ANALISIS DAN IMPLEMENTASI SPAM FILTERING MENGGUNAKAN METODE GRAMMATICAL EVOLUTION

Harun Al Rosyid[1], Agung Toto Wibowo[2], Fazmah Arif Yulianto[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Menyingkirkan e-mail yang tidak diinginkan (spam) dalam jumlah banyak secara manual akan sangat melelahkan. Namun dengan mekanisme spam filtering, hal itu dapat diatasi. Spam filtering akan mem-blok spam secara otomatis berdasarkan rule-rule tertentu. Semakin berkembangnya teknik spamming dari waktu ke waktu akan meningkatkan variasi rule. Penentuan rule secara manual tentu saja akan memakan waktu.

Oleh karena itu diperlukan suatu mekanisme pembangkitan rule secara otomatis berdasarkan data-data yang sudah ada. Metode yang paling cocok untuk masalah ini adalah dengan menggunakan teknik learning. Teknik ini dapat menemukan rule secara otomatis dan diharapkan dapat berlaku umum untuk datadata yang belum diketahui. Salah satu teknik learning yang dapat dipakai adalah metode Grammatical Evolution (GE).

GE dapat menghasilkan solusi berupa fungsi/program yang berbasis grammar. Grammar yang dibangun akan disesuaikan dengan masalah tersebut. Dengan kemampuan ini, GE digunakan untuk menghasilkan rule yang berupa program untuk kemudian dijadikan acuan untuk proses klasifikasi.

Dalam tugas akhir ini, dibuat perangkat lunak untuk mengimplementasikan metode GE untuk proses pembangkitan rule dan klasifikasi pada spam filtering. Hasil yang didapat menunjukkan bahwa penentuan grammar yang lebih kompleks memberikan akurasi yang lebih tinggi, begitu juga dengan meningkatkan parameter ukuran kromosom dan crossover rate. Sedangkan perubahan ukuran populasi dan mutation rate tidak mempengaruhi akurasi.

Kata Kunci : spam filtering, rule, grammatical evolution, klasifikasi.

Abstract

Eliminating unwanted e-mail (spam) in large amount manually can be exhausting. However, spam filtering mechanism can solve this problem. Spam filtering blocks spam automatically based on specified rules. More developed spamming technique over time will increased rule variation. Specify rules manually surely can be time consuming.

Therefore need an automatic rule(s) generation mechanism according to existing data. Suitable method for the problem is learning technique. The technique can found rules automatically and expect it to be generally valid for unclassified data. One of the technique can be used is Grammatical Evolution (GE) method.

GE can produce solutions in function/program form based on grammar. Evolved grammar suited to the problem of interest. With this capability, GE used for producing rules in program form then it is referable for classification process.

In this final project we made software to implement GE method for rule generation process and classification on spam filtering. The results show that defined more complex grammar has given higher accuracy as well as increasing chromosome size and crossover rate. However varied population size and mutation rate doesn’t give much effect on accuracy.

Keywords : spam filtering, rule, grammatical evolution, classification.

1. Pendahuluan

1.1 Latar belakang

Saat ini, membaca *e-mail* sudah menjadi keperluan bagi banyak orang mulai dari kalangan profesional sampai kalangan pribadi. Alasan banyak orang menggunakan *e-mail* antara lain karena cepat, murah, dan efisien. Tetapi adakalanya kita menerima *e-mail* dari sumber yang tidak jelas dan yang tidak kita inginkan. *E-mail* inilah yang disebut *spam* [4].

Pada dekade yang lalu, *worms* menjadi masalah utama bagi para pengguna *e-mail*. Tetapi dalam beberapa tahun terakhir ini, *spam* 'menyerbu' layanan internet paling banyak digunakan ini (*e-mail*) [4]. Pada Desember 2003, BBC News memperkirakan sekitar 40% dari semua *e-mail* yang terkirim diidentifikasi sebagai *spam*, dan proses identifikasi dan penghapusan *spam* oleh perusahaan-perusahaan di Inggris memakan waktu rata-rata 1 jam/pekerja/hari. Para *spammer* melakukan penyebaran spam dengan cara tiap pesan dikirim sekaligus dalam jumlah yang sangat besar. Hal itu dilakukan dengan harapan sekian presentase dari pesan (yang mayoritas berisi iklan suatu produk) yang terkirim akan merespon balik dan membeli produknya. The Wall Street Journal memperkirakan bahwa tingkat respon sebesar 0.0001% sudah dapat menghasilkan *profit* bagi *spammer* tersebut. Oleh karena itu, tidak heran jika *spam* dapat dijadikan sarana promosi produk yang efektif, cepat, dan tentu saja murah.

