doi:10.3969/j.issn.1000-436x.2014.z2.031

面向舆情监测的微博转世账户研判模型

卜俊丽，彭灿，郑毅，黄九鸣，周斌
（国防科技大学 计算机学院，湖南 长沙 410073）

摘 要：微博社交网络在在线社交平台中扮演着重要角色，微博言论对网络舆论的贡献越来越大，网络舆论监测存在巨大挑战。转世账户是在网络舆论监测过程中出现的一类特殊的账户。加强这些账户的监测力度对于监测网络舆论有着很大的意义，实施监测的首要前提是发现这些账户。针对转世账户的特点进行模型设计，提出了一种基于时序和相似性的转世账户研判模型，并基于新浪数据进行了有效性的验证。

关键词：微博；转世账户；时序；相似性
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)Z2-0228-05

Model to find incarnated accounts in micro-blogging for public opinion supervision

BU Jun-li, PENG Can, ZHENG Yi, HUANG Jiu-ming, ZHOU Bin
(Department of Computer Science,National University of Defense Technology, Changsha 410073, China)

Abstract: Micro-blogging has played an important role in online social networks, and it has become a valuable source of public opinions. So there is a huge challenge to supervise public opinions in micro-blogging network. Incarnated accounts is a special kind of accounts that appeared during the supervision of public opinions. Strengthening the supervision of these accounts has great significance. The precondition for supervision is to discover them. In order to find incarnated accounts, a model targeted their characteristic based on the approach of timing and similarity is proposed, then the effectiveness of the model in Sina micro-blogging is verified.

Key words: micro-blogging; incarnated accounts; timing; similarity

1 引言

中文微博平台出现以后，微博网民规模持续增长，2010 年至 2011 年中国微博活跃用户数爆发式增长，2012 年进入相对平稳的增长期。据统计，截至 2013 年 6 月底，微博用户规模达到 3.31 亿，占网民比例的 56.0%[1]。2013 年下半年活跃用户数量虽然有所下降但截至 12 月，微博用户规模仍达 2.81 亿，占网民比例的 45.5%[2]。如此庞大的用户规模使微博的社会影响力加强，目前微博已经远远不局限于单纯的社交性网站，而成为网民获取信息的重要途径之一，演变成大众化的舆论平台。

微博是一柄双刃剑，它不仅吸引了普通用户、公众人物和传统媒体通过微博来获取新闻、传播新闻、发表意见，也成为很多试图借助网络平台传播恶意甚至不法言论的人捏造事实、散布反社会言论、制造社会舆论的有力武器。因而微博成了网络舆情监测的主战场之一。平台的后台管理者通过主动删除一些账号来防止恶意言论的快速传播，但由于微博的开放性，这些用户可以再次注册新的账号，转世账户就此出现，发现和监测转世账户是微博舆情监测的有效手段之一。

本文首先对人工识别出的新浪微博转世账户进行了分析，针对该类账户的特点进行模型设计，提出了一种基于时序和相似性的转世账户研判模型。该模型首先根据账户名相似性构建前世账号候

收稿日期：2014-07-07
基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（2012AA013002）
Foundation Item: The National High Technology Research and Development Program of China (863 Program) (2012AA013002)

选集，然后在候选集上进一步地筛选验证，符合模型条件的研判为转世账户。

2　相关工作

微博的转世账户是一类具有特定特点的账户，其与被删掉的账号本身属于同一人账号，2 个账号之间有更多相似之处甚至是相同之处，因而转世账户的研判属于账户同一性研判的范畴。文献[3]研究显示，77%的用户在不同的社交网站上会选择相同的账户名，5%的用户会选择增加后缀的形式改变账户名，1%的用户选择增加前缀的方式改变账户名。文献[4]中的研究通过用户的行为分析关联用户。研究中受人类自身限制的行为包括受限的时间和记忆、受限的知识。研究表明，在受到自身记忆和知识的限制以及习惯影响的情况下，用户在命名自己的账户名时会选择相同或者相似的用户名。本文对转世账户的研究中也利用了用户账户名相似性的特点。

