垃圾邮件过滤系统的评估模型研究

项涛　龚俭　丁伟
（东南大学计算机科学与工程学院，南京，210096）
（江苏省计算机网络技术重点实验室，南京，210096）

摘要：在分析现有垃圾邮件过滤系统评估存在的各指标值不一致的情况的基础上，提出一个综合评估垃圾邮件过滤系统过滤效果的评估模型。依据该评估模型，设计和实现了一个评估系统；并使用该评估系统评估了多个开源的垃圾邮件过滤系统。实验结果表明本文提出的评估模型能够有效的综合评估垃圾邮件过滤系统的过滤效果。
关键字：评估；垃圾邮件过滤系统；评估模型；评估指标；评估方法
中图分类号：TP393

A Research on the Evaluation Model of Anti-Spam Filtering

Xiang Tao　Gong Jian　Ding Wei
(School of Computer Science and Engineering, Southeast University, Nanjing 210096,china)
(Key Laboratory of Computer Network Technology of Jiangsu Province, Nanjing 210096, china)

Abstract: The differences among the value of existent metrics, which are used to evaluate Anti-Spam Filtering, are analyzed in detail. An evaluation model is proposed to evaluate the filtering effect of Anti-Spam Filtering. Based on this model, an evaluation system is designed and implemented which has been used to evaluate many open-source Anti-Spam Filtering. The results demonstrate that the evaluation model proposed in this paper has a good effect on evaluating the Anti-Spam Filtering.
Key words: Evaluation; Anti-Spam Filtering; Evaluation Model; Evaluation Index; Evaluation Method

0　引言

[1]近些年来，垃圾邮件的危害越来越大，给网民们的生活和工作带来极大的不便。对垃圾邮件的防范的研究是当前的一个研究热点，有很多的垃圾邮件过滤技术，有些是基于规则过滤的，比如：Spamassassin 规则库；有些是基于统计学方法过滤的，比如：Bayes 方法；有些是采用机器学习的方法的，比如：支持向量机，K 邻近法，神经网络，模式匹配。此外，还有很多其他的方法，比如：最小墒，黑白名单，加密算法等等。目前市面上存在大量的垃圾邮件过滤系统，有基于客户端的，有基于服务器的，还有专门为企业开发的垃圾邮件防范的系统解决方案，他们大都是采用一种或多种以上过滤技术制作而成的，也都宣称他们的过滤系统能够获得很好的过滤效果。

然而，面对众多的垃圾邮件过滤系统，如何选择一个过滤准确度高的而又符合用户需求的过滤系统却依然没有一个好的依据。垃圾邮件过滤系统的评估研究就是旨在从第三方的角

作者简介：项涛（1983-），男，江西省新余市人，硕士研究生，方向为网络应用。龚俭（1957-），男，教授，博士生导师，研究方向为网络安全，网络管理，网络行为学。丁伟（1962-），女，教授，博士生导师，研究方向为网络测量，网络行为学。

度对垃圾邮件过滤系统展开评估；根据评估的结果，一方面，用户可以从中选择符合自己需求的过滤系统；另一方面，厂商也可以评估他们的各个不同版本的产品。

当前，国外有一些研究者对垃圾邮件过滤系统的评估做了相关的研究。Ion Androutsopoulos[1]使用加权正确率，加权错误率以及总代价比率（TCR）等代价敏感指标，以 Ling-Spam 为标准邮件集评估了基于简单贝页斯和基于关键字的垃圾邮件过滤系统。Gordon Cormack[2]使用 ROC 曲线下的面积，以及随邮件数量增长，误报率的大小作为评价指标，以 8 个月的个人邮件作为邮件测试集，评估了 CRM114、DSPAM、Bogofilter、SpamProbe、SpamAssassin 等多种邮件过滤系统。

