社交网络账号的马甲关系辨识方法*

樊茜[1,2]，许洪波[1]，梁英[1]

[1]中国科学院计算技术研究所 网络数据科学与技术重点实验室，北京，100190

[2]中国科学院大学，北京，100049

E-mail: fanqian@software.ict.ac.cn

摘要：正确辨识网络账号的马甲关系，能够维护网络环境的安全与和谐，抑制网络中不法行为和虚假信息。基于文本挖掘的作者身份识别一直受到广泛关注，但对社交网络中文本作者关系鉴别的研究较少，本文提出了一种社交网络账号的马甲识别方法，基于网络语言的风格和账号关系，分别提取网络文本特征和账号之间的回复关系频次两组特征构成特征集合，同时基于账号组合构建训练样本向量空间，鉴别网络账号的马甲关系。结合论坛数据对所提方法进行了实验验证，准确率达到 80%，结果表明该方法具有较高的马甲辨别准确率。

关键字：马甲识别；语言风格；关系特征；社交网络

中图分类号：TP391　　**文献标识码：**A

A Sock-puppet Relation Detection Method on Social Network

Fan Xi[1, 2], Xu Hongbo[1], Liang Ying[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190

[2]University of Chinese Academy of Sciences, Beijing, 100049

E-mail: fanqian@software.ict.ac.cn

Abstract: Real name registration suffers great difficulties in social network and it is not completely universal. Some users use multiple IDs (usually called “sock-puppet”) to publish disharmonious views in order to reach illegal attempt such as to start or spread a rumor. It’s important to figure out a way to identify these users. In this paper, we propose a feature extraction method, which combined text data and social relation data together, and build a novel vector-space-model based on the combination of different IDs, deriving an effective sock-puppet relation detection algorithm. In the experiment of the forum data, we achieved 93% of classify precision. The result verified the effectiveness of the proposed method.

Keywords: Sock-puppet identify; Writing style; Relation feature; Social network

1. 引言

当前，全球超过 15 亿人使用社交网络，全球社交网络的月活跃用户数量超过 20 亿。在社交网络中，同一人拥有多个账号的情况十分常见。某人在同一网站注册多个账号时，常用的账号为主账号，而其余账号称为马甲账号，简称马甲。马甲功能中一部分是负面的，比如，利用不同账号为自己所开的讨论冲人气或推文；在主账号已有固定的朋友圈或形成固定形象时，使用马甲反对甚至诋毁他人或发表另类见解；注册成千上万个账号来发布不良信息、散布谣言、炒作或者通过卖等级高的马甲账号获益等等。这样的行为既浪费网络资源，又影响网络的安全性和公平性。

社交网站的后台实名注册实施困难，目前在国内还没有完全普及；即使网站后台是基于实名制的，但是网络言论在网站前台大都是匿名的，不易知道网络上的言论所属网络用户的真实身份。当用户在社交网络中发表不和谐言论，如造谣、诽谤他人、宣传不良思想等危害民众甚至国家安全的状况发生时，将社交网络中属于同一人的账号（马甲）进行同一性认定，有利于协助政府相关部门打击犯罪行为。

目前基于语言风格进行文本挖掘识别作者身份的研究工作受到广泛关注[1],但缺少针对网络账号的马甲关系识别方面的研究。由于网络中的账号相关信息少、噪音大，真实用户信息难以获取，使得对社交网络中账号马甲关系的标注十分困难，现有研究中缺少能够有效验证其所提出辨识方法准确性的权威数据与方法。如何更好的实现网络文本的挖掘，并充分利用网络账号的其他相关数据，以及如何有效验证方法准确性等问题，都有待解决。

* **收稿日期：**　　**定稿日期：**

基金项目：国家重点基础研究发展计划(973 计划)项目(2012CB316303, 2013CB329602)；国家自然科学基金重点项目(61232010)；国家自然科学基金面上项目(61173064)；国家科技支撑计划(2012BAH39B04)

