一种基于人工免疫的多层垃圾邮件过滤算法

张泽明, 罗文坚, 王煦法
(中国科学技术大学计算机科学技术系, 安徽合肥 230027)

摘 要: 随着电子邮件日益广泛的使用, 如何有效地避免和防范垃圾邮件的侵扰已成为一个亟待解决的问题. 受生物免疫系统自我保护机制的启发, 本文提出了一种基于人工免疫的多层垃圾邮件过滤算法, 利用分层检测的思想来过滤垃圾邮件. 文中给出了针对多层过滤算法中获得性免疫层的垃圾邮件过滤测试实验, 实验结果表明本算法在垃圾邮件过滤中能得到较高的召回率、精确率和正确率. 文中也指出了可以通过合理地设置各检测器层之间的与或关系来得到更好的垃圾邮件过滤效果.

关键词: 人工免疫; 获得性免疫; 垃圾邮件; 多层垃圾邮件过滤
中图分类号: TP18 文献标识码: A 文章编号: 0372-2112 (2006) 09-1616-05

A Multilevel Spam Filtering Algorithm Based on Artificial Immunity

ZHANG Ze-ming, LUO Wen-jian, WANG Xu-fa
(Dept. of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China)

Abstract With the growing use of Email, it is urgent to resolve the severe problem of Spam. Inspired by the self-protection mechanism of biological immune system, a multilevel Spam filtering algorithm is proposed. Experiments on the acquired immune layer have been presented and the results show that the algorithm can get high grade in recall, precision and accuracy of Spam filtering. It is also noted that better performance can be achieved by properly arranging the relations between the detector layers.

Key words artificial immune system; acquired immunity; spam; multilevel spam filtering

1 引言

目前, 以电子邮件为媒介的信息交流方式越来越普遍, 但随之而来的垃圾邮件 (Spam)却严重干扰了正常的信息交流. 所谓垃圾邮件, 一般是指不请自来的 (Unsolicited)、带有商业性或政治性目的大批量的邮件. 对用户而言, 垃圾邮件严重干扰了个人的正常信息交流, 浪费大量时间和精力; 对网络而言, 垃圾邮件占用大量的传输、存储和运算资源, 造成网络资源的浪费, 同时也对系统安全构成威胁.

垃圾邮件问题已经引起人们的高度重视, 目前已有不少垃圾邮件过滤技术的研究工作. 根据邮件系统的角色结构, 垃圾邮件过滤可以分为两类: 一类是基于服务端的过滤, 一类是基于客户端的过滤. 本文所提出的基于人工免疫的多层垃圾邮件过滤算法是基于客户端的垃圾邮件过滤. 从垃圾邮件过滤技术上划分, 目前主要有三类:

第一类是基于地址的过滤, 主要是黑名单 /白名单技术. 黑名单 /白名单是一种已经被广泛采用的垃圾邮件过滤技术[1]. 黑名单是一组邮件服务器的 IP地址、域名或 Email地址列表, 来自黑名单列表中的任何邮件都被认为是垃圾邮件. 相对于黑名单, 来自白名单列表中的任何邮件都被认为是合法邮件. 这种方法简单易行, 但垃圾邮件的漏报率和误报率都会很高.

第二类是基于规则的过滤, 包括: 信头分析、群发过滤和关键词精确匹配等. 这类方法效率较高, 但不足之处在于规则需要用户手工创建和维护, 人为因数较多, 对用户要求较高[1, 2].

最后一类是基于内容的过滤, 包括: Bayesian分类算法[2, 3]、基于实例的方法、Boosting方法、支持向量集[4, 5]等. 文献[6]中采用人工免疫系统来进行邮件的分类, 提出了 AISEC算法来连续地将邮件分为感兴趣和不感兴趣两类. 文献[7, 8]也采用人工免疫的方法, 以正则表达式为抗体来检测垃圾邮件. 文献[9]中采用人工神经网络的方法来进行垃圾邮件的过滤. 这类基于学习的算法性能较好, 但计算复杂度很高, 同时也有很大的不确定性.

