SPAM DETECTION WITH GENETIC OPTIMIZED ARTIFICIAL IMMUNE SYSTEM

ALIREZA MEHRSINA

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Computer Science (Information Security)

Faculty of Computer Science and Information Systems
UniversitiTeknologi Malaysia

JANUARY 2013

This project report is dedicated to my family for their endless support and encouragement.

ABSTRACT

Spam has become one of the most serious universal problems, which causes problems for almost all computer users. These problems such as lost productivity, wasting user's time and occupying network bandwidth, causes a big problem for companies and organizations. This study presents a hybrid machine learning approach inspired by the Artificial Immune System (AIS), and Genetic algorithm for effectively detect the Spams. The Clonal Selection Algorithm (CLONALG) is one of the famous implementations of the AIS, which is inspired by the clonal selection theory of acquired immunity, which has shown success on broad range of engineering problem domains. This algorithm is quietly similar to Genetic Algorithm in terms of architecture and behavior. In this study, Comparisons are drawn with AIS and GA-AIS classifiers and it is shown that the proposed system performs better results than the original AIS.

ABSTRAK

Spam telahmenjadisalahsatumasalah yang paling seriussejagat, yang menyebabkanmasalahbagihampirsemuapenggunakomputer. Masalah-masalahsepertikehilanganproduktiviti, membuangmasapenggunadan bandwidth rangkaianpenjajah, menyebabkanmasalahbesarbagisyarikat-syarikatdanorganisasi.Kerjainimembentangkanpembelajaranmesin hybrid pendekatan yang diilhamkanolehSistemBuatanimun (AIS), danalgoritmagenetikuntukberkesanmengesan spams. AlgoritmaPemilihanklon (CLONALG) adalahsalahsatudaripadapelaksanaanterkenal AIS, yang diilhamkanolehpemilihanklonteoriimunitidiperolehi, yang telahmenunjukkankejayaan di pelbagai domain masalahkejuruteraan.Algoritmainiadalahsecarasenyap-senyapserupadenganAlgoritmaGenetikdarisegisenibinadantingkahlaku.Dalamkaryain iPerbandingandilukisdengan AIS dan GA-Pengelas AIS daniamenunjukkanbahawasistem yang dicadangkanmelakukanlebihbaikkeputusandaripada AIS asal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background of the Problem

Recently, almost all the Internet users communicate through a common and important medium, which is E-mails. Because of the globalization fact in the Internet, the time and distance of communication are not considered as limitations by E-mails. Also the conveniences of E-mail are more and more concerned by the users [1]. Since E-mails are simple, cheap, and a fast type of communicating for almost all computer users, they can be vulnerable of so many attacks and threats. One of the most important threats for the E-mail systems is Spam. This is one of the problems that almost everybody has experience of receiving it. The word “Spam” usually denotes the variety of E-mails that annoy Internet users to receive them. The spam e-mails, are sent to receivers’ mailbox without their permission, especially the large amount sending and annoying E-mails. It is now possible to send hundreds of Spam messages to thousands of users around the world with approximately no cost. Because connecting to the Internet is cheap and criminals have access to potentially several infected computers so that they are able to send their Spam E-mails form the legitimate sources [2]. Since nowadays we see that it is very common that the users around the world are seeing their Inbox have become full of these unwanted messages everyday [3].

Receiving these kinds of messages will cause so many problems for the user in terms of technical and personal views. These spam emails not only consume users' time and energy to identify and remove the undesired messages, but also cause many problems such as taking up limited mailbox space, wasting network bandwidth and engulfing important personal emails that needs to take more attention. Another problem caused by Spam E-mails is that they can easily spread malicious contents specially viruses through their propagation. The problems is because of the mass of unwanted bulk E-mail messages, propagating throughout the Internet, which is familiar to practically every e-mail user, whose mailboxes are filled by these messages daily. Moreover to the time spent for their removal, network bandwidth is wasted for its delivery. Some studies have been proposed the effects of Spam E-mails in the users experience with the Internet. A study presented that over 70% of today's business E-mails are spam. Spam E-mails are significantly different in content and most of times they belong to the following categories: money making scams, fat loss, improve business, sexually explicit, make friends, service provider advertisement, etc, [4].