本文利用某论坛泄露的账号信息数据，确定了一个已知相互马甲关系的账号集合，并提出了一种基于支持向量机，将社交网络中属于同一用户的账号的关系进行辨识的方法。通过研究人物的语言风格，挖掘论坛帖子文本特征，结合账号的回复关系特征，组合账号构造特征权值向量空间表示，利用支持向量机判别账号的马甲关系。实验结果证明这种方法能够有效辨识账号的马甲关系。

本文后续组织结构如下：第 2 部分主要介绍文本挖掘领域中作者识别方面的相关工作，第 3 部分详细描述本文提出的方法，第 4 部分是实验结果与分析，最后是总结。

2. 相关工作

社交网络上与账号相关的数据中文本数据最为丰富[2]，通过挖掘文本的语言风格进行作者识别的研究工作在国外很早以前就有了，但中文文本在这方面的相关研究较少，尤其在网络文本上的研究更少。

识别作者的关键问题是从其已知作品中统计出能代表其独特风格的识别特征，如：词汇总量及其特色词汇构成的数量比例、标点符号的使用频率、词语频率、句子长度、句式的使用分布、辞格的运用、声调和韵律分布等。根据文本是否规范，基于语言风格的研究方法有很大的差异。

对于文学作品等规范文本，单词是常用的文本挖掘的特征，但在语言风格分析中往往会结合其他的特征。王少康等[3]基于对句长的统计，构建段长的序列组合分析写作风格，利用不同作者写作时在文章语句节奏控制方面的特点，对 10 位作家进行识别分类，平均准确率约为 60%；孙晓明等[4]基于停用词使用的规律，使用文章中虚词频率分布作为特征，通过模式匹配，使用 SVM 和 K-means 对 13 位作家进行识别，正确率达 93.58%；日本学者金明哲[5]基于词性组合的统计分析，使用字符为单位的 unigram 和词性为单位的 n-gram 作为特征，其正判别率可达 95%。

相比文学作品，网络文本的特点是大多为短文本，语法不严谨，并且有许多的网络用语。短文本的特点是样本的特征稀疏，如何更好的利用短文本为数不多的特征是一个难点。网络用语的不规范使得流行语及奇异短语日异增多，识别流行词语或避免非正式语言的干扰也是一个难点。针对短文本的特征稀疏问题，武晓春[6]等基于语义扩展文本，取得了一定效果，但语言风格更多是形式上的分类，语义扩展的假设并不是很合理。DeVel 等人[7]从电子邮件中抽取了语言特征和结构特征作为作者的写作特征，采用支持向量机等机器学习方法，对电子邮件作者身份进行分类识别，但该方法基于电子邮件的相关格式特征，不能普遍应用到社交网络中的文本。Abbas 等[8]为有效监控互联网上的非法信息，提出运用文体学的方法对网络论坛发布信息的作者身份进行识别，抽取词汇、句法、结构、内容等特征，采用 SVM 和决策树分类算法，该方法只能区分论坛中发布恐怖信息的不法分子和普通用户两种类别，不能区分到具体真实用户。

除了文本数据，社交网络中账号之间的回复关系也具有很好的可利用价值。在现有的研究中，利用账号之间回复关系的研究工作主要用于社区发现、传播分析等领域。基于社交网络中账号之间的特性相似度与交互信息[11]，如兴趣、观点的相似度、互相关注关系、发言回复关系等构建的网络进行社区发现，可以用于发现共同兴趣的社会团体，识别水军团伙，对恶意水军进行防范等。但是，社区发现相较本文的账号马甲关系辨识是宏观层面的账号关系聚类的工作，不能够准确地发掘单个账号的马甲关系。

3. 基于账号组合的马甲辨识方法

目前，在社交网络中辨识账号的相关研究主要是基于文本挖掘账号的语言风格。在社交网络的文本挖掘分类任务中，经常面临网络语言不规范、文本长度短、无固定格式特征，待分类类别多等问题。而现有的对文本语言风格的研究工作多数基于文学作品文本和有格式的网络文本，如电子邮件，对于论坛帖子这种无格式的短文本的处理方法较少，文献[6]基于电子邮件的特殊格式进行作者识别研究的方法不适用论坛帖子。而在分类学习方面，前人工作的分类类别大多只有几十个，若将本文涉及的社交网络中的账号所属的真实用户作为类别进行分类，类别数量则会达到几百甚至数千，很容易导致分类算法运行效率降低，分类效果变差。为此，本文提出了基于账号组合的马甲辨识方法。

