文章编号:1001 - 9081(2005)11 - 2562 - 04

基于人工免疫的反垃圾邮件系统模型

周念念,冉蜀阳,曾剑宇,钟 响
(四川大学 计算机学院,四川 成都 610064)
(znn_326@tom.com)

摘 要:人工免疫是对生物免疫系统的工作机理及所具有各种优良特性的模拟。在邮件系统中引入免疫的思想,构建了一种基于人工免疫的反垃圾邮件模型。该模型将非法邮件、垃圾邮件看作侵入系统的病源,模拟机体消灭病源的机理,使系统在正常使用的过程中自动学习各种垃圾邮件的特性,从而自动识别、删除、过滤掉这些有害邮件。对模型的设计思想、关键技术等多方面进行了深入的讨论,并在实验中验证了其可行性。

关键词:人工免疫;邮件系统;否定选择算法;克隆选择算法
中图分类号: TP393.08 **文献标识码**:A

Anti-trash E-mail system based on artificial immunity model

ZHOU Nian-nian, RAN Shu-yang, ZENG Jian-yu, ZHONG Xiang
(*College of Computer Science, Sichuan University, Chengdu Sichuan* 610064, *China*)

Abstract: An artificial immune system studied the principle and some fine features of a biological immune system. The idea of immunology was adopted to build an anti-trash E-mail model. All kinds of illegal E-mail and trash E-mail would be considered as viruses invading the system in this model, and to be recognized and eliminated automatically according to the mechanism of eliminating viruses by organism. The design idea and the key techniques of the model were discussed intensively. Its practicability was proved by experiments.

Key words: artificial immunity; E-mail system; nagative selection algorithm; clonal selection algorithm

0 引言

电子邮件是人们在网上进行信息交流的一种必不可少的工具。但是一些垃圾、非法邮件甚至携带病毒的邮件日益猖獗,困扰着人们正常使用 E-mail 进行通信交流。如何有效地自动过滤、删除这些有害邮件成为一个热点问题。现有的反垃圾邮件系统大多是基于概率统计的原理对邮件进行过滤,这种方法需要较多的人工干预,在对垃圾邮件的多样性、变异性上反应较慢,缺乏自适应性和自动学习的功能。本文提出了一种全新的方法——基于免疫模型来构建反垃圾邮件系统,并在实际环境中验证了它的优良特性。

当生物系统受到外来病毒或异物的入侵时,能够自动识别这种入侵,并自动激发自身免疫细胞捕捉,消灭入侵者。在这一过程中生物免疫系统所表现出的耐受性、免疫记忆、自组织、自学习、自适应等许多特性引起了研究人员的普遍关注,逐渐形成了一门基于生物免疫学、人工免疫、以及计算机科学等的交叉学科——计算机免疫学。

1 设计思想

邮件系统和免疫系统具有很多相似性。邮件服务器接收的所有邮件可以看作是全体细胞的集合,其中正常邮件为自体细胞;而所有的垃圾、非法邮件被看作是有害的病毒。每封邮件都可被分解成若干特征分子(即将所有邮件提呈为若干特征字符串),然后用系统中存在的记忆细胞群及成熟抗体群去检测这些分子。若同一邮件产生的特征分子被抗体识别的数目超过某一阈值,那么这一邮件被认为是有害的,系统将其删除,并将它的所有特征分子的特征字符串加入抗体基因库,作为产生下一代抗体的材料。

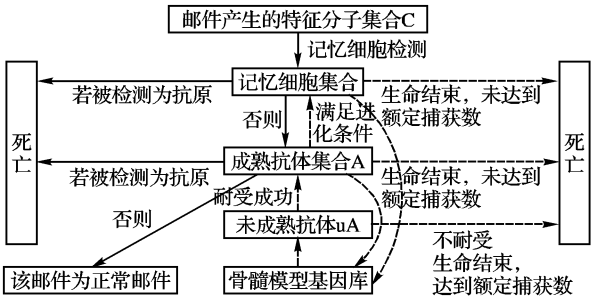

图 1 邮件系统免疫模型

(实线代表特征分子的移动方向,虚线代表抗体的移动方向)