Spam levels have been increased in many places of the world. Another study showed that in some countries spam accounted for 85% of email traffic in 2007. From the economic view a worldwide cost estimate of Spam for lost productivity and IT infrastructure investment was over 10 billion dollars in 2005. Solutions for solving this problem is described in law and technical point of view. Some countries have recently work on their law in this area, But there are some technical approaches remain essential. Since Spam are sent from anywhere in the world virtually, and tracking the actual sender of messages is so difficult [6]. Spam is incredibly annoying, especially in large quantities. If you have a public e-mail address you can receive hundreds of spam messages for every legitimate message that arrives. Even with good filters, some of the spam makes it through. And filters can sometimes delete messages that you really do want to receive. So as to facing this serious problem, Anti-Spam techniques by determining whether or not an incoming E-mails are Spam, has become an important issue. Many different approaches have been proposed for automatically detecting or filtering Spam emails. Many techniques such as [2] rely on building a database for blocking Spam E-mails whose addresses have

been reported as blacklists. Another approach is to looking into the message bodies for specific words or phrases, which are threatening terms. Among to the other techniques, many researchers proposed some machine-learning based approaches for context investigation. These machine-learning based techniques produce rules or models with weighted scoring about the positions, frequencies and context associations of terms or phrases used in spam and estimate the likelihood that an incoming email is spam or legitimate E-mail accordingly. Techniques based on these content filtering, or keyword-based filtering, are effective, if keywords are explicitly given. However, Spammers usually trying to make their messages undetectable from legitimate email as possible and change the patterns of spam to foil the filters. Some spams are customized by some programs to make them look like normal messages, which may not contain any specific keywords. From the point of view of machine-learning, the key to success of applying machine-learning based methods is the correctness of features which warded by illegal permission or accounts, delivered with a series of the same message repeatedly and unauthorized too many different recipients, and so on.

Using specific keywords is only a class of these behaviors. Although, spam emails are changing their forms, human beings can easily recognize them no matter how they are generated (for example, image spam) and distributed. Spam filtering which claims that such behaviors can be used for identifying spams since they have better resistance with respect to the change of time. Emails to be investigated by the neural network are described in terms of their spamming behaviors, not keywords them contains. The spamming behaviors of emails are first identified by a rule-based pre-processor. Next, the identified features are encoded as three-valued vectors and processed by the proposed neural network. Since spamming behaviors change statistically, in comparison with the changing frequency of keywords, so that classification of spams using behavior-based features may be more robust than keyword-based methods. Experimental results show that spam classification using behavior-based features is more robust [7].

Also AIS is relatively a new approach and it is very interesting to understand of biological models and mechanisms. The view of AIS to the problem of Spam is based on similarities between microorganisms. In comparison to the biology, Spam messages are also in evolution. This evolution is taking apart through different features. Apart from the evolution concept, Spams can be identified by their content similar to what we have in biology, which is called “Pattern Matching”.

1.2 Statement of the Problem

There are many techniques, which will try to stop or reduce the amount of Spam E-mails. Some techniques are based on using network information and IP addresses so as to detect if the message is Spam or not. Another approach is to filtering E-mails is content filtering and characteristics of the message itself. Although there are so many techniques proposed to filter the Spam filters, we are facing with the large number of Spams everyday [7]. With the use of content-based filtering, Spammers also have employed new tools to overcome and pass these filters. One of the tools that the spammers using is to obscuring the text in their own format that are very common in Spam messages for example the word “F r 3 3” instead of “free”. So as to prevent these problems, Machine-learning approaches have been proposed. These techniques provide automated and adaptive approaches. Machine-learning techniques are capable of classification of the new message by means of extracting knowledge from the previous reviewed messages. Moreover to the previous facts Spam filters needs to consider to the user feedback too. This feedback will help the text processing techniques.

The introduction of new techniques in Spam filtering such as Artificial Neural Networks (ANN) and Artificial Immune System (AIS) and Genetic Algorithms (GA) can now get benefit form the Machine-learning's capabilities and they can now play an important role in the fights against the Spams. Artificial Immune System (AIS), which is new paradigm, is classified as a Knowledge-Based

technique that use Machine-learning concept. There are some problems with the current techniques, which are the speed, accuracy and optimizations.