3.1 方法概述

针对网络文本挖掘中面临的短文本质量低的缺点，我们提取账号发言文本的 n-gram 与账号之间的回复关系频次，两组特征相结合进行账号马甲辨识，一方面最大程度提高文本特征的质量，另一方面加入账号之间的关系特征，扩展账号在文本内容之外的信息。

由于网络文本的长度短，为保证文本特征的数量，适合选择字、词、词组作为特征，不适合选择段落、句子作为特征。如果选择字表示文本特征，会丢失原始文本的大量信息；选择词组虽然会保留一些字、词丢失的有用信息，但使得特征向量更加稀疏，增加了分类的难度。因此我们采用分词后的词语和字 n-gram 作为特征，在保证文本特征数量的同时，保留了文本作者大量语言风格相关的信息。由于网络文本中的语法大多不严谨，并且夹杂许多网络用语，如果单纯利用中文分词的方式提取词语特征，由于分词性能十分依赖所用的词典的规模和质量，需要不断更新补充词典来保证分词效果，而采用 n-gram 特征提取方式可以不用考虑随语言领域和时间的变化而不断对分词词典进行修正扩充的工作。因而 n-gram 更适合作为网络文本的特征。另外，账号之间的回复关系是在作者身份识别领域，社交网络中账号特有的数据，能够体现有马甲关系的账号共同回复他人的规律。将文本特征与关系特征综合能够更好的体现账号的特性。

针对马甲关系辨别存在待分类类别多的问题，本文将账号两两组合构成新样本，而不是直接将单个账号作为分类实例，这样做的好处是可以将多类别分类问题转化成二类分类问题（详见 3.3 节分析），大幅简化分类算法的训练与测试的复杂度，也可以避免因为类别太多造成无法得到有较大区分度的分类结果。

假设 $\vec{w}_i$ 是用户 i 的特征向量表示，包括 n-gram 特征和账号回复关系特征，则账号间的马甲关系可以通过二分类的方法来判定。给定账号 i 和 j，如果 i 和 j 互为马甲关系，则 $\mathrm{f}(\vec{w}_i,\vec{w}_j)->1$，否则 $\mathrm{f}(\vec{w}_i,\vec{w}_j)->0$，f 为分类函数。通过账号两两组合构造出新的样本向量，使账号马甲识别算法的效率与效果都有很大的提升。

本文提出的马甲关系辨识方法主要步骤如下：

步骤一，提取账号的发言文本特征与回复关系特征，统计相应的频次，得到账号的特征权重向量；

步骤二，基于账号两两组合，得到账号组合的特征权重向量；

步骤三，由已知马甲关系的账号构成的账号组合构建训练样本，利用 SVM 算法训练得到相应的分类模型；

步骤四，测试含有待辨识马甲关系的账号组合，由账号组合的类别确定账号之间的马甲关系。

下面重点介绍上述步骤中的特征提取方法以及基于账号组合的向量空间构造方法。

3.2 特征提取和向量空间构造

3.2.1 网络文本特征提取

将社交网络中各个账号的所有发言内容分别汇总成一个文本，采用向量空间模型(VSM)来表示，一个文本可以表示为一个向量 $\overrightarrow{w\_doc}=(w_1,w_2,...,w_n)$，其中 w_i 为文本中第 i 个特征项的权重。将单个文本中所有字的 bigram 作为特征，统计相应的词频，筛除低频词，计算权重，得到该文本对应的向量。计算各个特征的权重采取的是 TF-IDF 方法，TF(Term Frequency)表示词频，IDF(Inverse Document Frequency)表示逆向文档频率。TF-IDF 公式的计算方法见公式（1）。

$$\mathrm{w}(t,\vec{d})=tf(t,\vec{d})\times\log(\frac{N}{n_i}+0.01) \qquad (1)$$

$w(t,\vec{d})$ 为词 t 在文本 $\vec{d}$ 中的权重，$tf(t,\vec{d})$ 为词 t 在文本 $\vec{d}$ 中的词频，N 为训练文档总数，n_i 为训练集中出现 t 的文本数。