Berbagai macam strategi telah dicoba untuk menyelesaikan masalah ini. Secara garis besar strategi-strategi tersebut dapat dibagi menjadi 2 teknik, yaitu teknik *prevention* dan *cure*. Teknik *prevention* bertujuan untuk mencegah terkirimnya *spam* dengan mengimplementasian sejumlah kendali dan pemeriksaaan pada sistem *e-mail* global yang akan menyulitkan para *spammer* dalam mengirim pesan dalam jumlah besar. Tetapi masalah utama teknik ini adalah implementasi kendali pada sistem global yang kompleks. Teknik lainnya yaitu *cure* yang bertujuan mencegah masuknya *spam* ke *inbox* si penerima. Teknik ini biasa dikenal sebagai *spam filtering*. Cara kerjanya yaitu dengan mem-*filter spam* menggunakan sejumlah *rule* (aturan). Jika kita sudah memiliki aturan yang jelas untuk membedakan *spam* atau bukan, maka kita tinggal menggunakan aturan tersebut dengan memanfaatkan pencocokan pola. Tetapi, tentu saja akan merepotkan jika kita harus memperbaharui aturan tersebut agar tetap berlaku untuk menyaring pesan baru dengan strategi *spamming* yang semakin bervariasi dan "kreatif".

Oleh karena itu, diperlukan metode *Artificial Intelligence* (AI) yang memungkinkan kita membuat daftar aturan secara otomatis berdasarkan data-data yang dimiliki. Metode AI yang paling cocok untuk kasus ini adalah dengan menggunakan teknik *learning*. Dengan teknik ini, kita dapat secara otomatis menemukan aturan yang diharapkan dapat berlaku umum untuk data-data yang belum pernah diketahui [1].

Salah satu *learning algorithm* yang dapat dipakai untuk membangkitkan aturan ini adalah *Grammatical Evolution* (GE). GE adalah metode yang dapat merepresentasikan solusi berbasiskan *grammar*. *Grammar* dalam metode ini

menggunakan notasi *Backus Naur Form* (BNF) [3]. BNF dapat dikodekan dengan mudah dan mewakili semua bahasa [2]. *Grammar* yang dibangun akan berisikan elemen-elemen yang dapat membentuk aturan-aturan untuk kemudian digunakan dalam *spam filtering*.

1.2 Perumusan masalah

Dari uraian di atas dapat dirumuskan beberapa permasalahan utama, yaitu:
1. Bagaimana menentukan atribut yang akan dimasukkan ke dalam daftar aturan.
2. Bagaimana menentukan *grammar* yang cocok untuk kasus *spam filtering*.
3. Bagaimana memanfaatkan metode GE untuk membangkitkan aturan.
4. Bagaimana menentukan fungsi fitness yang tepat.
5. Bagaimana melakukan pre-processing terhadap data uji.

Hipotesa yang akan dibuktikan yaitu bahwa *grammar* yang mengandung elemen yang lebih kompleks menghasilkan akurasi yang lebih tinggi.

Batasan masalah yang digunakan dalam menyelesaikan permasalahan di atas antara lain sebagai berikut:
1. Data uji menggunakan *corpus SpamAssasin* yang berisikan *spam* dan *ham* yang berasal dari http://spamassassin.apache.org/publiccorpus/.
2. Aturan yang digunakan terdiri dari beberapa atribut statistik, seperti frekuensi kemunculan kata/karakter, jumlah resipien, dsb.
3. *Spam* dan *ham* diletakkan di direktori yang berbeda agar dapat digunakan sebagai acuan untuk menghitung akurasi klasifikasi.
4. Parameter performansi diukur berdasarkan:
   - Akurasi yaitu presentase dari jumlah klasifikasi data yang cocok dibagi jumlah sampel data total. Semakin tinggi akurasi semakin baik.
   - *False positive* yaitu kesalahan identifikasi ham sebagai spam. Semakin rendah *false positive* semakin baik.