利用文本内容对用户身份进行分析由来已久，最早的应用是电子邮件作者身份的检测[5~7]。在针对微博的研究中，博文内容也是研究者研究用户不可忽视的信息来源。文献[8,9]中都用到了用户的博文内容作为信息的来源之一。本文在对候选账户集进行进一步筛查验证时利用编辑距离弥补了余弦相似性算法计算账户名相似性的不足，在满足时序的前提下将先后账号所发表内容的延续性进行验证，以进一步确定是否转世账户。

就研究方法而言，本文借鉴了文献[9]中研究问题的方法，对新浪上人工识别出的转世账户进行分析，然后根据分析进行针对性的模型设计，下面是具体的数据分析。

3　数据分析

对微博用户进行研究，可利用的信息包括账户基本信息、博文内容等。本文还利用了微博发表的时间信息。在对人工识别出的转世账户进行分析时采用假设提问数据解答的形式。

Q1：账户名是否具有相似性。

文献[3]的研究指出，83%的用户在选择自己不同账号的账户名时会使用原名或者通过增加前缀或后缀的方式改变账户名。转世账户的账户名也具有这种相似性，另外转世账户的账户名一般由文字字符和数字组成。

为了方便问题表述，进行如下解释。

解释 1　现世账号为用户在微博账号被删除之后目前正在使用的新账号。

解释 2　前世账号为现世账号出现之前，该用户已经使用过的账号，可能是一个或者多个账号。

Q2：前世账号和现世账号之间在时间上有什么样的联系。

时间信息相比于账户基本信息和博文内容信息而言，具有客观性和不可更改的特性，不会随人为的目的进行欺骗性的设计或更改，因而具有最高的可信度。在对转世账户的时间属性进行分析时选取了 2 个时间：第一是账号注册时间，第二是博文发表时间。

在对账号注册时间进行分析时发现，前世账号之间和现世账号的注册时间没有明显的关联。分析其原因可能有：1) 为了抢占账户名现时账户名在使用前很久甚至是与前世账号同一时间注册；2) 前世账号被封之后临时注册；3) 在闲暇时间注册账号备用。

每一条博文都对应一个时间信息。在对转世账户的博文发表时间进行分析时发现，现世账号发表的博文时间一般不早于前世账号。可以以此作为验证条件之一。

4　模型及方法

4.1　形式化定义

定义 1　第 i 个账号博文数据流

$$D_i=\left\langle d_i^{t_1},d_i^{t_2},d_i^{t_3},\cdots\right\rangle$$

其中，$d_i^{t_j}=\{w_{i1},w_{i2},w_{i3},\cdots\}$ 表示第 i 个账号的博文，t_j 表示博文发表时间，w_{in} 表示博文特征词。

定义 2　初始账户集

$$U=\{C_1,C_2,C_3,\cdots\}$$

其中，$C_i=(CoN_i,D_i)$ 表示账户集中账户信息元组，CoN_i 是第 i 个账户的账户名，D_i 是第 i 个账户的博文数据流。则设待查账户组元为 $C_*=(CoN_*,D_*)$，待查账户的候选账户集为 $I=\{C_1,C_2,C_3,\cdots\}$，初始账户名集合可表示为 $U_{CoN}=\{CoN_1,CoN_2,CoN_3,\cdots\}$，待查账户的候选账户名集合表示为 I_{CoN}。

