计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(10):184−189 [doi: 10.15888/j.cnki.csa.006014]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多特征的垃圾微博检测方法①

邹永潘[1,2], 李　伟[1], 王儒敬[1]

[1](中国科学院 合肥物质科学研究院 合肥智能机械研究所, 合肥 230031)
[2](中国科学技术大学, 合肥 230026)

摘　要: 随着微博平台的快速发展, 垃圾信息检测与过滤也面临着巨大的考验, 实时精确地识别垃圾信息对于提高用户的体验以及微博平台的可持续发展意义重大. 本文根据新浪微博的真实数据, 提出了一种基于多特征的垃圾微博检测方法. 首先, 提取微博的显式特征 (用户特征、内容特征); 然后利用文档主题生成模型 (LDA) 提取微博中的隐含主题特征; 最后根据所提取的微博特征利用支持向量机 (SVM) 构建分类器. 实验结果表明, 该方法相比于现有方法在准确率和 F1 值方面都有一定的提升.

关键词: 垃圾微博检测; 隐含狄利克雷分布; 支持向量机

引用格式: 邹永潘,李伟,王儒敬.基于多特征的垃圾微博检测方法.计算机系统应用,2017,26(10):184–189. http://www.c-s-a.org.cn/1003-3254/6014.html

Detection Method of Spam Based on Multi-Features of Micro-Blog

ZOU Yong-Pan[1,2], LI Wei[1], WANG Ru-Jing[1]

[1](Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China)
[2](University of Science and Technology of China, Hefei 230026, China)

Abstract: With the rapid development of micro-blog, spam detection and filtering is faced with enormous challenges. It is significant to realize realtime and accurate detection of spam, which is important to improve user experience and the sustainable development of micro-blog platform. In this paper, a spam detection method based on multi-features of micro-blog is proposed. The main procedures are: first, the features of user and content are extracted. Second, LDA is applied to extract latent topic features. Finally, the features above are fused and a proper classifier is trained based on SVM. Experimental results show that the precision and F1 get increased while adopting the method proposed in this paper compared to the pervious methods.

Key words: spam detection; latent Dirichlet allocation; support vector machine

1　引言

当今社会, 社交网络已经成为人们日常交流的一种重要方式. 大量的用户通过在社交网络发布消息与朋友进行互动, 用户也可以通过关注热门用户、热门话题来了解名人和最新的新闻动态. 诸如新浪微博等社交网络便捷、自由的传播方式使得其用户量和信息量获得了爆炸式的增长, 这也为垃圾用户和垃圾信息的产生提供了土壤. 因此, 检测社交网络中的垃圾信息对于增加用户的体验度以及对社交网络平台的可持续发展具有重要的意义.

微博等社交网络的垃圾信息检测过滤方法一直以来都是一个研究的热点. 现有的研究工作主要集中在识别垃圾信息制造者而非直接识别垃圾微博信息. 研究者们通过将特征统计方法与机器学习算法相结合, 提出了大量用于检测垃圾用户的方法. 如 Fabrício Benevenuto 等使用机器学习的方法通过 twitter 的内容

① 基金项目: 中国科学院战略性先导科技专项 (XDA08040110)
收稿时间: 2017-01-16; 采用时间: 2017-02-23

属性和用户的行为属性来识别 twitter 中的垃圾用户[1]; Zeng Z 等通过观察垃圾用户的行为统计规律和传播方式, 利用支持向量机对标注的微博用户集构建分类器, 取得了不错的效果[2]. 此外, 还有一些学者通过检测微博相似度和对用户的社交网络建立图模型来进行垃圾用户的检测[3,4].