前人的研究（文献[1][2][3][4][5][6]）大都是通过罗列多个评价指标的值来判定垃圾邮件过滤系统过滤效果的好坏，然而当各指标值出现不一致的情况时（一些过滤系统在某几个指标上优于其他系统，而在其他指标上又不如他们），就给判定垃圾邮件过滤系统过滤效果带来了困难，用户也难以做出正确的决定。因此，本文提出一个综合评估模型，对多个评估指标计算加权和来综合反映垃圾邮件过滤系统的过滤效果。

本文将在第二节论述影响垃圾邮件过滤系统评估的因素，于第三节给出垃圾邮件过滤系统过滤准确度的综合评估模型；依据此模型，第四节设计和实现了过滤准确度的评估系统，第五节给出几个开源的垃圾邮件过滤系统过滤准确度的评估结果，最后是本文得出的结论。

1 影响评估的因素

影响垃圾邮件过滤系统评估的因素有很多，包括主观的，客观的，甚至一些外在的因素。在它们当中最重要的因素有评估模型，标准邮件集以及测试方法。

标准邮件集：提供给垃圾邮件过滤系统训练和测试，且带有标准答案的邮件集合。由于涉及到隐私，且一封邮件是垃圾邮件还是正常邮件对不同的用户可能有不同答案，故标准答案通常是由特定用户反馈得到的。常见的标准测试集有：Spamassassin-Corpus,PU-Corpus, Ling-Spam.

Ling-Spam：它是由 Ion Androutsopoulos 收集的一组语言学家之间互相交流的 e-mail 和一些已知的垃圾邮件的混合。它包括 481 封垃圾邮件和 2412 封正常邮件，其中邮件中不包括附件，标点符号，HTML 标记等，同一天收到的多封相同邮件也只选取一封。

PU-Corpus：它由单个用户收集其自身的邮件构成的,属于私人邮件。它包括 481 封正常邮件和 618 封垃圾邮件。其中的正常邮件中的内容(包括单词,数字,标点等)都被替换掉了。可以从站点 http://www.iit.demokritos.gr/~ionandr 获得该标准邮件集。

Spamassassin-Corpus：它包括 6047 封邮件，其中 1897 封垃圾邮件，4150 封正常邮件。其中正常邮件分为两部分：其中有 250 封属于难以判定的邮件，它们具有很多垃圾邮件中常有的 HTML 标记，加颜色字体等等特点，另 3900 封属于一般的正常邮件。

测试方法：垃圾邮件过滤系统使用标准邮件集进行测试和训练的方法。最基本的测试方法是十字交叉法，将标准邮件集平均分成十等份，然后按顺序选择其中一份用作测试，其他九份用作训练，每次都计算相应的各个指标值。最后对十次获得的指标值求平均。还有其他的测试方法，比如：逐步增大训练集的大小，比较各指标值变化情况；调整用于训练和测试的垃圾邮件和正常邮件的比例等等。

评估模型：垃圾邮件过滤系统评估的最关键的因素，它包括评估指标和评估方法。评估指标指反映垃圾邮件过滤系统某方面的参数。有些评估指标可以通过其他指标计算出来，所以需要选择合适的评估指标。评估方法是对评估指标采取的一系列操作，包括确定指标权重系数，指标合成等等。

除以上之外，垃圾邮件过滤系统的评估系统还包括待评估的邮件过滤系统，它由训练器和分类器两部分组成。训练器对邮件集训练，训练完后使用分类器测试过滤系统的过滤效果。

2 过滤准确度的评估模型

过滤准确度是评估垃圾邮件过滤系统过滤效果好坏最重要的一个角度，也是用户和开发者最关注的一个角度。它反映了垃圾邮件过滤系统对到来的垃圾邮件判定的正确程度和完备程度。本模型首先给出反映过滤准确度的评估指标及相应的计算方法，然后对这些评估指标设计了一个综合评估方法，依据该评估方法综合判定垃圾邮件过滤系统的过滤效果的好坏。

