doi:10.13582/j.cnki.1672-9102.2014.01.016

基于空间向量模型的垃圾文本过滤方法

吴玮

(苏州工业职业技术学院 软件与服务外包学院,江苏 苏州 215104)

摘 要:针对垃圾文本识别计算的需求特性,应用VSM文本聚类算法思想,综合现有TFIDF算法特点,提出一种基于VSM和改进的TFIDF特征项提取算法.本方法在对垃圾文本高聚类特征项权值进行放大的同时,有效减小由二类数据样本数量偏差对计算结果带来的影响,提高了垃圾文本过滤识别效率和准确率.为垃圾文本识别提供了一种新的改进算法选择.

关键词:空间向量模型;垃圾文本;分类;过滤
中图分类号:TP391 **文献标志码**:A **文章编号**:1672-9102(2014)01-0078-06

Garbage text classification filtering method Based on VSM

WU Wei
(Department of Software and Service Outsourcing, Suzhou Institute of Industrial Technology, Suzhou 215104, China)

Abstract: A feature item extraction algorithm was proposed that based on VSM and improved TFIDF, according to the demand characteristics for the recognition and calculation of spam text by applying VSM's text clustering algorithm and summarizing features of existing TFIDF algorithm. The algorithm not only zoomed in weighted value for feature item of spam text clustering but also effectively reduced the impact on the result affected by the difference of sample number of second-class data and improve identification efficiency and accuracy in filtering spam text. It provided a new improved algorithm selection for identification of spam text.

Key words: VSM; the garbage text; classification; filtering

进入21世纪以来,我国国际互联网以前所未有的惊人速度发展,据中国互联网络信息中心(CNNIC)发布的《第31次中国互联网络发展状况统计报告》显示,截至2012年12月底,我国网民规模达5.64亿,全年新增网民5 090万人,互联网普及率42.1%.Internet的飞速发展使得网络上的信息资源成指数形式增长,这为广大网民带来了异常丰富的网络信息资源.但同时网络中传播的广告、色情、暴力、商业欺诈等不良信息内容也日益增多,这些信息通过BBS,E-mail,QQ等平台传播,在影响网络用户正常使用的同时,也在消耗有限的网络资源.

近年来,大批研究人员进行了大量基于词汇链、向量空间模型的文本处理方法的研究.如文献[1]提出一种通过构造多条词汇链来表达文本的叙事线索,再通过相互比较识别变异垃圾文本.该方法着重在对变异垃圾文本的识别,而非针对无参照对象的垃圾文本的识别.又如文献[2]通过对垃圾文本流各种特性的研究,提出一种条件概率集成方法,设计实现了分类模型的在线训练算法和在线分类算法,这种方法无需对文本进行向量表示,同时也就无需向量计算,但其SPAM会随时间逐步增加,从而影响执行效率.还有文献[3-5]也提出了基于VSM的文本处理方法,但未针对垃圾文本的特性进行处理研究.

单一垃圾文本识别与普通文本聚类在算法需求上有较大差别,主要表现在:1)垃圾文本识别

收稿日期:2013-09-29
基金项目:国家自然科学基金资助项目(61305088)
通信作者:吴玮(1978-),男,江苏苏州人,硕士,讲师,主要从事嵌入式系统开发研究.E-mail:wuw@siit.edu.cn

中,文本类仅有二大类,且二大类内文本聚合度有限;而普通文本聚类需求中文本类数量较多,各类内部文本聚合度较好;2)垃圾文本篇幅有限,结构简单;普通文本篇幅较长,特征项提供相对容易;3)垃圾文本样本库随时间的变化会出现数量偏差,通常正常文本数量会远大于垃圾文本数量. 而这种偏差存在于二类分类计算中,对结果影响更为明显.

针对垃圾文本的特点,本文提出一种基于 VSM 和改进的 TFIDF 特征项提取算法,将文本类样本偏斜和特征项在文本类内外分布偏差问题综合考虑. 整合文献[6]中 TFIDF 改进算法思想,使文本识别算法在垃圾文本识别运算中更具有针对性,提高了过滤运行效率.