定义 3　用户 i 的博文生存区间

$$is^i=\left\langle CoN_i,t_{first}^i,t_{last}^i\right\rangle$$

其中，t_{first}^i 表示账户 i 发表第一篇博文的时间，t_{last}^i 表示发表最后一篇博文的时间，TS 是当前数据集所

有用户博文生存区间的集合。

4.2 模型框架

转世账户的发现模型框架分为 2 个大的模块，如图 1 所示，模块 1 产生候选账户集，模块 2 对候选集合进行筛选验证，确定账户是否为转世账户。

4.2.1 产生候选集合

如前面的数据分析，本文基于账户名相似性产生候选账户集，由于之后的账户筛选验证都是建立在候选账户集合的基础上，因而候选账户集是否足够大以包含前世账户将会直接影响实验结果，为提高处理效率将帐号名抽取出来单独处理。由于账户名字符串一般比较短，为了实现高效的匹配，本文采用正则表达式进行部分匹配，将包含待查账户名任意一个文字字符的账户名匹配出来，存入候选账户集 I。算法如下。

算法 1 产生候选账户集

输入：待查账户组元 C_*和初始账户集 U

输出：候选帐户集 I

1) 抽取待查账户名 CoN_*和初始账户名集合 U_{CoN}；

2) 构建正则表达式，对 U_{CoN} 逐项进行匹配，

IF（CoN_i 包含文字字符，文字字符∈CoN_*），THEN

将账户名 CoN_i 加入候选账户名集 I_{CoN}；

3) 利用初始账户集 U 和 I_{CoN}，生成 I。

经过上述算法，生成了候选账户集，其中包含所有候选账号的账户名和账户博文数据流。上述算法在最大程度上将相似账号名的账号加入候选账户集，为后续的筛选验证提供了足够大的候选集。

4.2.2 筛选验证

本文对 I 做以下 2 种筛选验证：1) 基于时序的筛选验证；2) 基于相似度的筛选验证。

1) 基于时序的筛选验证

现世账号的博文发表时间早于前世账号，并且其发表第一篇博文时间不会远远晚于其前世账号发表最后一篇博文的时间。因而，为了缩小查找范围，提高模型效率，对 I 进行时序的筛选验证，得到如图 2 所示的账户时序树。具体筛选过程如下。

将待查帐号 C_*发表第一篇博文的时间 t^*_{first} 与 I 中的帐号 C_i 发表最后一篇博文的时间 t^i_{last} 进行差值运算，若差值大于 0，即表示 C_i 中所有的博文时间都早于待查账户 C_*，则将 C_i 加入 C_*的时序候选账户集 I_{time*}。设置时间阈值 T，将 I_*中所有与 C_*的差值不大于 T 的账户加入 C_*的疑似前世账户集 I^T_{time}（即图 2 时序树的第一层），并将 $I_{time}I^T_{time}$ 作为 I^T_{time} 中每一个账户 C_j 的候选账户集 I_j。

对 I^T_{time} 中每一个账户 C_j，继续按照上面的方法从其候选集 I_j 中生成 C_j 的疑似前世账户集 I^T_{time} 以及 I^T_{time} 中的每一个账户的候选账户集（即时序树的第二层）。如此循环操作直到某一层的候选账户集为空。这样就形成了如图 2 所示的不相交的账户时序树。图 3 是时序树的一个分支时间序列，即账户 1 是待查账户的疑似前世账户，账户 9 是账户 1 的疑似前世账户。得到时序树以备后续进一步验证。