收稿日期: 2005-02-03; 修回日期: 2006-06-14
基金项目: 国家自然科学基金 (No. 60404004)

本文借鉴生物免疫的多层保护机制, 将人工免疫模型和多种现有垃圾邮件过滤技术有机地结合起来, 提出了一种基于人工免疫的多层垃圾邮件过滤算法 (MSFA-AI: a Multilevel Spam Filtering Algorithm based on Artificial Immunity). 本文其余部分的组织如下: 第二节简要介绍生物免疫系统的多层防护机制; 第三节详细说明算法 MSFA-AI; 第四节是实验及讨论, 最后是结束语.

2　生物免疫系统

生物免疫系统的自我防护机制是多层次的, 由分布在几个层次的防御子系统组成. 第一层是皮肤等物理屏障, 阻止大多数大分子病原体; 第二层是生理屏障, 即温度、PH 值之类的条件; 第三层是固有免疫系统, 由噬菌细胞等组成; 最后一层是获得性免疫系统, 由淋巴细胞等组成. 生物免疫系统的主要功能是区分自我和有害的非我[8-10].

获得性免疫系统 (又名自适应免疫系统) 主要由淋巴细胞组成. 淋巴细胞是免疫系统中起主要作用的微小的白细胞, 它有两种主要类型: T 细胞和 B 细胞. 为防止自体免疫的发生, 未成熟 T 细胞和 B 细胞分别在胸腺和骨髓中被耐受化, 经非选择而成熟. B 细胞与抗原结合后被激活, 在淋巴结中经历克隆扩增与体细胞高频变异, 以得到与抗原决定基更高亲和度的 B 细胞, 达到亲和度成熟. 新 B 细胞与病原体抗原决定基结合成功后, 离开淋巴结, 分裂为浆细胞与记忆 B 细胞. 浆细胞分泌抗体来消灭外来病原体. 记忆 B 细胞使得免疫系统在再次遭受类似病原体入侵时能快速反应并反击抗原, 这个过程称为二次免疫应答[10].

3　基于人工免疫的多层垃圾邮件过滤算法

在生物免疫系统结构及自适应免疫系统工作机制的启发下, 本文提出了一种基于人工免疫的多层垃圾邮件过滤算法 (MSFA-AI). 算法框架如图 1所示.

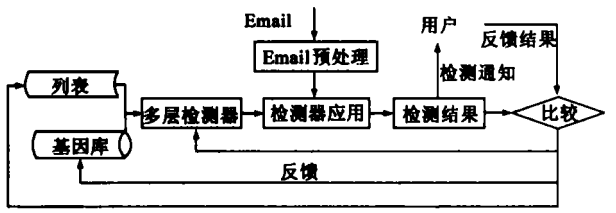


图 1　算法框架

算法流程主要分为五个阶段: Email预处理阶段、检测器的产生阶段、检测器的应用阶段、检测器的进化阶段和检测结果通知. 以下将具体说明算法的主要流程.

3.1　Email预处理

Email预处理就是将邮件转化为系统易于处理的模式的过程. 文中将每封 Email预处理为三部分, 表示如下:

(Sender Address) (< Sender, Subject, Content>) (AttachmentMD5)

其中, 第一部分为发信人地址; 第二部分由三个无序的可变长的子矢量组成, 分别是发送者、主题和内容子矢量. 子矢量的处理是一个自然语言的处理过程, 主要有三步: 分词、停用词和词根化. 分词是将一篇文本分解为单词的过程; 停用词删除一些高频但无意义的词, 如: "of"、"to"等; 词根化是将各种形式的单词还原为基本单词的过程, 如: "stopping"还原为"stop"等. 作为子矢量的处理的扩展, 本文采用 TF-IDF算法来评估每个单词的权重. 第三部分计算附件的数字签名, 本文采用的是 MD5算法.

