ภาคผนวก ก
ผลงานวิจัยที่ได้รับการตีพิมพ์

Adaptive Spam Mail Filtering Using Genetic Algorithm

Usarat Sanpakdee, Aranya Walairacht and Somsak Walairacht
Department of Computer Engineering, Faculty of Engineering,
King Mongkut's Institute of Technology Ladkrabang
Ladkrabang, Bangkok, Thailand
som_usarat@yahoo.com, {kwaranya, kwsomsak}@kmitl.ac.th

Abstract **— In this paper, we propose a mechanism for filtering incoming spam mails by generating spam mail prototypes using genetic algorithm. Firstly, words from e-mails are extracted and are categorized by their relating meaning into 7 groups. Then, we compose a string of chromosome having 7 genes, i.e., groups of words. Each gene, represented words in each group, is encoded into binary value. The genetic algorithm and its operations are applied to create varieties of spam mail prototypes which inherit from old spam mails. It saves time for preparing training sets and need no large training set for learning like other methods. The spam mail prototypes are the result of this learning mechanism. The experimental results show that the proposed system has efficiency. When testing with both spams and hams, the accuracy is about 85% in average.**

Keywords **— spam mail, spam prototype, spam filtering, genetic algorithm**

1. Introduction

Over the last decade, e-mail or electronic mail has become a popular method of communication because it has more convenient and cheaper than normal postal mail. On the other hand, e-mail has become a victim of abuse. Some business use e-mail for their benefit such as work at home business group, commercial business group. These businesses send a lot of e-mails which context concern about convince, offer prize for something, give assistant, mortgage, medicine, and etc. This kind of e-mail is called, "spam" mail. [1]

There are several meanings of spam, such as, unwanted, junk e-mail message, unsolicited commercial e-mail (UCE), unsolicited bulk e-mail (UBE), and so on. Spam mail does not only annoy e-mail users, it also increases the load of e-mail server and waste of bandwidth. Thus, Internet Service Providers (ISP) must pay more cost for bandwidth and storage. And, e-mail users feel uncomfortable, lose more time because of slower internet and need to pay more cost to use internet too. [2]

Moreover there are a lot of viruses which disguise in spam mail. When the users read their e-mails, they will receive virus attentively. Virus may be a little disturb system but it continually uses victim's computer for distribute spam. Furthermore, spam may create a way for offender to attack and rob benefit from the system. [3]

There are several techniques to classify spam such as header analysis, address list, keyword list, signature analysis, content statistical analysis. But the popular technique is machine learning. Thus, in this paper, we propose a technique for filtering incoming spam mails by generating spam mail prototypes using genetic algorithm. Taking the advantage of evolution mechanism of genetic algorithm, the system can adaptively generate spam prototype for filtering automatically.

2. Spam Mail Prototype

Our system consists of 2 major processes as shown in Figure1. An input e-mail is passed into a process of keyword extraction. Within the process of genetic algorithm, a chromosome represented that e-mail is constructed from the extracted words. The evolution mechanism creates spam mail prototypes as the output of the system.

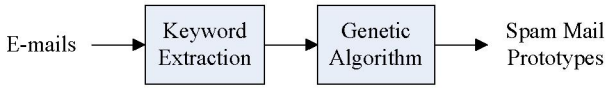

Figure 1. System Block Diagram

The collection of e-mails that considered as spam is called corpus. Spam mails from corpus are encoded to chromosomes and undergo with the genetic operators, i.e., crossover and mutation, and are evaluated by a fitness function. Resulting from the genetic algorithm, rules set or spam mail prototypes are obtained. Figure2 shows a flowchart of spam mail prototypes construction.

