计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140804
52(11):2527-2534,2015

微博自媒体账号识别研究

刘金宝　盛达魁　张　铭
(北京大学信息科学技术学院网络与信息系统研究所　北京　100871)
(shengdakui@pku.edu.cn)

Study on We Media Account Detection in Microblog

Liu Jinbao, Sheng Dakui, and Zhang Ming
(*Institute of Network Computing and Information Systems, School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)

Abstract　As an outcome of Web 2.0 era and a rising social media, microblog service has been playing a more and more important role in people's daily life. It serves as not only a bridge of communication and information sharing, but also a crucial way to acquire information. As a mixture of social network and information media, microblog has a diverse ecological environment. We media accounts as a component of microblog, have been taking rapid development. In this paper, we creatively introduce the we media account detection problem and illustrate its meaning, then we propose a comprehensive feature set from account profile, posting behavior and posting content. Based on these features, we perform a supervised learning method to detect we media account. Experimental results show that: 1) we media accounts distinct from general accounts in the environment of Sina Weibo, and the difference is mainly on the behavior of publishing microblogs and the topic of microblogs. 2) The proposed three feature sets are effective for we media account detection, and they complement with each other as well, achieving an impressively high accuracy of 96.71%.

Key words　microblog; we media account; classification; support vector machine(SVM); supervised learning

摘　要　随着 Web 2.0 时代的发展,微博作为新兴的社交网络媒体在人们的日常生活中扮演着愈发重要的角色.它不仅是用户交流与分享信息的桥梁,也是获取信息的重要方式.微博同时具有社交网络与信息媒体双重性,其生态环境中仅具有媒体属性,用于发布信息给公众的自媒体账号(we media account)发展迅速.首次提出微博自媒体账号识别这一研究问题,阐述了自媒体账号识别对分析微博生态环境、用户兴趣建模、优质内容挖掘的重要意义,提出了结合个人信息、账号行为及微博内容 3 类特征的有监督识别方法.研究结果表明:1)自媒体账号与普通的微博账号有着较明显的不同,主要体现在微博发布行为的规律性以及话题分布特性之上.2)提出的 3 类特征能够有效识别自媒体账号,不同类别的特征也能够相互补充,预测准确率高达 96.71%.

关键词　微博;自媒体账号;分类;支持向量机;有监督学习

中图法分类号　TP311

收稿日期:2014-09-03;**修回日期**:2015-01-13
基金项目:国家自然科学基金项目(61272343);教育部高等学校博士学科点专项科研基金项目(20130001110032)
通信作者:张　铭(mzhang@net.pku.edu.cn)

在近年来,随着 Web 2.0 时代的发展,以用户生成内容(user generated content)作为经营模式的网络应用蓬勃发展.博客、社交网络、微博等逐渐走进人们的视线并在人们的日常生活中扮演着越发重要的角色,成为人们生活不可缺少的一部分.其中,微博作为新兴的社交网络媒体,更受人们的青睐.截至目前,微博的注册用户已超过 5 亿,月活跃用户 1.29 亿,是用户交流与分享信息的桥梁,也是用户获取信息的重要方式.Kwak 等人[1]研究了国外的社交网络媒体 Twitter,认为相比一般的在线社交网络具有的社交属性,Twitter 具有更显著的媒体属性.新浪微博作为国内最大的社交网络媒体,和 Twitter 类似,不仅具有社交属性,媒体属性也非常鲜明.在这样的生态环境中,出现了大量仅具有媒体属性且用于向公众发布各类信息的自媒体账号(we media account).与在线社交网络中的"公共主页"类似,微博自媒体账号为普通用户获取信息提供了有效途径,是微博中信息的重要来源.然而微博本身并没有对自媒体账号的和普通账号加以区分,基于此,本文首次提出微博自媒体账号识别这一研究问题.

所谓自媒体账号,是指在微博中作为媒体存在、仅具有媒体属性的账号.自媒体账号向公众推送有价值信息,是典型的内容提供者.自媒体账号的范畴较宽泛,从组织机构(例如"中央人民广播电台"、"平安北京"等)到个人运营的"公众账号"(例如"美食为王"、"微博小百科"等)都属于自媒体账号.由于发布的信息面向公众,自媒体账号一般不会发布个人感受,公众也不关心其个人属性(例如年龄、性别、教育程度等).