3.2　多层检测器的产生

3.2.1　检测器的组成与表示

当前垃圾邮件的形式越来越多, 手法也越来越隐蔽, 因此单一的垃圾邮件过滤系统很难得到较高的检出率和正确率. 受生物免疫系统工作机制的启发, 本算法中的免疫检测器分为四层:

(1) Ps是第一层检测器, 对应于物理 (Physical) 屏障, 采用黑名单 /白名单技术, 由两部分组成: < Flag, Address>. 其中, Flag用来标志检测器是黑名单还是白名单检测器, Address是相应黑名单 /白名单列表中的项.

(2) Pb是第二层检测器, 对应于生理 (Physiological) 屏障, 采用针对附件的数字签名技术, 表示为: < Signature>, 其中 Signature为数字签名列表中的项.

(3) Ii是第三层检测器, 对应于先天免疫 (Innate Immune) 层, 采用关键词匹配技术, 由主题和内容两个子矢量组成, 表示为: < Subject, Content>. 其中每个子矢量都是一个无序的、可变长的数组, 数组中的元素来自相应的关键词数据库.

(4) B是第四层检测器, 对应于获得性免疫层中的 B 细胞, 采用基于内容的垃圾邮件过滤技术, 由发送者、主题和内容三个子矢量组成, 表示为: < Sender, Subject, Content>. 与 Ii检测器类似, 每个子矢量都是一个无序的、可变长的数组, 但数组中的元素来自相应的表示词数据库.

在生物免疫系统中, 记忆 B 细胞使得免疫系统在再次遭受类似病原体入侵时能快速反应并反击抗原, 借鉴此二次免疫应答的思想, 本算法中也包含了两种免疫记忆检测器, 分别是记忆 Ii检测器和记忆 B 检测器.

各检测器层之间的组合可采用"OR"或"AND"的关系. 当采用"OR"关系时, 算法的计算效率和垃圾邮件检出率较高, 但也容易将合法邮件误认为垃圾邮件; 当采用"AND"关系时, 计算量较大, 准确率较高, 但容易漏报垃圾邮件. 合理的安排各检测器层之间的关系可以使算法得到较高的过滤检出率和正确率.

3.2.2　检测器的产生与衰亡

Ps检测器由黑名单 /白名单列表产生, Pb检测器由数字签名列表产生. 每个检测器代表列表中的一项, 当列表中增加新项时, 就增加相应的检测器; 当检测结果与用户反馈相矛盾时, 删除相应检测器以及列表中相应的项. 黑 /白名单列表初始值可以从一些有信誉的组织获得, 也可以

为空, 经一段时间的训练后完善; 数字签名列表初始时为空, 当用户确认一封邮件为垃圾邮件时, 计算附件的数字签名并加入列表.

初始时, Ii检测器和 B检测器分别由关键词和表示词基因库随机产生. 基因库中每个基因都有一定的初始浓度, 当检测结果与用户反馈相反时, 相应检测器被删除, 同时相应的基因浓度衰减, 当衰减为 0时删除该基因, 同时删除所有含该基因的检测器; 当检测结果正确时, 相应基因浓度增加. 每个检测器被赋予一定的初始重要度, 重要度随时间衰减, 当检测器正确检测到垃圾邮件时, 重要度增加; 当重要度衰减到 0时, 删除相应的检测器.

记忆检测器为相应检测器转化而来. 当 Ii检测器或 B检测器被激活后, 经历克隆扩增以及体细胞高频变异, 达到亲和度成熟. 与抗原有较高亲和度的检测器转化为记忆检测器. 每个记忆检测器也被赋予一定的初始重要度, 该重要度也随时间衰减. 当新增一个记忆检测器时, 为了防止记忆检测器覆盖区域的过度重叠, 记忆检测器集中所有能检测到该检测器的记忆检测器都进行一定程度的重要度衰减. 当记忆检测器的检测结果与用户反馈矛盾时, 该记忆检测器被删除, 同时所有能检测到该检测器的记忆检测器也进行一定程度的重要度衰减.