2.1 评估指标

目前常用的评价过滤准确度的指标是误报率和漏报率。除此之外，结合文本分类评估，医疗诊断评估等研究领域的成果。在本论文中，可以使用正确率 Acc，垃圾邮件查全率（Spam Precision），垃圾邮件查漏率（Spam Recall），正常邮件查全率（Ham Precision），正常邮件查漏率（Ham Recall），F1 值[7]等指标来评估邮件过滤系统的过滤准确度。计算方法如下：

表 1：复合矩阵

过滤器 \ Golden Standards	Spam	Ham
Spam	A	B
Ham	C	D

其中 A 代表过滤器判定为垃圾邮件且“Golden Standards”也判定为垃圾邮件的数量
B 代表过滤器判定为垃圾邮件但“Golden Standards”判定为正常邮件的数量
C 代表过滤器判定为正常邮件但“Golden Standards”判定为垃圾邮件的数量
D 代表过滤器判定为正常邮件且“Golden Standards”也判定为正常邮件的数量

正确率(Accuracy) $Acc=\frac{A+D}{A+B+C+D}$，　垃圾邮件查全率(Spam Recall) $SR=\frac{A}{A+C}$

垃圾邮件查对率(Spam Precision) $SP=\frac{A}{A+B}$　正常邮件查全率(Ham Recall) $HR=\frac{D}{B+D}$

正常邮件的查对率(Ham Precision) $HP=\frac{D}{C+D}$　F_1值 $F1=\frac{2*SR*SP}{SR+SP}=\frac{2*A}{2*A+B+C}$

误报率（Misclassification） $Mis=1-HR=\frac{B}{B+D}$　漏报率 $Loss=1-SR=\frac{C}{A+C}$

误报率和漏报率可以通过垃圾邮件和正常邮件的查全率来表示，故在本论文中，将使用前面六个评估指标来评估垃圾邮件过滤系统的过滤准确度。

2.2 评估方法

各评估指标存在数值大小，单位等差异，为使它们在同等地位下参与评估，需要对评估指标进行无量纲化处理。本文将采取极值处理法对评估指标进行无量纲化处理。其原理如下：

设 $x_{ij}\,(i=1,2,\mathbf{L},n;\ j=1,2,\mathbf{L},m)$表示系统 j 的第 i 个指标值，$M_i=\max\{x_{i1},x_{i2},\mathbf{L},x_{im}\}$

$m_i=\min\{x_{i1},x_{i2},\mathbf{L},x_{im}\}$；则 $x_{ij}^*=\frac{x_{ij}-m_i}{M_i-m_i}$ 是无量纲的，且 $x_{ij}^*\in[0,1]$。由于各系统的同一指标值 $x_{ij}\,(j=1,2,...,m)$ 之间的差别很小，极值处理法可以将这种差别扩大到$[0,1]$之间。

过滤准确度的高低可以通过计算评估指标的加权和来描述，见公式（1）。

$W=\sum_{i=1}^{n}w_ix_i$ （1）。x_i 为评估指标无量纲化后的值，w_i 为指标 i 的权重系数。w_i 可以采

用G_1法[8]进行计算。其原理是：首先相对于某评价准则给出各指标的重要性程度的大小关系，记为$x_1 \ \mathbf{f}\ x_2 \ \mathbf{f}\ \mathbf{L}\ \mathbf{f}\ x_n$。再给出评估指标$x_{k-1}$与$x_k$的重要性程度之比$w_{k-1}/w_k$的理性判断，分别为：$w_{k-1}/w_k = r_k$，$k = n, n-1, \mathbf{L}, 3, 2$。$r_k$参考赋值为$\{1.0, 1.2, 1.4, 1.6, 1.8\}$。然后依据给出的$r_k$的理性赋值，按照公式（2）计算权重系数$w_i$。

$$w_n = (1 + \sum_{k=2}^{n} \prod_{i=k}^{n} r_i)^{-1}, \quad w_{k-1} = r_k w_k, \quad k = n, n-1, \mathbf{L}, 3, 2 \qquad (2)$$

