Artificial Immune System Based Methods for Spam Filtering

Ying Tan, Guyue Mi, Yuanchun Zhu, and Chao Deng
Key Laboratory of Machine Perception (Ministry of Education)
Department of Machine Intelligence, School of Electronics Engineering and Computer Science
Peking University, China
Email: {ytan,gymi,ychzhu}@pku.edu.cn, chao667788@yahoo.com.cn

***Abstract*—To solve the spam problem, many statistical learning methods and AIS methods have been proposed and applied. In essence, statistical learning methods and AIS methods have quite different origins, and they try to find the solutions from distinct aspects. In recent works, we proposed several hybrid methods, which combined immune theory with statistical methods in spam filtering. In this paper, we briefly review and analyze these works and possible extensions, and demonstrate the rationality of building hybrid immune models for spam filtering. In addition, a generic framework of an immune based model is presented, and online implementation strategies are given. It is well demonstrated that how to apply the immune based model to building an intelligent email server.**

I. Introduction

Artificial Immune System (AIS) is an inter-discipline research area [1] that aims to build computational intelligence models [2], [3] by taking inspiration from Biological Immune System (BIS). BIS is an adaptive natural system [4], which possesses several interesting properties [5], [6] such as distributed detection, noise tolerance, and reinforcement learning. It can detect and react to invading pathogens based on signals and interaction among immune cells. Taking immune processes as good metaphors, many AIS models [4] have been proposed to solve engineering problems. Some prevalent ones are negative selection, clonal selection, immune network model, and danger theory algorithm. These models have been applied to a number of real-world problems [7]–[9] such as pattern recognition, data mining, spam filtering [10], and computer security [11].

Spam filtering is an important and typical pattern recognition problem, as spam cause many problems to our daily-communication life [12], [13]. In solving the problem, both classical statistical methods and AIS methods have been presented, and most of them focus on studying feature extraction methods and design of classifiers. The main function of feature extraction is to extract discriminative information from messages, and transform messages into feature vectors. The statistical feature extraction methods try to collect and analyze numerical characteristics of messages, such as term frequencies, and relation between terms and email categories. Some prevalent ones are Bag-of-Words (BoW) [14], Sparse Binary Polynomial Hashing (SBPH), and Orthogonal Sparse Bigrams (OSB) [15]. Different from the statistical ones, the AIS methods [16] construct feature vectors by mimicking the process of antibody creation in BIS. In design of classifiers, classical pattern recognition methods, e.g. Naive Bayes(NB) [17], [18], Support Vector Machine (SVM) [19], k-Nearest Neighbor (k-NN) [20], [21], and Artificial Neural Network (ANN) [22], [23] were proposed on the basis of statistical theory. On the contrast, AIS models were inspired by natural functions and mechanisms of BIS [16], [24].

These statistical methods and AIS ones are quite different in terms of both origins and principles, which endow them with quite distinct properties. Combining the strength of statistical approaches with the AIS ones may help achieve better performance. In this paper, we introduce and discuss several recent works of our laboratory [10], [25]–[30], which applied mixed principles to feature attraction, classifier combination, and classifier updating, so as to demonstrate the rationality of combining statistical and AIS methods for spam filtering. In addition, we present a generic framework of an immune based model for spam filtering, and online implementation strategies are given to demonstrate how to build an immune based intelligent email server.

The remainder of the paper is organized as follows. A generic framework of the immune based model is proposed in Section II. In the model, concentration based feature vectors are extracted from messages, immune signals are utilized for combining classifiers, and dynamic immune mechanisms are utilized for updating classifiers. In section III, we present possible extensions of the immune model. The conclusion and some future directions are given in Section IV.

II. Immune Based Model for Spam Filtering

A. The Framework of the Immune Based Model

There exist many explanations about the mechanisms of BIS. An explanation may be superior for analyzing some specific immune phenomena, but less persuasive for some other aspects. For AIS practitioners, it is not quite necessary to find which theory is better in explaining immune mechanisms. What matters most is the heuristic principles behind these explanations. By taking inspiration from two prevalent immune theory, we proposed several immune based spam filtering methods [10], [25]–[30], in which statistical information is also well considered.

