Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7):2533-2540

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A spam filtering model based on immune mechanism

Ya-ping Jiang, Yue-xia Tian and Xiao Mei

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, HeNan Zhengzhou, China

---

ABSTRACT

With the development of network, some mail business growing has become a pressing problem in the internet. The problem for the traditional method of spam filtering can not effectively identify the unknown and variation characteristics, artificial immune system exists diversity, immune memory, adaptive and self learning ability, adopt the idea of to mail filtering, and design an improved spam filtering model based on immune mechanism. The model describes the concepts of self, nonself, antibody, vaccine and antigen, and introduced the process of evolution of detector and antigen presentation, make the antibody in all kinds of detectors use vaccines as a medium to communicate with each other, share antibody, increases flexibility of detectors, and effectively extract information and variability of spam. Using CCERT mail dataset of the model was trained and tested, the model and other models by the comparative test results show that the proposed model has better performance, and effectively improve the spam precision, recall characteristics.

Keywords: Artificial immune; spam; Antigen Presentation; vaccine

---

INTRODUCTION

The network is becoming more developed today,email has become one of the main means of communication because of its characteristics of fast and convenient, economy. But spam (Spam)[1] are also increasingly serious. Spam not only hinder the transmission of normal network information, but also some contained virus trojan and created a great threat to network security. Therefore, control the flood of spam has become one of the hot research at present.

In practical application, filtering the mail content is considered to be the most effective method of spam filtering, it mainly includes the naive Bayesian[2] , nearest neighbor algorithm of K[3], support vector machine (SVM)[4] and neural network[5]text classification method. These methods limit the spread of spam to a certain extent, but for the variation of the characteristics or the emergence of new features are often not just as our wishes, lacking of dynamic, adaptive, so that in the practical application of the effect is not ideal. In recent years, the artificial immune system AIS (Artificial Immune System) as the mechanisms of the natural immune system has been successfully used to imitate[6], anti spam system constructed by the technology of artificial immune is self-adaptive, self-learning and memory ability is becoming a hot research. So, based on the work mechanism of biological immune system, proposed a spam filtering model based on immune mechanism, proposed model has better performance, and effectively improve the spam accuracy, recall characteristics.

FILTERING MODEL DESIGN

The model has three stages: the detector produces evolutionary stage (dotted line), antigen presentation phase (solid line), vaccine introduction phase (line). The model system structure as shown in figure 1.

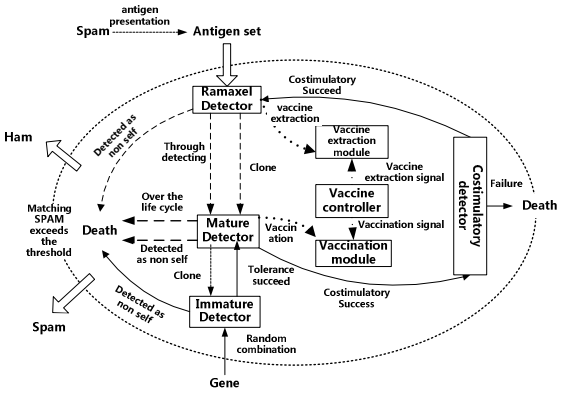

Figure 1 Model Architecture

A．Detector generation and evolution

In our model, the detector is divided into immature, mature and ramaxel detectors, respectively, which is defined as I、M and R，I, M, $R \subset D$。

After the spam samples were collected, then presentation, this will soon be converted into vector data set, and handed over to the detector for detection. If the data is recognized by memory immune detector, directly determined whether it is spam, issued a message to the user; If the memory is matched to the detector self data collection SP, the spam characteristics change, the memory can not be used to detect detector samples and will be deleted.

Mature detector (M) is in the middle stage of evolution of detector, in its life cycle matching enough mail sample data packet, evolving into a memory detector. The data after the memory detector transfer to the mature detector $d_M (d_M \in M)$ processing.

