ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 117－125
DOI:10.16337/j.1004-9037.2015.01.011

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于微博重复发送的垃圾用户甄别

吴 斌 李冠辰 刘 宇 张 雷 王 柏

(北京邮电大学计算机科学学院,北京,100876)

摘 要:针对微博平台上的垃圾用户甄别问题,本文提出了基于微博重复发送行为的垃圾用户行为建模和甄别算法。在真实微博垃圾用户数据分析的基础上,本建模方法综合考虑了微博垃圾用户的行为信息、社交网络信息和文本信息,从不同的角度对垃圾用户进行了分析和建模。在真实数据集上的实验证明了方法的有效性,并且对模型中若干参数进行了优化,同时也分析了垃圾用户行为信息、社交网络信息和文本信息对模型的影响程度。

关键词:垃圾用户检测;微博重复发送;主题模型

中图分类号:TP391　　**文献标志码**:A

Spammer Detection Based on Duplicate Microblog Post

Wu Bin, Li Guanchen, Liu Yu, Zhang Lei, Wang Bai

(School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, 100876, China)

Abstract: As the social networks (such as Sina Weibo) become more and more popular and significant, spammer′s behavior severely affects the credibility and readability on social network platforms. A spammer detection model along with the algorithm is proposed based on duplicate detection of microblog posts to detect spammers on Weibo platform. Based on analyses of real-world data, the model is built by considering user behavior information, user social network information and content information. Experiments on a collection of real Weibo data shows the effectiveness of the proposed model. Parameters′ impaction to the model is also studied. The improvement of incorporating behavior information, content information and network information has been analyzed, hence the model is promising.

Key words: spammer detection; duplicate microblog post; topic model

引　　言

近几年,新浪微博、Twitter 等微博平台已经成为了互联网用户重要的信息分享平台。在这里,人们关注时事新闻,并分享关于实事和自身事件的看法和评论。随着微博用户的增长,垃圾用户的出现不可避免。垃圾用户的评论和转发淹没了正常用户的意见;有些垃圾用户发布的微博带有虚假信息、链入含木马的钓鱼网站,对用户的计算机安全造成危害;有些垃圾用户伪装成微博大号来吸引粉丝以便发布广告和谣言;充斥着垃圾信息的社交网络还对有效的信息获取和网络舆情监测及演化分析[1]带来重大影

基金项目:国家自然科学基金重点(71231002)资助项目;北京市教育委员会共建项目专项基金资助项目。
收稿日期:2014-03-25;**修订日期**:2014-05-12

响。使用机器学习的方法对微博平台中的垃圾用户予以甄别,不仅可以提升微博的用户体验,还可以改善社交网络环境。一个基本的垃圾用户甄别思路是垃圾用户没有和真实用户相同的社交网络[2],垃圾用户会表现出两个特点:一是垃圾用户不能获得很多可信的真实粉丝[3],二是垃圾用户使用 link-farming[4,5]手段获取很多非正常粉丝。Boykin 等人[6]研究了社交网络垃圾用户中的杠杆效应,对社交网络垃圾用户做了全面的分析; Hull 等人[7]研究了社交网络中用户间的作用关系,完成了社交网络中的垃圾用户甄别;Yardi 等人[8]爬取了 Twitter 中的数据,利用 Twitter 中垃圾用户发布微博带有"#"话题,但是微博内容却和"#"的内容相悖的方法获取 Twitter 中的垃圾用户数据,抽取了大量属性对垃圾用户做了分析和甄别;Gao H 等人[9]分析了社交网络垃圾用户的分布和爆发两个方面,并设计了新的方法来探测和甄别垃圾用户;Gómez V 等[10]对社交网络中垃圾用户的特点进行了大量、全面的统计分析,并直观地展示出社交网络中垃圾用户的特点。虽然垃圾用户甄别工作在 Twitter[5],Facebook[11]和 YouTube[12]中的研究已经相当成熟,但是针对国内社交网络的垃圾用户研究工作尚未成型。本文主要研究新浪微博平台的垃圾用户问题,基于微博重复发送检测,提出了一个新的垃圾用户检测模型。