1.3 Research Questions

We review the possible questions that will be answered through the next chapters:

1. Where do the Spams come form?
2. Who are the senders of the Spam messages?
3. What are the objectives of the Spam messages?
4. How do the Spam filters will identify Spams from Legitimate messages?
5. How much are the Spam filtering methods accurate?
6. What is the difference between AIS and other Spam detections methods?
7. How is learning procedure in AIS Spam filtering?

1.4 Purpose of the Researcher

The purpose of the this research is to propose a optimum technique which can detect Spam messages automatically and also will learn from the past behaviors to optimize the detection procedure. Due to the fact that many Spam detection methods fail to detect messages that have already spammed, because of the learning problems. After the learning procedure, the classification part will be more accurate than we have in other techniques. We are going to use AIS in Spam detection and modify its parameters and use GA in order to optimize it and get more accurate results.

1.5 Objectives

This study will achieve some objectives at the end, which will result in proposing an efficient and hybrid method to detect Spams. We are going to achieve these objectives in this thesis:

1. Mapping AIS parameters to Spam features
2. Implementation of AIS to detect Spams
3. Design and Implementation of Hybrid AIS with the Genetic Algorithm
4. Test and validate the proposed approach

1.6 Research Scope and Assumptions

This research is limited to one of the machine learning approaches called Artificial Immune system and the data used in tis research (http://csmining.org/index.php/enron-spam-datasets.html) has been used by so many researches in the domain of spam detection by machine learning approaches.

1.7 Importance of the Research

Due to several problems that caused by Spam messages nowadays, Spam filtering has become more and more important and more researchers are interested to work on it. The economical losses and technical problems that have been caused by these unsolicited bulk E-mail messages, will make the countries to find a good solution to stop them. On the other hand Spammers are always trying to find new effective ways to break the contraindications and reach their goals from sending these Spam messages. Because E-mails are the most convenient and reliable medium for communicating through all over the world considering its simplicity and

cheapness, it seems that we need to provide its reliability by fighting against Spammers. And it should be considered that security experts should always be one step forward than the Spammers as they find new approaches simply. Based on comparison of different methods results, it is obvious that the technique including the AIS which has been hybridized by the Genetic Algorithm is the best one.

1.8 Organization of the Thesis

This thesis is organized into 4 chapters. Chapter 2 is go through the literature review and the related works that have been done in this area. Chapter 3 discuses on the framework for doing this research. And the chapter 4 represents some initial findings about applying AIS in spam detection. Figure 1.2 shows the overall view of the thesis organization.

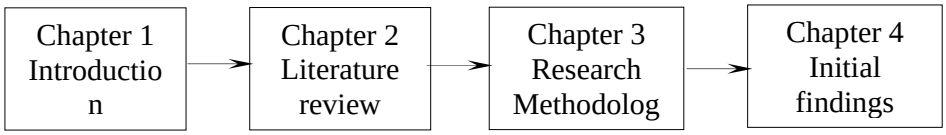

Figure 1.1. Thesis Organization

REFERENCES

[1] K.-ching Ying, S.-wei Lin, Z.-jung Lee, and Y.-tim Lin, “Expert Systems with Applications An ensemble approach applied to classify spam e-mails,” *Expert Systems With Applications*, vol. 37, no. 3, pp. 2197-2201, 2010.

[2] C. Lopes, P. Cortez, P. Sousa, M. Rocha, and M. Rio, “Expert Systems with Applications Symbiotic filtering for spam email detection,” *Expert Systems With Applications*, vol. 38, no. 8, pp. 9365-9372, 2011.

[3] C.-chin Lai, “An empirical study of three machine learning methods for spam filtering,” *Knowledge-Based Systems*, vol. 20, pp. 249-254, 2007.

[4] B. Yu and Z.-ben Xu, “A comparative study for content-based dynamic spam classification using four machine learning algorithms,” *Knowledge-Based Systems*, vol. 21, pp. 355-362, 2008.

[5] 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.,” *Bio Systems*, vol. 92, no. 3, pp. 215-25, Jun. 2008.

[6] C.-hung Wu, “Expert Systems with Applications Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks,” *Expert Systems With Applications*, vol. 36, no. 3, pp. 4321-4330, 2009.