研究自媒体账号识别有助于分析微博的整体生态环境、对用户兴趣建模以及发掘高质量内容.研究自媒体账号间的信息流向有助于分析信息传播特征,从宏观上分析微博生态环境.自媒体账号发布内容涉及的话题相对集中,普通用户关注自媒体账号的意图在于订阅优质信息,代表了用户偏好,因此有助于用户兴趣建模.高转发、评论、点赞量的自媒体微博意味着较高的质量,基于这些微博可以发现相似的高质量账号,推荐给感兴趣的用户,因此识别自媒体账号有助于挖掘高质量内容.

本文通过分析新浪微博自媒体账号和普通账号的区别,从账号个人信息、账号行为、账号微博内容 3 个不同角度提取特征,使用有监督学习模型识别自媒体账号.实验表明 3 类特征相互补充,结合使用可以有效地识别自媒体账号,预测准确率高达 96.71%.

1 相关工作

微博账号分类角度:Java 等人[2]将 Twitter 中的用户角色分为信息提供者和信息获取者,基于网络拓扑结构研究了用户在社区中扮演的角色.该工作仅从用户的粉丝数量这一维度对用户分类,粒度较粗.Wu 等人[3]使用基于规则的方法将账号分为明星、官方媒体、博客、机构组织及普通用户 5 类,从测量的角度研究了不同类别用户间的关注、转发关系.该工作研究中提到的官方媒体和组织机构账号对应于本文研究的部分自媒体账号,并且没有将个人运营的自媒体账号包括进研究范围,简单地将它们作为普通用户进行研究.除此之外,由于 Twitter 中分组默认公开,而新浪微博的个人分组默认关闭,简单地沿用基于规则进行统计原则上并不可行.Choudhury 等人[4]将 Twitter 中的账号分为媒体、组织与普通用户,使用账号拓扑属性、内容发布行为、实体识别与新闻话题等信息对 3 类用户自动分类,该工作中的媒体特指官方新闻媒体,与本文中的自媒体有所不同.Yan 等人[5]提出了将 Twitter 账号分为开放和封闭 2 类,并利用账号个人信息与拓扑信息分类.该工作中提及的开放账号专指用于营销自身公司、店铺、产品等的账号,分类角度有细微不同,并且也只使用了 2 类特征,没有充分考虑用户内容,预测准确率不高.

微博普通用户内在属性预测:Rao 等人[6]最早系统地研究了 Twitter 中用户属性预测的方法,使用账号拓扑信息、内容发布频率、N-gram、表情信息预测用户内在属性.Pennacchiotti 等人[7]提出了基于分类模型的通用用户属性预测框架,认定了账号发布内容相关特征对用户属性预测的重要性.Zamal 等人[8]基于同质性假设,利用 Twitter 用户好友提升了用户属性预测的效果.Siswanto 等人[9]使用词法结构信息在非英语环境下对 Twitter 用户进行属性预测.用户属性预测方面的工作面向微博中的普通用户,识别自媒体账号有助于清洗数据.除此之外,用户属性预测的研究方法值得借鉴.丁兆云等人[10]系统地总结了微博数据挖掘的诸多方向,其中也包括微博用户分类的研究方向.文章提及的研究方向主要包括普通用户的内在属性预测以及垃圾用户检测等,并未提及自媒体账号分类这一角度.

主题模型在微博中的应用:Hong 等人[11]提出了使用传统的主题模型对整个微博用户进行建模的方法.Ramage 等人[12]提出了 Labeled-LDA 模型,

为每一条微博内容引入可观测的主题标签. Zhao 等人[13]提出了 Twitter-LDA 模型,引入背景主题,在短文本环境中较 LDA 效果更好,本文使用了训练出的用户话题分布以及话题分布的方差和熵作为特征,并发现使用 LDA 对用户建模、在自媒体账号识别这一任务效果较好.