整个系统以 Perelson 和 Oster 提出的形态空间理论[1] 为其理论基础。

如图 2 所示,按照这一理论,形态空间内有一个体积为 V 的区域,V_ε 是半径为 ε 的抗体识别球体积。n 个抗体随机散布在形态空间中,如果每一个抗体有大致相同的体积 V_ε,则所有抗体所覆盖的体积总和是 $n * V_\varepsilon$。如果这个值和形态空间的总体积 V 相差不大,同时一般来说抗体有重叠的识别区域,因此能完全覆盖形态空间。事实上,每一个抗原决定基一般由 $n * V_\varepsilon / V$ 个不同的抗体识别,而不能识别的几率 P 是:

收稿日期:2005 - 05 - 25

作者简介:周念念(1981 -),男,四川广元人,主要研究方向:计算机网络安全、人工智能; 冉蜀阳(1962 -),男,四川崇庆人,教授,博士,主要研究方向:计算机测控、人工智能; 曾剑宇(1979 -),男,四川成都人,主要研究方向:计算机应用、并行分布式处理; 钟响(1981 -),男,湖南长沙人,主要研究方向:智能信息系统、数据库、计算机网络.

$$P = (1 - V_\varepsilon/V)^n \approx e^{-nV_\varepsilon/V} \quad (1)^{[2]}$$

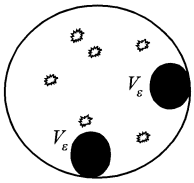

图 2　形态空间图

从(1)式可以看出,抗体的数量越多识别的几率越大,如果 $n = 10^6$, 那么 P 的值为 4×10^5,几乎能识别所有的抗原决定基。我们的系统正是要求用大量抗体群尽可能完备地覆盖由垃圾邮件的特征分子形成的抗原空间。

2　模型实现

2.1　自体/非自体[3]

为了清除有害物质,免疫系统所面临的首要问题是如何定义自体和非自体。在计算机内所有的信息将以二进制表示,所以问题域 $X \in \{0, 1\}^l$,定义所有正常邮件的集合 S 为自体,带有非法信息或者用户不愿收到的垃圾邮件组成非自体集合 N,满足 $S \subseteq X \wedge N \subset X \wedge S \cup N = X \wedge S \cap N = \varnothing$,此外我们的系统对于非法邮件的检测并不是以邮件本身为单位的,而是发生在更小的粒度上,因此我们另外定义两个集合 S', N' 分别代表 S 和 N 经过提呈后的特征分子形成的集合,它们同样满足 $S' \subseteq X \wedge N' \subset X \wedge S' \cup N' = X \wedge S' \cap N' = \varnothing$。对于一封邮件的检测过程就是对一个模式 $I \in X$ 的分类,看它是自体还是非自体:对一个 $I \in X$,一个检测器集合(记忆细胞,抗体等) D:$D = \{\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_i\}$, $\alpha \in \{\}^k$, $k \leqslant l$, $i \in N$, N 为自然数集。一个匹配函数 f: $f(I, \alpha) \rightarrow \{p \mid p \in R \wedge p \geqslant 0 \wedge p \leqslant 1\}$,其中,$R$ 为实数集,ε 为匹配阈值,由式(2)完成分类:

$$Classify(f, \varepsilon, I, D) = \begin{cases} \text{恶性}, f(I, \alpha) \geqslant 1 - \varepsilon \\ \text{良性}, \text{otherwise} \end{cases} \quad (2)^{[2]}$$

该检测过程可能会产生两种错误错误肯定和错误否定,分别用 ζ^+ 和 ζ^- 表示,其意义为:

$(I \in S \wedge match(f, \varepsilon, I, D) = \text{恶性} \rightarrow) \zeta^+$

$(I \in N \wedge match(f, \varepsilon, I, D) = \text{良性} \rightarrow) \zeta^-$ [2]