3.3 检测器的应用

3.3.1 亲和度的计算

本文采用如下公式来计算两个文本矢量之间的亲和度, 设 A、B 是两个无序的文本矢量则定义 A 和 B 的亲和度为:

$$\mathrm{Affinity}(A, B) = \frac{A \cap B}{\min(A, B)}$$

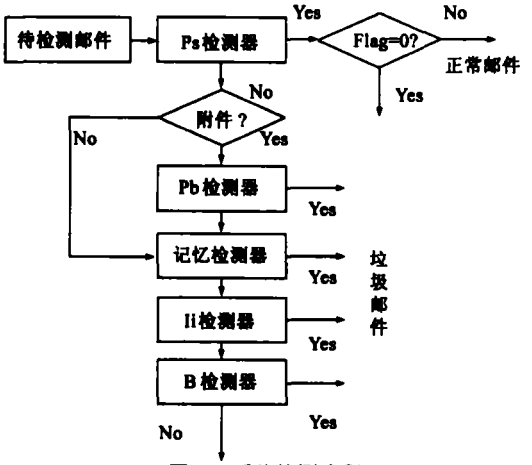


图 2 系统检测流程

其中, $A \cap B$ 表示矢量 A 与 B 中相同的表示词数目, $\min(A, B)$ 表示矢量 A 和 B 中较短的矢量的长度. 如上文所述, Ii检测器和 B 检测器分别由两个和三个子矢量组成, 在计算检测器和邮件的亲和度时, 可分别计算对应子矢量之间的亲和度, 然后取平均值或加权平均值. 由上述公式和计算方法可得亲和度的取值范围为 $[0, 1]$.

3.3.2 应用

假设各层检测器之间采用“OR”的关系, 那么对于待检测邮件, 首先由 Ps检测器进行检测, 如果检测器被激活, 则根据 Flag标志决定其后的处理方法; 如没有被激活, 则进一步判断待检测邮件是否存在附件, 若存在则交由 Pb检测器检测, 反之则交由记忆检测器检测. 若上述检测器都未被激活, 则待检测邮件交由 Ii检测器和 B检测器检测. 系统检测流程如图 2所示.

3.4 Ii检测器和 B检测器的进化

生物免疫系统中大约有 10^8 种不同类型的抗体, 相对于大约 10^{16} 种不同的外部抗原, 抗体的种类远远不足, 生物体通过抗体的动态更新机制和自适应学习机制来解决该问题[10]. 抗体的动态性主要是由 B细胞通过体细胞高频变异和受体编辑机制来实现的.

在本文的多层垃圾邮件过滤算法中, 检测器的进化主要是指 Ii检测器和 B检测器的进化, 其进化通过体细胞高频变异和受体编辑机制来实现[11], 如图 3所示.

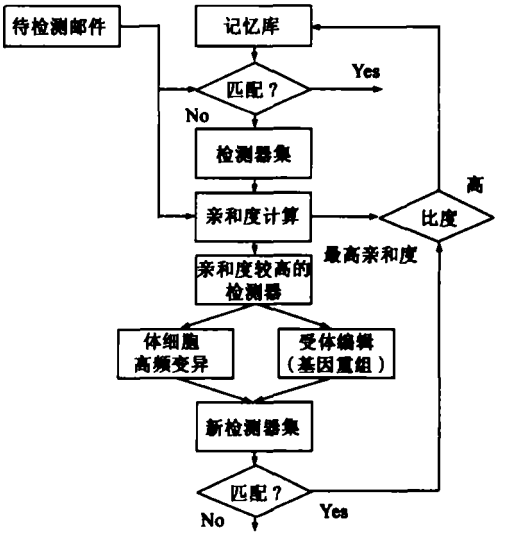


图 3 检测器进化流程

在检测流程中当记忆检测器未被激活时, 邮件交由 Ii和 B检测器进行检测. 具体进化步骤为:

(1)根据本文提出的亲和度计算方法, 依次计算邮件与检测器之间的亲和度.