2 识别自媒体账号的关键信息

媒体,是指传播信息的媒介,传统的媒体包括电视、广播、报纸、杂志等. 自媒体账号,顾名思义,是指微博生态环境中自身作为媒体存在,面向公众发布信息的账号. 在微博中它们扮演着类似于传统媒体的角色,具有鲜明的媒体属性. 自媒体账号涵盖的范围较广,大到各类官方正式的组织机构(例如"中央人民广播电台"、"平安北京①"),小到个人运营的"公众账号"(例如"美食为王②"、"微博小百科③"),只要面向公众发布有价值的信息,都属于自媒体账号.

自媒体账号具有如下特点:1)信息受众的广泛性. 自媒体账号发布的信息是面向用户群的,并不只面向私人,因此优秀的自媒体账号应有较多的听众. 2)微博内容的优质性. 普通用户关注自媒体账号的意图在于获取有用信息,因此优秀的自媒体账号的微博内容质量较高. 3)缺乏社交属性. 用户的关注意图决定了自媒体账号的粉丝并非基于朋友关系,自媒体账号的内在属性(例如性别、年龄、职业等)对关注用户来说并不重要.

根据自媒体账号的特点可以看出,研究自媒体账号能够更好地了解微博的生态环境、通过关注关系理解普通用户的兴趣以及挖掘高质量的微博内容及内容源.

本节根据自媒体账号的特点,提出了用于识别自媒体账号的 3 类关键信息:1)账号的个人信息(PROFILE),主要是基于用户基本属性与个人描述的统计量;2)账号行为(BEHAVIOR),即账号发布微博的行为特点,例如用户发微博时间的熵;3)微博内容(CONTENT),主要指账号发布微博的内容特点,例如账号的话题分布. 对于每一类特征,本文在现有工作的基础上,都提出了新的维度详细刻画用户的特点.

2.1 账号的个人信息

账号的个人信息包括描述账号固有属性的一系列特征. 本文主要使用了基本信息、网络拓扑特征、权限设置以及共现特征. 基本信息方面沿用了以往工作中广泛使用的用户属性,包括昵称的长度、账号自我介绍的长度、是否具有微号、是否具有个性域名、性别、发布微博数量、收藏数量;网络拓扑特征包括关注数量、粉丝数量、双向关注数量、关注数量与粉丝数量之比、双向关注数量占关注数量的比例、双向关注数量占粉丝数量的比例;除了已有工作中用到的用户属性,本文创新性地提出用户权限设置和共现特征. 用户权限设置,包括是否开放私信、是否分享地理位置、是否允许陌生人评论. 下面主要介绍共现特征.

共现特征(使用 PROFILE_LEXICAL 表示,下同):微博提供了自我描述和账号认证原因作为账号的个人描述. 为了精确描述自媒体账号的个人描述与普通账号的差异,本文提出了一系列衡量共现程度的特征. 具体地,共现特征描述了用户昵称和个人描述之间的相似程度. 由于用户的昵称和个人描述都很短,为了减少汉语分词带来的不确定性,本文采用了"字袋"模型的观点,即将账号昵称、账号自我描述、账号认证原因看成单字构成的多重集合,分别用 S_{nick},S_{desc} 和 S_{reason} 表示.

式 1 给出了 2 种不同的相似度计算方法,其中 Jaccard 相似度是用于搜索引擎网页查重的重要度量指标. 表 1 列出了本文提出的共现特征.

$$Jaccard(S_1,S_2)=\frac{|S_1\cap S_2|}{|S_1\cup S_2|},$$
$$Sim(S_1,S_2)=\frac{|S_1\cap S_2|}{|S_1|}. \tag{1}$$

Table 1 Co-occurrence Length and Co-occurrence Ratio

表 1 共现长度与共现比例

Formula	Description
$\lvert S_{\text{nick}}\cap S_{\text{desc}}\rvert$	The number of words that occur in both the nick and the self description
$\lvert S_{\text{nick}}\cap S_{\text{reason}}\rvert$	The number of words that occur in both the nick and the verified reason
$Sim(S_{\text{nick}},S_{\text{reason}})$	The similarity of the nick and the verified reason
$Sim(S_{\text{nick}},S_{\text{desc}})$	The similarity of the nick and the self description
$Jaccard(S_{\text{nick}},S_{\text{desc}})$	The Jaccard similarity of the nick and the self description
$Jaccard(S_{\text{nick}},S_{\text{reason}})$	The Jaccard similarity of the nick and the verified reason

① 系北京市公安局官方微博.