According to Self-Non-Self (SNS) theory [31], two typical immune processes, namely the primary response and the

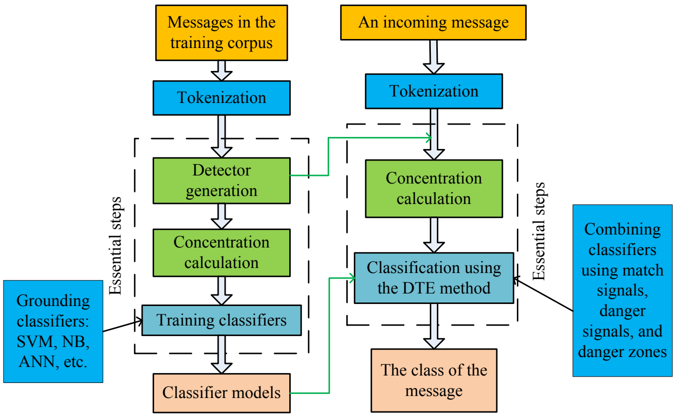
978-1-4673-5762-3/13/$31.00 ©2013 IEEE

secondary response, play important roles in the BIS. The primary response occurs when a type of pathogen appears in the body for the first time. As the BIS is not familiar with the pathogen, the antibodies with affinity to the pathogen are produced slowly. However, when the same pathogen is encountered next time, the secondary response is aroused, and the concentration of relevant antibodies increases rapidly. It is worth of noticing that the concentration of antibodies is a key point in recognizing the corresponding pathogen. A high concentration of antibodies reflects that the BIS detects the pathogen with high confidence. From another perspective, concentrations of antibodies characterize the detection of pathogens in BIS. Based on this observation, we proposed immune concentration based feature extraction methods for spam filtering, namely global concentration based methods [26], [27] and local concentration based methods [29], [30].

Danger Theory (DT), proposed by Matzinger [32], is a novel biological paradigm for explaining mechanisms of BIS. According to DT, the cells of the body interact with each other through match signals, danger signals, and danger zones. In BIS, antibodies bind to (match) antigens when the affinity between them is higher than a certain threshold. However, the antigens would not be culled from the body until the antibodies receive a danger signal from distressed cells. Danger signals indicate that the distressed cells are infected to death, and the cells release the signals to antibodies nearby (within the danger zone) just to activate immune response. Thus, the danger signals can be regarded as a confirmation of the match signals between antibodies and antigens. The interaction using the signals ensures the robustness of BIS. We find that the signals is quite helpful in combining classifiers, and design an DT based Ensemble (DTE) method [10] in the classification phase of spam filtering.

Besides detecting antigens, dynamics of immune cells is also one of the most important properties possessed by BIS. Antibodies can evolve to recognize emerging antigens, and the recognition memory will be preserved to detect antigens more effectively next time. In addition, there are some ways in measuring the importance of antibodies, such as lifespan and weights. The dynamic change of lifespan and weights ensures that the existing antibodies give the best protection to the body. Mimicking the mechanism, we presented dynamic strategies for classifier updating in [25], [28].

Based on these previous works, we present a generic framework of an immune based spam filtering model, as depicted in Fig. 1. According to the model, concentration based feature vectors are extracted from messages by computing match concentration of detections. Classifiers are then built on the concentration vectors of training corpus. Finally, incoming messages can be classified by using the DTE method. In addition, classifiers are updated at all times based on the drift of messages and classification performance. In the following subsections, we briefly introduce and discuss the principles of these methods, and analyze the rationality of combining statistical principles with AIS ones.