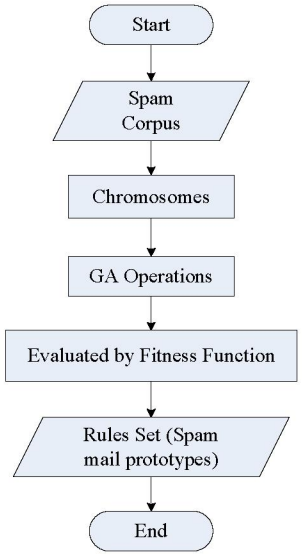

Figure 2. Constructing rules set (spam mail prototypes)

3. Genetic Algorithm

A genetic algorithm (GA) is an algorithm used to find approximate solutions to difficult-to-solve problems through application of the principles of evolutionary biology to computer science. Genetic algorithms use biologically-derived techniques such as inheritance, mutation, natural selection, and recombination (or crossover). Genetic algorithms are typically implemented as a computer simulation in which a population of abstract representations (called chromosomes) of candidate solutions (called individuals) to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but different encodings are also possible. The evolution starts from a population of completely random individuals and happens in generations. In each generation, multiple individuals are stochastically selected from the current population, modified (mutated or recombined) to form a new population, which becomes current in the next iteration of the algorithm.

3.1 Building Representation

An e-mail consists of header, subject and body, for our proposed technique, we extract words within the body part of e-mail only. The extracted words in the form of prepositional, article, and number are discarded. There are 416 words which have been categorized into 7 groups by their relating meaning. Table1 shows all words in each group.

A spam mail is encoded to a binary string. One spam mail represents one chromosome with 7 genes. The binary representation of each gene is computed from words extracted from a spam mail. Figure3 shows a pattern of spam mail chromosome.

	G1	G2	G3	…	G7
Chromosome	weight of G1	weight of G2	weight of G3	…	weight of G7

Figure 3. A pattern of spam mail chromosome

Let's show an example how a chromosome of a spam mail can be constructed. Let an e-mail contains 4 words, namely, "sex", "adult" belong to group G1, and "free", "special" belong to group G2, as shown in Figure 4. We look up table of the weight of these words in data dictionary. The aforementioned weight of word can be calculated by accumulating the frequency of all words and divide by the total number of words in data dictionary. In this case, we have the minimum weight of keyword equal to 0.013 and the maximum weight of keyword equal to 3.31. Then, the minimum and maximum range after normalization becomes 0.004 to 1, or 0000000100 to 1111111000 in binary. Table2 shows computation of the weight of words for this example.

Group	Word	Frequency	$\frac{\text{Frequency}}{\text{Total word(416)}}$	Weight of word	Weight of group
G1	sex	157	0.377	0.114	0.075
G1	adult	49	0.118	0.036	
G3	free	800	1.923	0.581	0.364
G3	special	201	0.483	0.146	

Table 2. Example of calculating weight of word in an e-mail

G1	G2	G3	…	G7
Sex, adult	---	Free, special	…	---

Figure 4. A chromosome before represented to binary string

After we calculate weight of group which average from weight of words that found in the same group. A chromosome, shown in Figure 4, shows weight values in each gene as illustrated in Figure5.

G1	G2	G3	…	G7
0.075	0	0.364	…	0

Figure 5. Weight of each gene in chromosome

The weight of each gene can be encoded into binary string in the following patterns.

Binary 0000000000 represent weight 0.000
Binary 0000000001 represent weight 0.001
Binary 0000000010 represent weight 0.002
.
.
Binary 1111100111 represent weight 0.999
Binary 1111111000 represent weight 1.000

Therefore, weight of G1 gene, 0.075, can be represented by binary value as 0001001011. In the same manner, weight of G3 gene, 0.364, can be represented by binary value as 0101101100. While the rest of genes which has no weight, are represented by binary value 0000000000, as shown in Figure6.

G1	G2	G3	…	G7
0001001011	0000000000	0101101100	…	0000000000

Figure 6. A chromosome representation in binary string

3.2 Genetic Operations

3.2.1 Crossover

For our proposed system, the crossover is allowed for bits of gene within the same group only. We use multiple-point crossover and randomly select the position to cross. In each generation, 15 percent of chromosomes are crossed.

3.2.2 Mutation

Mutation is doing for guarantee that some data will be not disappear. Mutation is done by changing bit in the position which gets from random. In each generation, 2 percent of chromosomes are mutated.