(2)选择亲和度较高的一组检测器集, 采用体细胞高频变异的方法来产生一组新的检测器; 同时采用基因重组的技术, 用轮赌法重组一批新的检测器, 共同组成新检测器集.

(3)计算邮件与新检测器之间的亲和度. 如果新检测器发生匹配, 则选取其中的最高亲和度与原检测器集中的相比较, 具有较高亲和度的检测器加入记忆检测器集.

此流程体现了免疫细胞的亲和度成熟机制, 以及先天免疫和获得性免疫相结合完成进化学习的机制.

4 实验及讨论

4.1 实验

实验采用 PU1语料库来验证算法的有效性. PU1为英文语料库[2], 来源于提供者一段时间内的真实邮件. 基于预处理方式的不同, PU1提供了四种方式的语料, 分别为: bare, lemm, stop, lemm_ stop 每种方式的语料均由 1099封邮件组成, 其中垃圾邮件 481封, 合法邮件 618封. 为减少无用信息的干扰, 实验中采用 lemm_ stop形式的语料库, 语料由 10部分组成, 每部分约 110封邮件. 为保护隐私, PU1语料库进行了 "加密", 实际的单词都由相应的数字表示.

由于语料库仅含主题部分和内容部分, 因此本文仅验证第四层 B检测器的有效性, 每个 B检测器由主题和内容两个子矢量组成. 试验开始时, 假设一个表示词在一封垃圾邮件中出现, 则其浓度增加, 反之如在合法邮件中出现, 则其浓度降低, 通过对所有训练集的学习后, 删除浓度低于一定阈值的表示词就可得到表示词基因库; 接着按本文第三部分所述的检测器产生、删除与进化算法就能得到一组有效的 B检测器来过滤垃圾邮件. 实验采用 K次交叉验证的方式, 即将语料库中 9个部分作为训练集, 剩余一部分作为测试集, 如此交叉做 10次取平均值. 算法性能主要由 Recall, Precision和 Accuracy这三个参数来衡量. 其中 Recall为召回率, 即垃圾邮件检出率; Precision为精确率, 即垃圾邮件检对率; Accuracy为正确率, 即所有邮件的检对率.

表 1为激活阈值为 0. 6时, 算法在不同 B检测器数目下的性能. 从表 1可以看出, 随着检测器数目的增加, Precision下降, Recall上升.

表 1　B检测器数目对算法性能的影响 (激活阈值为 0. 6)

Number of B	Recall(%)	Precision(%)	Accuracy(%)
100	55. 30	99. 25	80. 25
200	57. 38	98. 92	81. 26
300	63. 83	99. 03	83. 89
400	81. 29	91. 14	88. 35
500	79. 42	93. 17	88. 44

表 2为 B检测器数目为 400时, 算法在不同激活阈值下的性能. 从表 2可以看出, 随着激活阈值的增加, Precision上升, Recall下降.

表 2　检测器激活阈值对算法性能的影响 (B检测器数目为 400)

Threshold	Recall(%)	Precision(%)	Accuracy(%)
0. 55	70. 89	87. 44	81. 07
0. 6	81. 29	91. 14	88. 35
0. 65	58. 42	99. 29	81. 61
0. 7	37. 63	98. 37	72. 52

Naïve Bayesian算法是目前垃圾邮件过滤中应用较广泛且性能较好的一种算法, 其基本原理是计算文本属于某个类别的概率, 将文本分到概率最大的类别中. 文献[2]中给出了 Naïve Bayesian算法用于 PU1语料库中 lemm_ stop 形式测试数据的实验结果, 如表 3所示 (文中 Accuracy为 Weighted Accuracy, 为可比性将其转化为 Accuracy).