Messages in the training corpus
Tokenization
Detector generation
Concentration calculation
Training classifiers
Classifier models
Essential steps
Grounding classifiers: SVM, NB, ANN, etc.
An incoming message
Tokenization
Concentration calculation
Classification using the DTE method
The class of the message
Essential steps
Combining classifiers using match signals, danger signals, and danger zones

(a) Training phase of the model

(b) Classification phase of the model

Fig. 1. Training and classification phases of the immune based model

B. Concentration Based Feature Extraction Method

The essence of the feature extraction method lies in the construction of concentration feature vectors. In [26], [27], we presented Global Concentration (GC) based feature extraction methods for spam filtering. In [29], [30], Local Concentration (LC) based feature extraction methods were proposed. In these methods, statistical term selection methods [33] are utilized to remove uninformative terms. Then a tendency function are well designed to generate two detector sets [29], [30]. The tendency of a term t_i is defined in Eq. 1. $T(t_i)$ measures the difference between the term's occurrence frequency in two types of messages. Terms are added to corresponding detector sets according to their tendency. Detector concentration, which corresponds to antibody concentration in BIS, are then extracted from messages by using the detector sets. In addition, a sliding window is utilized to slide over a message to extract position-correlated information from messages. By using a sliding window, a message is divided into local parts. At each movement of the window, a spam detector concentration S_i and a legitimate detector concentration L_i are calculated with respect to the two detector sets and the terms in the window according to Eqs. 2 and 3.

$$T(t_i) = P(t_i|c_l) - P(t_i|c_s), \tag{1}$$

where $P(t_i|c_l)$ denotes the probability of t_i's occurrence, given messages are legitimate emails, and $P(t_i|c_s)$ denotes the probability of t_i's occurrence estimated in spam.

$$S_i = \frac{\sum_{j=1}^{w_n} M(t_j, D_s)}{N_t}, \tag{2}$$

$$L_i = \frac{\sum_{j=1}^{w_n} M(t_j, D_l)}{N_t}, \tag{3}$$

where N_t is the number of distinct terms in the window, D_s denotes the spam detector set, D_l denotes the legitimate email detector set, and $M()$ denotes the match function, which measures the number of terms in the window matched by detectors.

Each sliding window defines a specific local area in a message. To explore the effects of a sliding window, we design two strategies—using a sliding window with fixed-length (FL) and using a sliding window with variable-length (VL). When a fixed-length sliding window is utilized, messages may have different number of local areas (corresponding to different number of feature dimensionality), as messages vary in length. To handle this problem, we may either expand a short message by reproduce the existing features, or reduce the dimensionality of long messages by discarding uninformative features. In VL strategy, the length of a sliding window is designed to be proportional to the length of a message, and there is no need for specific process of feature dimensionality. Preliminary experiments showed that both the two strategies are effective in extracting discriminative features. In the circumstance that the size of a window is set to infinite, a message is taken as a whole for getting concentration features, GC feature vectors are extracted. When the window size is smaller than the message length, the window divide a message into individual local parts, and LC features are extracted from each window.

Experiments were conducted on real-word corpora Ling, PU1, PU2, PU3, PUA, and Enron-Spam[1] using cross validation to investigate the performance of the concentration based method. Meanwhile, four benchmark criteria, namely spam precision, spam recall, accuracy, and F_1 measure were adopted in analyzing the results. Among them, accuracy and F_1 were more important as they indicated the overall performance of approaches. From these experimental results, it can be seen that the combination of statistical information and immune characteristics helps achieve the best discriminative performance. The success lies in the following aspects: 1) By using term selection methods, noise and uninformative terms can be removed, which reduce computational complexity and enhance effectiveness of detectors. 2) The concentration principle helps obtain feature vectors with lower dimensionality. 3) The sliding window strategies provide effective ways of defining local area in messages, and extracting position-correlated information.

C. Danger Theory Inspired Ensemble Method

Mimicking the DT theory, we defined artificial signals and danger zones, and classifiers were combined using them [10]. First, two types of artificial signals, namely, signal I (match signals) and danger signals, were respectively generated using two independent classifiers. Depending on the classification results, negative or positive signals would be generated. After the production of the signals, the two classifiers were interacted through the transmission of the signals. Mimicking the DT mechanism, the transmission of the signals was designed to be different. An activated signal I would be sent only to the specific sample, upon which the signal was arisen. However, an activated danger signal would be sent to all the test samples within the danger zone, besides the specific sample. Finally, the result was acquired based on the interaction among classifiers.