3.2.2 账号回复关系特征提取

统计各个账号回复其他账号的频次，将被回复的账号作为特征，回复的频次即为特征的权值，得到关系特征的向量。假设账号 A 的回复账号集为 S_A（S_A 中有账号 $\mathrm{id}_{A_1},...,id_{A_i},...,id_{A_k}$），其中各账号被回复的频次为 $(\mathrm{f}_{A_1},...,f_{A_i},...,f_{A_k})$，则账号 A 的关系特征有为 $\{\mathrm{id}_{A_1},...,id_{A_i},...,id_{A_k}\}$，各个特征的权值为 $(\mathrm{w}_{A_1},...,w_{A_i},...,w_{A_k})$ $=(\mathrm{f}_{A_1},...,f_{A_i},...,f_{A_k})$。将所有账号的回复关系统计得到与文本特征平行的一组关系特征，表示为一个向量 $\overrightarrow{\mathrm{w\_relation}}=(\mathrm{w}_{A_1},...,w_{A_i},...,w_{A_k})$。

3.2.3 文本特征与关系特征融合

将文本、关系两组特征集汇总成一个特征集合，两组特征的权值向量$\overrightarrow{w_doc}$和$\overrightarrow{w_relation}$融合方法如公式（2）和公式（3）：

$$\vec{\mathrm{w}} = \overrightarrow{w\_doc} + \overrightarrow{w\_relation} \quad (2)$$

$$(w_1, w_2, ..., w_n) + (w_{A_1}, w_{A_2}, ..., w_{A_k}) = (w_1, w_2, ..., w_n, w_{A_1}, w_{A_2}, ..., w_{A_k}) \quad (3)$$

3.3 基于账号组合的向量空间构造

利用机器学习算法解决社交网络中账号马甲关系的辨识问题即是解决对所有账号的分类问题，目标是将拥有马甲关系的一组账号分为同一个类别，没有马甲关系的账号不分在同一类别。由此可知，如果将单个账号作为分类实例，分类问题即为将账号所属的真实用户作为类别进行分类的问题，待分类类别极多，远远超出常用分类算法待分类类别的适用数量。因此，本文提出将账号两两组合构成新样本，通过对账号组合类别的判断，确定两个账号之间的关系是马甲或非马甲。这样做的好处是可以将多类别分类问题转化成二类分类问题，大幅简化分类算法训练与测试的复杂度，也可以避免因为类别太多造成无法得到有较大区分度的分类结果。

社交网络中账号分为有马甲的账号和没有马甲的账号，而有马甲的账号分别属于各自马甲组。定义账号组合$\mathrm{pair} < i, j >$为账号i与j构成的账号组合。对于一个账号组合$\mathrm{pair} < i, j >$，可归纳为 5 种类型：

1）账号 i 与 j 均有马甲，且属于同一马甲组；
2）账号 i 与 j 均有马甲，但属于不同马甲组；
3）账号 i 有马甲，账号 j 无马甲；
4）账号 i 无马甲，账号 j 有马甲；
5）账号 i 与 j 均无马甲。

两两组合后的新样本的向量是将两个账号的向量合并，即将账号 i 和 j 的特征权值向量$\vec{\mathrm{w}}(i)$和$\vec{\mathrm{w}}(\mathrm{j})$，其中$\vec{\mathrm{w}}(i) = (x_1, ..., x_i, ..., x_k)$，$\vec{\mathrm{w}}(\mathrm{j}) = (y_1, ..., y_i, ..., y_k)$，组合成新样本$\mathrm{pair} < i, j >$的向量$(x_1, ..., x_i, ..., x_k, y_1, ..., y_i, ..., y_k)$。

在新样本的向量构成的向量空间模型中进行分类，账号组合类型 1）设为正例，账号组合类型 2）和 3）设为负例，账号组合类型 4）和 5）不参与计算，其中负例的可用样本数量比正例数量多很多，为了保证样本数量的平衡，实验时对负例的可用样本进行随机筛选，使正负两例的样本数基本一致。

