Study Minimum Risk Spam Filter Algorithm Based on Cognition-learning

Lu QingMei
School of Eletronics and Computer Science and Technology
North University of China
Taiyuan ,China, 030051
wk_xiaolu@126.com

Liu Hong Min
School of Eletronics and Computer Science and Technology
North University of China
Taiyuan ,China, 030051

Abstract—**The thesis puts forward anti-spam model and deficiency on spam filter based on minimum risk, and of austerity algorithm and through the introduction of cognition learning. The experiment proves that the forward of the method increased the recognition percentage of the spams, especially solved the problems of the spam filter.**

Keywords-Minimum Risk, Cognition-learning, Bayesian

I. INTRODUCTION

In the mail filtering and classification, legitimate mails have different characteristics from spams. It will likely to bring greater loss if the legitimate mails were convicted of junk mails. In this reason, not only should we consider making the right judgment as far as possible, but also must we think over the consequence of misjudgment.

At present spam filtering technology is developing fast, mainly based on rules and based on statistical method of filtering technology. The advantage of technology based on the rules is easy to understand , easy to amend and easy to popularize. For example , we can use technology of spam list rules and Spam Assassin tools after suitable parameter adjustment to filter 90% of spams. However its defect is that the effect in the application field of the not obvious regularity is poorer. The technology based on statistical method is a popular method and has a broad prospect of application with high accuracy and filter speed， such as bayesian SVM KNN and neural networks. But they also have some shortcomings like that the early learning time is long ,the parameter selection by experience is strong and the misjudgment risk is high and so on.

The spam filter algorithm based on cognition-learning and minimum risk is the very decision rule which is put forward by considering all kinds of loss caused by different errors.

II. ALGORITHM BASED ON MINIMUM RISK [1]

Minimum Risk brings in the loss factor $\lambda(a_i,c_j),i=1,2,\cdots,n;j=1,2,\cdots,m$ considering the risk of misjudgment based on the NaiveBayes. c_j is a natural state in the state space, a_i is a decision in the decision space, λ means the loss the decision brought, and it's a function of the natural state c_j and the decision . So we can use a decision table to present the relationship.

Although the algorithm based on the Minimum Risk can solve the problem of spam misjudgment well, it also has some shortage as following:

The system does not have a special knowledge base itself and its learning method was too simple, so it cannot really learn new knowledge effective and the learning efficiency is low. When $d\in C_k$ is mistaken for $C_k^{'}$ by the system, the system only simply increase or decrease the mail vector in different categories by learning . However, it cannot guarantee $P(C_k^{'}/d)>P(C_k/d)$ .This way, the adjusted system still cannot recognize the mails which have the same characteristic with d .

The system does not consider the risk of misjudgment. It has higher risk to look upon the legitimate mails as spams than to regard the spams as legitimate mails. This will lead to a high frequency of intervene by users and decrease the practical applicability of the system[2].

III. SPAM FILTER ALGORITHM BASED ON COGNITION-LEARNING AND MINIMUM RISK

On the basis of above analysis, we put forward a new algorithm based on cognition-learning and Minimum Risk for spam filter.

A. Definition 1:

Given Ω is a discrete sample space of mails, $s_1,\cdots,s_n$, are words or phrases which have their own significance included by Ω ,we call them knowledge . If we give every knowledge a order to define $S=\{(s_1,1),\cdots,(s_n,n)\}$,we call that S is the knowledge base of Ω.

According to the define 1, with the mail $d\in\Omega$, we make the vector space model vectoring with the following ways . Suppose the position in S of the phases $\omega_1,\cdots,\omega_m$ in $d\{\omega_1,\cdots,\omega_m\}$ is $i_1,\cdots,i_m$. So we can express d as $d=(\cdots,w_d^1,\cdots,w_d^2,\cdots,w_d^m,\cdots)$, in the equation , w_d^i is the times of the phase ω_i appears in d . The position

Published by Atlantis Press, Paris, France.

$w_d^1,\cdots,w_d^m$ appear in d is also $i_1,\cdots,i_m$. And the other positions in d , which are in S and are not belong to $\omega_1,\cdots,\omega_m$, their value are 0.

In order to decrease the situation that words in the text are too few to make the numerator or denominator to be 0 so that the system cannot operate well. We try to correct the conditions probability factor $P(\omega_j / C_k)$ as following :

$$P(\omega_j / C_k) = \frac{1+\sum_{i=1}^{c(k)}(\omega_j \otimes d_i^k)\times N(\omega_j, d_i^k)}{c(\Omega)+\sum_{k=1}^{c(k)}\sum_{i=1}^{c(k)}(\omega_j \otimes d_i^k)\times N(\omega_j, d_i^k)} \qquad (1)$$

In the equation d_i^k is the mail vector in C_k , $\omega_j \otimes d_i^k$ is the condition factor, that is to say if $\omega_j \in d_i^k$, it is w_d^j, else it is 0. $N(\omega_j, d_i^k)$ represents the value of ω_j in the vector d_i^k , it is the time of the ω_j appears in the mail d_i^k . $c(k)$ means the amount of the vectors of the total mails in the class C_k . $c(\Omega)$ is the whole categories of the mail in Ω , and Ω is a discrete mail sample space[3] .