然而, 在使用微博的过程中发现, 垃圾信息的发布者并不完全来源于垃圾用户, 大量的垃圾信息来自于普通用户. 因此, 对单条微博进行检测过滤从而净化社交网络显得十分必要. Ma Y 等人将英文微博客的垃圾检测方法应用到了中文微博客中, 并对支持向量机、随机森林、朴素贝叶斯三种方法进行了对比, 实验结果显示了支持向量机在处理文本分类中的优势[5]. 于然等人通过从微博的结构和内容两个视角建立规则, 再与分词结果进行融合构造符合特征, 并以此进行垃圾微博过滤[6]; 王琳等人提出了一种基于相似微博检测和 URL 链接、字符串、高频词等特征判别的垃圾微博检测方法, 取得了较好的效果[7]; 刁宇峰等人提出了一种博客垃圾评论发现方法, 通过 LDA 对博客的评论进行隐含主题提取, 结合博文主题信息进行垃圾评论判别[8].

针对垃圾微博的来源不一定是垃圾用户这一事实, 将工作重点放在了垃圾微博博文的检测上, 利用微博的显式特征、隐含主题特征来对微博进行垃圾检测. 首先, 根据微博自身的文本特点和结构特点, 提取微博的作者粉丝数关注数之比、微博的评论数、URL 数目等显式特征; 针对垃圾微博和正常微博中词的分布特点, 构建垃圾微博特征词库, 计算每一条微博的垃圾微博特征词比例作为微博内容特征的扩展; 由于微博具有正文文本较短, 用词自由等特点, 仅仅从统计特征入手, 忽略文本的语义特征不能满足过滤的需求, 因此考虑引入 LDA 概率主题模型, 通过抽取微博的隐含主题并计算其对该主题的隶属度来从语义的角度对微博的特征进行进一步扩展; 最后, 利用之前构建的特征向量采用支持向量机对微博数据建立分类模型. 通过实验对比, 验证了该方法的可行性.

2 基于多特征的垃圾微博检测方法

2.1 微博显式特征

通过对大量微博的对比分析发现, 正常微博和垃圾微博在显式特征上面存在有比较明显的差异. 例如, 对于一条用于做广告推广的微博, 它的被转发数、被评论数一般比较低, 而含有 URL 链接的可能性却很大; 对于一条关于人生感悟的微博, 一般不会含有特殊字符、URL 链接等内容特征. 为了准确的描述二者之间的差异, 本文从人工标注的 8000 条新浪微博数据 (其中垃圾微博 3193 条) 入手, 分别对用户特征和微博内容特征进行统计分析, 具体分析结果如下.

2.1.1 用户特征

在垃圾微博检测的工作中, 经常被考虑到的用户特征有用户的粉丝数、用户的关注数、微博的被赞数、微博的被评论数、微博的被转发数. 表 1 展示了正常微博和垃圾微博在各个属性上的平均值.

表 1 各个用户特征的统计平均值

微博类型	正常微博	垃圾微博
作者粉丝数	5605.38	4239.74
作者关注数	896.35	1167.62
微博被赞数	90.76	40.89
微博转发数	88.63	30.25
微博评论数	61.45	37.67

从表 1 中可以看出, 微博被赞数、转发数、评论数在正常微博和垃圾微博中区别较大, 而作者的粉丝数和关注数区别则不是很大, 因此考虑将粉丝数关注数之比作为一个用户特征应用到垃圾微博过滤当中. 通过以上分析可知, 将用户特征用于垃圾微博的过滤具有一定的实际意义.

2.1.2 内容特征

研究人员通过各个角度对微博的内容特征进行分析提取, 目前比较流行的内容特征有: 微博长度、链接数、链接占比、代词长度、“我”出现的次数、有没有“@”“#”等特殊符号、数字符号占比等, 本文通过计算所选特征之间的相关性并将部分属性融合为一个属性之后, 选择了微博正文长度、URL 数、非汉字字符比例、垃圾微博特征词比例作为用户的内容特征. 关于垃圾微博特征词比例 (*ratio*) 的定义如公式 (1) 所示. 其中 *len*() 函数表示求字符串的长度, $D_1(i)$ 表示微博正文 *Doc* 中的出现在垃圾微博特征词典中的词.

$$ratio = \frac{\sum len(D_1(i))}{len(Doc)} \quad (1)$$

经过特征选择之后, 再次对各个内容特征进行分

析, 结果如表 2 所示. 统计结果表明, 考虑微博的内容特征是必要的.