② 与美食杂志、美食节目职能类似.

③ 与科普类电视节目职能类似.

2.2 账号行为

基于账号行为的特征是刻画用户特点的关键维度,不同账号的行为反映了微博作为社交网络还是信息媒体的不同用途.例如,普通账号使用手机终端的比例要高于自媒体账号使用手机终端的比例.本文将基于账号行为的特征$|S_{nick} \bigcap S_{reason}|$分为评价特征、终端特征、时序特征和内容模式特征4类.

1) 评价特征(BEHAVIOR_EVALUATION).是指账号发布内容之后,其他关注者给予的评价行为统计,这些评价行为包括转发、评论、点赞等.这部分特征包括微博转发数的均值、方差与熵,评论数的均值、方差与熵,以及点赞数的均值、方差与熵等.

2) 终端特征(BEHAVIOR_DEVICE).包括账号使用不同微博客户端(包括手机终端、浏览器终端、第三方应用发布)的概率和熵.

3) 时序特征(BEHAVIOR_SEQUENCIAL).是衡量账号发微博的时间特性,包括周末发布微博的比例,每天发布微博数量的均值、方差与熵,微博发布间隔(以min为单位)的均值、方差与熵.

4) 内容模式(BEHAVIOR_PATTERN).用于衡量账号发布微博内容中的一些规律性,但并未涉及对内容的理解.例如账号发布的微博在hashtag、URL、图片上的特点,自转发的频繁程度,转发其他账号提及自己微博的情况等.表2列出了这部分特征及详细描述.

Table 2 Feature on Pattern of Content

表2 账号内容模式特征

Statistics	Description of the Feature	Calculation
POST_LEVEL	The repost level of every tweet	Mean, Variance, Entropy
POST_LEN	The length of the tweet	Mean, Variance, Entropy
HASHTAG_NUM	The number of hashtags that every tweet contains	Mean, Variance, Entropy
URL_NUM	The number of URLs that every tweet contains	Mean, Variance, Entropy
AT_NUM	The number of @ that every tweet contains	Mean, Variance, Entropy
PIC_NUM	The number of pictures that every tweet contains	Mean, Variance, Entropy
HASHTAG	The content of the hashtags that every tweet contains	Entropy
URL	The content of the short urls that every tweet contains	Entropy
URL_DOMAIN	The domain name that the short urls of every microblog correspond to	Entropy
AT	The nickname of every tweet @	Entropy
ORIGINAL	Whether the tweet is an original tweet	Mean
POST_SELF	Whether the reposted tweet is posted by the same user	Mean
POST_AT_SELF	whether the reposted tweet is @ by the user itself	Mean
CONTENT_OVERLAP	The proportion of the words that occur in both the nickname and every original tweet	Mean, Variance, Entropy
HASHTAG_OVERLAP	The proportion of the words that occur in both the nickname and every hashtag	Mean, Variance, Entropy

2.3 账号微博内容

相对于用户个人信息和行为信息,微博内容可以更好地从语义层面理解账号特点,特征也更难抽取.本文从不同层次上抽象账号微博内容特征,包括最直接的用词层面和更高的话题层面.

特征词与特征标签(CONTENT_WORDHASH):直观而言,由于自媒体账号不具有社交属性,在微博用词上会与普通账号有显著区别.Pennacchiotti等人[7]提出了一个基于条件概率的特征词与特征hashtag的抽取方法.这种方法假定不同类别用户微博数量以及长度差异不大.统计分析表明,自媒体账号微博的平均数量更多,且微博的平均长度也更长,因此这一假定并不适用于自媒体账号识别任务.针对这种情况,本文针提出了基于概率似然比的特征词和特征hashtag的提取方法.