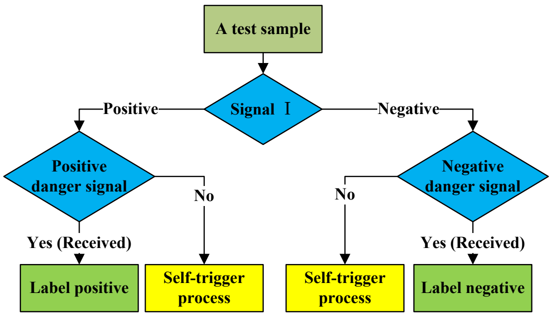


Fig. 2. The framework of the DTE method

The framework of the DT based Ensemble (DTE) method is depicted in Fig. 2. A test sample gets labeled by the first two classifiers if the two signals agree with each other. Otherwise, a third classifier (self-trigger process) is utilized to solve the conflict and get the test sample classified. According to the method, three classifiers are combined in a cascade way. Similar to other cascade method, the order of classifiers can be determined according to classifier performance on training corpus. The characteristics of the DTE method lie in the interaction among classifiers by using the danger zone and the signals.

The interaction between the first two classifiers is expressed as Eq. 4.

$$E(x_i) = \sum_{x_j \in D} \delta(c_1(x_i), c_2(x_j)) K(d(x_i, x_j)), \tag{4}$$

where x_i and x_j are test samples, D denotes the test set, $c_1(x)$ and $c_2(x)$ are the two classifiers, $d(x_i, x_j) = \| x_i - x_j \|$ is the distance between two samples, $K(z)$ is defined in Eq. 5, and $\delta(y_1, y_2) = 1$, if $y_1 = y_2$, and 0 otherwise.

$K(z)$ defines the effect of the danger zone as follows.

$$K(z) = \begin{cases} 1 & if\ z \leqslant \theta \\ 0 & otherwise \end{cases}, \tag{5}$$

where θ is the size of the danger zone.

After obtaining the weighted result $E(x_i)$, the sample x_i can get its class label using Eq. 6.

$$L(x_i) = \begin{cases} c_1(x_i) & if\ E(x_i) \geqslant 1 \\ f(x_i) & otherwise \end{cases}, \tag{6}$$

where $f(x)$ denotes the class label given by the third classifier.

The performance of the DTE was investigated on four real-world corpora, namely PU1, PU2, PU3, and PUA using 10-fold cross validation. In the experiments, SVM, NB, and Nearest Neighbor (NN) were utilized as three grounding classifiers. SVM was utilized to generate match signal, NB was utilized to generate danger signal, and NN was utilized in the

[1] The PU corpora and Enron-Spam are available from the web site: http://www.aueb.gr/users/ion/publications.html.

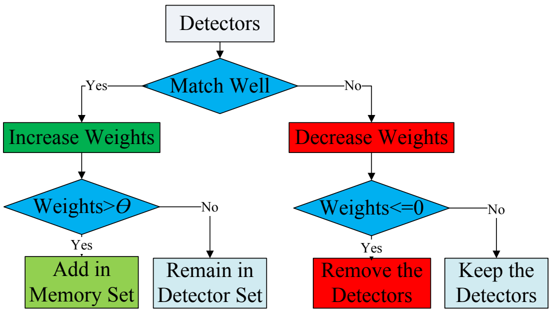


Fig. 3. Updating SVs according to their performance

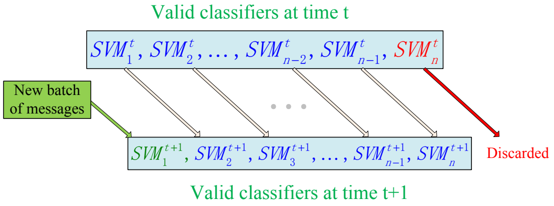

Valid classifiers at time t
$SVM_1^t, SVM_2^t, \ldots, SVM_{n-2}^t, SVM_{n-1}^t, SVM_n^t$
New batch of messages
$SVM_1^{t+1}, SVM_2^{t+1}, SVM_3^{t+1}, \ldots, SVM_{n-1}^{t+1}, SVM_n^{t+1}$
Discarded
Valid classifiers at time t+1

