http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 7 (1): 165-169, 2008
ISSN 1812-5638

Research on Spam Classifier Based on Features of Spammer's Behaviours

Liu Zhen, Tan Liang and Zhou Ming-Tian
Westone United Lab, College of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu, 610054 SiChuan, China

Abstract: Traditional spam filtering techniques based on email' s content used to implement text-related machine learning and classification. Nevertheless, the uncertainty of message's content causes a performance bottle for machine's discrimination and classification. Aiming at the deficiencies of traditional spam filtering methods, this study brings forward a vector supported classifier model based on the features of spammer's behaviours. The evaluation results for real spam testing set show that the spam classifier based on features of spammer's behaviour is capable of discriminating email's type with pretty high accuracy. Meanwhile, the classifier model is of robust performance on noise data.

Key words: Spam, features of spammer's behaviour, support vector, discrimination boundary

INTRODUCTION

Currently, common anti-spam techniques include key words filtering, black list or white list filtering, rule-based filtering, etc. Nevertheless, these methods mentioned above have many limitations and drawbacks such as existing high false positive ratio and false negative ratio, limited application conditions, no training or learning capability and so on (Zhen *et al*., 2005). Therefore, facing with more and more severe situation for anti-spam, to work on anti-spam algorithms with high performance is asked very urgently.

Figure 1 shows a spam's life cycle model. According to this model, we induced that the key point to eliminate spam is how to destroy the step of spam's delivery at server end or common users end. The most difficult point is how to distinguish spam on MTA or MUA correctly. As an information carrier, it is no difference for machine that an email is good or bad, real or faked. As for human being, people usually depend on email's content to judge whether an email is spam or not. However, to identify an email's type based on content is often biased by each people. Let's have an example to illustrate this. Some popular commercial emails may be normal advisements and also may be faked ones in some cases. Those people who don't need them would take them as spam, but others who do need them would never mind about them. Therefore, to identify an email's type might be implemented easily by people but very hard for machine. Based on our former analysis, the subjective uncertainty to identify an email's type would bring great challenges to classifiers and filters which are based on email's content. No matter text cluster, content mining or other similar techniques, the common characteristic of these methods is to use key words or key sentences to discriminate email's type. Generally speaking, these key elements are not only appearing in spam but also appearing in legitimate ones sometimes. This characteristic would lead to filter's false positive problem. On the other hand, some spammers who want to escape filtering would insert ASCII characters into key words or key sentence randomly. Such a content cheating trick is simple but effective which would lead to filter's false negative problem. So, blur and redundant information which key words and key sentences might possess would be a big problem needed to solve in connection with content based filters. In Sarah Jane Delany's paper, she points out that concept drift existed universally in emails would result in big changes in the hidden context which can induce changes in the target concept (Delany and Ham, 2005). Obviously, concept drift may cause semantic uncertainty and incapability for key words learning under specific context. This problem would introduce further

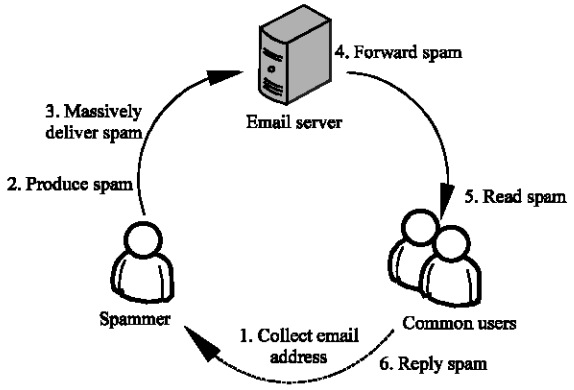

Fig. 1: Spam's life cycle model

Corresponding Author: Liu Zhen, Westone United Lab, College of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu, 610054 SiChuan, China

difficulties to content based filter. Another problem which content-based filter has to cope with is key words dictionary's refreshment and updating. As it is well known that content based filter would be out of date gradually alone with new key words' emergence. So, the key words dictionary must be maintained and updated from time to time. Now, some anti-spam organizations like MAPS Spamhaus have provided key words stop list downloading service. China anti-spam organization, CCERT anti-spam team, also provides this kind of service (Ying, 2005). But the list requires users to maintain personally. It's not a very convenient work for every common user. In a word, to improve content based filter's performance needs to overcome numerous drawbacks and bottlenecks.

