Design and Implementation of the Anti-spam System Based on AIS

Jue Huang and Changwu Liao

Abstract With the wide use of Email on the internet, it is urgent to resolve the severe problem of spam. Inspired by the biological immune system, it proposes an algorithm that can classify Spam without re-training. The result of experiments showed that the algorithm, compared with Bayesian algorithm, can improve recall rate and has a higher dynamic adaptability and stability.

Keywords Artificial immune system • Spam • Bayesian • Algorithm

1 Introduction

With the growing popularity of email, it plays a more and more role for people to communicate with each other. However, as the carrier of commercial advertising, viruses or content-sensitive, the spam occupies space and network bandwidth, people have to spend much time to pick normal mail from lots of spam. Spam even makes serious damage for our life and network. Anti-Spam has become a global problem. As spam is dynamic and various, we designed and implemented an anti-spam system based on the artificial immune.

2 Current Status

Because spam has became serious, there are many anti-spam system has been developed. They can be classified to based on email server or based on email client, and the email detect technical can be classified to based on rules

J. Huang (✉) • C. Liao
Department of Computer and Software, Nanjing Institute
of Industry Technology, Nanjing, China
e-mail: huangj@niit.edu.cn; liaocw@niit.edu.cn

B. Xu (ed.), *2012 International Conference on Information Technology and Management Science (ICITMS 2012) Proceedings*, DOI 10.1007/978-3-642-34910-2_90,

(Carreras and Marquez 2001), based on statistic(Sahami et al. 1998), based on connection(Cao et al. 2004).

The detect technical based on rules, such as decision tree, interact rule, they are used mainly in some special field, they need domain knowledge of the field and rule lib to support and the rule's design, update, self learning is very difficult. So they are not used very popular and the performance is not very good.

The detect technical based on connection, such as Artificial Neural Network. The method is good in some aspect, such as self learning. But it needs lots of sample to train system and the train need lots of time. So the method is hard to be used.

The detect technical based on statistic, it adopt pure statistic principle and system self learning method. It builds up characteristic set by learning known sample to detect new sample such as Bayesian algorithm. Most of the current anti-spam systems use the technical.

But the Bayesian algorithm is not perfect. First, it based on independence assumption. The technical doesn't consider the connection between words or phrase, so it can't analyze email content especial Chinese. Second, it can't continuously self learning. If the mail content has much change, the correct rate will downgrade. And the technical has worse anti-interference performance. The Bayesian classify system classify email based on known sample, It has high detect rate. But the detect performance depends on training sample. So for the unknown sample especially mutation sample, its detect result is not very good. Unfortunately, more and more spam sender use the method to escape detect system.

For that reason, the paper adapt to the protect mechanism of biology immune, use artificial immune system to build up anti-spam system. Give a new anti-spam system module based on artificial immune system.

3 Artificial Immune System Introduction

3.1 Biology Immune System VS Anti-spam System

Artificial immune system is a new research direction after artificial neural network and evolvement calculate. Recently, artificial immune system has become a research hot spot for message security and network security.

Biology immune system is a complex system consist of immune cell, organ. The major function of Immune system is distinguishing self and Non-self. When antigens begin to invade, immune system will generate lymph cells to resist antigens. Immune system can self learning, recognize antigens, immune remember and so on (Mo 2003).

Biology immune system and anti-spam system are similar. The following Table 1 is their compare relation.

Table 1 Natural system and email system

Natural system	Email system
Self cells	Normal emails
Non-self cells	The emails to be detected
Native Cells	Native detector
Memory Cells	Memory detector
Antigens	Spam
Co-stimulation	Feedback message from user
Affinity	Similarity between emails
Cells lifecycle	Detector lifecycle

3.2 Immune Remember

When immune system meet antigens at first time, It will generate immune response, when it meet the same antigens once more, because associate member, the member cells will be wake up, immune system will give response fast then last time and generate lots of immune body to destroy antigens. During anti-spam system running, remember detect set will be build up according to immune cells' remember characteristic.

3.3 Affinity

Affinity was defined to match rate between antigens and immune body. The spam detect process just is match process between detect set and unknown characteristic and judge whether the unknown characteristic is self or not.

The vector length of detect set Ta, The vector length of unknown mail Tb, statistic the same word between two vector and assign to Tsame, so we can get.

$\mathrm{Tsame} = \mathrm{Ta} \cap \mathrm{Tb}$, $\mathrm{Tmin} = \mathrm{MIN(Ta, Tb)}$, $\mathrm{Affinity} = \mathrm{Tsame}/\mathrm{Tmin}$.

When affinity greater than threshold, the mail will be judge to spam and require customer response for the detect result.

3.4 Detect Set Generate Algorithm

Detect set is the collection of immune body. Extract cell section randomly to generate un-mature detect set, then un-mature detect set match with normal email, if match can't success, the un-mature detect set will evolve to mature detect set. If the count of match spam in limit time exceed algorithm, mature detect set will evolve to remember detect set. If mature detect set can't match in a long time, it will be delete from mature detect set.

4 Chinese Spam Filter System Design

4.1 Mail Pre-Process

When Chinese email incoming, first parse email content, then split words from mail content and filter stop words.