根据公式（1）计算得到的W值越大，则说明垃圾邮件过滤系统的过滤准确度就越高。即过滤效果越好。

3 评估系统的设计

依据以上的评估模型，设计垃圾邮件过滤系统的评估系统结构图如下：

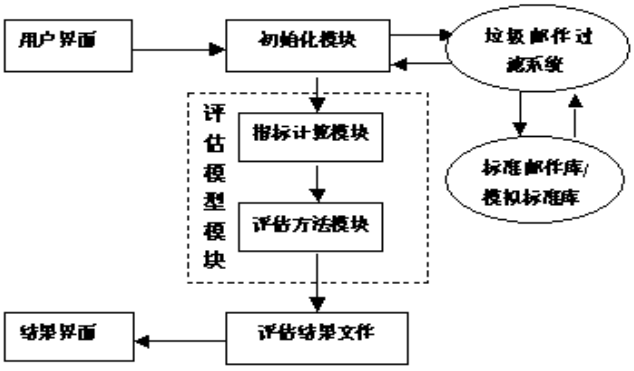

图 1 评估系统结构图

系统的工作流程为：

1．用户通过用户界面配置参数（包括垃圾邮件过滤系统存放路径，标准邮件库存放路径，测试方法等）。

2．初始化模块接收用户的配置，然后驱动垃圾邮件过滤系统对标准邮件库的训练和测试。

3．垃圾邮件过滤系统对标准邮件库进行训练和测试，将结果反馈给初始化模块。

4．当所有垃圾邮件过滤系统都完成训练和测试后，初始化模块将它们

5．评估模型模块接收初始化模块的的测试结果，依据测试结果计算指标，通过评估方法模块计算出对垃圾邮件过滤系统评估的最后结果，并存放到评估结果文件中去。

6．将结果显示先结果界面上。

4 实验结果与分析

依据以上的评估系统的设计，本文实现了一个垃圾邮件过滤系统的评估系统。本节中使用该评估系统评估 Bogofilter，Dbacl，CRM-114，SpamProbe，SpamBayes 五个开源邮件过滤系统的过滤效果。采用的标准邮件集是 Ling-Spam，PU-corpus，以及 Spamssassin-Corpus。测试方法将采用十字交叉法。

本文将计算 *Acc*，*SR*，*SP*，*HR*，*HP*，*F*1六个指标的权重系数，然后通过计算它

们的加权和来判定垃圾邮件过滤系统的过滤效果。实现流程为：首先确定一个标准测试集，并将五个垃圾邮件过滤系统对该标准测试集进行训练和测试，分别计算过滤系统的各个指标值。其次，将计算好的各过滤系统的指标值按上述极值处理法无量纲化。再对无量钢化后的指标值计算加权和。权重系数确定使用上述的 G1 法，具体如下：考虑到正常邮件被误报的代价要远大于垃圾邮件被误报的代价，所以可以认为以上六个指标重要程度的偏序关系为：$HR \succ SP \succ Acc \succ HP \succ SR \succ F1$。再根据前面建立的模型，以及指标之间的重要程度的赋值参考表，本文认为：$r_2 = \frac{w_1}{w_2} = 1.2$，$r_3 = \frac{w_2}{w_3} = 1.2$，$r_4 = \frac{w_3}{w_4} = 1.2$，$r_5 = \frac{w_4}{w_5} = 1.0$，$r_6 = \frac{w_5}{w_6} = 1.2$。

则根据公式 $w_k = \left(1 + \sum_{k=2}^{6} \prod_{i=k}^{6} r_i\right)^{-1}$，$w_{k-1} = r_k * w_k, (k = 6,5,...,3,2)$。由此可以推出