3.3 Evaluation

After e-mails from corpus had been encoded to chromosomes and underwent the operations of genetic algorithm, they are evaluated by the fitness function. The fitness value obtained and used for ranking spam prototypes can be computed from Eq. 1.

$$FitnessFunction = \sum_{i=1}^{i=n} \frac{\text{number of keyword i} \times W_i}{\text{total keywords in an e-mail}(n)} \quad (1)$$

Where the training weight (W_i) is the summation of weight of any word (w_i) found in each spam mail divided by total e-mails in corpus which we use for training. And w_i is calculated by count number of any word i in each e-mail and divided by total words in that e-mail.

3.4 Selection

After all chromosomes had been evaluated by fitness function, the system selects appropriate chromosomes for filtering incoming e-mails. The selection method used is roulette wheel technique.

4. Rules set for classifying e-mails

The weight of words of gene in testing mail and the weight of words of gene in spam mail prototypes are compared to find match gene. In this proposed system, we specify that if the number of matched gene is greater or equal to 3 then that spam mail prototype will receive one spam score point. The mentioned classification process for spam mails is show in Figure7.

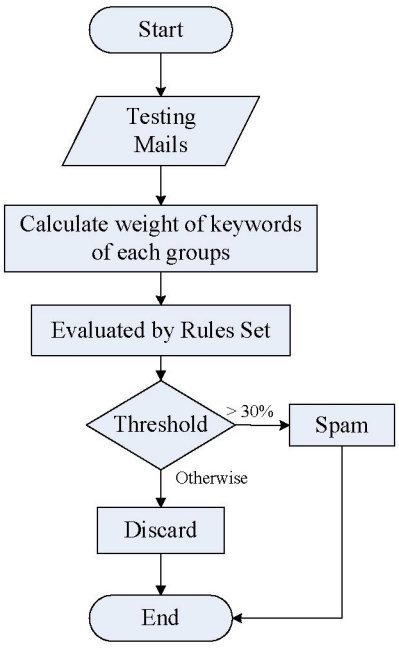

Figure 7. Show the classification process for emails

By comparing testing mails with all spam mail prototypes and the sum of spam score point of all prototypes, if the percentage of spam score point is greater than the percentage of the threshold, then this testing mail is defined to be spam mail. In the experiments, we set the threshold value at 30% in which this threshold value can also be manually adjusted to the appropriate value for optimal result.

4. Experimental Results

We collected spam corpus from [4, 5] and ham corpus from [6]. In the experiments, we use 1,097 of spam mails and 300 of ham (not spam mail).

4.1 Experiment 1

We divided e-mails for training by 80% and for testing by 20% of the total e-mails. The experimental result shows that the average accuracy obtained is 85.53%, with the average values of precision 89.83% and recall of 75.71%. Figure 8 shows the accuracy of training set and validation set in each generation.

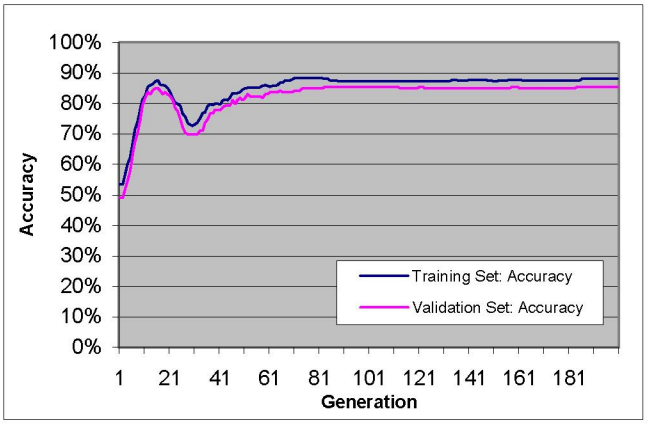

Figure 8. Results of experiment 1: The accuracy of training set and validation set in each generation

4.2 Experiment 2

We divided e-mails for training by 60% and for testing by 40% of the total e-mails. The result shows that the average values for the accuracy is 84.77%, the precision is 90.74% and the recall is 73.13%. Figure 9 shows results of this experiment.