[7] T. S. Guzella and W. M. Caminhas, “Expert Systems with Applications A review of machine learning approaches to Spam filtering,” *Expert Systems With Applications*, vol. 36, no. 7, pp. 10206-10222, 2009.

[8] V. Metsis, “Spam Filtering with Naive Bayes – Which Naive Bayes ?,” 2006.

[9] A. Hamdan and R. Abu, “Application of genetic optimized artificial immune system and neural networks in spam detection,” *Applied Soft Computing Journal*, vol. 11, no. 4, pp. 3827-3845, 2011.

[10] B. Yu and Z.-ben Xu, “A comparative study for content-based dynamic spam classification using four machine learning algorithms,” *Knowledge-Based Systems*, vol. 21, pp. 355-362, 2008.

[11] Y. Hu, C. Guo, E. W. T. Ngai, M. Liu, and S. Chen, “Expert Systems with Applications A scalable intelligent non-content-based spam-filtering framework,” *Expert Systems With Applications*, vol. 37, no. 12, pp. 8557-8565, 2010.

[12] B. Yu and Z.-ben Xu, “A comparative study for content-based dynamic spam classification using four machine learning algorithms,” *Knowledge-Based Systems*, vol. 21, pp. 355-362, 2008.

[13] R. I. Idovh et al., “An Artificial Immnity-Based Spam Detection System,” pp. 3392-3398, 2007.

[14] T. Oda, “A Spam-Detecting Artificial Immune System by Master of Computer Science A Spam-Detecting Artificial Immune System submitted by,” 2005.

[15] J. Brownlee, “CLONAL SELECTION THEORY & CLONALG THE CLONAL SELECTION CLASSIFICATION ALGORITHM (CSCA),” no. 2, 2005.

[16] D. Dasgupta, Z. Ji, and F. Gonzalez, “Artificial immune system (AIS) research in the last five years,” *The 2003 Congress on Evolutionary Computation, 2003. CEC ’03.*, pp. 123–130, 2003.

[17] D. Dasgupta, “Advances in Artificial Immune Systems ©,” no. November 2 006, pp. 40–49.

[18] J. Greensmith, A. Whitbrook, and U. Aickelin, “Artificial Immune Systems,” pp. 1–29.

[19] S. a Hofmeyr and S. Forrest, “Architecture for an artificial immune system.,” *Evolutionary computation*, vol. 8, no. 4, pp. 443–73, Jan. 2000.

[20] a. Secker, a. a. Freitas, and J. Timmis, “AISEC: an artificial immune system for e-mail classification,” *The 2003 Congress on Evolutionary Computation, 2003. CEC ’03.*, pp. 131–138.

[21] A. Sharma and D. Sharma, “Clonal Selection Algorithm for Classification,” pp. 361–370, 2011.

[22] R. A. Zitar and A. Hamdan, “Genetic optimized artificial immune system in spam detection: a review and a model,” *Artificial Intelligence Review*, Nov. 2011.

[23] I. Idris and A. L. I. Selamat, “OPTIMIZED SPAM CLASSIFICATION APPROACH WITH,” vol. 39, no. 1, 2012.

[24] G. Algorithm, “Genetic Algorithm and Artificial Immune Systems · A combinational Approach for Network Intrusion Detection Ii,” pp. 494–498, 2012.

[25] S. P. Koh, K. H. Chong, and D. F. W. Yap, “Hybrid Artificial Immune System-Genetic Algorithm optimization based on mathematical test functions,” *2010 IEEE Student Conference on Research and Development (SCOReD)*, no. SCOReD, pp. 256–261, Dec. 2010.

[26] C. a. C. Coello and N. C. Cortés, “Hybridizing a genetic algorithm with an artificial immune system for global optimization,” *Engineering Optimization*, vol. 36, no. 5, pp. 607–634, Oct. 2004.

[27] M. Korayem, W. A. Hamad, and K. Mostafa, “A hybrid genetic algorithm and artificial immune system for informative gene selection,” vol. 10, no. 7, pp. 76–83, 2010.

[28] R. A. Zitar and A. H. Mohammad, “Spam Detection Using Genetic Assisted Artificial Immune System,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 08, pp. 1275–1295, Dec. 2011.