1　向量空间模型

进行文本处理分析的前提是将文本以数学形式加以表示,目前最为常用的是由 Salton 提出的向量空间模型[7]. 将文本看成由多个特征项组成的序列,每个特征项对应向量中的一维,文本就可以转化为一个高维的向量,文本与文本间的相似度,可以由 2 个文本向量间的夹角来表示. 由向量数量积公式:

$$a * b = |a| * |b| * \cos\theta.$$

以数量积的坐标表示及向量的模的坐标表示式代入上式,变换可得:

$$\mathrm{Sim}(D1,D2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{(\sum_{k=1}^{n} W_{1k}^2)(\sum_{k=1}^{n} W_{2k}^2)}}. \quad (1)$$

式中,W_{1K},W_{2K} 表示 2 个文本第 K 个特征项权重,n 为维度.

通过特征项提取,确定垃圾文本类向量簇中心向量,再确定 $\cos\theta$ 阈值,通过训练不断修正垃圾文本类中心向量和阈值. 可见特征项的提取直接影响算法的效果,理想的特征词集,既向量簇中心向量需要具有完全性和区分性这二大特征. 所谓完全性,既特征词应该可以如实的标识文本内容;所谓区分性,既特征项应该具有区分垃圾文本与非垃圾文本的能力. 也可以认为,特征项应该尽可能多的出现在垃圾文本中,尽可能少的出现在非垃圾文本中.

2　特征项提取

2.1　原型

在众多特征项选取算法中,TFIDF 因其算法简单,准确率、召回率高而被大量采用,其公式为

$$f(t) = TF(t) * IDF(t);$$

$$f(t) = TF(t) * \log(\frac{N}{DF(t)}). \quad (2)$$

式中,t 表示词项,$TF(t)$ 表示词项 t 在文档 d 中出现的频次,$DF(t)$ 表示词项在文本集中包含词项 t 的文本数,N 表示文本集的文本数.

2.2　第一次改进

在进行垃圾文本过滤时,文本仅划分为垃圾文本与非垃圾文本二类,这是属于典型的二元分类. 文本集分为 2 个子类,在垃圾文本类中一些特定词项的词频会非常高,覆盖本类中的大多数文本,且在非垃圾文本类中极少出现,这些词项本因是高权重项,但根据 TFIDF 算法的定义,将文本集作为一个整体来计算时,由于出现词项的文本较多,将导致 IDF 项变小,从而导致权值变小. 当用户进行算法训练时,还有可能由于提供垃圾文档过多进一步降低 IDF 项值,从而在后续计算中淘汰关键词项. 因此在计算时应对 TFIDF 项进行改进.[8]

$$f(t) = TF(t) * \log(\frac{m}{m+k} * N). \quad (3)$$

式中,m 为垃圾文本类中包含词项 t 的文档数,k 为非垃圾文本类中包含词项 t 的文档数,通过改进使垃圾类中包含词项 t 的文档数量大,而在非垃圾类中包含词项 t 的文档数量小时,t 保持较高的权值.

2.3　第二次改进

由于垃圾文本识别中仅有二类分类,使用 TFIDF 改进方式进行特征项提取后,随着训练文本的增加,一方面特征向量维度将不断提高,另一方面由于应用环境的不同,样本库将会出现数量偏差,将直接影响相似度判断,降低垃圾文本的识别率. 因此,对垃圾文本进行分类识别时,增加函数项 $\lambda(t)$,放大垃圾文本特征项权值,进一步改进 TFIDF 算法.[9]

$$f(t) = TF(t) * \log(\frac{m}{m+k} * N) * \lambda(t). \quad (4)$$

$$\lambda(t) = \log(\frac{|c|}{cf(t)}). \quad (5)$$

式(5)中,c 代表类别数,$cf(t)$ 为包含词项 t 的类别数,当聚类计算中,类别较多时,$\lambda(t)$ 随 $cf(t)$ 的递增而递减,从而增加高聚类特征项权值. 由于垃圾文本过滤计算中,仅有二类,c 为常值 2,但当 $cf(t)$ 为 2 时,$f(t)$ 将归零,不利于特征项权值计算,因此再对公式(5)进行改进.