1 垃圾用户分析

1.1 数据集

本文使用网页爬虫来爬取新浪微博举报大厅(http://service.account.weibo.com/)中已被证实的垃圾用户。从举报大厅直接爬取一定数量的垃圾用户;随机选择一个用户,按照其粉丝关系爬取用户扩充为随机用户集;使用 Weibo API 获取用户发布的微博,按照微博转发关系,爬取垃圾用户转发微博的原用户及他们的微博。数据信息统计如表 1 所示。

表 1 本文使用的数据统计

Table 1 The outline of the Weibo dataset

用户类型	用户数量	微博数量
随机用户	10,394	1,310,940
随机用户转发源用户	54,023	65,231
骚扰他人者	1,747	203,390
不实信息者	2,813	401,231
冒充他人者	366	49,093
垃圾用户转发源用户	9,870	12,865

1.2 微博垃圾用户微博发布行为分析

本节对微博垃圾用户的微博发布行为特点进行分析。直观的推断如下:(1)垃圾用户经常重复转发或发布同一条微博,因此相邻微博的相似度很高。(2)垃圾用户连续转发或发布微博,相邻微博的发布时间临近。基于这两个假设,对数据集进行分析,为模型的建立提供方向。

(1)分析用户的文本重复度。基于相邻微博 m_1 和 m_2 的最长公共子串(Longest common subsequence,LCS)长度,建立某用户所有微博的相似度评价,并按照公式进行统计计算。

$$LCS-\text{Multiplicity}=\frac{\sum_{i=1}^{i=n-1}\left(\frac{\text{Length}[LCS(m_1,m_2)]}{[\text{Length}(m_1)+\text{Length}(m_2)]/2}\right)}{n-1} \tag{1}$$

从图 1 的统计计算结果可以看出垃圾用户的连续微博间的相似度高于随机用户(正常用户)。从表 2 的统计结果,无论是平均值还是中位数,垃圾用户的连续微博文本相似度远高于随机用户。垃圾用户和随机用户的连续微博的文本重复度有很大差异,垃圾用户的行为特点可以作为模型的切入点。

(2)对垃圾用户和随机用户的相邻微博发布时间间隔进行分析。图 2 是垃圾用户和随机用户的连续

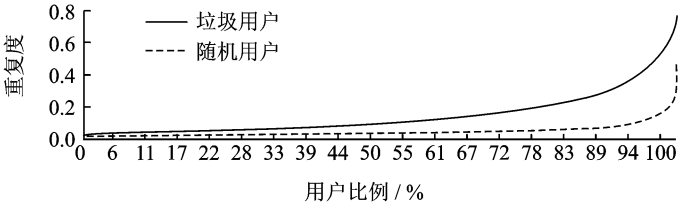

图 1　随机用户和垃圾用户的 LCS 重复度比较

Fig. 1　LCS duplication comparison between random users and spam users

表 2　随机用户和垃圾用户的 LCS 重复度比较

Table 2　LCS duplication comparison between random and spam users

基于 LCS 的微博文本重复度	随机用户	垃圾用户
平均值	0.042 1	0.123 9
中位数	0.031 5	0.073 8

微博的发布时间间隔的比较,图中 y 轴是微博用户发微博间隔(min)的倒数,此值越大,微博的间隔越短。

图 2 显示了随机用户和正常用户的发送时间间隔对比。可以看出,垃圾微博用户会短时间内发布大量微博,对微博平台和社交网络环境造成影响。

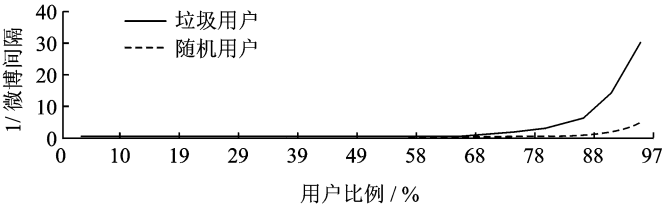

图 2　随机用户和垃圾用户的微博发送间隔比较

Fig. 2　Comparison of posting time gap between random and spam users

表 3 给出了随机用户和垃圾用户的微博发布间隔时间的对比,只有不超过 0.6%的随机用户发微博的平均间隔小于 10 min,而在相同微博发布间隔时间内的垃圾用户的比例却高达 7.4%。这说明在间隔时间较短的情况下,垃圾用户的发布微博行为特征非常明显,即与随机用户相比,垃圾用户更加倾向于连续地发布微博。

