# 由向量空间相关模型识别博客文章的垃圾评论

何海江,凌 云
(长沙大学计算机中心,湖南 长沙 410003)

**摘 要**:博客作者往往允许读者在文章后发表评论,许多评论充斥着形形色色的垃圾信息,破坏了博客社区的和谐.在向量空间的基础上构造了一个相关模型,将博客的文章和评论分别分词后,根据模型计算评论和文章的相关度,来判断是否为垃圾评论.该模型不需要训练样本,在一个中文博客测试集上,召回率和准确率分别达到82%和91%.

**关键词**:向量空间模型;博客;垃圾评论;相关度
**中图分类号**:TP391 **文献标识码**:A **文章编号**:1008-4681(2008)02-0063-04

博客(blog),是互联网上日记形式的个人主页,近年来快速增长,成为Web2.0的一个重要应用形式.综合国内外重要的博客社区来看,博客具有这样一些技术特征:每一篇日志或文章(post)都有时间戳,由博客工具软件维护,所有文章按反向的时间顺序排列;读者可以在博客上留言,发表评论(comment);可通过RSS或Atom聚合;允许使用BlogRoll和BlogTrack与其他博客建立链接.

博客工具软件屏蔽了具体的技术细节,各种商业的或非商业的博客网站都采用免费注册、免费维护的政策,吸引了众多网民开设博客.他们在博客上记录个人的日常生活及事务,抒发感情、释放情绪,发表对新闻事件及人物的看法和意见等.一些企业和组织也开办博客,向外发布信息,跟踪读者的反应.一般情况下,博客作者都允许来访的读者在文章后发表评论,Marlow就认为评论是博客作者和读者群之间一种简单而有效的交流途径[1],读者的真诚评论能促使大多数作者发表更多的文章.博客网站,也就是博客软件服务商也鼓励网站的浏览者在博客文章后发表评论,没有身份认证、也无需注册为用户,读者可匿名随意提交文章评论.发表评论的读者能带动博客网站的网络流量,吸引人气.特别需要指出的是,网站还会将评论多的博客文章列入热门文章排行榜.

自然地,博客的发展吸引了垃圾制造者的目光,许多评论夹杂着令人生厌的垃圾信息.综合来看,评论所带的垃圾信息主要有三类:(Ⅰ)广告信息,内容涉及产品推销、网站(包括博客)推介、信息发布等;色情、反动、暴力等不良信息.有的包含链接指向目标Web地址,有的则不包含(许多博客工具软件自动删除链接).(Ⅱ)超链接.评论中的内容看似正常,其实隐含了指向某Web地址的超链接.博客社区的繁荣,促使各种商业搜索引擎改进算法,他们纷纷将博客内容纳入搜索范围,博客所包含的词语和超链接也成为搜索引擎算法的考察对象.由于商业利益的驱使,垃圾制造者并非要引起网民们注意其评论上发表的内容,而是广泛发布这些包含超链接的评论,提高其链接指向目标Web地址在搜索引擎的评分.(Ⅲ)漫骂、下流、人身攻击等言论.一些人为发泄自己的不满情绪,不顾别人的感受,对博客文章的作者或者其他读者漫骂、人身攻击,这些文字出现在博客中,令人厌恶.

包含垃圾信息的评论称为垃圾评论,不清除这些垃圾评论,博客社区将是不健康的.因此,研究如何识别垃圾评论是一件非常有意义的事情,研究成果可协助博客软件服务商准确识别垃圾评论,过滤、删除垃圾评论,使博客社区更干净、更和谐.

文章其余部分这样安排,第一节介绍相关领域内的研究工作;第二节提出基于向量空间模型的文章和评论相关度模型;第三节用实验结果说明模型的有效性;第四节给出结论及以后的工作展望.

## 1 相关研究工作

垃圾评论的识别可看作二值(合法评论、垃圾评

**收稿日期**:2007-11-16;**修回日期**:2008-01-11
**基金项目**:长沙大学科研基金(批准号:CDJJ-07010110)资助项目.
**作者简介**:何海江(1970-),男,湖南望城人,长沙大学计算机中心副教授.研究方向:Web挖掘、数据仓库.

论)分类问题.二值分类技术在Web信息对抗领域受到广泛研究,方法的主要内容是:先收集一些垃圾对象和合法对象作为训练样本,再找出它们的显著特征,最后运用分类算法鉴别.