In the life cycle of α, when the num value reaches the matching threshold δ, D_M will be converted into the memory detectors, that is.:

$$\begin{cases} M = M - D_M \\ R = R + D_M \end{cases} \quad D_M.num \geq \gamma \wedge D_M.num.T_{-period} < \delta \quad (1)$$

If d_M can not reach the match threshold δ in the life cycle, then died. In order to improve the diversity of the detector, selected a certain proportion of the mature and ramaxel detectors, then generate a new group of immature detectors by colonel selection, or random manner. d_I may detected normal *self* data set SP ,then execute self-tolerance. Self-tolerance as follows:

$$I_{tolernace}(d_I, SP) = \begin{cases} success, & Afitness(d_I, SP) < \delta \\ \forall E_{sp} \in SP, d_I.T\text{-}period \geq \beta \\ fail, & otherwise \end{cases} \quad (2)$$

In the formula, SP is the self data set; δ is match threshold; β is the threshold of tolerance phase; E_{sp} is an element of the self data set. Only succeed through self-tolerance of D_I can continue to survive and have the qualification convert into mature immune detector; otherwise, D_I death.

The model is main to filter content-based, consider MI mutual information of the text of the word in the model for the impact of the message classification, in the process of new *self* libraries and detector generate, by calculating MI mutual information to select features words composed of *self* library and the gene library to control their size[7]. For each word, calculated the word and the average mutual information between the categories as shown in equation (3):

$$MI_{avg}(W, K_i) = \sum_{i=1}^{|k|} P(K_i) log \left| \frac{P(W \mid K_i)}{P(W)} \right| \quad (3)$$

In the formula $P(K_i)$ is the probability of the first i class text appear in the corpus; $P(W)$ is the probability of feature words W appeared in the corpus; $P(W \mid K_i)$:in the case of documents belonging to the first i category, the probability of occurrence of characteristic words W [8].

Taking into account the mail processing system can be divided into normal mail and e-mail spam, the difference between the use of mutual information as the evaluation function in Equation (4):

$$W(t) = MI_{max1} - MI_{max2} \quad (4)$$

Using of formula (4) calculate the improved mutual Information: MI_{max1} is the maximum value for the mutual information, MI_{max2} is the second largest value. Set a threshold, greater than this threshold value is characterized by the word (selection and training samples off threshold).

B. Antigen Presentation

Mail samples, after the antigen-presenting data to form an antigen collection, filtered through a high-frequency mode, each time a certain amount of antigen selected for testing. To achieve the spam filter, you need to convert the message into a computer tested recognizable form, which is antigen presentation.

Standard text email by envelope and content components. Envelope including the completion of all the necessary messaging and delivery, content there is a two-part letter header and body composition. Start message including sender, recipient, sender time, and subject information. Generally the content is text, ASCII code with composition. The paper said that following these three parts:

1. *From:Jim@163.com*
2. *To:Mark@163.com*
3. *Subject.:Explaination of mail format*
4. *Date:Sat,12 April 2014.10:00:00 GMT*
5.
6. *Hi,Mark*
7. *This mail is to explain you the mail format*
8. _____
9. *Thanks*
10. *Jim*

Lines of 1-4 which called letter letterhead (Message Header) , lines of 6-10 which is the main content of the letter, called the letter body (Message Body). The fifth line is blank, according to RFC822 requirements must be separated with a blank line between the header and the message body. Header typically contains fields From, To, Subject, and Date and other fields

The model of antigen presentation major extraction *Sender add* , *Subject* and *First e–mail server address* relevant information. Mail title generally includes the main contents; the sender address contains the information related to the sender, the two pieces of information are mainly composed of text words, is the model to determine whether the important standard of spam.

For the antibody information decision table, namely the letter body keyword system $S=(U,R,Y,f)$, in the system, $U=\{x_1,x_2,\cdots,x_n\}$ represent antibody object nonempty finite sets[9]. $R=C\cup D$, in it $C=\{a_i \mid i=1,2,...,m\}$ is the condition attribute antibody object. $D=\{d\}$ said antibody decision attribute nonempty finite sets, of which $R=C\cup D; C\cap D=\varnothing$; Y is a set of attribute values antibodies, V_a is the attribute range of $a\in C\cap D$;

f is a function of the information of $U\times(C\cap D)\to Y$, that is the attribute value of each antibody object. That is $\forall a\in R$, $x\in U$, $f(x,a)\in Y$.