3.4 基于账号组合的马甲关系辨识方法

从社交网络中账号发言的数据中提取特征，利用账号组合构造向量，然后利用 SVM 算法训练分类模型。预测时，判断一个未知账号 $\mathrm{id_x}$ 是否与某个账号集合$\mathrm{S} = \{s_1, s_2 ..., s_n\}$中的账号有马甲关系的步骤如下：

1） 提取该账号 $\mathrm{id_x}$ 在社交网络中的文本数据和关系数据，使用前述特征提取方法得到相关特征以及特征权重；

2） 使用账号组合向量空间构造方法，将账号集合 S 中的所有账号$\{s_1, s_2 ..., s_n\}$分别与该未知账号 $\mathrm{id_x}$ 组合，构成$\mathrm{pair} < s_1, id_x >$，$\mathrm{pair} < s_2, id_x >$, …, $\mathrm{pair} < s_\mathrm{n}, id_x >$，分别计算各个账号组合的特征权重向量；

3） 基于训练得到的支持向量机分类模型，测试步骤 2 得到的每一组向量，判断其中的各个账号组合$\mathrm{pair} < s_\mathrm{i}, id_x >$是正例还是负例，若为正例，则判定账号$s_\mathrm{i}$与$id_x$是互为马甲的关系，若为负例，则判定这两个账号没有马甲关系。

4. 实验与分析

4.1 实验数据

本文的实验数据是从某中文论坛上采集的帖子，由于该论坛数据庞大，仅采集了 2005 年 12 月到 2006 年 2 月之间“杂谈”板块内的所有帖子以及相关信息，包括发帖文本、回复关系、发帖时间等。

同时从网络上收集了该论坛账号信息三千多万条（根据该论坛账号信息数据库泄露出的文件整理），从中找到有发言文本数据的账号 19,883 个。在有文本的账号中，人工比对注册邮箱地址和密码，将注册邮箱地址一致或地址相似密码相同的账号标注为一个马甲组。一共整理出 1693 个账号，分为 551 个马甲组，每个马甲组中包含 2 到 33 个账号不等。

4.2 评价指标

本文采用的实验结果评价指标为准确率（precision，P）、召回率（recall，R）和 F1 值。首先计算每一类别的评价指标，为了表示分类器在全部类别上的综合分类性能，有宏平均和微平均两种方法。本文使用宏平均进行实验结果评价，即对所有类别的单项评价指标求取平均值。对账号组合所属类别的分类结果计算如下：

$$P = \frac{a}{a+b} \times 100\% \quad (4)$$

$$\mathrm{R} = \frac{a}{a+c} \times 100\% \quad (5)$$

$$F_1 = \frac{2PR}{P+R} \times 100\% \quad (6)$$

其中，公式 4,5 中 a 代表正例账号组合被判别正确的组合个数，公式 4 中 b 代表负例账号组合被判成正例的组合个数，公式 5 中 c 代表正例账号组合被判成负例的组合个数。

4.3 实验结果

4.3.1 基于账号组合的马甲辨识方法的分类结果

提取账号发言内容和关系特征，按照 3.3 节的方法，将属于同一马甲组的两个账号的组合向量作为正例，不属于同一马甲组的两个账号和一个账号有马甲而另一账号无马甲的组合向量作为负例，随机筛选负例样本使其数量与正例保持一致（即正负例样本数量相当），利用 liblinear(SVM 分类器)进行训练和分类测试。实验数据中共有 8080 个样本，每个样本是由一对账号组成的，其中 4040 个正例，4040 个负例。由于负例的 4040 个样本是从所有负例账号组合中随机采样出来的，为了避免数据采样的偶然性导致实验结果出现偏差，我们采用 5 次实验取平均值的方法，每次实验随机选择 4040 个负例样本跟 4040 个正例样本组成测试集，进行十折交叉验证，5 次实验的结果如表 1 所示。