这里肯定错误是指将自体认为是非自体,否定错误是指将非自体认为是自体。

我们用以下结构体来定义一封邮件:

```
typedef struct
{
  // -- 亲和力系数 -
  define APPETENCY 0.7
  // -- 由 Mail 提呈后得到的特征分子的数目 --
  int numOfCells;
  // -- 亲和力阈值,即抗原积累阈值,它等于分子总数乘上一个系数 --
  int critical = numOfCells * APPETENCY;
  // -- 一个指针,指向所有由此 Mail 提取出的特征分子的链表
  Cell * head;
  // -- 自体或非自体的标识,0 代表非自体,非 0 代表自体
  int status;
  // -- 被检测为抗原的特征分子数,如果大于亲和力阈值则被认为是非法邮件
  int numOfDetected;
} Mail;
```
(3)

2.2　抗体/抗原

对于系统中的每一封邮件我们将按如下算法将它提呈为若干特征分子(特征提呈):

1) 将其发件人地址、收件人地址和信件主题三个字符串依次排列作为特征分子 β_0 ;

2) 将邮件正文每一段的首句作为特征分子 β_i ;

3) 将每一段的末句作为特征分子 β_{i+1} ;

4) 若段落数不大于 1,那么每隔三句抽取一句作为特征分子 β_i 。

特征分子除了抽取邮件的子串作为其主体以外,还包含一些其他的属性,用结构体的形式描述如下:

```
typedef struct
{
  // -- 指向产生此特征分子的 Mail --
  Mail * matrix;
  // -- 指向同一 Mail 产生的下一个特征分子 --
  Cell * next;
  // -- 特征字符串 --
  char * string;
  // -- 正常特征分子或抗原的标识,0 代表抗原,非 0 代表正常分子
  int status;
} Cell;
```
(4)

于是整个邮件系统产生了一个特征分子的集合,它们便是系统要去检测和识别的最小单元。在这个集合当中,那些带有非法信息的特征分子我们把它定义为抗原,(也即是上文提到的集合 N' 的元素,剩下的正常分子为 S' 的元素),这些抗原就是系统要去识别的目标。

我们用另外一组结构去识别抗原,这就是我们要定义的抗体集合 A(为简单起见本文将抗体,B 细胞的概念合而为一),其结构体定义如下:

```
typedef struct
{
  // -- 捕获率系数,它规定了匹配时需要多少比例的位数相同就被认为识别发生
  define MATCHRATE 0.7
  int life;                          // -- 抗体的生命周期 --
  char [] identifyString;            // -- 特征字符串 --
  // -- r 位匹配规则的位数,它是一个根据特征字符串的长度而变化的动态值 --
  // -- length(string) 表示一个字符串的长度 --
  int r = length(identifyString) * MATCHRATE;
  // -- 捕获抗体的数目 --
  int numOfCapure;
  // -- 捕获抗体所包含于的邮件最终被确认为非法邮件的数目 --
  int numOfSeccess;
} Antibody;
```
(5)

抗原与抗体所包含的特征字符串具有可匹配性,并且每一抗原必须能够识别一定范围类的抗原。需要注意的是,识别过程发生在抗体 - 抗原即特征分子这一粒度上,那么如何将对抗原的识别转换为对特定邮件的识别,我们将在克隆选择一节做详细的讨论。

对于抗体与抗原的匹配,即识别,我们采用 Forrest 提出的 r 连续位匹配规则[4],针对两者的特征字符串进行匹配,从而计算抗体和抗原间的亲和力,决定是否发生识别。即对于

一个特征分子 Cell，如果至少存在连续 r 位相同，那么它们就是匹配的，即识别成功。式(6)定义了字符串 b 对 v 的识别。

$$match(b, v, r) = \begin{cases} 1, \text{if } \exists i \exists j, x_i \in b \wedge y_i \in v \wedge x_k = y_k \wedge \\ \quad k = i, \cdots, j \wedge j - i \geqslant r \wedge i, j, r, k \in N \\ 0, \text{otherwise} \end{cases} \quad (6)$$