$$\lambda(t) = \log(\frac{|c|+1}{cf(t)}). \quad (6)$$

将式(6)代入式(4):

$$f(t) = TF(t) * \log(\frac{m}{m+k} * N) * \log(\frac{|c|+1}{cf(t)}). \quad (7)$$

式中,$c+1$,在保持函数特性的基础上,既保留低聚类特征项在分类计算中的作用,又放大高聚类特征项权值,克服了样本数量偏差带来的中心向量偏移情况,从而增强垃圾文本识别能力.

3 实验及其分析

实验算法采用 Visual2010 和 SQL2008 实现,实验环境如下:CPU 为 Intel 酷睿 i3 3220 处理器,内存为 3G,操作系统为 Windows 7.

3.1 样本选取

随机分别获取文本数为 50,100,150 和 200 的 4 个样本训练集,各样本集中垃圾文本约占 43% 至 47%. 样本文本篇幅长短不一,长度在 18 至 135 个字符,内容涉及宽泛,无明确类属. 例如:"美国内申大学 MBA/DBA 上海 11 期交大开班,清华北大名师任教,毕业证 + 学位证,免试入学,学费最优,7 月 30 日报名截止 021 - 51620325"、" 第二期江苏省职业教育教学改革研究课题申报工作开始了,申报截止时间为 2013 年 5 月 8 日,有申报意向的老师请按照申报要求将材料提交给我."、"我是 21 岁的 XX,手机配对游戏说:你是我的梦中情人?真的吗? 直接回复短信我们聊聊吧,我等你,希望你就是那个梦中情人. 9578105515."等.

3.2 特征项提取

在对文本进行特征项提取前首先进行文本的预处理,将文本中出现的例如"我、你、的、啊、了……"等无意义字词以及数字、标点剔除. 然后按照分词字典对待处理文本进行分词.[10] 为了较好地比较 3 种不同文本特征项提取计算公式在不同数量的样本集中的效果,本文分别采用式(2)、式(3)、式(7)在前文所述的 4 个数量不同的样本集中提取文本特征项并计算权值. 为了降低计算复杂度,在特征项提取计算后,对取得的特征项进行降维处理[11].

3.3 中心向量计算

根据文本特征项及其权值计算垃圾文本类中心向量,本实验 4 个样本集得到的垃圾文本中心向量前 10 项分布如下表:

表 1 50 样本集垃圾文本中心向量前 10 项

排名	式(2)特征项	式(2)值	式(3)特征项	式(3)值	式(7)特征项	式(7)值
1	优惠	35.811 97	优惠	16.118 100	优惠	17.707 540
2	短信	25.579 98	短信	10.601 320	回复	10.118 590
3	回复	22.085 84	回复	9.210 340	奖励	7.588 945
4	太湖	20.723 27	国际	6.907 755	旅游	7.588 945
5	减肥	18.643 82	咨询	6.907 755	国际	7.588 945
6	奖励	18.643 82	太湖	6.907 755	咨询	7.588 945
7	旅游	18.643 82	奖励	6.907 755	商业	7.588 945
8	商业	18.643 82	减肥	6.907 755	太湖	7.588 945
9	国际	17.427 43	商业	6.907 755	减肥	7.588 945
10	直接	16.564 38	旅游	6.907 755	水岸	5.059 297