式(2)给出了词在用户集合中的典型度、典型度似然比、特征词w_i在用户u中出现的概率,以及用户u中一类特征词出现的概率.

$$
\begin{aligned}
p(w_i, S) &= \frac{|w_i, S|}{\sum_{w_j \in W} |w_j, S|},\\
r(w) &= \frac{p(w, S_+)}{p(w, S_-)},\\
f(w_i, u) &= \frac{|w_i, u|}{\sum_{w_j \in W} |w_j, u|},\\
c(u) &= \frac{\sum_{w_i \in F} |w_i, u|}{\sum_{w_i \in W} |w_i, u|},
\end{aligned}
\tag{2}
$$

其中,w 是词,S 表示账号集合,S_+ 为正例集合(也即自媒体账号集合),S_- 为负例账号集合,$|w,S|$ 为词语 w 在 S 中所有用户微博构成的词袋中出现的次数.所有的词进行长尾过滤,再根据典型度似然比 $r(w)$ 进行排序(正例特征词从大到小排序,负例特征词从小到大排序),各取排序结果的前 K 个形成 2 个特征词集合 F,W.

对每一个账号,使用其微博文本估计特征词出现的概率 $f(w_i,u)$ 作为用户特征.除单个特征词之外,再分别计算所有正例特征词和所有负例特征词在账号微博文本出现的比例 $c(u)$.每个用户共计算 $2K+2$ 个特征,分别来源于 $2K$ 个特征词以及 2 个特征词集合.特征 hashtag 的特征计算方法与特征词的特征计算方法完全一致.本文选取特征词中的 $K=300$,对于特征 hashtag,由于数据的不平均,最终正例特征 hashtag 选取了 178 个,负例特征 hashtag 最选取了 58 个.

特征词与特征 hashtag 只能从词义的角度理解微博内容.为了获取话题层次上的抽象,本文使用了 LDA 模型[14]和 Twitter-LDA 模型对用户话题建模.

1) 基于 LDA 模型的用户话题分布(CONTENT_LDA).为从更高层次上抽象微博文本,本文采用了 LDA 模型估计账号的话题偏好.传统的 LDA 模型将一篇文章看成一个文档,本文将一个账号发布的所有微博看作一个文档,假定该文档可以被表示成关于主题的多项分布.实验采取了基于吉布斯采样的 LDA 实现,对所有用户的微博进行话题抽取.话题数量为 50①,迭代次数为 500 次,词典大小为 98 189.对每一用户,计算在 K 个话题下的概率分布以及该分布方差和熵作为特征.

2) 基于 Twitter-LDA 模型的用户话题分布(CONTENT_TLDA).除了传统的 LDA 模型,本文也使用了 Zhao 等人[13]针对微博短文本特性提出的 Twitter-LDA 模型.Twitter-LDA 模型引入了背景话题,生成微博中词的过程不仅可以从当前选定的话题中选取,也可以从背景话题中选取,使用哪个话题生成该词取决于一个伯努利分布.实验采取了基于吉布斯采样的 Twitter-LDA 实现,对用户话题建模.话题数量 $K=50$,迭代次数为 500,词典大小为 98 189.对每一用户,计算在 K 个话题下的概率分布及该分布的方差和熵作为特征.

3 实验评测

3.1 实验设置

本文实验数据的采集使用了新浪微博提供的 API 接口.先人工选取了 150 个种子账号,基于关注关系和宽度优先搜索策略采集候选账号集合.实验共采集 3 032 796 个候选账号,2 名数据标注人员从候选账号集合中随机选取账号进行二值标注,最终收集到 2 人标注结果一致的账号 7 697 条,其中正例(自媒体账号)3 892 条,负例 3 805 条.实验采集了标注集合中所有账号的个人信息、最近发布的 500 条微博数据,实验共采集到 2 892 733 条微博数据.

在抽取微博内容相关特征过程中,实验对微博文本内容进行了繁简转换、切词处理、停用词过滤及标点符号过滤等预处理步骤.

实验采用支持向量机(support vector machine,SVM)分类模型②,使用十折交叉验证方式对数据训练与预测,采用精确度(precision)、召回率(recall)、F 值(F-measure)以及分类准确率(accuracy)③作为评测指标.

3.2 实验结果与分析

3.2.1 整体效果分析

表 3 展示了使用第 2 节描述的特征集合的分类效果,表格分为 4 个比较组.

第 3 组结果显示,CONTENT_LDA 的效果最好,但与 CONTENT_TLDA 相比优势不大,反而 Twitter-LDA 模型的效果比 LDA 要差,这是由于自媒体账号发布话题较为统一,因此将账号的所有微博看成一篇文档也是合理的;CONTENT_WORDHASH 效果不明显,说明高层次的文本抽象具有更强的区分度.