表 3　随机用户和垃圾用户的微博间隔时间

Table 3　Weibo posting time gap of random and spam users

微博间隔时间/min	随机用户	垃圾用户
	少于 X 的比例	
1	0.16	3.09
2	0.27	4.54
3	0.34	5.26
4	0.38	5.77
5	0.43	6.16
6	0.45	6.54
7	0.49	6.81
8	0.51	7.02
9	0.53	7.22
10	0.56	7.40

2 基于微博重复发送的垃圾微博甄别模型

2.1 重复检测

重复检测主要应用在数据库中重复和相似记录的检测，避免主键不同的数据出现重复。最近一段时间，重复检测被越来越多的研究者应用到垃圾行为检测中。Nitin Jindal 等人[13]利用重复检测技术甄别评论系统中的垃圾评论；Zhe Wang 等人[14]将重复检测应用在在线图片系统中，侦测重复和相似的图片上传；Draisbach U 等人[15]通过自适应时间窗口来检测文本列表的重复度。赵斌等[16]利用重用检测来甄别微博垃圾消息；Q Zhang 等[17]基于 Hadoop 平台和重复检测算法，甄别了社交网络平台中的垃圾消息。

已有的重复检测方法在社交网络中的应用存在如下的几个问题：

(1)目前的国内外工作主要从微博文本的相似度出发[17]，或者加上时间窗口，考虑用户的行为特点[16]。而微博垃圾用户所发布的微博文本本身就具有明显的特征[8,16]。本文将利用微博文本信息提高垃圾用户检测模型的有效性。

(2)文献[16]只考虑了用户转发的微博，而文献[17]只考虑了用户的原创微博。而实际的微博平台中，垃圾用户不仅转发他人微博，还会有自行原创微博来进行广告等垃圾活动。本文综合考虑转发微博和原创微博，并使用参数调整两者之间的重要程度，可以提升垃圾用户甄别算法的效果。

针对以上的问题，本节提出了基于重复检测的微博垃圾用户检测模型(Spammer detection model, SDM)，此模型不仅考虑了微博用户的行为信息，还结合了文本信息和社交网络信息，增强了模型的效果。

2.2 SDM 模型流程

对一个微博用户的所有微博，按照转发的源用户进行分类，每一条被转发微博形成一个微博列表，并对此微博列表应用重复检测。SDM 模型计算流程如下：

(1)如果用户 u 所转发的微博为用户 v 所发布，则用户 v 构成转发源用户集合 V。对每一个 u 及 $v \in V$，构建按照时间排序的用户原创微博序列 L_{uu} 以及 u 转发 v 的微博序列 L_{uv}。

(2)对序列和中每一条微博进行预处理，去掉“＃＃”话题标签，“@”符号，短链接和表情符号等之后进行分词处理。

(3)对步骤(1)中得到的 L_{uv} 和 L_{uu} 中每一对相邻的微博，计算微博重复度函数 $F_m(m_i, m_{i+1})$。其中包含微博文本相似度 $Sim(m_i, m_{i+1})$ 和微博文本垃圾程度度量函数 $F_s(m_i, m_{i+1})$。

(4)对 L_{uv} 中每一对微博的重复度求平均值再求和，并与 L_{uu} 的微博重复度平均值组合，使用参数 α 调整两者比例，并除以总转发用户数加一进行归一化。