Schneider[2]将电子邮件文本视为词的集合,并假设词之间没有依赖关系,采用朴素贝叶斯分类,比较了多项式模式和多变量贝努利模式的性能差异.赖均等[3]采用概率统计方法计算邮件样本词条权值,选定概率最大的20个词条作为基本特征元,结合贝叶斯和遗传算法过滤垃圾邮件.陈蔚然等[4]将Email文本视作生物信息,设计了一个基于生物序列模式提取技术的垃圾邮件过滤算法.Paranam等[5]选择网址的词语、文章的N元短语等作为博客的显著特征,使用支持向量机的机器学习算法分类垃圾博客和合法博客.Lin等[6]在文章内容、文章发表时间、文章包含的链接等属性上构造自相似性测度,再运用支持向量机分类算法,取得了90%的垃圾博客识别准确率.然而,运用这些算法只能识别Ⅰ Ⅱ类垃圾评论,却无法识别没有显著特征的Ⅲ类垃圾评论.

将与文章文本内容无关的评论都归类为垃圾评论,这样,垃圾评论的识别可参考文本分类、文本过滤、文本检索等文本挖掘技术[7],考察文章和评论的语义内容相关度.

文本检索的搜索算法在搜索引擎中得到深入研究和应用,文本分类、文本聚类等技术也一直是研究者的兴趣所在.Yang等[8]评估了文档词频、信息增益、$\chi^2$统计量等五种特征在文本分类算法K最近邻、线性最小二乘中的性能.黄萱菁等[9]基于对数互信息量,使用加权公式,计算初始模板和批量文本的相似度,实现文本过滤.

博客的研究大多聚焦于文章,研究评论的报道较少.Gilad和Natalie采集了一个包含685976篇文章、645042个评论的文集[10],详细分析了评论的Zipf分布频度、评论文本长度、评论内包含的链接等统计特征.Gilad等[11]提出一种基于语言模式的算法,能有效识别第Ⅱ类垃圾.该算法比较文章、评论、评论内链接所指向的Web页面这三类文本的语言模式,计算字串的概率分布交叉熵(Kullback-Leibler Divergence)来判别垃圾链接.Yuan Niu等[12]基于上下文分析博客的文章和评论,识别那些网址重定向、伪装成合法博客评论的垃圾信息,取得很好的效果.

## 2 文章和评论的相关度模型

向量空间模型,将文本看作由若干向量(词)组成的对象,由训练样本计算出的词频或其它度量定义为向量权,在文本挖掘及Web垃圾信息对抗中广泛运用[2,3,5,9,12].本文的模型也基于向量空间,但不需要训练样本,向量的权皆为1.这样做,计算速度快,算法也不容易受到攻击.如果采用有训练样本的机器学习分类算法.每当垃圾制造者发表具有新特征的评论时,要在博客社区停留一段时间,等待垃圾评论特征更新后,才能被识别.

评论大多是较短的文本,合法评论者经常使用一些常用词或者网络用语表达自己的观点和情绪,比如"沙发"、"顶"、"加精"等,这一类词称为常用词集合,令为Ths,共收集了十五条.垃圾评论中往往出现"免费"、"兼职"、"激情"、"浪货"等词语,称为不良词集合,令为Tb,共收集了六十一条.另外,令文章包含的词语集合为Tpost,评论包含的词语集合为Tcm,并记录评论中每个词出现的次数$\{FRTM_i\}$.模型只实验于中文博客,依据中文的语言特性,名词、动词、形容词最能表达文本语义特征,而大多数虚词和部分实词的特征很弱.为简便说明问题,将名词、动词、形容词称为nvj词.NA、NB、NC分别是评论与文章、评论与常用词集、评论与有效不良词集(不良词集删除文章中出现的不良词)之间相同词在评论中出现的nvj词总个数.有:

$$NA = \sum_{word\in(Tcm\cap Ths)\ and\ word's\ tag\ is\ nvj} FRTM_{word} \tag{1}$$

$$NB = \sum_{word\in(Tcm\cap Tpost)\ and\ word's\ tag\ is\ nvj} FRTM_{word} \tag{2}$$

$$NC = \sum_{word\in(Tcm\cap(Tb-Tpost))\ and\ word's\ tag\ is\ nvj} FRTM_{word} \tag{3}$$

其中$FRTM_{word}$表示词$word$在评论中出现的次数.