This study tries to avoid many difficulties that content based filters have encountered and research on the spam categorization method from another point of view. We consider investigating the email categorization method based on spammers' behaviors mode. After carefully model design and simulation tests, we found that the new categorization model is accurate and robust for spam discrimination.

BEHAVIOR FEATURES ANALYSIS

We collect a big volume of spammers' behavior features from the Internet and organize them into a behavior features library. Among these behavior features, we divide them into three categories, namely e-mail's time anomaly behaviors, e-mail's sending anomaly behaviors and e-mail's format anomaly behaviors. E-mail's time anomaly behaviors represent delivering spam on an important date like holiday, the day of a virus out-breaking, etc. E-mail's format anomaly behaviors are mainly concerned with the e-mail addresses like forged delivery address and over large carbon copy addresses, etc. e-mail's format anomaly behaviors are related to e-mail's layout and style like html formatted email body, null email title, etc. Table 1 shows part of these behaviors and corresponding instance.

According to RFC822 and the later revised RFC2822, Table 2 shows some message fields either in message's head or message's body mapped by those behaviours features shown in Table 1.

BEHAVIOR FEATURE SELECTION

To ensure online updating's efficiency for the discrimination boundary, choosing features vector with over-high dimensionality is not suitable, Meanwhile, in view of that spammers would change their behaviors to avoid filtering, the behavior features library would be expanding and growing. Aiming at this paradox, to meet the precondition of computing efficiency and trusted features extraction, we use Mutual Information (MI) features extracting algorithm (Pelletier *et al*., 2004). MI can well represent the statistic essentiality of a behavior and therefore can beutilized as a trusted approach for feature selection. To obtain a better discrimination boundary, MI algorithm can adapt to select the most important behaviors and discard less important behaviors dynamically.

Table 1: Category list for spammer's behavior features

Category	Feature description	Value	Instance
Time anomaly	Important date	D	Date: Sun, 1 May 2007 21:27:02 +0800
	…	…	…
Sending anomaly	Over-big Carbon Copy	G	CC: = ?gb2312?B?sqnKvzM=?= " <24?~LIST~?@mci.uestc.edu.cn>
	Forged sender address	C	From: " = ?GB2312?B?wfXV8A = = ?= "< ovMZ2MS4w @123.com>
	Forwarded email	F	Received: from gcpcmail.gh-ca.com (unknown [61.129.69.28]) by mx10 (Coremail) with SMTP id wI1tfGEEjUMqUsk_.1
	…	…	…
Format anomaly	Null title	T	Subject:
	Html format	H	Content-Type: multipart/alternative; boundary="1101353164304.MimeBoundarY"
	Short message	S	–
	Hyperlink	L	–
	Attachment with Activex or javascript	A	–
	…	…	…

Table 2: The mapping relation between behaviours and messages fields

	D	G	C	F	T	H	S	L	A
From		√	√	√					
To		√	√	√					
Reply-To		√	√	√		√	√	√	√
Deliver-To		√	√	√					
Return-Path		√	√	√					
Received		√	√	√		√	√	√	√
Subject					√				
Data						√	√	√	√
Date	√								

MI behaviors feature selection algorithm is described as follows:

- Establish a message samples set.
- For each message m to create binary behavior feature vector $V_i = \{v_1, v_2..., v_N,\}$ where setting v_j as 1 represents that m contains the behavior feature f_j and otherwise setting v_j as 0.
- For each f_j to compute the mutual information value I.

$$I_{f_j}(X;Y) = \sum_{i=1}^{2}\sum_{j=1}^{2} p(x_i, y_j)\log\frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

Where, stochastic variable X and Y are defined as follows.

$$X = \begin{cases} x_1 \text{ denoting that a massage is ham probability } p(x_1) \\ x_2 \text{ denoting that a massage is spam probability } p(x_2) \end{cases}$$

$$Y = \begin{cases} y_1 \text{ denoting that a behaviour feature is in} \\ \text{probability with } p(y_1) \\ y_2 \text{ denoting that a behaviour feature is not} \\ \text{contained in a massege with probability with } p(y_2) \end{cases}$$

- According to all of the I_{fj} (X, Y) values sorted by descend order, pick out the first n behavior features to be mapped as a n-dimension feature vector.