表 3　Naïve Bayesian算法性能

l/Number of attr	Recall(%)	Precision(%)	Accuracy(%)
1/100	79. 60	97. 96	90. 34
9/100	75. 86	97. 91	88. 72
999/600	49. 45	98. 31	55. 86

由上述实验结果可以看出, 本文提出的算法中第四层 B检测器的综合性能与 Naïve Bayesian算法的性能相当. 但是, 对于用户而言, 更多情况是希望在维持高 Precision的前提下追求高 Recall. 由表 3, Naïve Bayesian算法在 Precision取最大值 98. 31% 时, Recall仅为 49. 45%; 然而, 对于本文算法, 由表 1和表 2, 当 Precision 为 99. 29% 和 99. 03%, Recall分别为 58. 42% 和 63. 83%. 因此, 本文算法与 Naïve Bayesian相比, 具有一定的优势. 此外, Bayesian算法在当用户兴趣发生转移即垃圾邮件特征发生突变时不能立即做出反应, 需要一段时间来学习适应, 不能做到实时性[16]; 而本算法可以通过 B检测器的删除和新增来对垃圾邮件特征的突变做出及时的反应, 因此本算法 MSFA-AI较之 Naïve Bayesian分类算法有较好的自适应性能.

4.2 讨论

本文所提出的基于人工免疫的多层垃圾邮件过滤算法由四层检测器组成, 下面将分别讨论各层检测器训练及检测时的时间复杂度. 设 N 表示训练集中所包含的样本数, F 表示特征数目, T 表示表示词数目, K 表示关键词数目, 显然, F、T 和 K 是一个量级的, 一般有 $K < F < T$. 对于 Ps层检测器, 其采用黑名单/白名单技术, 故训练的时间复杂度为 $O(N)$, 设黑/白名单列表长度为 L, 则检测的时间复杂度为 $O(L)$. 发件人地址也是邮件的一个特征, 故 $L \ll F$. 对于 Pb层检测器, 采用针对附件的数字签名技术, 与 Ps层检测器类似, 设附件签名列表长度为 L_a, 则其训练时间复杂度为 $O(N)$, 检测时间复杂度为 $O(L_a)$且 $L_a \ll F$. Ii层检测器与 B层检测器类似, 其训练时间复杂度分别为 $O(KN)$ 和 $O(TN)$, 这两层检测器在检测过程中会进行进化, 设进行进化的检测器比例为 α, 每个检测器高频变异生成 M 个检测器, 则其检测时间复杂度分别为 $O(\alpha MK)$ 和 $O(\alpha MT)$. 故算法训练时的时间复杂度为 $O_{MSFA-AI} = O(N) + O(N) + O(KN) + O(TN) = O(TN)$. 由于算法在检测过程中可以合理的安排各层检测器之间的 OR 和 AND 关系, 故算法检测时的时间复杂度 $O(L) \leqslant O_{MSFA-AI} \leqslant O(L) + O(L_a) + O(\alpha MK) + O(\alpha MT) = O(\alpha MT)$. Naïve Bayesian算法训练时的时间复杂度为 $O(FN)$, 检测时的时间复杂度为 $O(F)$[12]. 可见本文算法与 Naïve Bayesian算法训练时的时间复杂度相当. 本文算法可以通过调整各层检测器之间的关系来调整检测时的时间复杂度, 最坏情况下稍高于 Naïve Bayesian算法.

5 结束语

本文借鉴了生物免疫系统的多层防御机制,提出了一种基于人工免疫的多层垃圾邮件过滤算法.文中详细描述了多层检测器的定义、产生与删除方法,亲合度计算方法以及各层检测器的应用方法,多层检测器中 Ti检测器和 B检测器的进化机制,并基于 PU1语料库进行了测试实验.实验表明本文算法能有效地过滤垃圾邮件.通过合理的设置各层检测器之间的关系还能得到更好的垃圾邮件过滤效果.