By the equation (1) , we can get that whether the characteristic information included by all categories is complete and whether the amount of knowledge in S is sufficient have the directly impact on the correctness if we use the Bayes algorithm to filter mails. Because the amount of knowledge in S will affect the vectoring of the mail text. And the characteristic of included by the categories will have directly influence on the result of the equation (1).

B. *Definition 2 :*

For mail $d\{\omega_1,\cdots\omega_m\}\in\Omega$,we call the binary vector $((w_d^1,c(\omega_1)),\cdots(w_d^m,c(\omega_m)))$ the vector expression of d . In the expression , $\omega_1,\cdots,\omega_m$ are the check tags of d . w_d^i is the times of phase ω_i appears in d . And $c(\omega_i)$ means the position serial amount that ω_i appears in S .

From the Definition 2 ,we can get that vector d in the primary super high dimension vector space is turned into a vector in a low dimension space. So we can save some storage space for vectors. And it reduces the computational complexity .

C. *Definition 3 :*

Given the all the mails in the sample space Ω have been divided into categories $C_{t1},\cdots,C_{tn},C_{f1},\cdots,C_{fn}$, in this expression $C_{t1},\cdots,C_{tn}$ is the legitimate category and $C_{f1},\cdots,C_{fn}$ is the spam category . For any category $C_k(k=t1,\cdots,tn,f1,\cdots,fn)$ we have:

1) $C_k=\{d_1^k,\cdots d_{c(k)}^k\}$ means the vector set in C_k , in which $c(k)$ means the rank of the mail vector in C_k .

2) $S_k=\{(c_k^1,f(c_k^1)),\cdots,(c_k^{|k|},f(c_k^{|k|})\}$ means the set of knowledge in C_k , in which $f(c_k^j)$ means the position amount C_k^j appears in S . And $|k|$ means the NO.j mail vector in C_k .

3) $C_k=((w_k^1.f(\overset{\to}{c_k^1}),\cdots,(w_k^{|k|}.f(c_k^{|k|}))$ is the vector expression of C_k , in the equation w_k^j is the times of C_k^j appears in C_k .

From Definition 3, we can get that for mail $d\{\omega_1,\cdots\omega_m\}\in\Omega$ we can simplify the Bayes formula as following:

$$P(\omega_j / C_k) = \frac{1+\sum_{i=1}^{|k|}(\omega_j \otimes c_k^i)\times w_k^i}{c(S)+\sum_{k=1}^{c(S)}\sum_{i=1}^{|k|}(\omega_j \otimes c_k^i)\times w_k^i}$$

In this formula $\omega_j \otimes c_k^i$ is the condition factor , and if $\omega_j = c_k^i$ it is w_d^i , else it is 0 [4].

D. *Definition 4 :*

Given the divided categories $C_{t1},\cdots,C_{tn},C_{f1},\cdots,C_{fn}$ of all the mails in the sample space Ω . In this expression $C_{t1},\cdots,C_{tn}$ is the category of legitimate mails and $C_{f1},\cdots,C_{fn}$ is the category of spam. Also we define some symbols : Q^T is the category space of legitimate mails, Q^F is the category space of spam, and Q^N is the category space of the system cannot recognize .

In the definition 4 we have :

① $Q^T=\{x\mid\forall x\in S,\exists C_k\in\{C_{t1},\cdots,C_{tn}\},$**to make** $P(C_k/x)>\alpha\}$

$Q^F=\{x\mid\forall x\in S,\exists C_k\in\{C_{f1},\cdots,C_{fn}\},$**to make** $P(C_k/x)>\beta,$

② $and\forall C_{k'}\in\{C_{t1},\cdots,C_{tn}\},P(C_k/x)\le\alpha\}$

$Q^N=\{x\mid\forall x\in S,\exists C_k\in\{C_{f1},\cdots,C_{fn}\},$**to make** $P(C_k/x)\le\beta,$

③ $and\ \forall C_{k'}\in\{C_{t1},\cdots,C_{tn}\},$ we have $P(C_k/x)\le\alpha\}$

We call α,β the probability threshold . Based on the consideration of minimum risk , we have $0.5<\alpha<\beta<1$.

When we judge the category of mails by the Bayes formula, for any mail d , if $P(C_{t1}/d)=\max_{j=1}^{n}(P(C_{t1}/d))>\alpha$ we have $d \in C_{t1}$. Else if $P(C_{t1}/d) \le \alpha$ and $P(C_{f1}/d) = \max\limits_{j=1}^{n}(P(C_{f1}/d)) > \beta$, we have $d \in C_{f1}$. Else we have $d \in C_N$.