第 4 组对比了三大类不同特征集合的分类效果.BEHAVIOR_ALL 的分类效果最好,其次是 CONTENT_ALL.ALL 相比每一类特征集合都具有更好的效果,并且高出很多(分别是 11.7%,6.2%,8.6%),说明多个不同的角度提取的特征之间相互补充,能够有效识别自媒体账号.

① 本文在话题数量分别为 20,50,100 的情况下,评测了 LDA 和 Twitter-LDA 实验效果与账号识别分类效果,结果表明话题数量为 50 时效果最好.

② 本文使用了一系列不同的分类模型进行训练,包括 K-近邻、决策树、RBF 神经网络、支持向量机、随机森林和 AdaBoost,实验发现采用支持向量机模型的效果最好,分类效果以支持向量机模型举例展示.

③ 本文提到的分类准确率是指所有被分类正确的账号比例(包括自媒体账号和非自媒体账号),而精确度和召回率则是以识别自媒体账号的角度定义的.

Table 3 Classification Performance

表 3 整体分类效果

Feature Set	Precision	Recall	*F*-measure	Accuracy
PROFILE_CLASSIC	0.718 9	**0.799 6**	0.757 1	0.743 4
PROFILE_LEXICAL	**0.820 5**	0.743 3	**0.780 0**	**0.790 3**
BEHAVIOR_EVALUATION	**0.874 5**	0.792 9	0.831 7	0.839 5
BEHAVIOR_DEVICE	0.757 8	**0.876 9**	0.813 0	0.798 3
BEHAVIOR_SEQUENCIAL	0.813 6	0.758 2	0.784 9	0.792 3
BEHAVIOR_PATTERN	0.852 3	0.862 8	**0.857 5**	**0.856 6**
CONTENT_WORDHASH	0.674 6	**0.956 6**	0.791 2	0.747 6
CONTENT_LDA	**0.885 0**	0.850 2	**0.867 3**	**0.869 9**
CONTENT_TLDA	0.854 8	0.852 8	0.853 8	0.853 9
PROFILE_ALL	0.867 7	0.825 8	0.846 2	0.849 9
BEHAVIOR_ALL	**0.910 9**	**0.898 5**	**0.904 7**	**0.905 3**
CONTENT_ALL	0.886 8	0.873 3	0.880 0	0.880 9
Total	**0.974 4**	**0.959 4**	**0.966 9**	**0.967 1**

虽然整体分类效果较好,研究分类失败的特例发现,自媒体账号识别的主要失败原因有:1)账号原本是个人账号,后来改头换面成为自媒体账号,而遗留的微博部分刻画了成为自媒体账号之前的个人属性;2)账号是个人账号,除了发布特定内容,也发布人心情等,但是更倾向于发布特定内容,这部分账号在标注过程中认为是个人账号,但实际上也具有较强的自媒体属性.

3.2.2 特征重要性分析

实验使用了信息增益对分类特征重要性排名.

表 4 给出最具有区分度的 20 个特征,其中 PROFILE_ALL 中的特征有 3 个,均来自 PROFILE_LEXICAL;BEHAVIOR_ALL 中的特征有 10 个,可以看出虽然设备相关特征和时序特征单独用于分类时效果并不好,但也有着比较重要的权重;CONTENT_ALL 中的特征有 7 个,其中负例特征词的比例和负例特征 hashtag 的比例、LDA 的 2 个话题分布都从有效刻画负例的角度识别自媒体账号,这也说明账号发布内容特征与其他 2 类特征是相互补充的.

根据特征排名的重要程度可以发现,自媒体账号更加倾向于使用第三方定时应用发布微博(例如皮皮时光机等),发布微博更倾向于具有规律性(时间间隔固定),倾向于发布原创长微博(原创微博比例高于一般用户、微博长度高于一般用户),转发内容一般层数较少(原创或一层转发),并且发布微博中更容易与昵称相关(原创微博中出现昵称比例的平均值较高).