其中 1 表示匹配 0 表示不匹配，N 为自然数集。这里我们系统选取的 r 值并非一个固定值，而是根据特征字符串的长度乘上匹配系数而动态变化。

2.3 记忆细胞

在生物系统中，初次进入机体的抗原必须足够的多，而且有一个积累的过程，才能刺激免疫细胞达到足够的亲和力发生特异性克隆扩增产生免疫应答。此后会有一部分识别此类抗原的抗体进化为记忆细胞，若机体再次受到相同或相似抗原的刺激，记忆细胞会很快检测到这些抗原并引起迅速激烈的反应，如图 3。

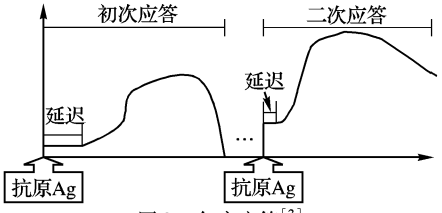

图 3 免疫应答[3]

正是为了模拟在初次应答中积累亲和力的过程以及在二次应答中记忆细胞的作用，我们才将一封邮件提呈为若干特征分子。在初次应答时，这些特征分子的匹配数量需达到一定的阈值才能确认邮件为非自体，这样有效地降低了错误肯定事件的发生。然后，从参与此次反应的抗体中找出满足进化函数(见式(7))的抗体进化为记忆细胞，若下一次有某一封邮件的特征分子的匹配发生在记忆细胞上，则直接确认该邮件为非法邮件。在我们的系统中，记忆细胞的结构和抗体定义相同，只是具有更长的生命周期值，并且把它们归属于另一个集合 M。

$$evolute(Antibody) = \begin{cases} 1, \text{ if } (Antigody.numOfCapture/ \\ \quad Antibody.numOfSuccess) \geqslant EVOLUTERATE) \\ 0, \text{ otherwise} \end{cases} \quad (7)$$

引入记忆细胞的概念可使系统的运行效率得到显著提高。

2.4 自体耐受与否定选择[4]

抗体除了应该尽可能多地识别抗原外，还必须保证不能错误地将正常细胞(即正常邮件产生的特征分子)误认作抗原(错误肯定)。为了达到这种要求由骨髓模型产生的抗体必须经过自体耐受。我们采用基于 r 连续位匹配的否定选择算法——线性检测器生成算法[5]来产生成熟抗体集合。算法大致分为两步：

1) 利用式(8)，进行一个有限的递归运算以得到一定数量不与自体集合匹配的字符串：

$$C_i[s] = \begin{cases} 0, & \text{if } t_{i,s} \text{ 在 } S \text{ 中存在匹配} \\ C_{i+1}[\hat{s}.0] + C_{i+1}[\hat{s}.1], & \text{otherwise} \end{cases} \quad (8)$$

其中对一个长度为 r 的串 s，$1 \leqslant i \leqslant (l - r + 1)$，定义 $C_i[s]$ 为 $t_{i,s}$ 的不与自体集合中任意串匹配的右填充的数量。

2) 利用式(9)，用枚举法从候选检测器中随机选出检测器(即成熟抗体)：

$$P_1 = \sum_{s<s_1} C_1 < k < Q_1 = \sum_{s \leqslant s_1} C_1[s] \quad (9)$$

整个算法的空间复杂度为 $O((l-r)^2 \cdot 2^r)$，其时间复杂度为 $O((l-r) \cdot N_s) + O((l-r) \cdot 2^r) + O(l \cdot N_R)$。

2.5 克隆选择

克隆选择原理被用来描述免疫系统是怎样与抗原作战的。当外部细菌或病毒侵入机体后，免疫细胞开始大量克隆并消灭入侵者，那些能够识别抗原的细胞根据识别的程度繁殖后代，与抗原亲和力越高，该细胞就能产生更多的后代进入骨髓模型，并且经历较小的变异。同样，克隆选择也是本文的核心内容，它使我们的系统具有了多种良好的特性，如自适应，自学习，对未知类型的非法邮件反应迅速等。根据邮件系统的特点以及吸收动态克隆选择算法[6]的优点，本文提出了以下克隆选择算法，应用在系统模型当中。