2.3 SDM 模型定义

首先定义垃圾用户检测函数

$$F_u(u) = \frac{\sum F_{uv}(v) + \alpha F_{uu}(u)}{|V| + 1} - \theta \tag{2}$$

由用户 u 转发的微博的源用户 v 形成一个集合 V，针对每一个 $v \in V$，计算 u 转发 v 的微博序列的重复度 $F_{uv}(v)$。$F_{uu}(u)$ 是用户发布的原创微博的重复性评估函数。$F_{uu}(u)$ 和 $F_{uv}(v)$ 的不同是 $F_{uu}(u)$ 没有用户重要性的度量因子，只有微博的重复度。α 是控制用户原创微博重复度和转发微博重复度的参数，默认值为 1。θ 是判别函数的阈值，当用户的微博重复相似性大于 θ 时，模型认为此用户是垃圾用户，反之则为正常用户。此 SDM 模型主要考虑了用户的转发微博和原创微博的重复性和文本信息等。

2.4 用户网络信息建模

定义函数 $F_{uv}(v)$，见式(3)。$F_{uv}(v)$ 针对用户 u 的每一个被转微博源原用户 v，都将有一个函数来评

价用户 u 对于用户 v 的转发行为的垃圾程度。

$$F_{uv}(v)=\frac{\sum_{i=1}^{N_m-1}F_m(m_i,m_{i+1})}{N_m-1}\times F_v(v) \tag{3}$$

$$F_{uu}(u)=\frac{\sum_{i=1}^{N_m-1}F_m(m_i,m_{i+1})}{N_m-1} \tag{4}$$

式中：$\frac{\sum_{i=1}^{N_m-1}F_m(m_i,m_{i+1})}{N_m-1}$ 称之为微博序列重复度评估函数，是对用户 u 的所有转发自用户 v 的微博集合的重复度的评估。针对 u 转发 v 的每两条连续微博，计算两条微博的重复度 $F_m(m_i,m_{i+1})$，并除以微博数减一来进行归一化。

定义源用户评估函数为 $F_v(v)$，是对用户 v 的网络信息重要度的评估。β 是重要度函数的参数，初始值设置为 1。

$$F_v(v)=(1+\frac{FC(v)}{\mathrm{Max}[FC(V)]}+\frac{SC(v)}{\mathrm{Max}[SC(V)]})^{\beta} \tag{5}$$

其中 $FC(v)$是用户 v 的粉丝数(Follower count)，$SC(v)$是用户 v 的微博数(Status count)。SDM 模型主要考虑微博用户的重要性，基于垃圾微博的情况，大量的广告垃圾微博是转发自粉丝很多的官方运营账号，是官方运营账号为了推广而雇佣的水军。式(5)中参数 β 是调整微博账号重要程度的参数。β 越大，说明源微博用户重要度对模型的影响越大。

2.5　用户行为和文本信息建模

由于微博文本，特别是本文所采集的新浪微博文本存在如下两个特点：

(1)文本短小，新浪微博的文本被限制在 140 个字符，计算标点符号和“@”“＃”等微博功能符号，有意义的文本很少。

(2)微博话题多变，新、热点事件很多，可能会造成文本特征维度过高。

针对微博短文本的这两种特点，特别是文本短小特点，学术和工业界的方法一般有两种：第一类是借助外部文本如搜索引擎结果，扩展短文本；第二类是借助知识库，如 WordNet 或 Wikipedia 等，挖掘短文本中词语的内在联系。

用于文本分类的机器学习算法主要有 SVM，Bayes，KNN，LLSF 和决策树等[18]。本文则利用 LDA 模型，从主题维度对微博短文本进行建模；首先，垃圾用户和正常用户的主题分布差别较大；其次，本文研究垃圾用户甄别，是一个二分类任务，不需要十几个甚至上百个的文本分类中需要的充足特征，LDA 模型建模的结果已经可以满足要求。

微博重复度函数 $F_m(m_i,m_{i+1})$ 中包含了微博用户的行为信息和文本信息。垃圾微博用户的行为信息和文本信息与正常用户差异较大。SDM 模型在两条连续微博的重复度度量中考虑了相似度和时间的因素来体现用户行为信息，并用 LDA 建模的微博垃圾程度度量函数 $F_s(m_1,m_2)$ 来定义微博的文本信息，通过对训练集的观测和训练，垃圾程度度量函数 $F_s(m_1,m_2)$ 可以较好地体现一条微博是垃圾微博的概率。