Fig. 4. Updating SVMs with time according to their lifespan

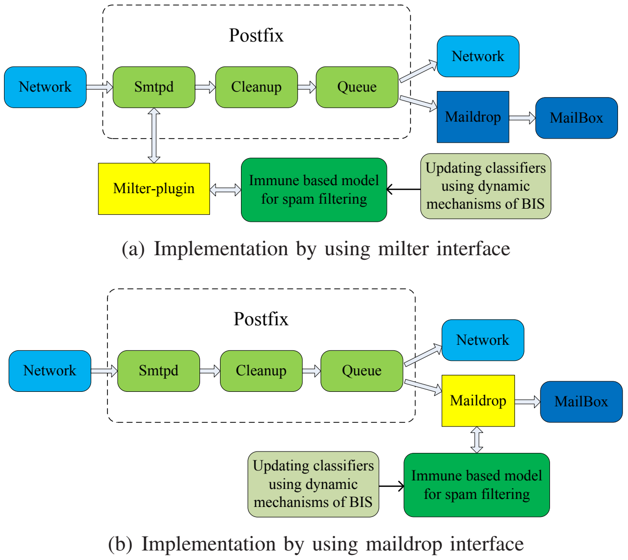
self-trigger process. The experimental results of [10] show that the danger zone provides a well defined interaction between the two types of signals, and classifier are combined through the interaction. By using the DTE method, the performance of classifiers can be effectively improved.

D. Immune Based Dynamic Updating Strategies

Mimicking dynamic mechanisms of BIS, we proposed several classifier updating strategies in [25], [28]. The updating process of SVMs is depicted in Fig. 3 and Fig. 4. Support vectors (SVs) of a SVM are used as detectors (antibodies) and SVs are updated according to their performance by mimicking the dynamic mechanisms of BIS. In measuring the importance of SVs, we assign weights to SVs, and build up two sets, a Detector Set and a Memory Set. According to [25], [28], the weight is increased when a SV correctly classify a sample (according to hamming distance), and vice versa. When the weight of a SV is above a pre-defined threshold, the SV will be add to the memory set and the weight will be increased significantly. On the contrary, when the weight of a SV is decreased to zero, the SV will be culled from the detector set. In addition to SVs, the whole SVM is also updated with time. The updating of a SVM is in a greater magnitude as most of the SVs will be changed in this process. In the process, a sliding window strategy is adopted, and the window size controls the lifespan of SVMs. When the updating moment is arrived, the oldest SVM is discarded and a new SVM will be built using the new arrival messages. The final classification decision will be made by the majority voting of the SVMs in effect.

E. Online Spam Filtering Implementation

In real-world application, we use Postfix as email server, and apply the immune based model as a plug-in unit for spam filtering. Two strategies can be adopted, either by using the milter interface or the maildrop interface. Figs. 5(a) and 5(b) respectively depict the two strategies. The immune based model is implemented as an intelligent filter in the processing emails. Besides, immune based dynamic updating strategies are adopted in order to build an adaptive system.


Postfix
Network
Smtpd
Cleanup
Queue
Network
Maildrop
MailBox
Milter-plugin
Immune based model for spam filtering
Updating classifiers using dynamic mechanisms of BIS
Postfix
Network
Smtpd
Cleanup
Queue
Network
Maildrop
MailBox
Updating classifiers using dynamic mechanisms of BIS
Immune based model for spam filtering

(a) Implementation by using milter interface

(b) Implementation by using maildrop interface

Fig. 5. Implementation of hybrid intelligent methods on Postfix server

III. Discussion

The immune based model is built by borrowing some ideas from mechanisms of BIS. The effect of the model is not limited by the implementation details. Thus, it is natural and easy to extend the model using different implementation strategies.