表 2 微博内容特征统计平均值

微博类型	正常微博	垃圾微博
URL数	0.8256	1.7232
非汉字字符占比	0.1384	0.3055
垃圾特征词比例	0.0536	0.1649
微博正文长度	46.77	99.36

2.2 微博隐含主题特征

2.2.1 LDA 模型描述

Latent Dirichlet Allocation 模型是由 Blei 等在 2003 年提出[9], 属于一种典型的概率主题模型. 作为一种产生式主题模型, LDA 已经广泛的应用于文本分类、信息检索等诸多领域[10,11]. LDA 主题模型本质上是一个"文档-主题-词"的三层贝叶斯网络, 文档和主题、主题和词之间均服从狄利克雷分布. LDA 的模型表示如图 1 所示[10], 关于图中各个参数的含义如表 3 所示.

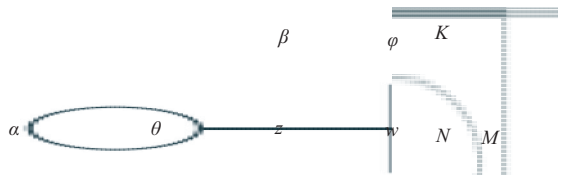

图 1 LDA 概率图模型表示

表 3 LDA 模型中各符号的含义

符号	含义	符号	含义
α	θ的超参数	W	词
β	φ的超参数	M	文档数
θ	文本—主题概率分布	N	词数
φ	主题—词概率分布	K	主题数
z	词的主题分布		

在图 1 的模型中, 词 w 在主题 z 上以及主题 z 在文档 d 上分别服从以 φ 和 θ 的多项分布, 而 θ 和 φ 又分别服从参数为 α、β 的 Dirichlet 分布. 整个 LDA 模型的对应联合概率分布为[12]:

$$p(w_d,z_d,\theta_d,\varphi|\alpha,\beta)= p(\theta_d|\alpha)\cdot\prod_{n=1}^{N_d}p(w_{d,n}|\varphi_{z_{d,w}})p(z_{d,w}|\theta_d)p(\varphi_z|\beta) \quad (2)$$

$p(\theta_d|\alpha)$和$p(\varphi_z|\beta)$分别表示 θ 和 φ 服从的参数为 α、β 的 Dirichlet 分布. 则:

$$p(\theta|\alpha)=Dir(\theta|\alpha)=\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k-1} \quad (3)$$

其中, $\Gamma(\cdot)$ 表示 Gamma 函数, 定义如下:

$$\Gamma(n)=(n-1)\Gamma(n-1) \quad (4)$$

通过整合 θ、φ, 使得公式 (2) 中仅仅保留可供观测的单词 w、已知的参数 α、β 和待推导的主题分布 z, 即:

$$p(w,z|\alpha,\beta)=p(w|z,\beta)p(z|\alpha) \quad (5)$$

对于模型中的参数, 通常设置参数 α=50/K、β=0.01、K=50. 可以通过对变量 z 进行 Gibbs 采样来近似估算 θ 和 φ, 计算公式如下[12]:

$$\theta_{m,s}=\frac{n_m^{(s)}+\alpha}{\sum_{j=1}^{K}n_m^{j}+K\alpha} \quad (6)$$

$$\varphi_{s,k}=\frac{n_s^{(k)}+\beta}{\sum_{s=1}^{N}n_s^{j}+N\beta} \quad (7)$$

其中, $n_m^{(j)}$表示文档 d_m 中赋予主题 j 的词的总个数, $n_s^{(j)}$表示词 v_i 被赋予主题 s 的总次数.