$F_s(m_1,m_2)$ 称为度量微博垃圾程度函数，定义为

$$F_s(m_1,m_2)=\Phi(m_1+m_2)\times\Psi(T_M) \tag{6}$$

其中 $\Phi(m_1+m_2)$ 是微博的文本的 LDA 模型主题向量，$\Psi(T_M)$ 是从训练集数据中统计出的每个主题维的垃圾微博条件概率，即针对 LDA 建模中的每一个主题维度，有如下统计计算

$$\Psi(t)=\frac{\sum_{u\in \text{spammer}} t_u/Num}{\sum_{u\in \text{all}} t_u/Num} \tag{7}$$

对于每一个主题，如果垃圾用户的概率较高，说明此主题多为垃圾用户转发或者发布；而如果是随机用户概率较高，则说明此主题较少被垃圾用户的微博所涉及。

表 4 随机用户和垃圾用户概率最高的主题分布

Table 4 Top 5 topic probability distributions

排名	随机用户	垃圾用户
1	孩子，妈妈，宝宝 男人，爸爸，儿子	机会，获得，活动 抽奖，粉丝，免费
2	自己，我们，没有 生活，人生，事情	婚纱，摄影，喜庆 新人，美丽
3	世界，旅行，自己 一起，这个，遇见	鼓掌，围观，开心 礼物，蛋糕，蜡烛
4	美国，市场，经济 银行，投资，公司	推荐，地址，详情 点击，链接，看看
5	中国，律师，新闻 社会，人民，国家	游戏，大家，一起 免费，快乐，投票

本文对数据集内各微博主题概率结果进行了统计，从表 4 可以直观地看出，垃圾用户和随机用户的微博文本主题具有较大差别，随机用户概率高的主题中，包含了亲子、旅行、经济时事等内容，而垃圾用户概率较高的主题中，包含了抽奖活动推广、婚纱摄影推广、地址推荐等内容。这些统计说明了垃圾用户和随机用户在微博文本的主题分布上有较大的区别，将微博文本信息特征用在模型 SDM 中可行。

函数 $F_s(m_1,m_2)$ 是 $F_m(m_i,m_{i+1})$ 函数中的一个因子，$F_m(m_i,m_{i+1})$ 定义如式(8)，其中 m_1，m_2 是两条连续的微博。

$$F_m(m_1,m_2)=\frac{Sim(m_1,m_2)\times F_s(m_1,m_2)}{\ln(t_2-t_1)} \tag{8}$$

其中 $Sim(m_1,m_2)$ 是两条微博文本用 LDA 主题模型建模的余弦相似度。而 t_1 和 t_2 分别是 m_1，m_2 的发布时间。

本文提出的 SDM 检测模型综合考虑了垃圾用户的微博发布行为信息、用户关系网络属性信息等，微博重复度度量函数 $F_m(m_i,m_{i+1})$ 评估垃圾用户的行为信息，即连续两条微博之间文本相似度越高、发布时间越相近，则这条微博是垃圾微博的概率越高。还通过因子 $F_s(m_1,m_2)$，即垃圾文本概率评估函数来评估垃圾用户的文本信息，统计计算微博文本的垃圾微博概率。$F_v(v)$ 作为源用户评价函数，主要评估用户的转发社交网络关系和 $F_m(m_i,m_{i+1})$ 微博相似度配合，即一个用户高频率转发一个大号的微博，就很可能是一个垃圾用户。

3 实验和结果分析

3.1 算法验证

在 SDM 模型 $F_u(u)$ 中，$\dfrac{\sum F_{uv}(v)+\alpha F_{uu}(u)}{|V|+1}$ 是对一个微博用户的垃圾程度进行评估的函数，θ 为判别是否为垃圾用户的阈值。其他两个参数在最优情况下，θ 从小到大按照步长 0.001 进行增大，观测对判定某用户是否属于垃圾用户集合的准确率和召回率的变化。