The essence of both LC and GC methods lies in the mechanism of concentration. It is concentration that endows the model with noise tolerance and robust properties. Besides characteristic terms, other attributes can also be taken as elements for calculating concentration. For instance, binary string or regular expression can characterize a message well. Thus, it is rational to use the concentration of them as messages' features. In classification phase, other classifiers, e.g. NB, ANN, can be applied instead of SVM. The possible extensions may help us learn the mechanisms of the concentration method better.

In the classification phase, match signals, danger signals and danger zones play important roles. A match signal indicates a primary recognition of the message type, and a danger signal is a confirmation to the match signal. A danger zone defines a way of utilizing neighborhood information. In extending the model, more danger signals can be brought in to define a cascade way of combining multiple classifiers. In the model, each classifier is confirmed by a subsequent classifier. As more classifiers are added in, the performance of the model may be further improved.

The updating mechanisms in II-D can be adopted to other classifiers, such as NB and ANN. Taking ANN for another example, we can gradually update the weights of neurons and update the structure of the ANN when the current one becomes overdue. Preliminary results show that the dynamic mechanisms are effective in catching the change of messages.

IV. Conclusion

In this paper, we briefly introduce our recent advances in immune based spam filtering methods, and put emphasis on combining immune theory with statistical methods. It is shown that combining immune ideas with classical statistical methods can effectively improve the performance of a spam filter. In addition, we present a framework of an immune based spam filtering model, which demonstrate how to utilize these methods in real-world applications. In the model, immune mechanisms are brought in different phases of spam filtering model. First, concentration concept is utilized for extracting feature vectors from messages, and it is demonstrated that the concentration method is more robust and accurate than the prevalent BoW method. Mechanisms of DT are then shown to be effective in combining classifiers. Besides, dynamic mechanisms of BIS are adopted to updating classifiers of the spam filter. Finally, implementation strategies of an immune based intelligent email server are also given. In future works, we seek to extend the model and apply the model to other pattern recognition problems. The research of immune based approaches is still young and very promising. Both the efforts in immune theory and statistical methods will facilitate its development.

Acknowledgment

This work was supported by National Natural Science Foundation of China (NSFC), under Grant No. 61170057 and 60875080, and partially supported by the National High Technology Research and Development Program of China (863 Program), with Grant No. 2007AA01Z453.

References

[1] J. Timmis, P. Andrew, N. Owens, and E. Clark, "An interdisciplinary perspective on artifical immune systems," *Evol. Intel.*, pp. 5–26, 2008.

[2] I.R. Cohen, "Real and artificial immune systems: computing the state of the body," *Imm. Rev.*, pp. 569–574, 2007.

[3] A. Freitas and J. Timmis, "Revisiting the foundations of artificial immune systems for data mining," *IEEE Transactions on Evolutionary Computation,* vol. 11, no. 4, pp. 521–540, 2007.

[4] D. Dasgupta, "Advances in artificial immune systems," *IEEE Comutational Intelligence Magazine*, pp. 40–49, 2006

[5] Leandro Nunes de Castro and Fernando José Von Zuben, *Artificial Immune System: Part I—Basic Theory and Applications*, Tech. Rep. TR-DCA 01/99, Dec. 1999.

[6] J. Timmis, "Artificial immune systems—today and tomorrow," *Nat. Comput.*, pp. 1–18, 2007.

[7] L. N. de Castro, J. Timmis, "An artificial immune network for multimodal function optimization," In: *Proc. 2002 World Congress on Computational Intelligence (WCCI2002)*, 2002, pp. 699–704.

[8] J. Hunt, J. Timmis, D. Cooke, M. Neal, and C. King, "Jisys:Development of an artificial immune system for real world applications", In: *Artificial Immune Systems and Their Applications*, 1998, pp. 157–186.

[9] U. Aickelin, S. Cayzer, "The danger theory and its application to artificial immune system,", In: *Proc. First International Conference on Artificial Immune Systems*, 2002, pp. 141–148.

[10] Y. Zhu and Y. Tan, "A danger theory inspired learning model and its application to spam detection", In: *Proc. International Conference on Swarm Intelligence*, 2011, pp. 382–389.