2.2.2 提取隐含主题特征

相比于目前大部分研究利用 LDA 主题模型的主题—词概率分布来解决特征稀疏问题, 文中主要利用 LDA 的文档—主题概率分布来预测文档的隐含主题作为文本的特征来进行特征扩展. 因此, 该部分的主要工作包括两部分: 首先, 利用一个外部微博文本集 *docs* 训练 LDA 模型, 即估计模型中的参数 θ 和 φ; 然后, 对于一个新的文档 doc_{new}, 利用训练得到的 LDA 模型来计算该文档的主题分布 θ_{new} 并选择对应的 TopN 个主题的主题编号以及以及对应的概率作为隐含主题特征, 具体流程如图 2 所示.

关于图 2 的几点说明:

(1) 图中的"外部文档集合"是通过网络爬虫在新浪微博平台爬取的涉及各个主题的微博文本集合; "实验数据集"是包含微博的作者、作者 ID、微博 ID、正文等一系列特征的结构化数据集.

(2) 相比于"外部文档集合"预处理中的分词、去停用词等操作, 对"实验数据集"要首先从结构化微博数据中提取中微博的正文.

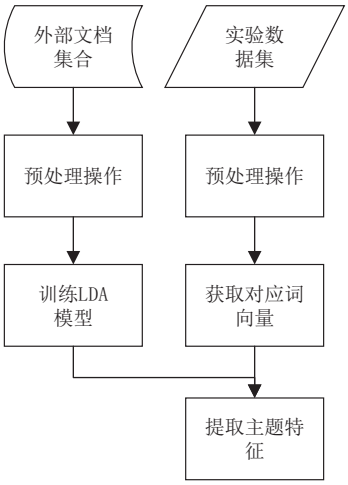

图 2 隐含主题特征提取框架

(3) 利用外部文本集训练出来的 LDA 模型提取每一条微博的 TopN 个主题的主题编号以及对应的概率, 其中 N 需要通过实验进行确定.

2.3 算法描述

文中提出的垃圾微博检测算法的思路是从微博的显式特征 (包括用户特征和内容特征) 和隐含主题特征入手构建特征向量, 然后再利用 SVM 进行分类器构建. 算法流程如图 3 所示.

具体实现过程如下:

输入: 微博原始训练集 D, $Model_{LDA}(\theta, \varphi)$

输出: 用于对测试集进行分类的最终分类器 C.

Step 1. 将每一条微博数据划分为微博文本和非文本数据, 并从非文本中提取出该条微博的用户特征;

Step 2. 引入垃圾微博特征词典, 对于微博文本, 分别利用正则表达式技术匹配微博正文中的 URL 标签、垃圾微博特征词、非汉字字符来提取微博的内容特征.

Step 3. 将通过 Step 2 处理过后的微博正文进行预处理操作: 去非汉字字符、分词、去停用词, 得到每一条微博对应的词向量.

Step 4. 利用 $Model_{LDA}(\theta, \varphi)$ 对每一条微博进行隐含主题特征提取.

Step 5. 将显示特征和隐含主题特征整合到向量空间中, 利用 SVM 算法构建分类模型.

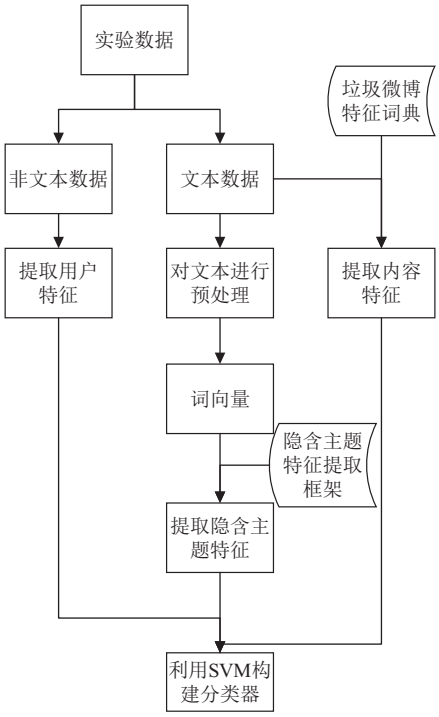

图 3 垃圾微博检测框架

3 实验

3.1 实验数据获取

利用网络爬虫从新浪微博爬取 1500 个用户 2016 年 7 月 1 日至 7 月 31 日期间发送或转发的 129648 条新浪微博正文, 数据集涵盖了体育、经济、娱乐、情感等各个领域, 经过去重处理之后作为外部文档集来训练 LDA 模型. 随机抽取 8000 条记录作为训练数据集, 经过人工标注后得到垃圾微博 3193 条, 正常微博 4807 条. 文中扩展的垃圾微博词典是通过收集网上微博常用广告词获得的 214 个词或短语.