Definition 1: Each of these antibodies attribute subset A decision of a binary equivalence relation P, expressed by the formula $IND(P)=\{(x,y)\in U\times U \mid \forall b\in P, f(x,b)=f(y,b)\}$, $IND(P)$ will be divided into k classes U, $X_1, X_2, \cdots, X_k$ is a different type of antibody, that is:

$$IND(P)=[X]_{IND(P)}=[X_1, X_2, \cdots, X_k] \quad (5)$$

Next on the set X is defined as the approximate P

$$\underline{PX} = U\{Y_i \mid Y_i \in U / IND(P) \wedge Y_i \in X\} \quad (6)$$

Q is denoted by P positive region, as shown in Equation (7):

$$POS(Q) = \bigcup_{X \in U/Q} \underline{PX} \quad (7)$$

If $POS_{IND(P)}(IND(Q)) = POS_{ind(P)-\{a\}}(IND(Q))$, according to the definition of P and Q reduction is equivalent to the family relationship, a_i is called unnecessary Of P in Q: non-self rule base, otherwise C is necessary of P in Q: self rule base.

Algorithm is described as follows:

Input: Decision Table antibody formation

Output: self rule base, non-self rule base
1) Calculate the equivalent condition attributes $X\{x_1, x_2, \cdots, x_n\}$ set of antibodies by the formula (5)
2) Calculate the equivalent antibody decision attribute set D
3) For each equivalent set do
4) Calculated by the following formula (6) antibody decision attribute approximations
5) Calculated by formula (7) $POSP(X, D)$
6) End for
7) For each attribute do / / Attribute Reduction
8) To calculate the equivalent condition attributes set of $X - x_i$
9) By the formula (6) computing antibody attribute equivalence set lower approximation set
10) By the formula (7) calculated $POS(X\{xi\}, D)$
11) *If* $(C-\{a\}, POS(C-\{a\}), (Q) = POSC(Q))$
12 Then $C' = \{C - \{a_i\}\}$ / / delete the column where U and merge duplicate rows;
13) *Else* $C' = C$

The remaining keywords removed to form a new vector into antigen feature vector. This can not only ensure the representative feature vector selection of message content, while limiting hyperinflation feature vector length.

C. Vaccine introduction

In the phase of introducing the vaccine, including vaccine extraction module and vaccination module, two modules by vaccine controller signal coordination operation. The detector operation process, including module working time. When the detector need vaccine inoculation or injected at any time, namely the dynamic extraction and vaccination. Working cycle module is received signal extraction vaccine inoculation or from the start, end the cycle until the vaccine extraction.

Antigen ag represent the feature vectors of all message, immune cells (antibody) represent the feature vectors of spam. Set $Ag = \{ag\}, ag \subset D, D=\{0,1\}^l, (l \in N, l>0)$, $Ab=\{ab\}$, β as a mature cell activation threshold. The definition of mature immune cells set $T_b=\{x \mid x \exists Ab, x.count < \beta\}$, memory immune cells set $M_b=\{x \mid x \exists Ab, x.count >= \beta\}$, $Ab = Mb \cup Tb$; The immature immune cells set I_b, $|T_b| + |I_b| = \sigma$, where A is a constant. $|T_b|$ Represent the number of mature cells; $|I_b|$ represent immature cells[10]. Where Ag a collection of antigens, D is represents the length of the binary string space 1 antigen ag review of the data on behalf of the values characteristic attributes of the binary encoding.

Vaccine extraction

Extraction means extracting the vaccine antigen stimulation into data with the resistance characteristics of the detector. In the model, the vaccine is the process of extracting the characteristics of the problem for the information extraction process.

In the model, the vaccine extracted using dynamic way that antibodies to extract data from the memory characteristics of the detector into a vaccine. Vaccines extraction step model:

Step 1: Vaccine controller sends a signal extraction module to work (the vaccine to extract the source from the memory detector set).

Step 2: Memory detector set for classification into classifier, divided into several parts;; then, each classifier in a certain number of random factors into immature vaccine antibody collection;

Step 3: Feature information extraction, to calculate vaccine affinity for immature collection, if affinity does not meet the requirements to enter the classifier, to meet the requirements, transforming into a mature vaccine.