$W^* = [0.2400, 0.2000, 0.1666, 0.1389, 0.1389, 0.1157]$。

依据以上步骤，使用 Ling-Spam 标准邮件集，评估五个开源垃圾邮件过滤系统的过滤效果。首先计算各过滤系统的指标值如下：

表 2：Ling-Spam 标准邮件集下，各垃圾邮件过滤系统的指标值

Metrics / Filters	ACC	SR	SP	HR	HP	F1
Bogo	0.909689	1.000000	0.457568	1.000000	0.902536	0.621637
Dbacl	0.988235	0.987707	0.941964	0.997510	0.988564	0.963487
CRM114	0.979585	0.948699	0.929592	0.989625	0.986069	0.937672
SpamProbe	0.974394	1.000000	0.846514	1.000000	0.970400	0.914968
SpamBayes	0.988927	0.997872	0.935714	0.999585	0.987333	0.965422

从上面可以看出，Ling-Spam 标准邮件集，CRM114 在 SP，$F1$，HP 上要明显好于 SpamProbe，而 SpamProbe 在 SR，HR 上较 CRM114 有较大优势。在其他指标上两者不相上下，若仅仅依靠以上指标值比较 Dbacl 与 SpamBayes 两系统过滤准确度显然是困难的。依据以上过滤准确度的评估模型，计算最终的过滤系统的加权和为 0.819253，0.978415，0.963855，0.951187，0.979243。依据该综合值判定过滤系统的过滤效果好坏的顺序为：SpamBayes，Dbacl，CRM114，SpamProbe，Bogo。由此可见综合评估可得 CRM114 略好于 SpamProbe。

同理，使用 Spamassassin-Corpus 标准邮件集，并计算各过滤系统的指标值结果如下：

表 3：Spamassassin-Corpus 标准邮件集下，各垃圾邮件过滤系统的指标值

Metrics / Filters	ACC	SR	SP	HR	HP	F1
Bogo	0.910282	0.999531	0.721238	0.999683	0.828034	0.830990
Dbacl	0.939635	0.924782	0.846003	0.988991	0.846939	0.878603
CRM114	0.962189	0.910319	0.936645	0.960875	0.882165	0.912377
SpamProbe	0.966501	0.987044	0.915366	0.978279	0.885686	0.948683
SpamBayes	0.969983	0.977958	0.922794	0.994573	0.882484	0.948671

同样可以发现，在 Spamassassin-Corpus 标准邮件集下，Bogo 的 SR， HR 两指标值最高，而其他指标值却最小，同样需要综合的评估值来反映其过滤准确度。依据综合的评估模型，计算 Spamassassin-Corpus 下过滤系统的加权和为 0.885819，0.910848，0.932778，0.948764，0.953032。故在 Spamassassin-Corpus 下各过滤系统的过滤效果好坏的顺序为 SpamBayes，SpamProbe，CRM114，Dbacl，Bogo。

同样，使用 PU-Corpus 标准邮件集，并计算各过滤系统的指标值如下：

表 4：PU-Corpus 标准邮件集下，各垃圾邮件过滤系统的指标值

Metrics / Filters	ACC	SR	SP	HR	HP	F1
Bogo	0.559633	0.0	0.0	1.0	0.559633	0.0
Dbacl	0.977064	0.975218	0.972917	0.980328	0.979281	0.973761
CRM114	0.949541	0.941286	0.945833	0.952459	0.958081	0.942824
SpamProbe	0.945439	0.999531	0.756689	0.999683	0.925730	0.852535
SpamBayes	0.977982	0.981406	0.968750	0.985246	0.976315	0.974633

在 PU-Corpus 标准邮件集下，很明显，Bogo 的过滤准确度最差，然而，其他的过滤系统的好坏难以判断，SpamProbe 在 SR， HR 两指标上非常好，而在 SP， $F1$， HP 指标上非常差。在 PU-Corpus 标准集下，各过滤系统的评估值为 0.410968，0.976785，0.948857，0.914829，0.977833。由此判定垃圾邮件过滤系统好坏顺序为 SpamBayes，Dbacl，CRM114，SpamProbe，Bogo。