表 1 基于账号组合的马甲辨识方法的分类结果

实验组号	准确率	召回率	F1 值
1	79.76%	79.79%	79.78%
2	79.68%	79.73%	79.71%
3	80.18%	80.24%	80.21%
4	79.92%	79.87%	79.89%
5	80.56%	80.56%	80.56%
平均	80.02%	80.04%	80.03%

从表 1 可以看出，5 次实验的平均准确率、召回率、F1 值等各项评价指标均超过 80%。实验同时表明，本文方法在不同的随机数据上都有较优的表现，实验结果能充分验证本文提出的方法的有效性。究其原因，虽然账号组合并没有增加两个账号各自的特征，但以账号组合为单位计算分类相似度，会显著放大两个账号中相同特征的影响，因此，两个账号相同的特征越多，其组合被判为正例的可能性越大，这样的账号组合是马甲的可能性也就越大。

需要指出的是，如果不采用账号组合构造向量的方法，而直接使用原始的单个账号构造向量，则需要分类 551 个马甲类别，如此多的类别对现有的多类分类算法是很大的挑战，基本都很难得到与我们的算法准确率相当

的实验结果。

4.3.2 特征有效性分析

为了测试不同特征的有效性，我们对文本分词后的词语、字 bigram、回复关系特征及其组合进行了实验对比，跟前面相同，每组实验均对负例样本进行 5 次随机采样取结果平均值，对比结果如表 2 所示：

表 2 使用不同特征的实验结果对比

使用特征种类	文本分词后的词语	字 bigram	回复关系	文本分词后的词语+回复关系	字 bigram +回复关系
准确率	72.31%	75.58%	78.69%	79.12%	80.02%
召回率	71.97%	74.99%	78.66%	79.16%	80.04%
F1 值	72.14%	75.28%	78.67%	79.14%	80.03%

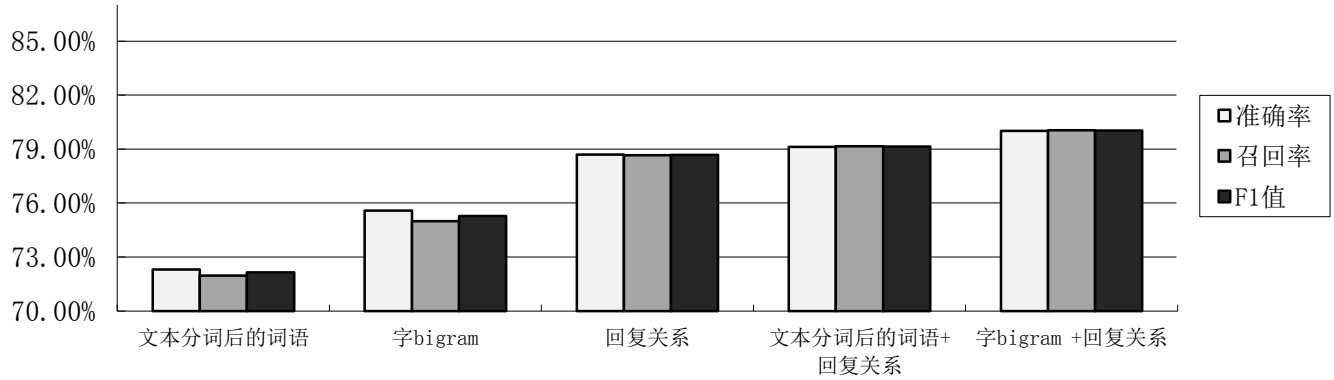


图 1 使用不同特征的实验结果对比

把表 2 的结果转换成图 1 所示的直方图，可以直观地看出本文选用的字 bigram 和回复关系两种特征较已知的其他方法只选用一种特征或选用分词后的词语的分类效果更优，准确率、召回率、F1 值各项评价指标均超过 80%。实验结果表明，文本特征中选用字 bigram 比文本分词后的词语效果更好；融合了社交网络关系特征与文本特征的方法比单用文本特征的效果更好，原因是有马甲的账号大部分都很活跃，而活跃的账号与其他账号有较多的回复关系，即有丰富的关系特征，因此融合关系特征后使有马甲的账号组合更容易被识别，提升了整体的实验效果。由此表明本文选用的字 bigram 更适合有不规范用语、网络流行新词的网络文本，提出的融合关系与文本特征的方法更适合进行马甲关系识别。