$f(x)$ represents the affinity functions. The real number value of the affinity between [0,1], said set A, B two disordered text vector, the text vector A, B is the affinity between the definition equation (8):

$$f(A,B)=\frac{A\cap B}{min(A,B)} \qquad (8)$$

Define 2: $M_k=\{U_{i,k}\,|\,i=1,2,..,n\}$ vaccine is expressed as K substituting immature populations, and n represents the number of vaccine mature population, I_b represents a collection of mature vaccine, immature and substituting the A K represents the best individual in the population vaccine vaccine. Extraction process described below:

T_{mature} represents use time of the mature vaccine, $T_{pick-up}$ represents the vaccine extraction cycle, $T_b(size)$ represents the size of the collection of mature vaccine. Mature vaccine extraction algorithm flow chart shown as Figure 2:

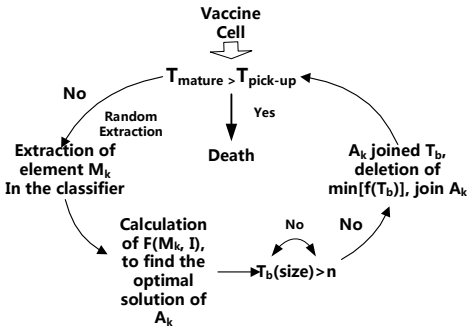

Fig. 2 mature vaccine extraction algorithm flow chart

The definition of 3: vaccine $V_a\in S, S=\{0,1,*\}^l(l\in N,l>0)$, V_a defined by the 0, 1, *. The length of the string is 1, V_a^m expressed the first m position on the gene encoding of V_a .Assuming an individual antibody population A is $a_1,a_2,...,a_n$, a_i^m represents the first M gene's coding in the first i antibodies. Fig. 2 mature vaccine extraction algorithm flow chart

Set $a_1,a_2,...,a_s$ ($s\le n$)is excellent individuals of the antibody population A under certain criteria, vaccine extraction operation is defined as:

$$V_i^m=\begin{cases}1, \dfrac{1}{s}\sum_{i=1}^{s} a_i^m(j) > \alpha\\ 0, \dfrac{1}{s}\sum_{i=1}^{s} a_i^m(j) < \beta\\ *, \quad otherwise\end{cases} \qquad (9)$$

The parameters α,β value is determined according to the actual situation.From definition 3 can be drawn: a vaccine represents an excellent model. Vaccine except the values for * gene, called it a good gene.

Vaccination

Vaccination: using the extraction vaccine (including qualified vaccines and excellent vaccine) to change the position of certain genes of antibodies, resulting in excellent genes to the next generation, increase the probability of excellent mode of reproduction to repair the good genes In the crossover and mutation

Vaccination module start to work from receiving a signal, the target of Vaccination is put the vaccine into the mature antibody detector set. Vaccination procedure described as follows:

Step 1: the vaccine controller to send the signal of vaccine vaccination, module start to work (vaccine extract from mature detector set).

Step 2: A certain number of mature detector to be inoculated target randomly selected, with mature vaccine set of injection, conducting vaccination after vaccination emitted signal

Step 3: Calculation of vaccine vaccination in the threshold, if the threshold is reached, to be inoculated into mature detector set, and added to the mature detector set; if did not reach the threshold of matching, exclude delete it.

Definition 4: the assumption that the antibody is a , vaccine is v_a , vaccination operation is $\hat{a} = a \Theta v a$, antibody a coded form is $\hat{a}$ after vaccination. Vaccination operation is defined as:

$$\hat{a_i^m} = a^m \Theta v_i^m \begin{cases} v_i^m, & v_i^m = 0 \, or \, 1 \\ a^m, & v_i^m = * \end{cases} \qquad (10)$$

In order to improve the sharing between the detector antibodies, using the fitness sharing niche technique, Design of a mechanism of sharing algorithm to access to a variety of immune memory cells, and its purpose is to enhance the diversity of antibody detector and restore the good individual, create a niche evolutionary environment for it. The sharing function is a function of the degree of close relationship about two individuals [14]. The antigen antibody as the problem (P), antibody Ab is the candidate solutions to the problem, Antigen's affinity for antibody Ab is $f(Ab)$,the population size A. When the relationship between the two antibodies very close, function value of sharing is larger, approaching in 1; otherwise, approach in 0. In the immune algorithm, affinity function equivalent to a fitness function.

Definition 5: Suppose groups is $A = \{Ab_1, Ab_2, \ldots, Ab_n\}$, let $S(d_{ij})$ represent a shared function between antibody Ab_i and Ab_j :

$$S(d_{ij}) = \begin{cases} 1 - (\frac{d_{ij}}{r_s})^a, d_{ij} < r_s \\ 0, \quad d_{ij} \geq r_s \end{cases} \qquad (11)$$

In the above formula: d_{ij} is the distance measure between individuals, using Euclidean distance; r_s is niche radius; a shared parameter adjustment functions .