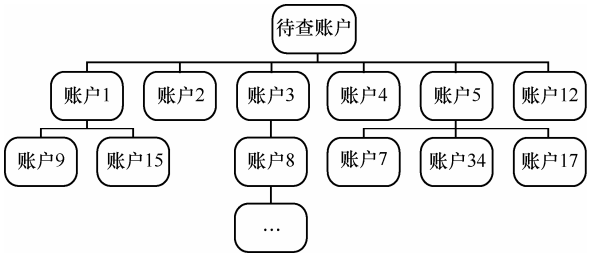

图 2 账户时序树

算法 2 基于时序的筛选验证

输入：待查账户组元 C_*，对应的候选账户集 I，阈值 T

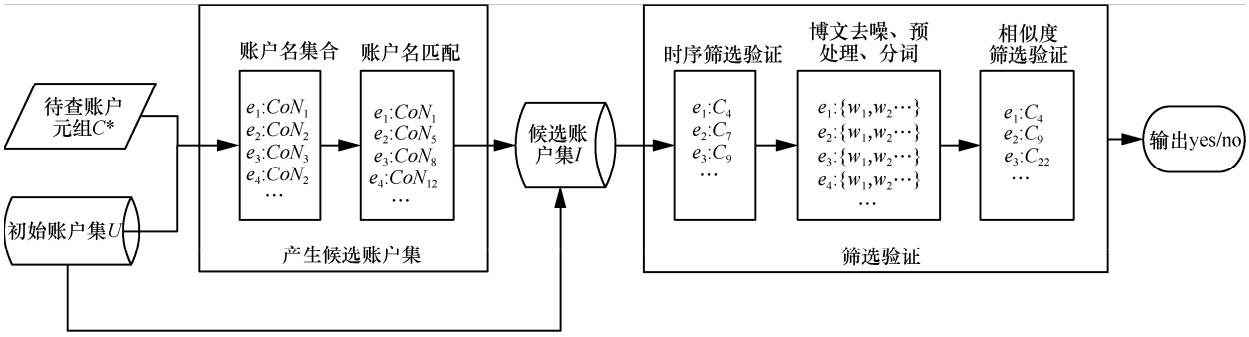

图 1 转世账户研判模型

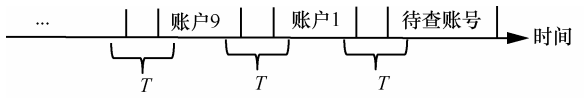

图 3　时序树的一个分支时间序列

输出：账户时序树

① 抽取 C_* 的 t^*_{first} 和 t^*_{last}，对 $\forall C_i \in I$，做同样的操作，构造博文生存区间集 *TS*。

② $\forall ts^i \in$ 候选集I，计算 t^*_{first} 与 t^i_{last} 的差值 sub_i，

IF sub_i>0，THEN

将 C_i 加入 C_*的时序候选账户集 I_{time*}

IF sub_i<=T，THEN

将 C_i 加入 C_*的疑似前世账户集 I^T_{time*} 并将 $I_{time*}-I^T_{time*}$ 作为 I^T_{time*} 中每一个账户 C_j 的候选账户集 I_j；

③ 对 $\forall C_j \in I^T_{time*}$，依次重置待查账户 C_j 为 C_*，I_j 为 I，调用步骤②；

④ 递归步骤③直到候选集为空。

2) 基于相似度的筛选验证

基于相似度的筛选验证包括 2 方面的相似度，账户名的相似度和博文内容相似度。

账户名相似度 $simi_c$ 采用余弦相似度算法 $simi_{cos}$ 和编辑距离相似度算法 $simi_{edit}$ 的加权平均，即

$$simi_c = 0.5simi_{cos} + 0.5simi_{edit} \tag{1}$$

其中，余弦相似度计算了用户名的字符相似性却忽略了字符出现的顺序，而编辑距离相似度计算了字符出现的顺序，弥补了余弦相似度的不足。

博文内容相似度基于时序筛选过后形成的时序树，设置博文数量 *n*，一一比较父节点 t_{first} 时间点后 *n* 条博文内容和子节点 t_{last} 时间点前 *n* 条博文内容相似性，选取最大值作为博文内容相似度的值。设置相似度阈值 *SIMI*，将账户名相似性与博文内容相似性进行加权，得到的相似度大于 *SIMI*，判断父节点是子节点的转世，否则不是。