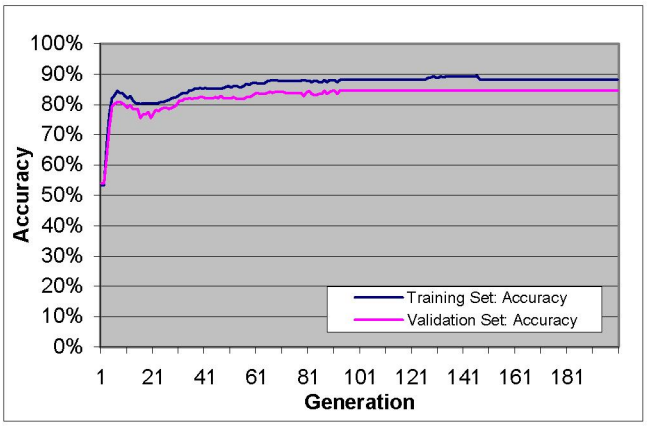

Figure 9. Result of experiment 2: The accuracy of training set and validation set in each generation

4.3 Experiment 3

When we use 100% of spam for testing, the average values of the accuracy obtained is 84.77%, the precision is 90.74% and recall is 73.13%. The results are shown in Figure10.

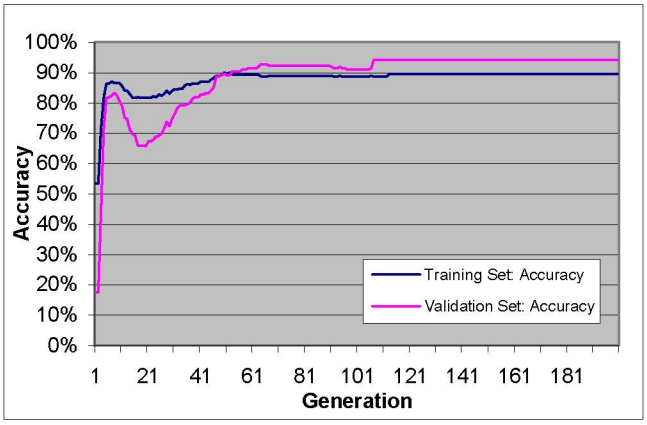

Figure 10. Results of experiment 3: The accuracy of training set and validation set in each generation

4.4 Experiment 4

When we used 50% of spam mails and 50% of ham for testing, the average values of the accuracy is 85.14%, the precision is 90.91% and the recall is 73.86%. The accuracy of training set and validation set in each generation are shown in Figure11.

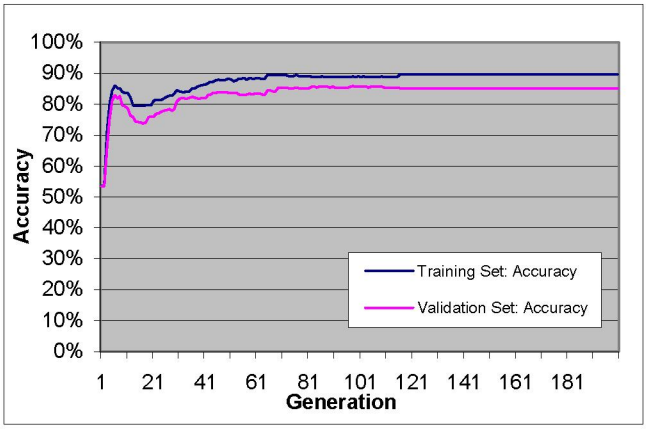

Figure 11. Result of experiment 4: The accuracy of training set and validation set in each generation

5. Conclusions

In this paper, we proposed an adaptive spam mail filtering which uses genetic algorithm and its operations, i.e., crossover and mutation, to create new varieties of spam mail prototypes. The experiments show that proposed adaptive spam mail filtering performs efficient results. In additional, the system allows the threshold value for matching rules set can be manually adjusted to the appropriate level of filtering. In the future, we plan to test our system with more different sets of corpus. The categories and keywords, that represent spam mail from the real situation usages, are to be included to cover broader filtering performance of the system.

