文章编号:1673-0062(2017)01-0077-04

基于模拟退火算法和发送行为的垃圾邮件检测模型

张　汛,刘朝晖*

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘　要:垃圾邮件给当今人们的生活带来严重的负面影响.虽然已经有很多过滤方法,但大多存在一定的不足之处,如检测时间长、召回率低等问题.本文提出了一种基于模拟退火算法和发送行为的垃圾邮件检测模型,旨在弥补已有检测方法的不足.模拟退火算法可能找到全局最优解,且收敛性强;而基于发送行为的垃圾邮件检测技术能显著提高服务器处理垃圾邮件的速度.本文尝试将二者相结合,以提高垃圾邮件的召回率及服务器处理能力.通过实验结果可以看出,该方法在垃圾邮件的召回率上有较大提升,较适于部署在小型邮件服务器上.

关键词:垃圾邮件;模拟退火算法;发送行为

中图分类号:TP393.4　　**文献标志码**:A

A Spam Detection Model Based on Simulated Annealing Algorithm and Sending Behavior

ZHANG Xun, LIU Zhao-hui*

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: Spam has brought serious negative impact on people's life today. Many filtering methods have been proposed, but less perfect. In this essay, a detection model based on simulated annealing algorithm and sending behavior is presented to cover some shortcomings of the existing methods. Simulated annealing algorithm can find the global optimal solution, and the convergence is strong, and the spam detection technology based on sending behavior can significantly improve the speed of the server to deal with spam. Through experiments, the recall rate of spam is significantly improved based on the model and it is more suitable for deployment in small mail server.

key words: spam; simulated annealing algorithm; sending behavior

收稿日期:2016-11-15

基金项目:南华大学研究生科研创新项目(2016XCX16)

作者简介:张　汛(1990-),女,硕士研究生,主要从事网络与信息安全的研究.E-mail:630226692@qq.com. * 通讯作者:刘朝晖,E-mail:10928478@qq.com

0 引 言

因特网的飞速发展给人们的生活带来便捷,电子邮件作为一种方便快捷的通信方式受到广大网民的青睐,很多企业和公司也把电子邮件作为内部员工的交流方法.然而,一些不法分子利用电子邮件向人们传递广告或者非法信息,既浪费用户的时间,又浪费网络流量,这些邮件给用户的正常工作和生活带来了极大的不便.因此,寻找一种可行且高效的反垃圾邮件技术变得尤为重要.

目前垃圾邮件的过滤算法主要分为三大类[1-3]:基于规则过滤,如 IP 地址过滤、黑白名单等技术;基于内容过滤,如贝叶斯算法,KNN 算法;基于发送行为的过滤.贝叶斯算法是目前最广泛采用的垃圾邮件过滤算法[4],该算法首先得出邮件内容中所包含的关键词,并分别计算这些关键词在垃圾邮件和正常邮件中出现的概率.当这些关键词出现时,求得当前邮件为垃圾邮件的概率,最后将这个概率与垃圾邮件的概率阈值做比较,从而判断该封邮件是否为垃圾邮件[5].贝叶斯算法需要先行备有大量的训练集,通过训练学习到一定的规则之后,才可自行判断邮件的属性.该算法的显著优点是稳定性强、判准率高,但是基于该算法的过滤器本身不具备增量学习的功能,导致其在召回率上所表现的性能不佳.

模拟退火算法的核心思想是,在寻找最优解的过程中并不只是接受当前最优解,对于非当前最优解,也会以一定的概率将这个非当前最优解加入到搜索最优解的路径中,从而得到最优解.本文拟通过这种方法提高垃圾邮件在检测中的召回率.

1 邮件特征提取与权值计算

邮件头中包含了发件人、收件人、邮件发送时间等重要信息,这些信息对判断一封邮件具有关键性作用.邮件在每次进行发送或转发时,都会被邮件用户代理添加相应的邮件头信息.

1.1 提取邮件头信息

结合邮件的行为特征及实际情况,可以建立表 1.

表 1 邮件属性表

Table 1 The table of mail attribute

特征	特征含义	具有该特征	不具有该特征	对应邮件中的字段
sender	发件人地址是否为空	1	0	from
receiver	收件人地址是否为空	1	0	to
sender_length	发件人地址长度是否合法	1	0	from
return_reply	邮件的回复地址是否和退回地址一致	1	0	return_path 和 delivered_to
sender_return	邮件的发件地址和退回地址是否一致	1	0	from 和 return_path

1.2 邮件行为特征权重计算

邮件头信息中提取出来的特征有很多,需要选择其中最具代表性的特征作为判断指标,从而减少判断时间和复杂度.利用主成分分析法可以将衡量一个样本的多个指标特征转换为较少的互相无关的综合指标.SPSS 是现在比较流行的进行主成分分析的软件,主成分分析方法的步骤为:

1)将样本矩阵进行标准化变换,得标准化矩阵 A;

2)对矩阵 A 求相关系数,得矩阵 A 的相关系数矩阵 B;

3)解矩阵 B 的特征方程,得 M 个特征根,确定主成分;

4)由特征根得出特征变量,并转换为主成分;

5)综合评价 M 个主成分.