余弦相似度是计算文本相似度的经典算法，经过分词之后，它将文本的每一个词都作为一个维度，以该词的权重为该维度的值，组合成向量代表该文件，设文件 *i* 表示成文件向量 $\boldsymbol{d}_i$=(w_{i1}，w_{i2}，…，w_{in})，文件 *j* 表示成文件向量 $\boldsymbol{d}_j$=(w_{j1}，w_{j2}，…，w_{jn})，则这 2 个文件的相似度计算为

$$simi_D(\boldsymbol{d}_i,\boldsymbol{d}_j) = \frac{\sum_{k=1}^{n} w_{ik}w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2}\sqrt{\sum_{k=1}^{n} w_{jk}^2}} \tag{2}$$

余弦相似度最小值为 0，最大值为 1，越接近 1，表明文本越相似。基于相似度的筛选验证的算法如下。

算法 3　基于相似度的筛选验证

输入：账户时序树，博文数量 *n*，相似度阈值 *SIMI*，加权系数 β

输出：研判结果 yes/no

① 对 $\forall$ 账号 $c \in I$，以单个字符作为一个向量维度，利用式(1)计算待查账号与 *c* 的相似度；

② 提取父节点和子节点对应时间点前后的 *n* 条博文数据流，去噪，分词；

③ 对 $\forall$ 父节点和子节点，利用式(2)计算父节点和子节点任意 2 条博文的内容相似度，取最大值作为博文内容相似度；

④ 按照加权公式 $simi_{sum}=\beta simi_c+(1-\beta)simi_D$ 得到最终的相似度；

⑤ 与 *SIMI* 进行比较，输出 yes/no。

5　实验结果与分析

5.1　算法性能分析

利用上述转世账户研判模型框架，设计系统验证。实验采用的数据是 2014 年 4 月 1 日到 4 月 30 日一个月的账户名以数字结尾的所有新浪微博账户共 7 393 313 个，从中经过人工研判选择了 50 个转世账户作为测试集进行实验。

模块 1 结束后得到的候选集包含账户数从几百到 1 万多不等。

这里将研判结果作为一个二分类，使用准确率、召回率和 *F*1 值等分类评价指标评价实验结果，具体公式如下。

$$Pr = \frac{TP}{TP+FP} \tag{3}$$

$$Re = \frac{TP}{N} \tag{4}$$

$$F1 = \frac{2PrRe}{Pr+Re} \tag{5}$$

其中，*Pr*、*Re*、*F*1 分别是转世账户的准确率、召回率和 *F*1 值。*TP* 是研判正确的账户数，*FP* 是研判错误的账户数，*N* 是测试集大小。

经过反复实验设定不同的时序筛选验证时间阈值 *T* 和不同的相似度阈值 *SIMI*，得到的准确率、召回率和 *F*1 值如表 1 所示。

表 1 *T* 和 *SIMI* 对 *Pr*，*Re*，*F*1 的影响

T/h	*SIMI*	*Pr*	*Re*	*F*1
16	0.6	0.766 66	0.479 16	0.589 743
32	0.6	0.717 94	0.583 33	0.643 678
48	0.6	0.744 18	0.666 66	0.703 296
16	0.7	0.84	0.437 5	0.575 342
32	0.7	0.823 52	0.583 33	0.682 926
48	0.7	0.833 33	0.625	0.714 285
16	0.8	0.913 04	0.437 5	0.591 549
32	0.8	0.937 5	0.625	0.75
48	0.8	0.939 39	0.645 83	0.765 432

由表中数据可以看出算法基本满足随着 *SIMI* 增大，准确率 *Pr* 增大；随着 *T* 增大，召回率增大。由于转世账户的账户名相似性比较高，算法可以达到比较高的准确性，由于转世间隔时间难以预料，因而在形成时间树时可能会删掉前世账号。因而算法的性能指标会受到测试集和实验基础数据的影响。