文章和评论的相关度定义为:

$$Corr_{(post,\ comment)} = \frac{1+NA+NB-\Psi\times NC}{|comment|\times\log_{10}(1+|comment|)} \tag{4}$$

分母中$|comment|$为评论的所有词语个数,不限于$nvj$词.而$\Psi$是不良词的影响因子,区分合法评论和垃圾评论的重要参数,实验环节将详细讨论.模型的基本思想是考察评论中体现合法性的词总个数占评

论总词个数的比例,也就是:

$$\frac{NA + NB - \Psi \times NC}{| comment |} \quad (5)$$

从大量博客观察到,长文本评论更倾向于是广告信息,所以在分母中加了对数,使得较长文本要比短文本体现合法性词的比例要更高,才能被判为合法评论.一些评论只有标点符号,词的个数为 0,可以认定为合法评论,模型不用考虑.一些评论只有一个词,对数处理后式 4 分母为 0,所以在对数中加 1.部分评论与与文章、常用词集、不良词集没有 nvj 词的交集,式 5 分子为 0,在分子加 1 与分母加对数一样,较长文本判为合法评论的概率比短文本小一些.由式 4 计算的相关度是一个数值,若大于阈值 δ,则判定为合法评论,否则为垃圾评论.

## 3　实验结果

从新浪博客、博客网等七个博客网站收集了 85 篇文章,共 815 条评论,采用人工标注方法,其中 315 条评论被标注为垃圾评论.文章包含文章作者、标题、内容三部分,因为很多时候评论者并非评论文章本身,只是和作者打招呼或者就标题而简短发言.针对所有文章和评论,使用中科院计算所汉语词法分析系统 ICTCLAS[13]分词后,记录词和词性.

为评价分类效果,采用两值分类问题最通用的性能评价方法:

$$召回率\ R = \frac{被正确判别为垃圾评论的评论条数}{实际存在的垃圾评论条数}$$

$$准确率\ P = \frac{被正确判别为垃圾评论的评论条数}{模型识别为垃圾评论的评论条数}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

图 1、图 2、图 3(如图 1、图 2、图 3 所示)分别是模型参数 Ψ 和 δ 对召回率、准确率、F1 的影响.图 1 显示,Ψ = 1…6,随着 δ 的增大,召回率越来越高;图 2 显示,Ψ = 1…6,随着 δ 的增大,准确率越来越小;图 3 显示,Ψ = 1…6,δ = 0.07 时 F1 最大.图 4 表示,Ψ 从 1 到 3,最大的 F1、次大的 F1、三个最大 F1 的平均这三项评价指标显著改善;而 Ψ 从 3 到 6,三项评价指标改善极小(如图 4 所示).Ψ 增大,将导致合法不良词(Tb 和 Tpost 的交集)较多的评论倾向于被判定为垃圾评论.综合考虑,式 4 的 Ψ = 3.将 δ 设为 0.07 时,召回率和准确率分别达到 82%和 91%.