Table 4 Rank of Feature Importance

表 4 特征重要性排名

Rank of Features	Meaning
1	The proportion of the negative words that occur in the tweet
2	The distribution probability of the 43rd topic got from LDA(users' personal feelings)
3	The average length of the posted tweets
4	The proportion of the words that occur in both the nick and the self description
5	The distribution probability of the 43rd topic got from Twitter-LDA(topics about IT, digital and science)
6	The entropy of the reposting level of the tweets
7	The entropy of the length of the tweets
8	The number of the words that occur in both the nick and self description
9	The mean of the proportion of weibos that the nick of the user occurs in its original tweet
10	The probability of posting tweets by using the iphone client
11	The distribution variance of the topic of LDA
12	The distribution probability of the 49th topic got from Twitter-LDA(topics about the emotions of users and wish)
13	The distribution variance of the topic of Twitter-LDA
14	The mean of the reposting level of the tweets
15	The probability of posting tweets by using the application called pipi time machine
16	Whether the verified reason of the account contains the words like official microblog
17	The distribution variance of the posting level of the tweets
18	The proportion of negative hashtags that occur in tweets
19	The proportion of original tweets
20	The entropy of the time interval of the posting of tweets

进一步地,表5列出了对分类效果有突出贡献的负例特征词.这些词语多是表达语气和情感的形容词、副词,说明非自媒体账号具有更强的社交属性.负例特征hashtag的影响很小但也远大于正例特征hashtag.典型的hashtag包括"让红包飞"、"带着微博去旅行"、"Weico拼图"、"随手拍"、"我在看新闻等",这些hashtag多是用户主动参与平台活动或是第三方个人应用发布微博提供的.

Table 5 Typical Negative Words

表5 典型负例特征词

Rank of Negative Words	Negative Words
1	泪
2	终于
3	回来
4	真
5	生日
6	死
7	真是
8	晕
9	开心
10	好久

4 总结与展望

本文首次提出了自媒体账号识别这一研究问题,阐述了这一问题的研究意义,并利用账号个人信息、账号行为以及账号微博内容3类信息和有监督学习方法识别自媒体账号.实验结果表明,不同的特征集合之间互相补充,有效结合可以较大程度地提升自媒体账号的识别效果.

下一步研究工作集中于结合自媒体账号的话题分布与社交网络拓扑结构,对非自媒体兴趣建模.

参 考 文 献

[1] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media? [C] //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 591-600

[2] Java A, Song X, Finin T, et al. Why we twitter: Understanding microblogging usage and communities [C] // Proc of the 9th WebKDD and the 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. New York: ACM, 2007: 56-65

[3] Wu S, Hofman J M, Mason W A, et al. Who says what to whom on Twitter [C] //Proc of the 20th Int Conf on World Wide Web. New York: ACM, 2011: 705-714

[4] De Choudhury M, Diakopoulos N, Naaman M. Unfolding the event landscape on Twitter: Classification and exploration of user categories [C] //Proc of the 2012 ACM Conf on Computer Supported Cooperative Work. New York: ACM, 2012: 241-244

[5] Yan L, Ma Q, Yoshikawa M. Classifying Twitter users based on user profile and followers distribution [G] // Database and Expert Systems Applications. Berlin: Springer, 2013: 396-403

[6] Rao D, Yarowsky D, Shreevats A, et al. Classifying latent user attributes in Twitter [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media. Menlo Park, CA: AAAI, 2011: 281-288

[7] Pennacchiotti M, Popescu A M. A machine learning approach to Twitter user classification [C] //Proc of the 2nd Int Workshop on Search and Mining User-Generated Contents. New York: ACM, 2010: 37-44

[8] Al Zamal F, Liu W, Ruths D. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors [C] //Proc of the 6th Int AAAI Conf on Weblogs and Social Media (ICWSM 2012). Menlo Park, CA: AAAI, 2012: 387-390

[9] Siswanto E, Khodra M L. Predicting latent attributes of Twitter user by employing lexical features [C] //Proc of the 5th Int Conf on Information Technology and Electrical Engineering. Piscataway, NJ: IEEE, 2013: 176-180

[10] Ding Zhaoyun, Jia Yan, Zhou Bin. Survey of data mining for microblogs [J]. Journal of Computer Research and Development, 2014, 51(4): 691-706 (in chinese)
(丁兆云, 贾焰, 周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展, 2014, 51(4): 691-706)