表 2 100 样本集垃圾文本中心向量前 10 项

排名	式(2)特征项	式(2)值	式(3)特征项	式(3)值	式(7)特征项	式(7)值
1	咨询	25.296 48	咨询	23.025 85	咨询	45.098 60
2	优惠	22.766 83	优惠	20.723 27	优惠	43.454 82
3	免费	20.237 19	免费	18.420 68	手机	37.684 25
4	提供	12.648 24	手机	15.176 96	免费	37.684 25
5	经理	12.648 24	联系	13.279 84	联系	32.973 71
6	联通	10.118 59	活动	11.796 68	活动	29.771 07
7	国际	10.118 59	提供	11.512 93	提供	26.491 59
8	机会	10.118 59	经理	11.512 93	经理	26.491 59
9	全场	7.588 945	机会	9.210 34	联通	23.236 57
10	情人	7.588 945	国际	9.210 34	短信	23.025 85

表 3　150 样本集垃圾文本中心向量前 10 项

排名	式(2)特征项	式(2)值	式(3)特征项	式(3)值	式(7)特征项	式(7)值
1	咨询	70.297 83	咨询	39.143 95	咨询	43.004 02
2	免费	49.608 46	免费	25.328 44	免费	27.826 13
3	优惠	47.105 31	优惠	23.025 85	优惠	25.296 48
4	联系	46.955 68	联系	21.627 24	电话	20.237 19
5	公司	42.686 98	公司	18.607 52	减肥	10.118 59
6	手机	40.588 74	电话	18.420 68	国际	10.118 59
7	电话	35.382 79	手机	16.655 40	机会	10.118 59
8	提供	33.798 20	提供	15.183 38	全场	10.118 59
9	经理	33.798 20	经理	15.183 38	联通	10.118 59
10	活动	30.959 94	酒店	12.890 61	手续	10.118 59

表 4　200 样本集垃圾文本中心向量前 10 项

排名	式(2)特征项	式(2)值	式(3)特征项	式(3)值	式(7)特征项	式(7)值
1	咨询	84.991 12	咨询	50.656 87	咨询	55.652 26
2	优惠	60.799 28	电话	34.538 78	电话	37.944 72
3	免费	60.799 28	优惠	32.236 19	免费	35.415 08
4	电话	60.260 75	免费	32.236 19	优惠	35.415 08
5	联系	56.243 37	联系	27.680 28	机会	20.237 19
6	公司	54.768 32	公司	26.205 23	收益	17.707 54
7	机会	39.694 76	机会	18.420 68	发票	17.707 54
8	手机	37.797 35	收益	16.118 10	开盘	15.177 89
9	收益	37.088 22	发票	16.118 10	手续	12.648 24
10	发票	35.811 97	酒店	15.183 38	联通	12.648 24

表 1 至表 4 中 3 种算法所取得的中心向量存在较大差异. 以 200 样本集为例,“联系”在式(2)和式(3)中均排名第 5,而到式(7)中已经跌出前 10. 同时,式(7)中出现了“联通”,体现出针对特定对象的个性化高权重项的提炼能力较强,即垃圾文本特征项权值被放大. 可见,从式(2)到式(7)所取得的中心向量正是垃圾文本识别所需要的预期结果.

3.4　θ 值计算

根据 3.3 中所得中心向量,分别针对不同样本集计算 3 种算法下文本与中心向量的夹角 θ 值,得到如下图结果.