从图 3 中可以看到,随着 θ 的增大,准确率一直增大到趋近 1,召回率则从接近 1 一直降低,当 $\theta=0.11$时,得到 F 值最大值为 0.918。结果说明绝大多数的垃圾用户的垃圾程度评估值高于正常用户,SDM 模型对垃圾用户甄别的建模是成功的。

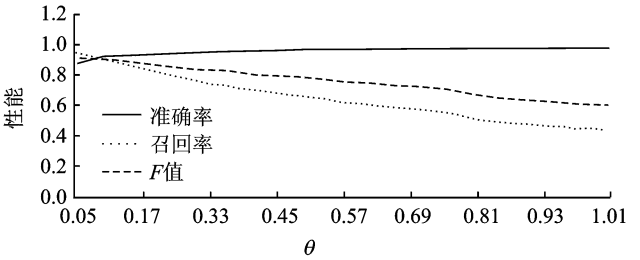

图 3　不同 θ 取值时算法性能评价

Fig. 3　Comparison of algorithm performance with θ variation of

3.2　参数的调优

式(2)中 α 是控制用户转发微博和原创微博重要性的参数,α 越大,表示垃圾用户的原创微博的危害程度高于转发微博。式(5)中 β 是控制转发源用户因子的参数,β 越大,表示垃圾用户转发源用户的越重要转发微博垃圾危害越大,反之则表示转发微博危害越小。

首先进行实验验证参数 α 对模型性能的影响,也就是探究垃圾用户和正常用户之间,原创微博的重复度的区分度。α 是调整用户转发微博重复度和用户原创微博重复度的一个参数。α 越大,说明用户原创微博的重复度在 *SDM* 模型中的影响越大。

本实验是将 α 从 0 开始增大,在每个 α 的取值下,再计算 F 值最优的 θ,得到算法性能如图 4 所示。经过分析可知,在 *SDM* 检测算法中,即使是垃圾微博用户,自己发布的微博的重复性也较低,在整体的算法中影响比转发微博小。统计结果显示,在 $\alpha=0.13$ 时,F 值最大为 0.918。

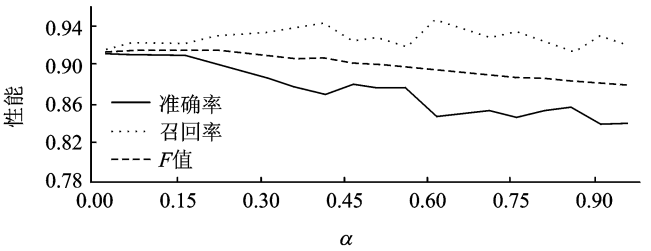

图 4　不同 α 取值的算法性能评价

Fig. 4　Comparison of algorithm performance with variation of α

经过对数据的审视之后,认为由于垃圾微博用户很少发布原创微博,原创内容较少,即使有,重复度也不高,因此 SDM 模型中,转发评估函数部分较为重要,而用户原创微博的重复度重要度较低。

β 是用户转发微博的源用户的重要度评估函数,越大,说明源用户的重要度在 SDM 中的影响越大。

令 $\alpha=0.13$,β 由 0 开始增加,在 β 取不同的值时计算算法的准确率、召回率和 F 值,结果如图 5 所示。从结果可以看出,如果 $\beta=0$,对结果的影响不大,随着 β 增大,算法的 F 值也之增大,其间 β 有一个峰值,之后开始降低。β 的精确最优值为 0.15,F 值最大为 0.918。

3.3　对比实验

通常的垃圾用户检测,通常为基于单一的用户行为,或基于单一的文本信息,无法全面利用用户的信息。而本文提出的 SDM 模型综合了微博用户的微博发布行为信息、微博内容文本信息等,为了能够