Model by using niche technology, so you can exchange information between the detector and sharing antibodies, improves the flexibility of the detector, and a distributed way to identify spam messages feature.

TEST AND ANALYSIS

A. Test and evaluation standard

Model uses information retrieval precision (precision), recall (Recall), precision (Accuracy) and false rate (Fallout) as the main measure of the efficiency of the model. Testing China Education and Research Network (CCERT) Chinese e-mail sample set (Jun 2005), 9272 regular mail, spam 25088. Check the accuracy of the simulation model and the recall rate, and compared with other algorithms (such as Naive Bayes, artificial immune).

Assuming a total of N messages test set, the definition of variables: spam (Spam), legitimate messages (Ham). $C_{S \to S}$expressed as the number of Spam judge as Spam; $C_{H \to S}$expressed as the number of Ham judge as Spam; $C_{S \to H}$ expressed as the number of Spam judge as Ham; $C_{H \to H}$ expressed as the number of Spam judge as Spam , $N = C_{H \to S} + C_{H \to H} + C_{S \to S} + C_{S \to H}$.

Then we can define the accuracy rate, false alarm rate, precision rate and recall evaluation index, to evaluate the mail classification system performance.

$Recall = C_{S \to S} / (C_{S \to S} + C_{S \to H})$, namely the spam detection rate. This indicator reflects the ability of the model to detect spam, the recall rate is high, the more the spam detection.

$Precision = C_{S \to S} / (C_{S \to S} + C_{H \to S})$, which is a spam probability of correct. The correct rate is higher; the misjudgment of legitimate messages as spam, the fewer the number of.

$Accuracy = (C_{S \to S} + C_{H \to H}) / N$, that is to judge all mail, and determine the probability of correct.

$Fallout = C_{H \to S} / (C_{H \to S} + C_{H \to H})$, the model will judge the normal mail as spam probability, false alarm rate is high, the system will judge the normal mail as spam may be greater.

B. Experimental results and analysis

The simulation of the model, the data set was divided into training set and test set, choosing 2360 emails (1272 letters and 1088 samples of normal mail spam) as the training set to get the initial detector. Using the remaining 8000 legitimate emails and 24000 spam emails were divided into 10 groups, consisting of the test set, then test, finally take the average of 10 experiments as the last experimental data, the average check the correct ratio and the average precision rate is and the average precision rate model.

Figure 3 is the model (threshold 0.60) statistical data obtained spam filtering experiment in a simulation environment, mainly for accuracy, precision, recall rate, false alarm rate of four indicators.

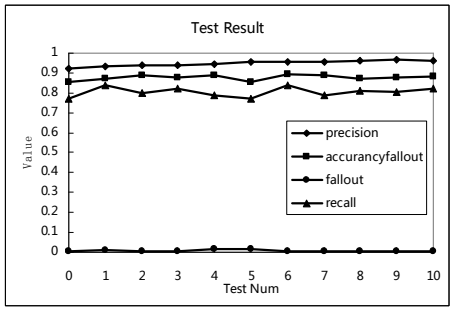

Figure 3 Mail Filtering Index Graph

As can be seen from Figure 3, the correct rate is relatively stable, the recall rate is rising, the basic can reach more than 90%, indicating that the model can determine the spam well. Precision rate also is rising, indicating that the antibody with the model of self-learning, the ability of distinguish spam also increased. After test data, the model has better recall in terms of the mode Identify spam features, with better learning and memory recognition. False rate is lower and relatively stable, in this case, since the model has a co-stimulatory mechanism, but the mail will not be accidentally deleted, indicating that the model has high reliability.

In order to verify the efficiency of the model, the paper in the same conditions (threshold 0.60) compare the with test based on Bayesian spam filtering model[11] and AIS spam filtering model[12], shown in Figure 4, the improved immune system model have greatly improved over the precision rate than the model based on Bayesian model and AIS methods, and with a stable trend, reduce the rate of false positives and less volatile.