在开始进行主成分分析时,将每个特征的有无分别用 1,0 来表示,同时设置一个标志 flag 字段,flag=1 表示该封邮件时正常邮件,flag=0 表示该封邮件是垃圾邮件,从而得出主成分分析法的初始样本矩阵输入.结合实际情况,最终提取出 sender,receiver,sender_length,return_reply,sender_return 等 5 个特征作为邮件判断指标,再分别抽取等数量的正常邮件和垃圾邮件,统计出每个行为属性在正常邮件和垃圾邮件中分别出现的次数.

权重是某个指标在某个评价体系中的贡献度,从指标权重的大小可知其在该评价体系中的地位高低,即将该指标在这个评价体系中的作用进行量化,从而可以知道该指标对这个评价体系贡献度.本文采用统计的方法来计算各个邮件特征对整个邮件属性的贡献度.设 weight[i]表示第 i

个特征的权重[6],则权重计算如公式(1)所示:

$$\text{weight}[i] = \frac{\sum_{j=1}^{n} h_j - \sum_{j=1}^{m} S_j}{\sqrt{\left(\sum_{j=1}^{n} h_j\right)^2 + \left(\sum_{j=1}^{m} S_j\right)^2}} \quad (1)$$

其中,m 表示训练集中垃圾邮件的个数,n 表示训练集中正常邮件的个数,h_j 表示第 i 个特征是否在第 j 封正常邮件中出现.分子表示第 i 个特征在正常邮件和垃圾邮件中出现的总次数之差,分母则对第 i 个特征的统计次数进行标准化,从而保证每个特征的权值分布在[-1,1]之间.

由公式(1)求得各个邮件特征对垃圾邮件和正常邮件的贡献度,将其作为该特征的权值.经计算得到的权值如表 2 所示.

表 2　特征权值集

Table 2　Feature weight set

特征	权值
sender	0.219
receiver	0.143
sender_length	-0.201
return_reply	0.510
sender_return	0.147

2　模拟退火算法

模拟退火算法思想最初由 Metropolis 等在 1953 年提出,该算法在组合优化领域有很好的表现.由于其良好的实用性和极佳的应用性能,引发众多专家学者对其进行深入的探索与研究.

模拟退火算法[7-8]的基本描述为:

(1)若 $J(Y(i+1)) \geq J(Y(i))$,即移动后的得到更优解,则总是接受该更优解;

(2)若 $J(Y(i+1)) < J(Y(i))$,即移动后的解比当前解要差,则以一定的概率接受该移动.随着时间的推移,此概率会逐渐降低,最后趋于稳定.整个过程从热力学原理中得到启发.在 T 温度条件下,降温概率计算如公式(2)所示:

$$P(\mathrm{d}E) = \exp\left(\frac{\mathrm{d}E}{\mathrm{K}T}\right) \quad (2)$$

其中,$\mathrm{d}E$ 为能量差,$P(\mathrm{d}E)$ 为降温概率,K 是一个常数.由公式可知,随着 T 值的逐渐减少,$P(\mathrm{d}E)$ 也会逐渐减少.

模拟退火算法在寻找最优解的过程中并不只是接受当前最优解,也会以一定的概率将非当前最优解加入到搜索最优解的路径中,从而得到最优解.

3　模型描述

基于模拟退火算法和发送行为的垃圾邮件检测模型如下:

1)以 weight_i 表示第 i 个特征的权值,β_i 表示是否具有第 i 个特征.若检测到邮件内容具有某种特征,则将该特征赋值为 1,若检测到邮件内容不具有某种特征,则将该特征赋值为 0.

则第 i 封邮件所具有的权值和为:

$$w_i = \text{weight}_1 \times \beta_1 + \text{weight}_2 \times \beta_2 + \cdots + \text{weight}_n \times \beta_n \quad (3)$$

其中,n 为邮件特征数.

2)以 sum_i 表示包含 i 封垃圾邮件权值的权值和,则权值和为:

$$\text{sum}_i = \sum_{i=0}^{n} w \quad (4)$$

3)由公式(3)和公式(4)计算所得结果可以看出,垃圾邮件的 w_i 几乎全为负数,而正常邮件的 w_i 几乎全为大于的数,故本文采取的方法为:

若 $\text{sum}_{i+1} \leqslant \text{sum}_i$,则将 w_{i+1} 加入到当前垃圾邮件集合中;

若 $\text{sum}_{i+1} > \text{sum}_i$,则以一定的概率将 w_{i+1} 加入到当前垃圾邮件集合中.