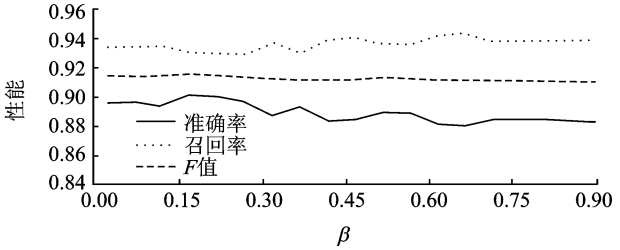

图 5 不同 β 取值的算法性能评价($\alpha=0.13$)

Fig. 5 Algorithm performance comparison with variation of β($\alpha=0.13$)

对比微博文本信息在模型中的作用,实验去除了 SDM 模型中 $F_s(m_1, m_2)$ 微博垃圾程度评估函数,在阈值 θ 从 0 开始增大时,计算算法的召回率、准确率和 F 值。并将实验结果和 3.1 节完整的 SDM 模型实验结果进行对比。

从图 6 中可以看到,如果不使用微博的垃圾程度评估函数,最优的结果明显低于带文本信息的模型,F 值为 0.78,小于完整模型算法的 F 值 0.918。而且在较为明显的、垃圾程度很高的垃圾微博用户的评估中,分数也明显低于带文本信息的模型。这个实验说明了微博垃圾程度评估函数对于模型的整体效果有较高的提升。

3.4 SDM 算法时间性能

在算法时间复杂度方面,SDM 主要耗时在 LDA 模型训练上。本文使用垃圾和随机用户各 20 万条微博进行基于 Gibbs 采样的 LDA 模型训练,在 AMD A-10 6800CPU/8GBMEM 下,训练需 230 min。LDA 模型建立后,算法的其他部分耗时很少,对 20 万垃圾用户微博和 120 万随机用户微博进行试验,在以上软硬件环境下,只需 7 min 即可完成。

4 结束语

本文研究了新浪微博中垃圾用户的检测问题。在实际数据的用户分析的基础上,本文提出了基于微博重复发送行为的模型 SDM,并在真实的用户数据集上验证了算法的有效性和不同信息对算法性能的影响。SDM 综合考虑了垃圾用户行为、社交网络信息和文本信息。经过分析,对模型影响最大的信息是转发微博文本信息和相似度,而原创微博的信息和转发用户的信息对模型性能的影响较小。

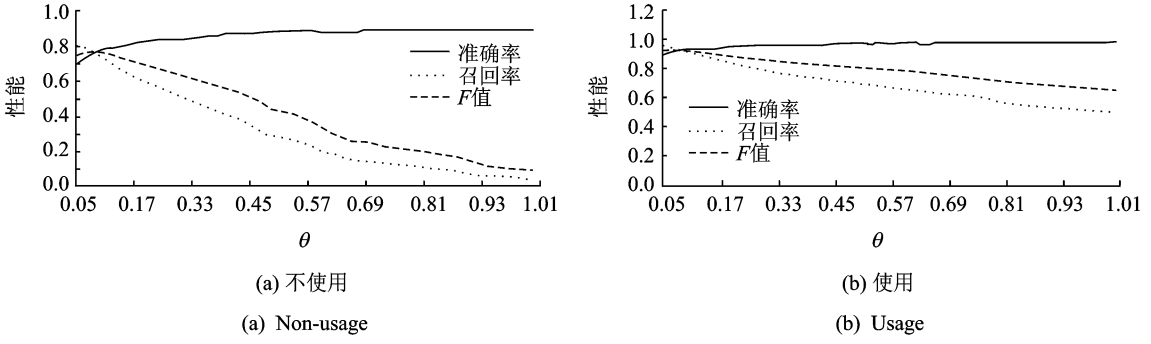

图 6 微博文本信息的算法性能的变化

Fig. 6 Algorithm performance with Weibo text characteristic

参考文献:

[1] 周耀明,李弼程.一种自适应网络舆情演化建模方法[J]. 数据采集与处理, 2013, 28(1):69-76.
Zhou Yaoming, Li Bicheng. Adaptive evolution modeling method of internet public opinions[J]. Journal of Data Acquisition and Processing, 2013, 28(1):69-76.