[11] Hong L, Davison B D. Empirical study of topic modeling in Twitter [C] //Proc of the 1st Workshop on Social Media Analytics. New York: ACM, 2010: 80-88

[12] Ramage D, Dumais S T, Liebling D J. Characterizing microblogs with topic models [C] //Proc of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM 2010). Menlo Park, CA: AAAI, 2010: 130-137

[13] Zhao W X, Jiang J, Weng J, et al. Comparing Twitter and traditional media using topic models [G] //Advances in Information Retrieval. Berlin: Springer, 2011: 338-349

[14] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022

Liu Jinbao, born in 1987. Master from Peking University. His main research interests include social network analysis and data mining on social media.

Sheng Dakui, born in 1991. Master candidate from Peking University. His main research interests include social network analysis and data mining on social media.

Zhang Ming, born in 1966. Professor and PhD supervisor in Peking University. Senior member of China Computer Federation. Her main research interests include text mining, data mining on social media, and data mining on massive open online courses.

2016 年《计算机研究与发展》专题征文(正刊)

——面向“互联网+”的未来网络理论、体系结构与应用

“互联网+”用互联互通的办法提高社会运行效率,以互联网为纽带、以网络思维为基础,结合生产、生活等社会运行的各个方面.“互联网+”已经获得了初步的成功,新的应用如神州出行、微银行等正在改变人们的生活.“互联网+”未来潜力不可限量.在这个背景下,不但我们需要新的“互联网+”的应用,而且也需要思考新的互联网理论、互联网体系结构.特别地,“互联网+”以互联网为基础,但是更强调互联网和其他网络、应用之间的关系,以及其他网络、应用在以互联网为纽带下之间的关系.

《计算机研究与发展》将于 2016 年 4 月出版面向“互联网+”的未来网络理论、体系结构与应用专题.以前网络编刊或着眼于综合型的互联网体系结构(如下一代互联网体系结构),或着眼于互联网中的性能问题(如软件定义网络),或着眼于互联网中某一部分(如边缘的物联网).本刊和上述在方法上虽有重叠,但是着重思考互联网+下的各种挑战,希望能系统地汇集一批专注于互联网+的学术文章.

特别地,2016 年《计算机研究与发展》面向“互联网+”的未来网络理论、体系结构与应用专题,将侧重于互联网下网络经济、工业互联网、跨学科网络、智慧城市、社交网络等方面的论文.“互联网+”同时也强调互联网怎样更方便、更合理地支持这些网络和应用,因此也欢迎这方面的文章.欢迎相关领域的专家和学者和科研人员踊跃投稿.

现将专题论文征集的有关事项通知如下:

征文范围(但不限于)

- 互联网+的新理论、网络经济、网络科学
- 互联网+的新体系结构、跨学科网络系统
- 互联网+环境下的工业互联网
- 互联网+的新应用:智慧城市、社交网络、工业网络、新医疗
- 互联网+环境中的隐私与安全
- 网络测量和数据科学

征文要求

1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值与推广应用价值,未在国内外公开发行的刊物或会议上发表,不存在一稿多投问题.作者在投稿时,需向编辑部提交投稿声明.
2) 论文一律用 Word 格式排版,论文格式体例参考近期出版的《计算机研究与发展》的要求(http://crad.ict.ac.cn/).
3) 论文通过期刊网站(http://crad.ict.ac.cn)投稿,投稿时提供作者的联系方式,备注中务必注明“网络技术 2016 专题”(否则按自由来稿处理).

重要日期

征文截止日期: 2015 年 12 月 10 日
录用通知日期: 2016 年 1 月 10 日
修改稿提交日期:2016 年 1 月 20 日
出版日期:2016 年 4 月

特邀编委:

王兴伟 教授 东北大学 wangxw@mail.neu.edu.cn
王 丹 副教授 香港理工大学 dan.wang@polyu.edu.hk
崔 勇 教授 清华大学 cuiyong@tsinghua.edu.cn

联系方式

编辑部: crad@ict.ac.cn, 010-62620696,010-62600350
通信地址: 北京 2704 信箱《计算机研究与发展》编辑部
邮政编码: 100190