By utilizing this algorithm, compared with content-based feature selection, the volume of the selected behaviors features would be much smaller. The smaller volume of features is able to improve the efficiency of training and classification.

SVM CLASSIFICATION MODEL

Based on SVM theory, email classification objective is about to solve a two-class quadratic optimization question which can be described as $f(x) = h(x)T\ \beta + \beta_0$. Subject to the bias and variance limitation, we need to achieve a fitted support vector discrimination boundary. The boundary should adapt to the classification model's online updating. An applicable classification model should be well-balanced between bias and variance,for the two factors might lead to over-fitting or under-fitting problem. Figure 2 shows the relationship among bias, variance and discrimination boundary.

Solution of the convex quadratic optimization question is described as expression (1)

$$\min_{\beta,\beta_0}\frac{1}{2}\|\beta\|^2 + \gamma\sum_{i=1}^{N}\varepsilon_i$$

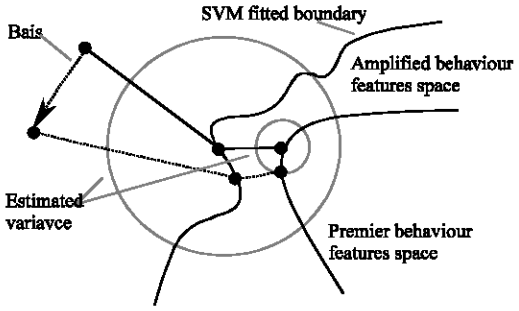

Fig. 2: Bias-variance influence on SVM fitted boundary

$$\text{Subject to } \varepsilon_i \geq 0,\ y_i(h(x_i)^T\beta + \beta_0 \geq 1 - \varepsilon_i\ \forall i \qquad (1)$$

Where:
ε_i = A relaxation factor,
γ = An adjustable parameter.

To ensure better testing error, we use normalized parameter γ. The normalized parameter can well balance bias and variance and ensure to achieve an optimum boundary.

By using basis function, we would obtain an expanded feature space. The classifier function can be formalized as follows:

$$\begin{aligned} G(x) &= \text{sign}[\hat{f}(X)] \\ &= \text{sign}[\sum_{i=1}^{N}\hat{a}_i y_i K(X, X_i) + \hat{\beta}_0] \end{aligned} \qquad (2)$$

In related SVM literatures (Duda *et al*., 2003), there are three kinds of popular kernel functions available: namely multinomial function, radial basis function and sigmoid function. All of them satisfy Mercer condition.

$$K(x, x') = (1 + < x, x' >)^d \qquad (3)$$

$$K(x, x') = \exp(-\|x - x'\|^2 / c) \qquad (4)$$

$$K(x, x') = \tanh(k_1 < x, x' > + k_2) \qquad (5)$$

We are going to make related performance tests by applying the three kernel functions into the classifier function, respectively. The testing results in next section suggest that the classifier who uses radial basis function has the best performance. Here, we use two sets of random 2D samples to testify the SVM classifier's performance preliminarily. After the considerate comparison and regulation, we choose normalized parameter $\gamma = 0.002$ and $\gamma = 0.006$ to have the test. From the Fig. 3, we found that the classification boundary is very smooth and not oscillating for the possible over-fitting.