[2] Jaime T, Ramage D, Morris M R. TwitterSearch: a comparison of microblog search and web search[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. [S. l.]: ACM, 2011.

[3] Liben N, David, Jon K. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.

[4] Wu B, Davison B D. Identifying link farm spam pages[C]//Special Interest Tracks and Posters of the 14th International Conference on World Wide Web. [S. l.]:ACM, 2005: 820-829.

[5] Saptarshi G, Viswanath B, Kooti F. et al. Understanding and combating link farming in the twitter social network[C] // Proceedings of the 21st International Conference on World Wide Web. [S. l.]:ACM, 2012:61-70.

[6] Oscar P, Roychowdbury V P. Leveraging social networks to fight spam[J]. IEEE Computer, 2005, 38(4): 61-68.

[7] Hull M , Farmer F, Perelman E. Selective electronic messaging within an online social network for SPAM detection[P]. US: 10/946,630, 2005-8-4.

[8] Sarita Y, Romero D, Schoenebeck G. Detecting spam in a twitter network[EB/OL]. First Monday, 2009.

[9] Gao H, Jun H, Wilson C. et al. Detecting and characterizing social spam campaigns[C] // Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. [S. l.]:ACM, 2010:35-47.

[10] Gómez, Vicenç, Kaltenbrunner Andreas, et al. Statistical analysis of the social network and discussion threads in Slashdot [C] // Proceedings of the 17th International Conference on World Wide Web. [S. l.]: ACM, 2008:645-654.

[11] Boshmaf Y, Muslukhov I, Beznosov K, et al. The socialbot network: when bots socialize for fame and money[C] // Proceedings of the 27th Annual Computer Security Applications Conference. [S. l.]: ACM, 2011:93-102.

[12] Sureka A. Mining user comment activity for detecting forum spammers in youtube[C] // 20th International World Wide Web Conference (WWW2011). Hyderabad, India:[s. n.],2011.

[13] Jindal, Nitin, Liu B. Review spam detection[C] // Proceedings of the 16th international conference on World Wide Web. [S. l.]: ACM, 2007:1189-1190.

[14] Zhe W, William J, Lv Q. Filtering image spam with near-duplicate detection[C] // Proceedings of the 4th Conference on E-mail and Anti-Spam. Mountain View, CA, USA:[s. n.], 2007.

[15] Draisbach U, Felix N, Sascha S, et al. Adaptive windows for duplicate detection[C] // Data Engineering (ICDE), 2012 IEEE 28th International Conference on. [S. l.]:IEEE, 2012:1073-1083.

[16] 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[J].南京大学学报:自然科学版, 2013, 49(4):456-464.
Zhao Bin, Ji Genlin, Qu Weiguang,et al. Detecting microblog spammers based on reuse detection[J]. Journal of Nanjing University:Natural Sciences, 2013, 49(4):456-464.

[17] Zhang Qunyan, Ma Haixin, Qian Weining, et al. Duplicate detection for identifying social spam in microblogs[C] // Big Data Congress, 2013 IEEE International Congress. [S. l.]:IEEE, 2013:141-148.

[18] 邸鹏, 段利国. 一种新型朴素贝叶斯文本分类算法[J]. 数据采集与处理, 2014, 29(1):71-75.
Di Peng, Duan Liguo. New naïve bayes text classification algorithm[J]. Journal of Data Acquisition and Processing, 2014, 29(1):71-75.

作者简介:**吴斌**(1969-),男,教授,研究方向:复杂网络、数据挖掘和智能信息处理,E-mail:wubin@bupt.edu.cn;**李冠辰**(1989-),男,硕士研究生,研究方向:数据挖掘和云计算;**刘宇**(1986-),男,博士研究生,研究方向:复杂网络和数据挖掘;**张雷**(1961-),男,教授,研究方向:软件与服务、智能信息处理;**王柏**(1962-),女,教授,研究方向:电信系统软件、分布计算技术、数据挖掘、智能信息处理。