4.3.3 不同分类方法的对比分析

为了进一步验证本文的方法在辨识马甲应用上的有效性，我们将基于账号组合的马甲辨识方法与经典的分类方法如逻辑回归法、朴素贝叶斯（NB 算法）方法进行了实验对比，每组实验同样对负例样本进行 5 次随机采样取结果平均值，实验结果如表 3 所示。

表 3 使用不同分类算法的实验结果对比

分类算法	逻辑回归法	NB 算法	SVM 算法
准确率	77.63%	65.32%	80.02%
召回率	77.65%	65.50%	80.04%
F1 值	77.64%	65.41%	80.03%

从表 3 可以看出，SVM 算法比逻辑回归和朴素贝叶斯分类算法在辨识马甲时效果更好。由于论坛短文本的稀疏性导致 NB 算法的概率估计偏差严重，因而效果相对最差，实验结果的准确率等评价指标远远低于 SVM 算法。逻辑回归法表现比较稳定，效果不错，但仍低于 SVM 算法。

5. 总结与展望

本文主要研究了社交网络中账号之间马甲关系的辨识方法。在特征选择方面，利用账号的发帖文本，选择适合网络文本的特征，并结合回复关系的信息，从中挖掘出马甲账号之间的相似性。在向量构造方面，通过将账号

两两组合构建新的向量空间，克服了多数分类算法不能有效对多类别数据进行分类的缺陷。此外，在社交网络中还有一些信息是很有价值的，比如发言时间、马甲账号的上网作息规律等，将在后续的工作中加以考虑。

参考文献：

[1] Smita Nirkhi, Dr.R.V.Dharaskar. Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.5, 2013.

[2] Zheng R, Li J, Chen H, et al. A framework for authorship identification of online messages: Writing - style features and classification techniques[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 378-393.

[3] 王少康, 董科军, 阎保平. 基于语句节奏特征的作者身份识别研究[J]. Computer Engineering, 2011, 37(9).

[4] 孙晓明, 马少平. 基于写作风格的作者识别[C]//中国中文信息学会第五届全国会员代表大会暨成立二十周年学术会议论文集. 北京: 清华大学出版社. 2001.

[5] 金明哲. 中文文章的作者识别[R]. 第二届中国社会语言学国际学术研讨会暨中国社会语言学会成立大会, 2003.

[6] 武晓春, 黄萱菁, 吴立德. 基于语义分析的作者身份识别方法研究[J]. 中文信息学报, 2006, 20(6): 61-68.

[7] De Vel O, Anderson A, Corney M, et al. Mining e-mail content for author identification forensics [J]. ACM Sigmod Record, 2001, 30(4): 55-64.

[8] Abbasi A, Chen H. Applying authorship analysis to extremist-group web forum messages [J]. Intelligent Systems, IEEE, 2005, 20(5): 67-75.

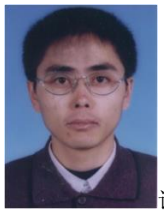
[9] Yu B. An evaluation of text classification methods for literary study [J]. Literary and Linguistic Computing, 2008, 23(3): 327-343.

[10] Diederich J, Kindermann J, Leopold E, et al. Authorship attribution with support vector machines[J]. Applied intelligence, 2003, 19(1-2): 109-123.

[11] Ge R, Ester M, Gao B J, et al. Joint cluster analysis of attribute data and relationship data: The connected k-center problem, algorithms and applications [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2008, 2(2): 7.

作者简介：

樊茜（1990 年生），女，硕士生，主要研究领域为 Web 搜索与挖掘，Email: fanqian@software.ict.ac.cn（通讯作者）；



许洪波（1975 年生），男，副研究员，主要研究领域为 Web 搜索与挖掘、大数据分析等， Email: hbxu@ict.ac.cn；

梁英，（1962 年生），女，副研究员，主要研究领域为大数据分析、服务计算、中间件等，Email: liangy@ict.ac.cn。