5.2 实验结果分析

对验证出来的转世账户设计实验进行统计分析，发现了转世账户博文的另一个特点：比正常用户包含更多的提及关系@，具体分析如下。

转世账户有如图 4 所示的博文，从统计表 2 和表 3 的结果来看，在正常用户的 2 173 条博文中，共包含提及关系@1 179 个，每条博文平均包含 0.542 567 85 个；而在这 2 173 条博文中，包含提及关系@的博文仅有 726 条，约占总数的 33.41%，每条包含@的博文中平均包含提及关系@1.623 966 9 个。而转世账户的 7 459 条博文共包含提及关系@高达 20 392 个，每条博文平均包含 2.733 878 6 个；在这 7 459 个博文中，包含提及关系@的博文就有 4 858 条，约占总数的 65.13%，每条包含@的博文中平均包含提及关系@的数目多达 4.197 612 3 个。

查看博文内容分析产生差异的原因发现：正常账户发表博文的目的是分享生活等，一般不会特意提醒别人去看，出现@是由于转发原因。而转世账户发博文是为了达到更好的传播效果，会以@的形式提醒其他用户查看其博文；有些转世账户在申请一个新的账号后为了找到之前的好友也会采取@的方式。因而，@的数量也可以考虑作为研判转世账号的条件之一，改进模型以提高研判效果。另外，从转世账户的提及关系@入手，也可以进行人物关联关系的挖掘分析，这都可以作为进一步研究的内容。

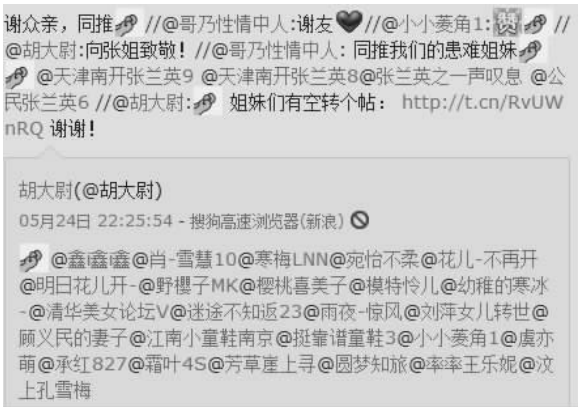

图 4 转世账户的一条博文

表 2 正常账户与转世账户包含@的博文数对比

账户类型	博文总数/个	包含@的博文数/个	包含@的博文数占博文综述的百分比
转世账户	7 459	4 858	65.13%
正常账户	2 173	726	33.41%

表 3 正常账户与转世账户每条博文@数对比

账户类型	平均每条博文@数/个	包含@的博文平均每条博文的@数/个
转世账户	2.733 878 6	4.197 612 3
正常账户	0.542 567 85	1.623 966 9

6 结束语

对于在网络上传播恶意言论的人，其转世账户账户名和博文内容有区分于正常账户的特征，本文针对这样的账户，提出了一种面向网络舆论监测的转世账户研判模型，取得了不错的研判效果。

参考文献：

[1] 第32次中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201307/P020130717505343100851.pdf

[2] 第 33 次中国互联网络发展状况统计报告[EB/OL].http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201403/P020140305346585959798.pdf

[3] ZAFARANI R, LIU H. Connecting corresponding identities across communities[A]. ICWSM[C]. 2009.354-357.

[4] ZAFARANI R, LIU H. Connecting users across social media sites: a behavioral-modeling approach[A]. KDD'13[C]. 2013.41-49.

[5] OLIVIER D V. Mining E-mail authorship[A]. KDD-2000 Workshop on Text Mining, ACM International conference on knowledge Discovery and Data Mining[C]. Boston, USA, 2000.

[6] OLIVIER D V, ANDERSON A, CORNEY M, *et al*. Mining E-mail content for author identification forensic[J]. SIGMOD Record, 2001, 30(4): 55-64.

（下转第 239 页）

protocol for vehicular ad hoc networks[J]. IEEE Transactions on Mobile Computing, 2013, 12(1):78-89.