[11] S. Forrest, A. Perelson, L. Allen, and R. Cherukuri, "Self-nonself discrimination in a computer," In: *Proc. IEEE Symposium on Research Security and Privacy*, 1994, pp. 202–212.

[12] Commtouch, Internet Threats Trend Report. Jul. 2011 [Online]. Available: http://www.commtouch.com

[13] Sophos, Sophos Security Threat Report 2011. Jul. 2011 [Online]. Available: http://www.sophos.com

[14] I. Androutsopoulos, G. Paliouras, and E. Michelakis, "Learning to filter unsolicited commercial e-mail", NCSR "Demokritos" Tech. Rep. No. 2004/2, minor corrections: October 2006.

[15] C. Siefkes, F. Assis, S. Chhabra, and W.S. Yerazunis, "Combining winnow and orthogonal sparse bigrams for incremental spam filtering", in *Lecture Notes in Computer Science*, vol. 3202/2004, 2004, pp. 410–421.

[16] T. Oda and T. White, "Developing an immunity to spam", in *Lecture Notes in Computer Science (LNCS)*, 2003, pp. 231–242.

[17] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A baysian approach to filtering junk e-mail", AAAI Technical Report WS-98-05, 1998, pp. 55–62.

[18] A. Ciltik and T. Gungor. "Time-efficient spam e-mail filtering using n-gram models," *Pattern Recogn. Lett.*, vol. 29, no. 1, pp. 19–33, 2008.

[19] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," *IEEE Trans. Neural Netw.*, vol. 10, pp. 1048–1054, 1999.

[20] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach," in *Proc. the workshop "Machine Learning and Textual Information Access", 4th European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD' 00)*, 2000, pp. 1–13.

[21] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos, and P. Stamatopoulos, "A memory-based approach to anti-spam filtering for mailing lists", *Inform. Retrieval*, vol. 6, no. 1, pp. 49–73, 2003.

[22] J. Clark, I. Koprinska, and J. Poon, "A neural network based approach to automated e-mail classification", in *Proc. IEEE Int. Conf. Web Intelligence (WI' 03)*, Halifax, Canada, 2003, pp. 702–705.

[23] C.-H. Wu, "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks", *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4321–4330, April 2009.

[24] T.S. Guzella, T.A. Mota-Santos, J.Q. Uchôa, and W.M. Caminhas, "Identification of spam messages using an approach inspired on the immune system", *Biosystems*, vol. 92, no. 3, pp. 215–225, June 2008.

[25] G. Ruan and Y. Tan, "Intelligent detection approaches for Spam," In: *Proc. Third International Conference on Natural Computation (ICNC2007)*, 2007, pp. 672–676.

[26] Y. Tan, C. Deng, and G. Ruan, "Concentration based feature construction approach for spam detection", In: *Proc. International Joint Conference on Nearal Networks (IJCNN2009)*, 2009, pp. 1344–1350.

[27] G. Ruan and Y. Tan, "A three-layer back-propagation neural network for spam detection using artificial immune concentration," *Soft Comput.*, vol. 14, pp. 139–150, 2010.

[28] G. Ruan and Y. Tan, "Uninterrupted approaches for spam detection based on SVM and AIS", *IEEE Transactions on Systems, Man, and Cybernetics—PartB: Cybernetics* (In revision)

[29] Y. Zhu and Y. Tan, "Extracting discriminative information from E-mail for spam detection inspired by immune system", In: *Proc. IEEE Congress on Evolutionary Computation (CEC2010)*, 2010, pp. 2491–2497.

[30] Y. Zhu and Y. Tan, " A local concentration based feature extraction approach for spam filtering," *IEEE Transactions on Information Forensics and Security*, pp. 486–497, 2011.

[31] J. Timmis, M. Neal, and J. Hunt, "An artificial immune system for data analysis," Biosystems, vol. 55, pp. 143–150, 2000.

[32] P. Matzinger, "The danger model: a renewed sense of self," Science's STKE, vol. 296, no. 5566, pp. 301–305, 2002.

[33] Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", in*Proc. Int. Conf. Machine Learning (ICML'97)*, 1997, pp. 412–420.