REFERENCES

[1] L. Pelletier, J. Almhana, V. Choulakian, 'Adaptive Filtering of SPAM' in Proceedings of the Second Annual Conference on Communication Network and Services Research (CNSR'04) , (2004)
[2] A. Boonyu, http://www.dss.go.th/dssweb/st-articles/files/sti_6_2546_spam_mail.pdf
[3] Mehmed Kantardzic(2003), Data Mining: Concept, Model, Medthod, and Algorithms, 'Genetic Algorithms', IEEE Press, 221-245 (2003)
[4] Ramesh Krishnamurthy, Constantin Orasan, A linguistic investigation of the junk emails, http://clg.wlv.ac.uk/projects/junk-email/
[5] Spam Assassin, http://spamassassin.org/publiccorpus/.
[6] Enron Email Data Set, http://www.cs.cmu.edu/~enron/

Table 1. Show words extracted from spam mails categorized in each group

Group	Content	Example of keywords in each group
G1	Adult	adult, aphrodisiac, big, cam, climax, company, cum, desire, erotic, fantasy, fuck, gay, girl, greate, guy, hard, hardcore, heaven, hot, huge, long, man, max, maxlength, nude, orgasm, penis, performance, pheromone, pill, porn, powerful, pussy, satisfy, sex, stamina, sweet, teen, viagra, webcam, x, xxx, xxx-porn, young
G2	Financial	Account, accountant, alert, analyst, attorney, bank, bankruptcy, benefit, bill, billing, broker, budget, building, cash, cheque, commission, consolidate, court, credit, creditor, currency, customer, debt, deposit, discover, economy, entrepreneur, estate, exchange, fee, finance, freedom, fund, help, high-risk, insurance, invest, investor, judgment, legal, legitimate, lender, loan, mastercard, mortgage, obligate, pay, payable, payable, paycheck, promote, purchase, rate, refinance, refund, rent, revenue, risk, service, statement, stock, support, tax, transaction, vat, visa, wealth, worth
G3	Commercial	agency, agent, arrival, bargain, better, brand, buy, camera, cdrom, celeb, chance, cheap, Christmas, collect, college, commerce, computer, cost, deliver, discount, especial, expensive, express, fantastic, free, furnishing, furniture, game, get, gif, gift, great, guarantee, inexpensive, invite, item, just, keyboard, license, lifetime, magazine, maintenance, mall, market, material, materials, mobile, motherboard, mouse, offer, online, only, order, palm, pamphlet, percent, premium, price, produce, product, program, recommend, refill, release, resell, reseller, retail, sale, save, save, sell, ship, shipping, shop, shopping, special, subscribe, supply, surprise, trade, trademark, upgrade, voucher, whole, wholesale, within
G4	Beauty & Diet	after, age, amaze, anti-aging, appetite, beauty, become, before, believe, blood, body, botanic, breast, build, burn, calorie, capsule, card, cell, change, chemical, cholesterol, confirm, course, diet, difference, dose, drug, effect, effective, eliminate, energy, enhance, exercise, eye, face, fast, fat, firm, fit, fitness, flexible, gary, grow, grown, growth, hair, health, healthcare, heart, height, herb, herbal, hormone, improve, inche, incredible, kidney, large, laser, life-changing, light, lose, loss, low, magic, medicine, metabolism, micro-cap, miracle, modem, move, muscle, nature, nutrient, old , over, overweight, permanent, plain, potential, pound, power, protect, reduce, remanufacture, repair, restore, retain, reverse, safe, satisfaction, secret, size, step, strength, strong, tablet, therapy, thin, toxin, treatment, under, virginia, vitamin, weight, woman, wonderful, wrinkle
G5	Traveling	book, deluxe, excite, guide, holiday, honest, hotel, luxury, meal, package, plan, problem, relax, relief, reserve, resort, summer, temple, ticket, tour, train, travel, traveler, trip, vacation,
G6	Home-Based Business	address, astonishment, base, broadcast, bulk, business, comfort, connect, demo, domain, downline, download, earn, email, emailing, ethernet, facemail, fresh, home, homebased, homeworker, host, income, interest, international, internet, investigate, job, list, lucrative, mail, mailbox, mailer, mailing, make, marketing, message, million, money-making, opportunity, part-time, people, private, profit, reach, receive, recipient, require, re-register, return, server, software, subscriber, success, teach, unsubscribe, user, visit, website, work, work-at-home, worker, working
G7	Gambling	action, award, bet, bonus, casino, challenge, extra, gambling, gold, hunt, las, lucky, millionaire, player, poker, prize, reward, rich, vegas, win