根据骨髓模型和自体耐受模型，我们不妨设当前时刻，系统中有一个记忆细胞集合 M，一个成熟抗体集合 A，一个未成熟抗体集合以及一个由 SMTP 协议收到的邮件队列。

```
for (队列中的每一封 Mail)
begin
  获得一个 Cell List;
  for (每一个 Cell)
  begin
    if (∃Antobody∧Antobody∈M(match()=1)) then
      /*首先用记忆细胞集合检测这个 Cell,如果有一个匹配成功*/
      Cell.status=0 and Mail.status=0
      /*该 Cell 为抗原,并且确认该 Mail 非法*/
      break
    else
      if (∃Antobody∧Antobody∈A(match()=1)) then
        Cell.status=0;
        Mail.numOfDected=MailbumOfDetected+1;
      end if
    end if
  end
  if (Mail.status=0) then
    /*如果是被记忆细胞检测到,并确认为非法邮件的*/
    把所有此 Mail 产生的所有 Cell.string 及其变异加入抗体基因库 AND
    识别成功的 Antibody.life=Antibody.life+1
    Continue;
    /* 进入下一封邮件的检测*/
  else
  if (Mail.numOfDected > = Mail.critical) then
    /* 如果抗原的累积达到亲和力阈值*/
    Mail.status=0;
    /*同样可以确认该邮件非法,并且*/
    把所有此 Mail 产生的所有 Cell.string 加入抗体基因库,并且
    for(刚才发生识别的所有 Antibody)
    begin
      Antibody.numOfCapture= Antibody.numOfCapture+1
      Antibody.numOfSeccess= Antibody.numOfSeccess+1
      If( envolute(Antibody)=1) then
        /*如果满足进化条件,进化公式 evolute() 见式(7)*/
        把此 Antibody 加入记忆细胞集合 M,并重设其生命值
      else
```

```
        Antibody. life = Antibody. life + 1
        /* 延长其生命值作为奖励 */
       end if
     end
     else
      /* 如果抗原的累积没有达到亲和力阈值 */
      Mail. status = 1;
      /* 邮件为正常 */
      for(刚才发生识别的所有 Antibody )
      begin
        Antibody. life = Antibody. life - 1
        /* 缩短其生命值作为惩罚 */
      end
      把此邮件产生的所有未被抗体识别的 Cell 加入自体耐受集合 S′
     end if
   end if
end
```

进化函数 evolute( )刻画了当一个抗体识别出一个 Cell 的条件下,包含这个 Cell 的 Mail 是非自体的条件概率有多大,因为记忆细胞具有完全激活性,为了有效避免错误肯定的发生,我们有理由要求这个概率要足够的大,即一旦某封邮件的 Cell 被记忆细胞检测到,那么就能肯定此邮件非法,因此我们一般将 EVOLUTERATE 设为 95% 或更高。

在把抗原的特征字符串加入基因库时,根据其与抗体的匹配情况应当在一定程度上进行变异,这是保持系统多样性的根源,在实际应用中这也是识别变异垃圾邮件的有效方法。

3　实验结果

在实验中,我们发现以下几个参数设置得是否合适将在很大程度上影响系统的性能以及错误肯定和错误否定的发生率,必须通过测试给出一个较为良好的取值:

1) Mail 的亲和力系数 APPETENCY,它描述了一封邮件有多少特征分子被检测为抗原就可确认该邮件非法。

2) Antibody 结构的捕获率系数 MATCHRATE,它描述了在特征分子识别这一粒度上抗体和抗原有多少位相同,就认为识别发生。

3) 进化函数 evolute( )中的进化系数 EVOLUTERATE,它刻画了当一个抗体识别出一个抗原,那么含有这个抗原的邮件有多大的几率为非自体的条件下,发生识别的抗体可以进化为记忆细胞。

4) 成熟抗体集合所拥有的抗体数量。

5) 记忆细胞集合拥有的细胞数量。

实验环境为:HP Unix Visualize B2000 WorkStation 作为邮件服务器,邮件用户为 1 000 人,将邮件队列中的邮件数量设为 3 000 封,其中正常邮件 2 700 封,垃圾邮件 300 封。