4.2 Pick Characteristic

The popular characteristic select method is word appear frequency, message gain, mutual information, statistic, expect cross entropy. They all calculate statistic message for a Chinese word, then set a threshold, delete the words their statistic message less than the threshold. Others are valid characteristics. The system will use mutual information as evaluate function.

$$\mathit{difference}(W) = \left\| \log \frac{P\ (W|C_1)}{P(W)} \right| - \left| \log \frac{P\ (W|C_2)}{P(W)} \right\| \tag{1}$$

P(W):the word W appeared probability in all documents;

P(W|Cj):the word W appeared probability in document j.

We sort the difference according to the calculation result, select some words as characteristics.

4.3 Build Up Vector Space Module

Phase email, then split words, select characteristics, generate words set, then describe email as a method which fit to calculate, and classify them.

4.4 The Judgment of Unknown Email

The following flow will imitate the characteristic of Biology, use member detect set, mature detect set analyze unknown email.

```
//unknown email detect algorithm
Read mail vector
{Read remember detect set;
   Match email vector with each remember detect set member;
If(affinity>MT) //if it can match, memory detector regard the mail as spam
User confirm the detect result
```

```
    If (user think it is a spam)//email.flag=1
    {Put EM into spam set;
      Put all cells of EM to cell set;
      Ab1.count=Ab1.count+1;}//the match count of immune body +1
    Else delete remember detect set;
  Else {Read mature detect set;
    Match mail vector with each member of mature detect set;
      If (affinity>MT) //if it can match, native detector regard the mail as spam
        User confirm the detect result;
        If(user think it is a spam)//email.flag=1
          {Put EM into spam set;
          Put all gene from EM into gene set;
          Ab2. count=Ab2. count+1;
          Ab2. age=Ab2. age+1;}
      Else delete mature detect;}
  Else User confirm the email;
    If(user think it is a spam)//email.flag=1
      {Put EM into spam set;
        Put all gene from EM into gene set;}
    Else Put EM into self set;}
```

5 Experiment and Result

The purpose of the experiment is compare artificial immune system and Bayesian algorithm. The operate system is Windows XP Professional, and language is C++. We select 2,000 Chinese emails as the sample.

5.1 Accuracy

Accuracy is the correct rate of system detect. From the Fig. 1, the accuracy of AIS is 85–90%, the accuracy of Bayesian is 82–90%, so we can see that AIS has a better stability robustness than Bayesian.

5.2 Recall

Recall stands for call back rate. The high recall rate mean less spam leak by system. From the Fig. 2, the recall rate of AIS is better than Bayesian.

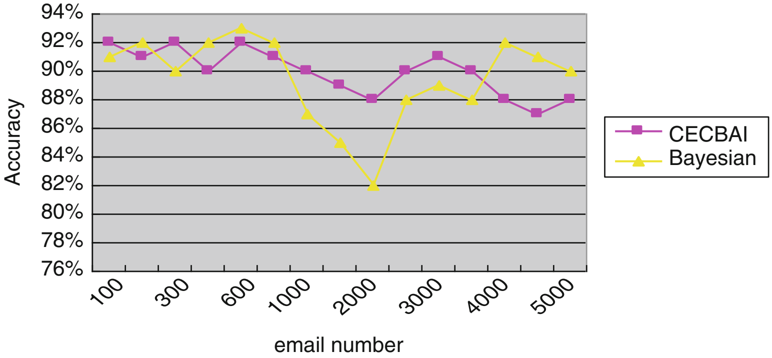

Fig. 1 The accuracy of AIS and Bayesian

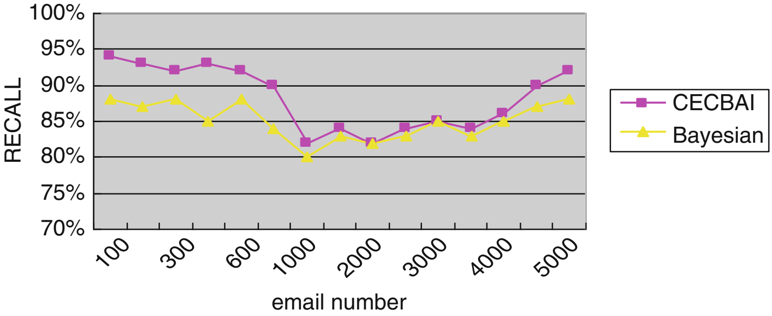

Fig. 2 The recall of AIS and Bayesian

6 Conclusion

The paper implements anti-spam system based on artificial immune algorithm. The system can handle mutagenesis and has better performance. In recall rate, it is better than Bayesian algorithm.

References

Cao YK, Liao XF, Li YF (2004) An email filtering approach using neural network, Lecture notes in computer science. Springer-VerlagGmbH, Heidelberg, pp 688–694

Carreras X, Marquez L (2001) Boosting trees for anti-spam email filtering. In: Proceedings of Euro conference Recent Advances in NLP (RANLP – 2001), pp 58–64

Mo HW (2003) The principles and application of artificial immune system. Harbin industry university publishing house, Harbin, pp 38–48

Sahami M, Dumais S, Heckerman D (1998) A Bayesian approach to filtering junk email. In: Proceedings of AAA'I workshop on learning for text categorization, pp 55–62