法能够有效且准确地对水军进行检测。

表 2　　检测结果对比

检测方法	准确率	召回率	综合评价
本文方法	0.999	0.956	0.977
文献[18]方法	0.90	0.72	0.80

对检测算法中各特征的重要性进行了分析，结果如图 9 所示。重要性按由高到低的顺序排名依次是 *FDT>DR>ADI>ADT>VDI>DST*。

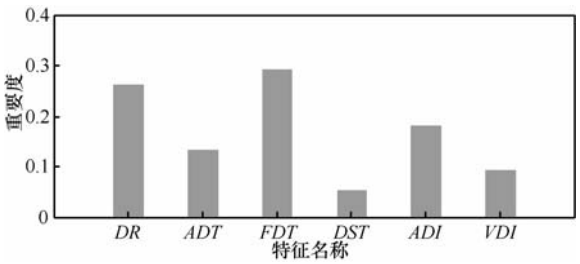


图 9　特征重要性评估

检测结果的高准确性证明了传播模型中特征选取的有效性，说明本文定义的特征能够准确描述网络水军和正常用户行为和传播过程的差异。正常用户可以根据关注关系实时获取更新提醒，而网络水军需要跟踪雇主发布的任务进行消息传播。正常用户对信息的访问和传播基于自己的日常习惯，而网络水军的消息传播依赖于任务发布时间和任务限额。

6　结束语

本文提出了基于交互行为的信息传播模型，从交互关系的角度定义了 3 种 6 个特征对传播行为进行量化。在此模型之下利用决策树算法对网络水军传播的信息进行检测。利用新浪微博的真实数据对传播模型进行分析并验证检测方法的有效性，结果表明本文的方法可以高效地检测出网络水军。尽管网络水军在种类功能方面各有差异，但传播行为上的共性使得本文的检测方法更具有通用性，可以适用于多场景下的水军检测。

参考文献：

[1] http://news.ifeng.com/opinion/special/wangluoshuijun/[EB/OL].

[2] http://zh.wikipedia.org/zh-cn/%E7%9B%B2%E6%8%A2%E7%9B%90%E4%BA%8B%E4%BB%B6[EB/OL].

[3] http://qcyn.sina.com.cn/news/ynyw/2011/1205/01134061411.html [EB/OL].

[4] RAYMOND Y K, STEPHEN L, LIAO S Y. Text mining and probabilistic language modeling for online review spam detection[J]. ACM Trans Management Inf Syst, 2011,2(4):25.

[5] GRIER C, THOMAS K, PAXSON V, *et al*. @spam: the underground on 140 characters or less[A]. Proceedings of the 17th ACM Conference on Computer and Communications Security[C]. Chicago, Illinois, USA, 2010. 27-37.

[6] IRANI D, WEBB S, PU C. Study of static classification of social spam profiles in MySpace[A]. ICWSM[C]. 2010.

[7] THOMAS K, GRIER C, SONG D, *et al*. Suspended accounts in retrospect: an analysis of twitter spam[A]. Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference[C]. Berlin, Germany, 2011. 243-258.

[8] SHIN Y, GUPTA M, MYERS S. Prevalence and mitigation of forum spamming[A]. IEEE INFOCOM 2011[C]. 2011. 2309-2317.

[9] BENEVENUTO F, RODRIGUES T, ALMEIDA V, *et al*. Identifying video spammers in online social networks[A]. Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web[C]. Beijing, China, 2008. 45-52.

[10] RAJADESINGAN A. MAHENDRAN A. Comment spam classi-

fication in blogs through comment analysis and comment-blog post relationships[A]. Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing-Volume Part II[C]. New Delhi, India: Springer-Verlag, 2012.490-501.

[11] HEYMANN P, KOUTRIKA G, GARCIA-MOLINA H. Fighting spam on social Web sites: a survey of approaches and future challenges[J]. IEEE Internet Computing, 2007, 11(6):36-45.

[12] BENEVENUTO F, MAGNO G, RODRIGUES T, *et al*. Detecting spammers on twitter[A]. CEAS[C]. 2010.

[13] WANG A H. Detecting spam bots in online social networking sites: a machine learning approach[A]. Data and Applications Security and Privacy, 25th Anunual IFIP WG11.3 Conference[C]. 2010. 335-342.

[14] 苏金树, 张博锋, 徐昕等. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 19(9):1848-1859.
SU J S, ZHANG B F, XU X, *et al*. Advances in machine learning based text categorization[J]. Journal of Software, 2006, 19(9):1848-1859.

[15] ZHANG X, ZHU S, LIANG W. Detecting spam and promoting campaigns in the Twitter social network[A]. The 12th IEEE International Conference on Data Mining[C]. 2012.1194-1199.

[16] GAO H, HU J, WILSON C, *et al*. Detecting and characterizing social spam campaigns[A]. The 10th ACM SIGCOMM Conference on Internet Measurement[C]. Melbourne, Australia, 2010.

[17] CHEN C, WU K, SRINIVASAN V, *et al*. Battling the internet water army: detection of hidden paid posters[EB/OL]. arXiv preprint arXiv:1111.4297v1[cs.SI]. 2011.

[18] 韩忠明等. 面向微博的概率图水军识别模型[J]. 计算机研究与发展, 2013, S2:180-186.
HAN Z M, XU F M, DUAN D G. Probabilistic graphical model for identifying water army in microblogging system[J]. Journal of Computer Research and Development, 2013, S2:180-186.

作者简介：

陈侃（1985-），男，陕西汉中人，国防科学技术大学博士生，主要研究方向为社会网络与网络安全。

陈亮（1984-），男，四川泸州人，国防科学技术大学博士生，主要研究方向为社会网络与网络安全。

朱培栋（1971-），男，山东兖州人，博士，国防科学技术大学教授、博士生导师，主要研究方向为网络安全、在线社会网络、网络科学。

熊岳山（1963-），男，湖南岳阳人，博士，国防科学技术大学教授、博士生导师，主要研究方向为图形和图像处理。