3.2 实验设置

1) 实验预处理

实验中对文本的预处理主要有 HTML 解析、分词、去停用词等操作, 其中分词过程使用了 HanLP 开源汉语言处理包, 并添加了用户词典, 使得分词具有更好的效果.

2) 确定隐含主题数 N

隐含主题数 N 是在利用 LDA 主题模型进行主题预测时选择该文本所属的主题的个数. 选择的主题个数过小会使得判断文本所属主题时存在很大的偶然性, 而选取的 N 过大又会导致隐含特征性质的下降. 为了研究选取的隐含主题个数 N 对实验结果的影响, 选取

4000 条经过标注的实验数据集, 以 F1 值作为评判标准, 实验结果如图 4 所示. 从图中可以看出, 随着 N 的增大, F1 值先增大后减小, 在 N=5 的时候得到最大的 F1 值, 此时分类器的性能是最好的, 故在分类器的训练中设定 N=5.

3) 算法选择及评价标准

根据预处理中获取的显式特征和隐含主题特征构建特征向量, 利用支持向量机 (SVM) 进行分类, 实验中采用 LIBSVM 软件包. 算法性能通过准确率 *P*(Precision)、召回率 *R*(Recall)、以及综合考虑指标 *F1* 来衡量.

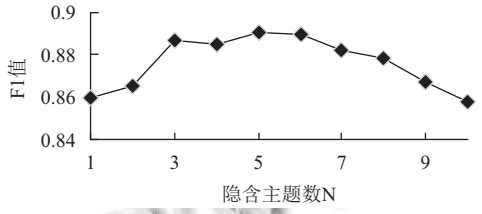

图 4 隐含主题数 N 对 F1 值的影响

$$P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN},F1=\frac{2P*R}{P+R} \quad (8)$$

以上公式中, *TP* 表示正确分类的正例数目, *FN* 表示错分为负例的正例数目, *FP* 为错分为正例的负例数目.

4) 对比实验分组

为了验证文中提出的垃圾微博检测方法的有效性, 文中利用相同的实验数据集, 分别实现了三种方法, 并将三种方法的实验结果进行对比. 其中:

方法 1: 利用文献[6]提出的垃圾微博过滤方法.

方法 2: 引入垃圾微博特征词特征, 利用上文中提到的显式特征作为特征向量构建分类器.

方法 3: 文中提到的垃圾微博检测方法.

3.3 实验结果及分析

利用人工标注的 8000 条实验数据 (垃圾微博 3193 条) 构建分类器, 并通过十折交叉验证的方法来对分类模型进行评估. 针对以上三种方法, 得到的实验结果如表 4 所示.

表 4 实验结果

	P	R	F1
方法1	0.874	0.846	0.860
方法2	0.903	0.864	0.883
方法3	0.927	0.860	0.892

通过方法 1 和方法 2 的实验结果对比可以发现, 引入垃圾微博特征词之后对于检测的准确率和召回率方面均有所提高; 通过对比方法 2 和方法 3 可以发现, 考虑了微博文本的隐含主题特征之后, 分类器的综合性能得到了一定的提高.

为了进一步验证分类器的泛化能力, 从 2016 年 8 月 1 日到 20 日的微博数据集中每天随机抽取 150 条分别利用以上三种方法进行测试, 得到的实验结果如图 5 所示.