图 4 是在 APPETENCY = 0.5, EVOLUTERATE = 95% 的情况下,调整 MATCHRATE 的测试,我们发现,当 MATCHRATE 降低时,错误肯定率几乎趋于 0,但是错误否定率有一定程度的上升,反之亦然。

而在 APPETENCY 和 MATCHRATE 固定的情况下,EVOLUTERATE 对系统性能和错误肯定率也有非常显著的影响。通过试验可以看到 EVOLUTERATE 值降低时系统的性能会成几乎指数级上升。但是相应的,在 EVOLUTERATE 降到 85% 以下时,系统的错误肯定率将会上升到让人无法接受的程度。

另外,抗体集合 A 与记忆细胞 M 所拥有的细胞数量也会影响到系统的错误率和性能。大体上来说两个集合的数量越多,错误否定会越少,识别率越高,但是记忆细胞的增多将使错误肯定率稍有上升。另外随着这两个集合的增大,对检测周期会有所影响,具体情况如表 1 所示。

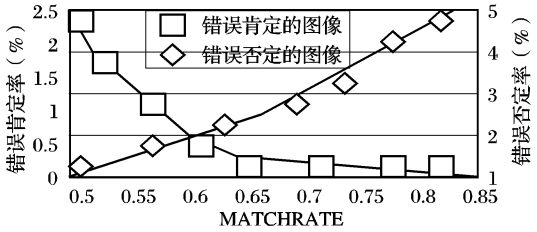

图 4　调整 MATCHRATE 的值获得的坐标图

表 1　免疫细胞数量与系统性能实验数据表

待测邮件数量	成熟抗体数量	记忆细胞数量	错误肯定率(%)	错误否定率(%)	检测周期
3 000	300	50	0	9.2	21:18
3 000	400	50	0	9	25:02
3 000	500	50	0	8.6	32:27
3 000	600	50	0	7.5	42:05
3 000	700	50	0	6	51:39
3 000	500	75	0	8.4	31:51
3 000	500	100	0.1	7.1	30:00
3 000	500	125	0.2	6	28:40
3 000	500	150	0.4	4.0	25:40
3 000	500	175	0.5	2.8	20:48

对于表 1 中的每一项,我们都是在相同的设定条件下随机变化邮件的具体内容,进行多次测试,然后取其平均值得到的数据,尽量避免了偶然性对系统带来的干扰。

最后我们将用户行为作为协同刺激,引入到系统之中,除了内存占用量稍有上升外,在不需任何额外工作的情况下可使错误否定平均下降 3 ~ 5 个百分点,周期时间缩短 5 ~ 7 分钟,并且对一些邮件在自体与非自体之间的动态变化作出迅速的反应。整个系统在无人干预,无人职守的情况下能稳定工作并表现出非常良好的性能。

参考文献:

[1] PERELSON AS, WEISBUCH G. Immunology for physicists[J]. Review of Modern Physics, 1997, 69(4).

[2] 李涛. 计算机免疫学[M]. 北京:电子工业出版社出版, 2004.

[3] HARMER PK, WILLAMS PD. An Artificial Immune System Architecture for Computer Security Applications[J]. IEEE Transacions on Evolutionary Computation, 2002, 6(3).

[4] FORREST S, PERELSON AS, ALLEN L, *et al*. Self-Nonself Discrimination in a Computer[A]. Proceedings of IEEE Symposium on Research insecurity and Privacy[C]. Oakland, 1994.

[5] HAESELEER PD, FORREST S. An Immunological Approach to Change Detection: Algorithm, Analysis and Implication[A]. Proceedings of IEEE Symposium on Research in Security and Privacy [C]. Oakland, CA, 1996.

[6] KIM J, BENTLEY P. An Investigation of Clonal Selection with a Negative Selection Operator[A]. Proceedings of the Congress on Evolutionary Computation(CEC)[C]. Seoul, Korea, 2001.

[7] HUNT J, TIMMIS J, COOKE D, *et al*. Jisys: The development of an artificial immune system for real world applications[M]. Springer-Verlag, 1999.