Now we put forward the definite procedure as following:

1)First we should divide mail d into $d\{\omega_1, \cdots \omega_m\}$, and add the new knowledge in $d\{\omega_1, \cdots \omega_m\}$ to the end of knowledge base , $S = S \cup \bigcup\limits_{i=1}^{k}\{(\gamma_i, n+i)\}$ in the equation γ_i is the new knowledge of $\omega_1, \cdots \omega_m$ which is not in S .

2)Vector the mail $d\{\omega_1, \cdots \omega_m\}$ into $d = ((w_d^1, c(\omega_1)), \cdots, (w_d^m, c(\omega_m)))$ by the binary vector model .

3)Judge the category which the mail belongs to .

① if the algorithm put the mail to the category $C_k \notin Q^N$, the algorithm will add the new knowledge in $d\{\omega_1, \cdots \omega_m\}$ to the end of C_k and put d into the category C_k dynamically , so $S_k = S_k \cup \bigcup\limits_{i=1}^{I}\{\bullet i\tau_i, |k| + i \bullet j\}$,

$$C_k = C_k \cup \{d\} \ , \ \overset{\rightarrow}{C_k} = \overset{\rightarrow}{C_k} + d$$

In these τ_i is the new knowledge of $\omega_1, \cdots \omega_m$ which is not in the category C_k

② if the algorithm put the mail to the category C_N , it will judge the category which it belongs to by users , and then put d into the category C_k dynamically .

For the misjudged mail $d\{\omega_1, \cdots \omega_m\}$,users can transfer the wrong category C_k to the right category $C'k$ by recognition themselves . System simulates the cognition-learning process of human being by the users' recognition in order that the system can recognize the mails who have the same characteristic with d next time.

IV. CONCLUSION

In the experiment , we chose 300 of known sample mails in which spams' amount is 250 and the legitimate mails' amount is 50. Figure 1 is the comparison between the algorithm based on Minimum Risk and the algorithm based on Cognition-learning . The algorithm can adjust the categories after recognition itself by importing the cognition-learning ability . And it can also increase the recognition speed of the system and recall ratio by decreasing the dimensionality of the vector space and adjusting the risk threshold

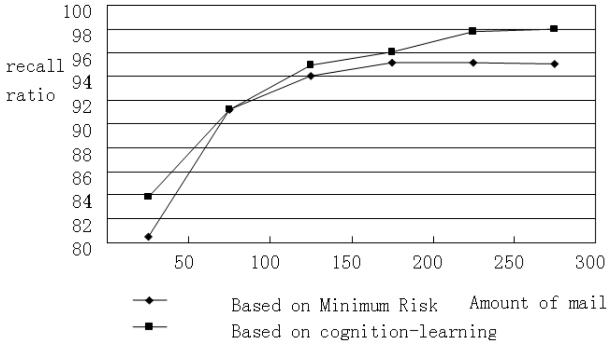

Figure 1. comparison of recall ratio

In order to check the self-learning ability of the improved Bayes algorithm we use the mail to test directly without using the samples . For the given sample, we give the different thresholds to α 、 β to test in order to check the filter effect with different threshold. The consequence is in Figure 2.

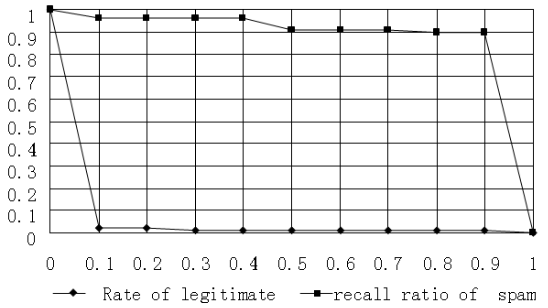

Figure 2. recall ratio and misjudgment ratio in different threshold

The experimental results show that the method based on cognition-learning is a new kind of spam filter algorithm, it can simulate the learning process of human beings and it has good learning ability and results. What's more , it also has good dynamic regulation and high recall ratio and precision ratio . So it can improve the function of spam filter. Combining with more sample test in advance and the right threshold choice , the algorithm will be more outstanding.

REFERENCES

[1] Ma Fengyun ,Liu Peiyu. Spam filter technology based on neural network. information security, 2005,4:76-78

[2] Li Peifan, Ma Heng. the world wide web client content instant filtration system based on neutal network , The Chinese university science and technology management institute, 2006 - 6

[3] Robertson S, Hull DA. The TREC-9 filtering track final report. In: Voorhees EM, Harman DK, eds. Proceedings of the 9th Text Retrieval Conference (TREC-9). Gaithersburg: NIST Special Publication, 2006. 25~40.

[4] Liu Zongren. The present situation and the problems of online content filtering technology. Shandong University of Technology, 2007-3

[5] David D. Lewis. Feature Selection and Feature Extraction for Text Categorization. In speech and Natural Language Workshop,1992

[6] G. Developments in automatic text retrieval. Science, 1991,253(5023):974~979.

[7] Zhang Yingjiang , Chen Chi. Study and application of WEB filter. Journal ofWuhan university of science and technology . 2008 , (2)

[8] Gartner, Market Share: URL Filtering, Worldwide, 2007

Published by Atlantis Press, Paris, France.