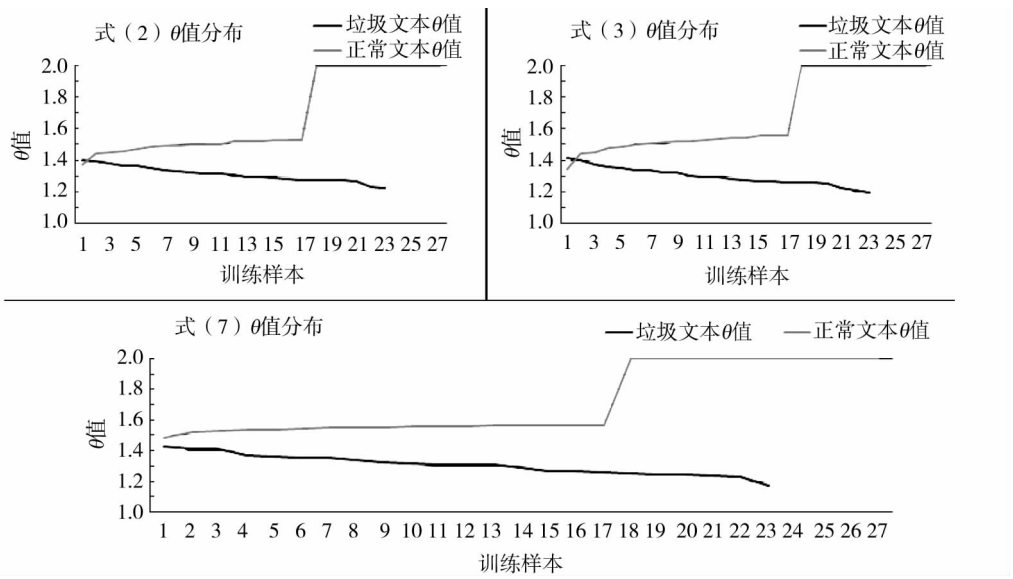

图 1　50 样本集分类比较

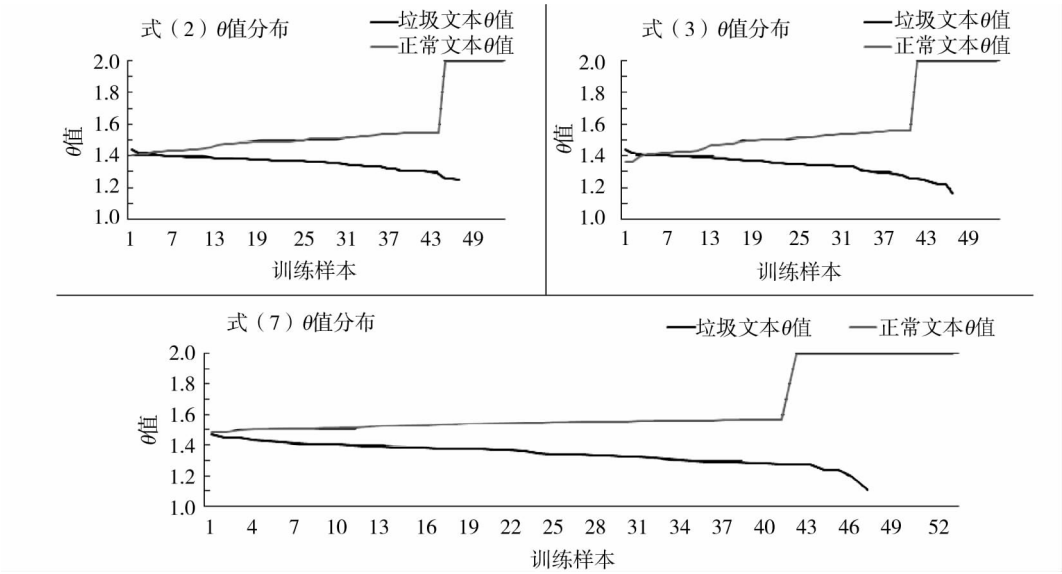

图 2 100 样本集分类比较

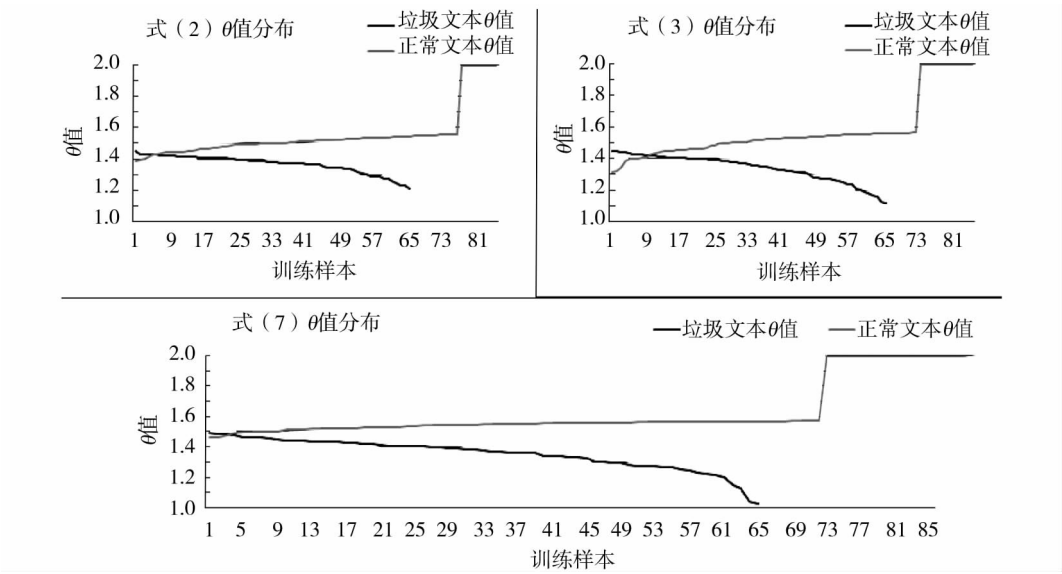

图 3 150 样本集分类比较

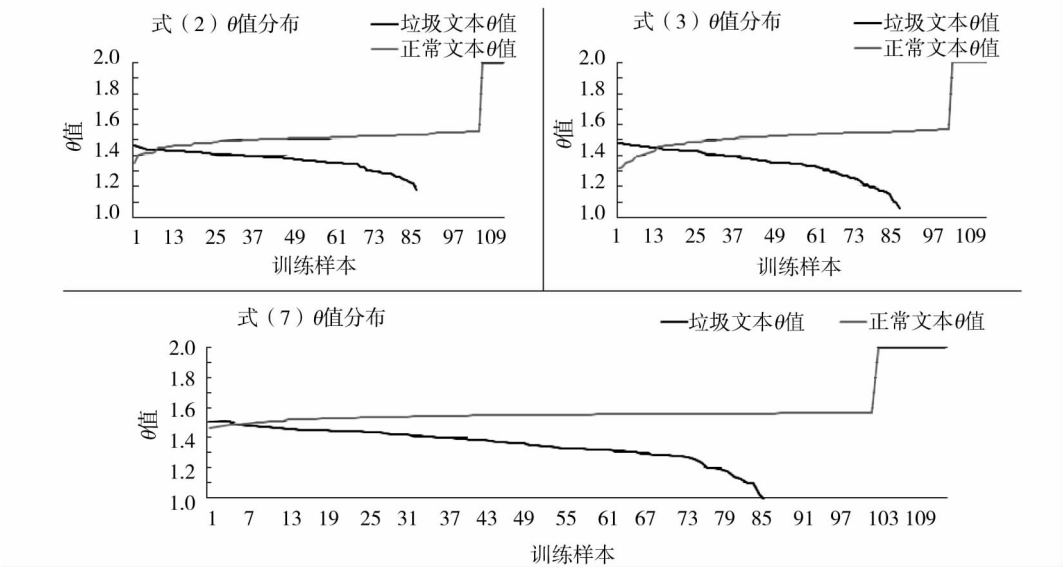

图 4 200 样本集分类比较

对 4 个样本集进行分类计算后发现,式(2)、式(3)和式(7)在不同样本中正常文本 θ 均值被依次放大. 分别为:50 样本集1.680 678,1.690 036 45,1.714 351 36;100 样本集1.575 345 598,1.606 135 88,1.642 712 829;150 样本集1.554 635 171,1.577,111,04,1.610 971 89;200 样本集1.537 697 897,1.556 835 672,1.589 275 5. 同时,3 种计算结果显示在不同数量样本集中,垃圾文本与正常文本交集区的垃圾文本数,式(7)的计算结果明显小于式(2)和式(3),详见表 6.