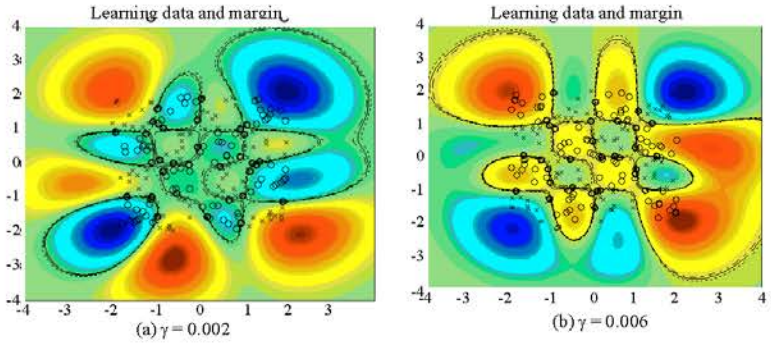

Fig. 3: 2D SVM discriminant boundary generated by RBF kernel

Table 3: Basic statistics for sample sets

Email set division	Training set	Evaluting set
Ham messages	963	645
SPAM messages	472	338
Total messages	1435	938

According to the testing result, the ratio of wrong classified points is 1.25% in Fig. 3a and 1.37% in Fig. 3b. The results basically prove that the discrimination boundary is precise and stable on 2D testing samples. As for standard email testing, we use a email testing set collected from an authentic mail server mci.uestc.edu.cn. The scale of the email set is shown as Table 3.

We use training set as initial samples set. After accomplishing the extraction of features matrix under the condition of pre-treatment, we put it into the spam classifier for training and then obtain a hyper-plane which supports vector in high dimensional space. The classifier maps the hyper-plane back to primitive space and forms an authentic discrimination boundary. In practice, the discrimination boundary trained by trusted behaviours vector is correspondingly stable but not static. Alone with the new emails arrived, the classifier's boundary should be self adapted. For satisfying this applicable requirement, we propose a boundary self-adjusting algorithm.

- Extracting new arrived email's behaviour features. Constructing new input vector and appending it into the initial vector set S. For each vector, let $\alpha_1 = 0$;
- Selecting b vector samples randomly from S to form a subset;
- Using G (β) to train vector set;
- Applying the model on all of the vector samples of S; If there exist samples which are not satisfied Karush-

Kuhn-Tucker condition, then using these samples to replace all of the samples in β and corresponded α_i; If it's not converged, then go back to step 3.

Table 4: Testing error comparisons under disturbing features input

	Testing error		
Classifier	No perturbation	3 Perturbations	8 Perturbations
SVM/polynomial Kernel	0.095	0.0990	0.102
SVM/RBF kernel	0.036	0.0365	0.037
SVM/sigmoid kernel	0.091	0.0940	0.093
BP/ANN	0.159	0.1620	0.185
Naïve Bayes	0.127	0.1430	0.191
Decision tree	0.068	0.7800	0.930

TESTING RESULTS AND ANALYSIS

The spam classifier's generalization performance is always involved with its predicting capacity for independent email testing samples. Testing error which is also called generation error is expected to predict error on independent testing samples (Zhang *et al*., 2004).

$$Err = E[L(Y,\hat{f}(X))] \quad (6)$$

Email filters based on machine learning are often concerned with its robust performance. For content related email filter, if disturbed words or sentence inserted, its classification performance would change evidently. But it would be no influence on behavior-based filter we proposed. So, we try to introduce some legitimate behaviors as disturbing factors to test the filter's robust performance. Table 4 shows that all SVM based filters outperform other filters. Among those SVM based filters, SVM/RBF kernel filter is the best one, for it is most stable under introduced disturbing behaviors. On the other hand, Naïve Bayes filter has the poorest performance under disturbing condition.

Aiming at two-class classification problem like spam category, using cross-validation approach (O'Brien and Carl Vogel, 2003) (data are shuffled and different parts are used for training and test in each iteration) could evaluate objectively how precise the spam filter is. Spam's misclassification problem can be divided into two types,