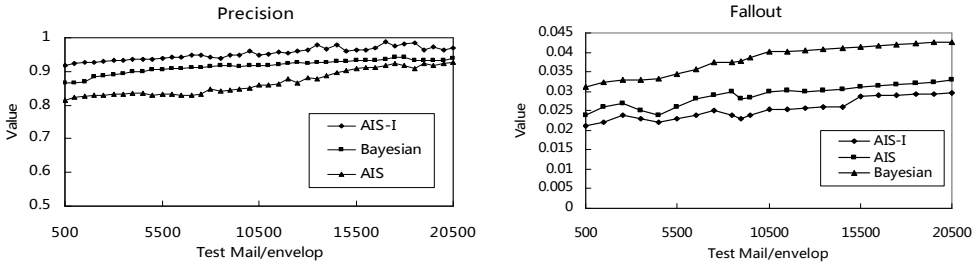

Figure 4 Efficiency comparison chart

Based on Bayesian probability models generally use a priori rules to test results, if the detection phase detection mail contains a lot of new training phase does not appear in the mail over the word, the model takes a long time to learn to adapt, the capacity of learning and memory is lower. However, the AIS model use a distributed manner to identify spam messages and to learning and memory features, but poor flexibility of the detector, and can not be shared between antibodies. In the test phase is not accurate to classify spam, memory antibody characteristics. The proposed model, produced antibody in the initial training phase. Vaccines and vaccination extraction module is introduced into the filter model, which not only allows the detector to make all kinds of vaccines as a medium to communicate with each other, sharing antibodies, and showing compared high efficiency.

CONCLUSION

Based on the related principle of artificial immune, proposed a filtering model based on immune mechanism spam. The model introduce into mutual information, control scale of autologous libraries and gene banks; designing a letter body keyword system in the course of the antibody proposed. The model can effectively classify spam e-mail feature extraction; in the stage of vaccine introduced, through vaccination and extraction increases the flexibility of the detector, and can make antibodies shared, effectively identify unknown and variability characteristics. In addition, Some of the parameters in the model is the interaction should maintain proper proportions. Experimental results show that the model by comparing with other models has better performance, but also effectively improve the efficiency of spam filtering.

Acknowledgements
National Natural Science Foundation (No.61272038); Henan Science and Technology Agency-funded science and technology research projects (No.0624220084); Henan Province Department of Education Program (NO.2010A520044)

REFERENCES

[1] Gansterer W, Ilger M, Neumayer P,et al. Anti-spam methodsstate-of-the-art[D].Vienna: Faculty of Computer Science.University of Vienna, **2005**: 1-99.
[2] MN Marsono,MW El-Kharash,F Gebali.Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification[J].*Computer Networks*,**2009**,53(6):835-848.
[3]Mehmet Aci,Cigdem İnan,Mutlu Avci.A hybrid classification method of k nearest neighbor,Bayesian methods and genetic algorithm[J]. *Expert Systems With Applications*,**2010**,37(7):5061-5067.
[4] Yu Bo, Xu Zongben. A comparative study for content-based dynamic spam classification using four machine learning algorithms[J]. *Knowledge-Based System* s,**2008**,21(4):355-362.
[5] Clark, J.;Koprinska, I.;Poon, J. A Neural Network Based Approach to Automated E-Mail Classification[C].Web Intelligence:Proceedings of the 2003 *IEEE/WICInternational Conference on Web Intelligence*. **2003**:13-17.
[6] Qing J J,Mao R L,Bie R F,et al.An AIS-based e-mail classification method[C]//The **2009** International Conference on Intelligent
Computing,Ulsan,Korea,**2009**:492-499.
[7] Jue Huang, Bing Chen, Changwu Liao. Artificial immune spam [J].*Computer engineering and application filtering algorithm*, **2011**,47 (30): 72-74.
[8] Shaorong He,Jinming Liang,He Zhiyong. Selection method for[J]. Computer engineering characteristics and mutual information based on the theory of the relationship between product, **2010**,36 (13): 257-259.
[9] Ling Zhang, Zhongying Bai, Luo Shoushan et al. The Integrated intrusion detection model based on rough set and artificial immune [J]. *Journal of China Institute of communications*, **2013**,34 (9): 166-176.
[10] Tao Li. Network monitoring model based on immune [J]. *Chinese Journal of computers*, **2006**, 29 (9): 1515-1522.
[14] Le Zhang,Zhu Jing-bo,Yao Tian-shun.An evaluation of statistical spam filtering techniques[J].ACM *Transactions on Asian Language Information Processing*(TALIP),**2004**,3(4):243-269.
[15] T.S. Guzella,T.A. Mota-Santos,J.Q. Uchôa,et al.Identification of SPAM messages using an approach inspired on the immune system[J]. *Biosystems*, **2008**, 92(3):215-225.