分析以上实验结果可以发现，在 Ling-Spam，PU-Corpus 两标准邮件集下，由综合结果判定的各垃圾邮件过滤系统的过滤准确度的高低顺序是完全一致的。在 Spamassassin-Corpus 下，评估模型判定的顺序与在以上两邮件集下的结果大体是一致的。说明本文建立的过滤准确度的评估模型能够有效的评估垃圾邮件过滤系统的过滤效果。但由于 Spamassassin-Corpus 下判定的顺序与 Ling-Spam，PU-Corpus 不完全一样，也说明垃圾邮件过滤系统的评估和标准邮件集有很大关系。分析 Spamassassin-Corpus 可以发现，除了普通的正常邮件外，它还包含 250 封包含许多垃圾邮件特点的难以判定的邮件，可以认为这些邮件对于过滤系统的过滤效果产生较大的影响。为此，本文选用 Spamassassin-Corpus 中的 150 封垃圾邮件，100 封普通正常邮件和 250 封难以判定的正常邮件作为标准邮件集，并使用它对 Dbacl，CRM114，SpamProbe 进行评估。计算综合评估值分别为：0.818826，0.904534，0.918826。说明对于这些难以判定的邮件，过滤效果按 Dbacl，CRM114，SpamProbe 越来越好。将这个结果与在 Spamassassin-Corpus 集的整体效果对照，可以认为 Dbacl，CRM114，SpamProbe 的评估效果的顺序与其他两个邮件集不同是邮件集的原因，而不在于模型本身。

5 结束语

垃圾邮件过滤系统的评估研究旨在为用户和研究者提供一个判定垃圾邮件过滤系统好坏的依据，具有很强的实际应用价值。影响垃圾邮件过滤系统的评估的因素有很多，目前的评估研究主要集中在提出新的评估指标，采用不同的测试方法和不同的标准邮件集来判定垃圾邮件过滤系统的好坏。本文分析了现有过滤指标在评估垃圾邮件过滤系统时存在不一致的情况，在此基础上提出了一个评估垃圾邮件过滤系统过滤准确度的综合评估模型，并使用依据此模型实现的评估系统评估了多个开源的垃圾邮件过滤系统。结果表明在不同的标准邮件集下，该评估模型都能取得一致的评估结果。

参考文献

[1] Androutsopoulos I., Koutsias J., Chandrinos K.,et al. An evaluation of naive bayesian anti-spam filtering[A].In:G.Potamias, V.Moustakis and M.van Someren. proceedings of the workshop on Machine Learning in the New Information Age[C].Barcelona,Spain:11th Eueopean Conference on Machine Learning,2000.pp.9-17.

[2] Gordon Cormack, Thomas Lynam, A Study of Supervised Spam Detection applied to Eight Months of Personal E-Mail[A], In Proceedings of Conference on Email and Anti-Spam (CEAS) 2004[C], Mountain View, CA, July 30 and 31, 2004.

[3] Jose Maria,Gomez Hidalgo.Evaluation Cost-Seneitive Unsolicited Bulk Email Categorization[A].In Proceedings of Conference of SAC[C].Madrid,Spain.ACM .2002.615.

[4] Harris Drucker,Donghui Wu,Vladimir N.Vapnik.Support Vector Machine for Spam Categorization[J]. IEEE TRANSACTIONS ON NEURAL NETWORKS, 1999. 10(5).1 048

[5] Andrew Tuttle, Evangelos Milios, Nauzer K..An Evaluation of Machine Learning Techniques for Enterprise Spam Filters[R].Canada: Faculty of Computer Science in Dalhousie University. 2004.

[6] Androutsopoulos I., Koutsias J., Chandrinos K.,et al.An Experimental Comparision of Navie Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Message[A].In Proceeding of the 23rd Annual International ACM SIGIR Conference on Research and Development in information Retrieval[C].Greece: ACM. 2000,160.

[7] 宋枫溪，高林. 文本分类器性能评估指标[J]. 计算机工程，30（13）：107.

[8] 郭亚军著. 综合评价理论与方法[M]. 科学出版社. 北京. 2002.. 30.

作者联系方式：项涛　江苏省南京市东南大学华东（北）地区网络中心　210096
手机 13770563775　E-mail：txiang@njnet.edu.cn