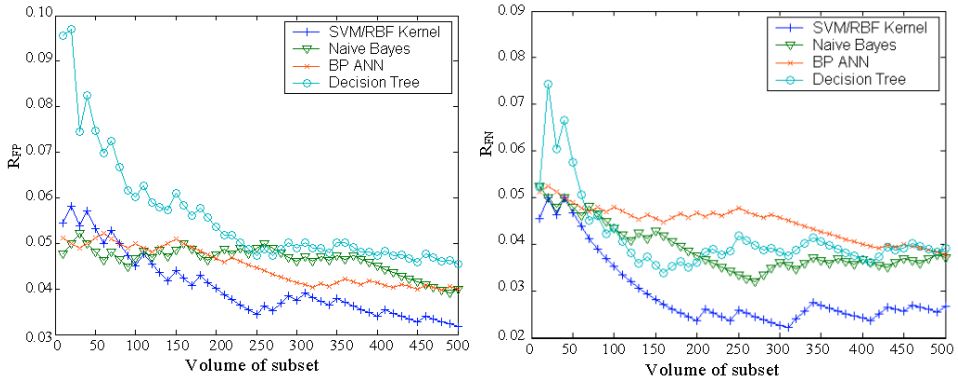

Fig. 4: Comparison for R_{FP} and R_{FN}

namely false positive and false negative. The former means the filter recognizes the ham as spam. The latter, on the contrary, means the filter recognizes the spam as ham. Here, we give the FP ratio and FN ratio's definitions (Androutsopoulos *et al*., 2000) as follows:

$$R_{FP} = \frac{n_{L\to S}}{N_S + n_{L\to S}} \quad (7)$$

$$R_{FN} = \frac{n_{S\to L}}{N_L + n_{S\to L}} \quad (8)$$

Where, N_L and N_S denote the volume of ham and spam, respectively. $n_{L\to s}$ and $n_{s\to L}$denote false positive emails and false negative emails separately. From Fig. 4, alone with the input testing samples increased, all of the filter's FP ratio and FN ratio is decreasing. SVM/RBF kernel filter's FP ratio and FN ratio is decreasing much faster than others which suggests it has comparative less misclassifications than other filters. BP ANN filter's FN ratio is less than FP ratio which suggests its performance is not well-balanced. Naïve Bayes filter's performance is stable and balanced both on FP ratio and FN ratio. But its precision is not as good as SVM/RBF kernel filter.

CONCLUSIONS

The main contributions we made in this paper can be concluded as follows:

- We investigated on spammers' behavior profoundly and utilized MI algorithm to extract trusted behavior features vector based on a behavior features library.
- We brought forward SVM-based spam filter with normalized parameters which ensure well balanced bias and variance. According to the comparison tests among several popular filters under authentic email testing set, we can draw the conclusion that new email filter has high precision in email category and is robust to disturbing data.

ACKNOWLEDGMENTS

We thank professor Tan Liang and doctor candidate Wang TieJun at Westone United Lab for their previous work on Chinese email testing set's collecting and organization. This work is supported by China national 863 subject foundation (Grant No. 2006AA01Z411).

REFERENCES

Androutsopoulos, I., J. Koutsias, V. Chandrinos, G. Paliouras and Spyropoulos, 2000. An evaluation of naive bayesian anti-spam filtering. In: Workshop on Machine Learning in the New Information Age, pp: 578-584.

Delany, S.J. and P.C. Ham, 2005. A case-based technique for tracking concept drift in spam filtering. Knowledge Based Syst., 4: 187-195.

Duda, R.O., P.E. Hart and D.G. Stork, 2003. Pattern Classification. 2nd Edn. China Machines Press.

O'Brien, C. and C. Vogel, 2003. Spam filters: Bayes vs. chi-squared; letters vs. words. Proc. Series-Proceeding-Section-Article, pp: 291-296.

Pelletier, L., J. Almhana and V. Choulakian, 2004. Adaptive filtering of spam, Proceeding of the second annual conference on communication networks and Service Research (CNSR'04), pp: 218-224.

Ying, C.G., 2005. http://www.ccert.edu.cn/spam/sa/Chinese _rules.htm.

Zhang, L., J. Zhou and T. Yao, 2004. An evaluation of statistical spam filtering techniques. ACM Trans. Asia Language Inform. Process, 4: 243-269.

Zhen, L., S. Kun and Zhou Mingtian, 2005. Research on advanced filtering algorithm for spam email based on bayes parameter estimation. J. Comput. Sci., 9: 55-58.