Abstract—Spam is any unwanted message, especially advertisement and fraud schemes. The average cost of spam per employee per year at 82 of the Fortune 500 companies is estimated to be $1934. This paper presents a way for detecting spam with automatically generated regular expressions which are subjected to evolution using a genetic algorithm consisting of breeding and mutation.

Keywords—RegEx, Genetic algorithm, spam, ham, chromosome, gene, mutation, cross-over.

I. INTRODUCTION

Spam is the use of electronic messaging systems to send unsolicited bulk messages, especially advertising and fraud schemes, indiscriminately.

The paper suggests a method for detecting spam via genetic regular expressions. The regular expressions are generated from a collection of ‘spam’ and ‘ham’ emails. These regular expressions are then evolved through generations using a genetic algorithm presented in the following sections.

II. REGULAR EXPRESSIONS

Regular Expressions (or RegEx) are a powerful tool for matching text. RegEx may contain metacharacters, character classes and grouping.

For example, the RegEx: /^He (is|was) running [. .]$/ matches lines beginning with “He ...”, allows a choice of “is” or “was”, and matches a space, period or comma after “running”.

A. Generating RegEx

1. Create two sets of emails – ‘spam’ and ‘ham’.
2. Break spam into tokens.
3. Generate a large number of RegEx based on the spam tokens.
4. Delete RegEx that match any ham.
5. Delete RegEx that match only one spam.
6. Rank the remaining RegEx, sorted from highest number of matching spam.

III. EVOLVING REGULAR EXPRESSIONS

In the field of Artificial Intelligence, a genetic algorithm is a heuristic that mimics the natural evolution process. Genetic algorithms use chromosomes made up of individual genes.

The RegEx generated, as described in the previous section, are considered to be chromosomes. These chromosomes consist of small components or pieces called genes.

For example: /Subject: [A-Z]+ [A-Z][a-z]+/ is a chromosome. Subject: ; [A-Z]+ and [A-Z][a-z]+ are individual genes.

The pseudo code for the genetic algorithm:

1. Loop until n\textsuperscript{th} generation
   a. Score chromosomes based on fitness function.
   b. Keep the fittest k chromosomes.
   c. Breed the survivors to obtain children.
   d. Mutate children.
2. End Loop
3. Print n\textsuperscript{th} generation chromosomes sorted according to their scores.

A. Fitness Function

The fitness function has two components: number of spam lines matched and the length of chromosomes. Shorter chromosomes that match a lot of spam lines receive greater score.

For example: 
\[ f = L \times (1 + ((200 - \text{len}) / 200)) \]

Where
\[ L = \text{number of lines of spam matched} \]
\[ \text{len} = \text{length of chromosome} \]

Here 200 is the threshold length of the chromosome. That is, any chromosome above 200 characters receives a penalty. Without the length ‘penalty’, chromosomes will grow longer and longer each generation (simply adding other winning genes via breeding).

B. Breeding

For choosing chromosomes from a breeding pool Roulette Wheel selection process can be opted. Here high scoring chromosomes have a better chance of being chosen than the low scoring ones.

Each survivor from the previous generation produces a child chromosome. The second parent is chosen by the
Roulette Wheel selection process. These two parents together create one child using crossover. The crossover options available are the OR method and the CAT method.

In the OR method the two parent chromosomes are broken down into their unique genes which are then OR together to create a child. For example: Crossover between (Apple|Orange) and (Banana|Apple) creates (Apple|Orange|Banana).

In the CAT method the two parent chromosomes are concatenated together to create a longer child chromosome. For example: Crossover between (chromosome1) and (chromosome2) creates (chromosome1| chromosome2).

X. Mutation

Mutation in a child chromosome brings about changes in it in a randomized way. The simplest way is to delete a random amount of genes. Mutation may help the fitness of a child as shorter chromosomes are considered more fit. Also it may remove some harmful genes.

IV. CONCLUSIONS

The final chromosomes obtained after the n\textsuperscript{th} iteration of the genetic algorithm will be used to detect spam.

The proposed strategy for generating RegEx and evolving them through generations is still a hypothetical concept requiring proof of concept.

REFERENCES