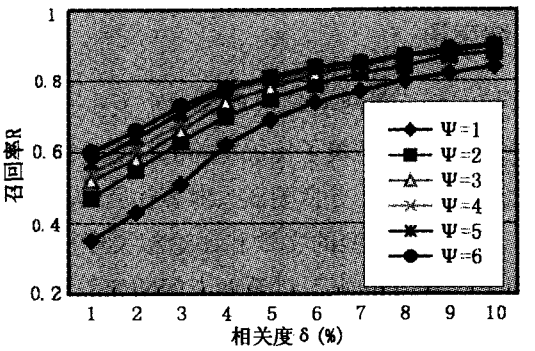


图 1　模型参数对召回率的影响

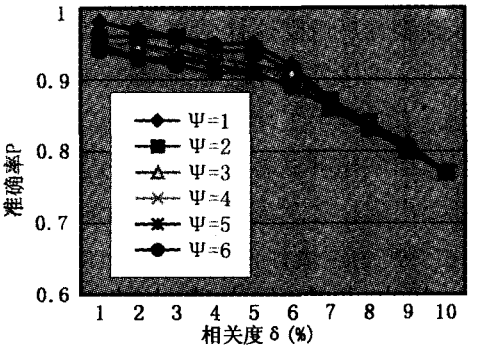


图 2　模型参数对准确率的影响

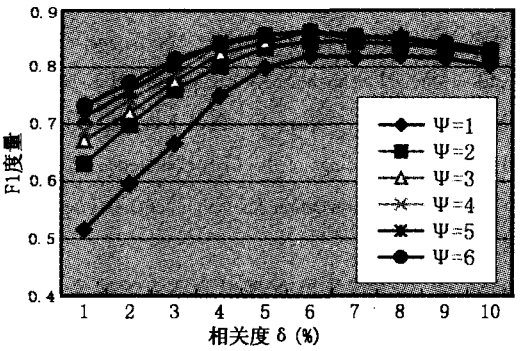


图 3　模型参数对 F1 的影响

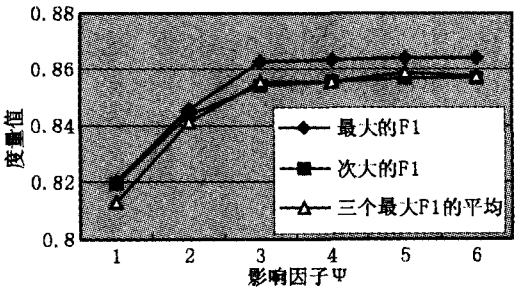


图 4　Ψ 对 F1 的影响

## 4 结束语

博客是互联网上日记形式的个人主页,被认为草根出版媒体,颠覆了传统媒体的信息传播模式.自然地,各种信息不分良莠涌入博客社区.设计了一个基于向量空间模型的文章和评论相关度模型,将评论和文章分别分词后,计算两者间的相关度,小于阈值的评论就判为垃圾评论.

常用词集和不良词集的维护并不难,毕竟网络用语更新速度并不快,漫骂、下流的词语也有限,只要不公开这两个词集,模型将保持健壮.

该模型容易遭受这样的攻击,垃圾制造者将文章的一段文本拷贝到评论中,掩饰其中的不良词.但垃圾制造者只能手工拷贝,显然提高了垃圾制作成本.另外,模型还有显著的弱点,若较长文本评论中没有那些出现在文章或常用词集的词,但却语义相近,将被判定为垃圾评论.以后进一步的工作中,作者将研究如何改进模型,消除这两个弱点.

**参考文献:**


[1] C. Marlow. Audience, structure and authority in the weblog community[C]. New orleans: In The 54th Annual Conference of the International Communication Association, 2004.

[2] Karl - Michael Schneider. A Comparison of Event Models for Naive Bayes Anti - Spam E - Mail Filtering[C]. Buelapest, Hungary: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL' 03), 2003.

[3] 赖均,等.基于遗传算法、贝叶斯学习的网段反垃圾邮件系统[J].计算机工程,2006,32(2): 189 - 193.

[4] 陈蔚然,等.基于生物序列模式提取技术的邮件过滤算法[J].清华大学学报(自然科学版),2005,45(5):1734 - 1737.

[5] Pranam Kolari et al., Detecting Spam Blogs: A Machine Learning Approach[C]. Boston, Massachusetts: In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006), 2006.

[6] Yu - Ru Lin et al., Splog Detection Using Self - similarity Analysis on Blog Temporal Dynamics[C]. Banff, Albertu, Canada in Proceedings of AIRWeb 2007, May 8, 2007.

[7] 袁军鹏,等.文本挖掘技术研究进展[J].计算机应用研究,2006,23(2):1 - 4.

[8] Yang, Y., Pedersen, J.O., A Comparative Study on Feature Selection in Text Categorization[C]. San Francisco Proc. of the 14th International Conference on Machine Learning ICML97,1997:412 - 420.

[9] 黄萱菁,等.基于向量空间模型的文本过滤系统[J].软件学报, 2003, 14(3): 435 - 442.

[10] Gilad Mishne and Natalie Glance ,Leave a Reply: An Analysis of Weblog Comments[C]. Edinburgh, scotlamel In Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference, 2006.

[11] Mishne, G., D. Carmel, et al., Blocking Blog Spam with Language Model Disagreement[C]. Chiba, Japan Proceedings of the 1st AIRWeb, 2005.

[12] Yuan Niu et al.,A Quantitative Study of Forum Spamming Using Context - based Analysis[C]. San Diego. CA Proceedings of the 14th NDSS, 2007.

[13] 中文自然语言处理开放平台[EB/OL].http://www.nlp.org.cn/,2007/01/12.


(作者本人校对)