邮件内容的预处理属于自然语言理解的范畴,目前还没有一个很完善的解决方案,在将内容划分为表示词时加入权重,有助于增强算法的性能;同时,为了使算法达到更好的性能,还需要做更多的实验来合理的设置各算法参数.在加上前三层检测器并合理的安排各检测器层之间的关系后,算法会有更高的召回率、精确率和正确率,并可以在一定程度上降低算法的时间复杂度.这是本文下一步的研究内容.

参考文献:

[1] L Pelletier, J Almhana, V Choulakian. Adaptive filtering of SPAM[A]. Communication Networks and Services Research Proceedings[C]. USA: IEEE Computer Society Press, 2004. 218-224.

[2] Ion Androutsopoulos, John Koutsias, Konstantinos V Chandrinos, Constantine D Spyropoulos. An experimental comparison of Naïve Bayesian and keyword-based anti-spam filtering with personal E-mail message[A]. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York: ACM Press, 2000. 160-167.

[3] Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz. A bayesian approach to filtering junk E-mail[A]. Learning for Text Categorization[C]. Madison Wisconsin: AAAI Press, 1998. 55-62.

[4] Jongsub Moon, Taeshik Shon, Jungtaek Seo, Jongho Kim, Jungwoo Seo. An approach for spam E-mail detection with support vector machine and n-Gram indexing[A]. Lecture Notes in Computer Science[C]. Heidelberg Germany: Springer-Verlag GmbH, 2004. 351-362.

[5] Harris Drucker, Donghui Wu, Vladimir N Vapnik. Support vector machines for spam categorization[J]. Neural Networks, IEEE Trans, 1999, 10(5): 1048-1054.

[6] Andrew Secker, Alex A Freitas, Jon Timmis. AISEC: an artificial immune system for E-mail classification[A]. The 2003 Congress on Evolutionary Computation[C]. California USA: IEEE Computer Society Press, 2003. 131-138.

[7] Terri Oda, Tony White. Increasing the accuracy of a spam-detecting artificial immune system[A]. The 2003 Congress on Evolutionary Computation[C]. California USA: IEEE Computer Society Press, 2003. 390-396.

[8] Terri Oda, Tony White. Developing an immunity to spam[A]. Lecture Notes in Computer Science[C]. Heidelberg Germany: Springer-Verlag GmbH, 2003. 231-242.

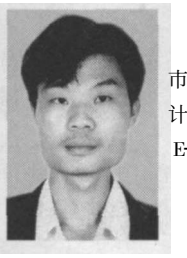
[9] Yukun Cao, Xiaofeng Liao, Yunfeng Li. An E-mail filtering approach using neural network[A]. Lecture Notes in Computer Science[C]. Heidelberg Germany: Springer-Verlag GmbH, 2004. 688-694.

[10] 莫宏伟.人工免疫系统原理与应用[M].黑龙江哈尔滨:哈尔滨工业大学出版社, 2003.

[11] 罗文坚.面向入侵检测的人工免疫模型和算法研究[D].安徽合肥:中国科学技术大学, 2003.

[12] I Androutsopoulos, G Paliouras, E Michelakis. Learning to Filter Unsolicited Commercial E-Mail[R]. Athens, Greece: National Centre for Scientific Research Demokritos, 2004.

作者简介:



张泽明 男,1980年11月出生于湖北省天门市.博士研究生.主要研究方向为人工免疫、自然计算以及硬件进化.
E-mail: zmzhang@mail.ustc.edu.cn

罗文坚 男,1974年11月出生于广东省梅州市.2003年获中国科学技术大学计算机应用专业博士学位.现为中国科学技术大学副教授,从事自然计算、人工免疫系统、硬件进化及网络安全等方面的研究.E-mail: wjluo@ustc.edu.cn

王煦法 男,1948年11月出生于江苏省盐城市.现为中国科学技术大学计算机系教授、博士生导师.主要从事智能信息处理、人工免疫系统及网络安全等方面的研究工作.
E-mail: xfwang@ustc.edu.cn