4　实验结果及分析

本实验数据取自 SpamAssassin 垃圾邮件公共语料库,其中 3 500 封正常邮件,3 500 封垃圾邮件.实验环境为 Windows 环境,利用 C 语言进行编程.将提取出来的头信息进行量化后作为程序输入,并对运行得出的分类集合进行统计与分析.

本文采用判准率、召回率和 F1 值作为垃圾邮件过滤算法的主要判断指标[9],判准率指算法是否能判别出垃圾邮件,召回率指算法能否将垃圾邮件全数检测出来,F1 值反映过滤算法的综合过滤能力.

(1)判准率 $P = \text{TS}/(\text{TS}+\text{FS})$

(2)召回率 $R = \text{TS}/(\text{TS}+\text{FL})$

(3)$F1 = 2PR/(P+R)$

表 3　判断指标

Table 3　Judgment index

	实为垃圾邮件	实为正常邮件
判定为垃圾邮件数	TS	FS
判定为正常邮件数	FL	TL

模型算法可以产生两个结果集:一个垃圾邮件集和一个正常邮件集.为使实验更为稳定、准确,迭代实验过程中所采用的邮件增量值为 1 000,同时对 P 值及 R 值进行计算.计算结果如表 4 所示.

由表 5 可知,基于模拟退火算法和发送行为的垃圾邮件检测模型在判准率和综合过滤效果上比贝叶斯算法稍差,但其在一定条件下的召回率却在贝叶斯算法之上.实验结果表明,在邮件数量较小的情况下,其召回率会有明显提升,因此更适于部署在小型邮件服务器上.在接受劣质解过程中,通过加入训练集的平均特征权值,可以有效提升算法运行的稳定性.实验过程中造成判准率较低的原因可能有:未精确细化邮件特征,权值计算出现偏差等.

表 4 不同邮件数量下的 P 值和 R 值

Table 4 Values of P and R in different number of mails

邮件数量	P 值	R 值
6 000	0.828 3	0.853 3
5 000	0.827 5	0.855 8
4 000	0.831 8	0.861 1
3 000	0.821 4	0.874 6
2 000	0.817 8	0.879 1

表 5 结果分析

Table 5 Result analysis

	特征权重	判准率 P	召回率 R	$F1$ 值
模拟退火算法和行为特征加权算法	有	0.82	0.86	0.82
贝叶斯算法	无	0.91	0.80	0.85

5 结 语

本文将模拟退火算法与垃圾邮件的发送行为相结合,通过不断搜索解空间进而找到全局最优解,由此提出一种基于模拟退火算法和发送行为的垃圾邮件检测模型.该模型充分考虑了多特征及权值选择过滤等因素的影响,有效提高垃圾邮件的召回率,且伴随邮件数量的减少,垃圾邮件召回率也会随之提高.下一步工作考虑在随机过程中加入记忆机制,对某个阶段的极值进行记录,以该极值作为判断依据,使邮件综合过滤效果获得更大程度的提升.

参考文献:

[1] 师文轩,殷爱茹.垃圾邮件的概念漂移及过滤技术研究[J].中国科技论文,2014,9(10):1111-1117.

[2] RODAN A,FARIS H,ALQATAWNA J.Optimizing feed-forward neural networks using biogeography based optimization for e-mail spam identification[J].International Journal of Communications Network & System Sciences,2016,9(1):20-25.

[3] SHAHI T B,YADAV A.Mobile SMS spam filtering for Nepali text using naïve bayesian and support vector machine[J].International Journal of Intelligence Science,2014,4(1):25-27.

[4] 赵晓丹,徐燕.垃圾邮件分类技术对比研究[J].信息网络安全,2014(2):79-80.

[5] 李雯,刘培玉.基于贝叶斯的垃圾邮件过滤算法的研究[J].计算机工程与应用,2007,43(23):174-176.

[6] 李璇.基于行为识别的垃圾邮件过滤技术的研究与应用[D].武汉:武汉理工大学,2013.

[7] METROPOLIS R,ROSENBLUTH A,TELLER A,et al.Simulated annealing[J].Journal of Chemical Physics,1953,21(161-162):1087-1091.

[8] 王雪梅,王义和.模拟退火算法与遗传算法的结合[J].计算机学报,1997,20(4):382-384

[9] 甘冬连,张永,刘博.基于 MapReduce 并行 SVM 的垃圾邮件分类[J].软件导刊,2016,15(6):11-12.