表 5　θ 均值一览表

	式(2) 垃圾文本 θ 均值	式(2) 正常文本 θ 均值	式(3) 垃圾文本 θ 均值	式(3) 正常文本 θ 均值	式(7) 垃圾文本 θ 均值	式(7) 正常文本 θ 均值
50 样本集	1.311 355	1.680 678	1.298 485	1.690 036	1.305 910	1.714 351
100 样本集	1.359 207	1.575 346	1.339 349	1.606 136	1.348 305	1.642 713
150 样本集	1.365 735	1.554 635	1.339 067	1.577 111	1.353 609	1.610 972
200 样本集	1.377 123	1.537 698	1.346 653	1.556 836	1.354 745	1.589 276

表 6　交集区垃圾文本数量一览表

	式(2)	式(3)	式(7)
50 样本集	3	4	0
100 样本集	5	6	0
150 样本集	29	45	6
200 样本集	64	62	10

实验数据结果表明在 50,100,150,200 等数量不同的样本集中,式(7)在垃圾文本过滤识别计算中,垃圾文本高聚类特征项权值被有效放大,即算法文本区分度增强. 同时,垃圾文本与正常文本交集区的正常文本数较之式(2)和式(3)有明显减少,式(7)在 4 个样本中的识别正确率可达到 90% 以上. 综合上述实验结果可见在不同样本训练集中公式改进效果均良好,基本达到预期要求.

4　结论

本文针对垃圾文本的特性,提出一种基于 VSM 的改进型文本过滤方法. 本方法针对垃圾文本二类分类特性,在综合现有 TFIDF 改进算法的基础上,对 TFIDF 特征项提取算法进行分类改进. 在减小样本空间偏差对计算结果的影响的同时,对垃圾文本类高聚类特征项权值进行有效放大,为垃圾文本的识别提供了一种高效、准确的改进算法.

另外,本文所提垃圾文本过滤算法由于未对垃圾文本进行分类,因此算法训练过程中提供更多数量,更全类型的垃圾文本样本对于提高算法识别正确率至关重要.

参考文献:

[1] 刘金岭,冯万利,高丽. 基于词汇链的中文变异垃圾短信文本语义识别[J]. 计算机工程与应用,2012,48(19):135 - 139.

[2] 刘伍颖,王挺. 适于垃圾文本流过滤的条件概率集成方法[J]. 计算机科学与探索,2010,4(5):445 - 454.

[3] 陈宝楼. K - Means 算法研究及在文本聚类中的应用[D]. 合肥:安徽大学,2013.

[4] Vahora S, Hasan M, Lakhani R. Novel approach: Naive Bayes with vector space model for spam classification [C]// 2011 Nirma University International Conference on Engineering. Ahmed - abad Gujarat: Nirma University Press,2011:1 - 5.

[5] 彭富明,张卫丰,彭寅. 基于文本特征分析的钓鱼邮件检测[J]. 南京邮电大学学报(自然科学版),2012,32(5):140 - 145.

[6] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009(29):167 - 180.

[7] 翟延冬,王康平,张东那,等. 一种基于 WordNet 的短文本语义相似性算法[J]. 电子学报,2012,40(3):617 - 620.

[8] 韩美灵,杨勇. 一种面向语义检索的向量空间模型改进方法[J]. 农业网络信息,2012(10):39 - 41.

[9] Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews[J]. Research Synthesis Methods,2011,2(1):1 - 14.

[10] 李慧,叶鸿,潘学瑞,等. 基于 SVM 的垃圾短信过滤系统[J]. 计算机安全,2012,13(6) :34 - 38.

[11] Bergholz A, Beer J D, Glahn S, et al. New filtering approaches for phishing email[J]. Journal of Computer Security,2010,18(1):7 - 35.