[5] GOLLE P, GREENE D H, STADDON J. Detecting and correcting malicious data in VANETs[A]. Proceedings of the First International Workshop on Vehicular Ad Hoc Networks[C]. Philadelphia, PA, 2004. 29-37.

[6] PETIT J, FEIRI M, KARGL F. Spoofed data detection in VANETs using dynamic thresholds[A]. 2011 IEEE Vehicular Networking Conference (VNC)[C]. Amsterdam, 2011.25-32.

[7] PETIT J, MAMMERI Z. Dynamic consensus for secured vehicular ad hoc networks[A]. 2011 IEEE 7th International Conference on Wireless and Mobile Computing, Networking and Communications[C]. Wuhan, 2011.1-8.

[8] DIETZEL S, PETIT J, HEIJENK G, *et al*. Graph-based metrics for insider attack detection in VANET multihop data dissemination protocols[J]. IEEE Transactions on Vehicular Technology, 2013, 62(4): 1505-1518.

[9] KIM T H J, STUDER A, DUBEY R, *et al*. VANET alert endorsement using multi-source filters[A]. Proceeding of the Seventh International Workshop on Vehicular Ad Hoc Networks[C]. Chicago, IL, 2010. 51-60.

[10] RAYA M, PAPADIMITRATOS P, AAD I, *et al*. Eviction of misbehaving and faulty nodes in vehicular networks[J]. IEEE Journal on Selected Areas in Communications, 2007, 25(8):1557-1568.

[11] HSIAO H C, STUDER A, DUBEY R, *et al*. Efficient and secure threshold-based event validation for VANETs[A]. Proceedings of the Fourth ACM Conference on Wireless Network Security[C].New York, NY, USA, 2011.163-174.

[12] LI X J, WANG L M. A rapid certification protocol from bilinear pairing for vehicular ad hoc networks[A]. Trust, Security and Privacy in Computing and Communcations[C]. 2012. 890-895.

作者简介：

刘怡良（1990-），男，江苏徐州人，江苏大学硕士生，主要研究方向为车联网、网络安全。

石亚丽（1992-），女，安徽芜湖人，江苏大学硕士生，主要研究方向为车联网安全。

冯嵩（1979-），男，河南新密人，国网三门峡供电公司工程师，主要研究方向为电力系统自动化、继电保护。

王良民（1977-），男，安徽潜山人，江苏大学教授、博士生导师，主要研究方向为物联网信息处理技术、物联网安全协议、车联网安全结构。

（上接第 232 页）

[7] OLIVIER D V, ANDERSON A, CORNEY M, *et al*. Multi-topic E-mail authorship attribution forensics[A]. ACM Conference on Computer Security Workshop on Data Mining for Security Applications, Philadelphia, 2001.

[8] PARIDHI J, PONNURANGAM K, ANUPAM J. @I seek 'fb.me': identifying users across multiple online social networks[A]. IW3C2[C]. 2013.

[9] ZI C, STEVEN G, HAINING W, *et al*. Who is tweeting on Twitter: human, bot, or cyborg?[A]. ACSAC '10: Proceedings of the 26th Annual Computer Security Applications Conference[C]. 2010.21-30.

作者简介：

卜俊丽（1990-），女，陕西大荔人，国防科学技术大学硕士生，主要研究方向为信息安全、数据挖掘与分析等。

彭灿（1982-），男，湖北武汉人，国防科技大学硕士生，主要研究方向为信息安全、人工智能。

郑毅（1989-），男，重庆人，国防科技大学硕士生，主要研究方向为数据挖掘。

黄九鸣（1981-），男，福建安溪人，博士，国防科学技术大学助理研究员，主要研究方向为文本挖掘、社交网络分析与大数据处理技术。

周斌（1971-），男，江西吉安人，博士，国防科学技术大学研究员，硕士生导师，主要研究方向为互联网数据分析与挖掘、信息检索等。