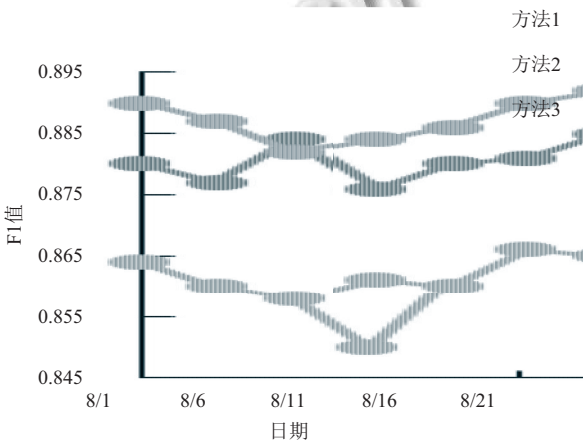

图 5 三种方法的实验结果对比

图 5 的实验结果表明, 文中提出的垃圾微博检测方法相比于方法 1 和方法 2 有较大的提高. 由于充分考虑了微博的显式特征和微博文本中的隐含主题特征, 从图中可以看出, 针对每一天不同的测试数据集, 分类器的分类性能相对比较稳定, 表明该算法具有实际的应用价值.

4 结语

文中通过对现有垃圾信息过滤方法的分析, 针对垃圾微博的特点, 引入了基于显式特征和隐含主题特征结合的方法来对微博进行特征扩展进而实现垃圾信息检测. 通过实验表明, 相比于仅仅通过微博的显式特征或微博的文本内容进行垃圾过滤, 综合考虑微博的显式特征和隐含主题特征在检测垃圾微博时会取得更好的效果. 事实上, 文中算法考虑的特征依然较少, 微博中还有很多待挖掘利用的信息. 比如, 垃圾特征词的位置信息、微博中的图片信息等. 此外, 微博用户的可信度也是非常有价值的待考虑特征, 当一个用户的可信度较低时, 其所发微博是垃圾微博的可能性就会变大. 可以先利用 PageRank 等算法对用户进行评分[13], 然后

再结合文中算法进一步提高垃圾微博的检测准确率.

参考文献

1 Benevenuto F, Magno G, Rodrigues T, *et al*. Detecting spammers on twitter. Proc. of the 17th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference. Redmond, Washington, US. 2010.

2 Zeng ZP, Zheng XH, Chen GL, *et al*. Spammer detection on Weibo social network. Proc. of the 6th International Conference on Cloud Computing Technology and Science (CloudCom). Singapore. 2014. 881–886.

3 Xu Y, Zhou Y, Chen K. Observation on spammers in Sina Weibo. Proc. of the 2nd International Conference on Computer Science and Electronics Engineering. Paris, France. 2013.

4 杨凯帆. 微博垃圾信息检测[硕士学位论文]. 合肥: 中国科学技术大学, 2015.

5 Ma YC, Niu Y, Ren Y, *et al*. Detecting spam on Sina Weibo. Proc. of International Workshop on Cloud Computing and Information Security. Paris, France. 2013. 404–407.

6 于然, 刘春阳, 靳小龙, 等. 基于多视角特征融合的中文垃圾微博过滤. 山东大学学报 (理学版), 2013, 48(11): 53–58.

7 王琳, 冯时, 徐伟丽, 等. 一种面向微博客文本流的噪音判别与内容相似性双重检测的过滤方法. 计算机应用与软件, 2012, 29(8): 25–29, 94.

8 刁宇峰, 杨亮, 林鸿飞. 基于 LDA 模型的博客垃圾评论发现. 中文信息学报, 2011, 25(1): 41–47.

9 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022.

10 吕超镇, 姬东鸿, 吴飞飞. 基于 LDA 特征扩展的短文本分类. 计算机工程与应用, 2015, 51(4): 123–127.

11 张志飞, 苗夺谦, 高灿. 基于 LDA 主题模型的短文本分类方法. 计算机应用, 2013, 33(6): 1587–1590.

12 Heinrich G. Parameter estimation for text analysis. Technical Report, 2005.

13 杨赫. 垃圾微博信息过滤技术的研究[硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2015.