1.3 Tujuan

Tujuan dari pembuatan TA ini adalah sebagai berikut:
1. Menerapkan metode GE untuk *spam filtering*.
2. Menganalisa pengaruh penentuan *grammar* terhadap akurasi *spam filtering*.
3. Menganalisa pengaruh penentuan parameter evolusi terhadap akurasi *spam filtering*.

1.4 Metodologi penyelesaian masalah

Ada beberapa tahap yang dilakukan dalam menyelesaikan masalah ini, yaitu:
1. Pengumpulan atribut
   Dalam tahap ini dilakukan studi data untuk menentukan atribut-atribut mana saja yang nantinya akan dimasukkan ke dalam rule.
2. Penentuan parameter dan grammar
   Parameter-parameter evolusi seperti jumlah populasi, generasi maksimum, crossover rate, mutation rate, dll ditentukan dalam tahap ini. Pembangunan grammar yang cocok dengan masalah juga dilakukan.
3. Pembangkitan rule dengan GE

Proses ini disebut juga dengan proses pelatihan, yaitu menggunakan sejumlah data latih dari corpus untuk kemudian dijadikan acuan pengukuran akurasi sementara. Akurasi tersebut dihitung dengan menggunakan acuan rule sementara pada data latih. Nilai ini digabungkan dengan nilai false positive sementara kemudian dijadikan sebagai nilai fitness. Tahap ini akan menghasilkan satu rule yang akan dijadikan acuan untuk tahap klasifikasi.

4. Klasifikasi
   Tahap ini membutuhkan data uji dan rule. Data uji berasal dari corpus dengan jumlah tertentu dan rule berasal dari tahap sebelumnya.
5. Analisa hasil
   Hasil klasifikasi kemudian dianalisa performansinya berdasarkan parameter evolusi dan grammar.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis di atas, dapat ditarik beberapa kesimpulan antara lain:

1. Kombinasi parameter terbaik adalah pada kombinasi uji ke-165, yaitu pada saat pengujian dengan grammar-III, ukuran populasi = 40, ukuran kromosom = 20, crossover rate = 0.6, dan mutation rate = 0.3, dengan akurasi sebesar 84. 75% dan false positive rate sebesar 6.103.
2. Dari ketiga grammar yang dibuat, grammar III lebih cocok diimplementasikan pada kasus spam filtering, dibuktikan dengan akurasi yang dihasilkan lebih tinggi dibandingkan grammar I dan grammar II pada akurasi rata-rata.
3. Secara umum, grammar yang memiliki kemungkinan solusi lebih beragam menghasilkan akurasi yang lebih tinggi.
4. Ukuran populasi = 20 menghasilkan akurasi rata-rata yang lebih tinggi dari ukuran populasi = 40.
5. Ukuran kromosom berpengaruh pada akurasi. Semakin tinggi dan ukuran kromosom, maka semakin tinggi akurasi.
6. Crossover rate yang lebih tinggi menghasilkan akurasi yang lebih tinggi pada akurasi rata-rata tetapi tidak pada akurasi tertinggi.
7. Perubahan mutation rate tidak terlalu mempengaruhi hasil akurasi baik pada akurasi rata-rata maupun akurasi tertinggi.

5.2 Saran

Beberapa saran untuk pengembangan Tugas Akhir ini antara lain:

1. Dapat dibuat grammar yang lebih cocok dan dengan variasi lain agar solusi yang dihasilkan lebih optimal.
2. Untuk mendapatkan nilai fitness yang lebih tinggi pada saat pelatihan, dapat menggunakan metode seleksi orang tua, crossover, mutasi, atau survivor replacement yang lain.
3. Perlu dikembangkan sistem pengklasifikasian spam dengan metode yang sama tetapi diimplementasikan pada basis web sebagai client-side spam filtering.

Referensi

[1] Suyanto, ST, MSc, 2007, *Artificial Intelligence: Searching, Reasoning, Planning, and Learning*, Bandung: Informatika.

[2] Suyanto, ST, MSc, 2008, *Evolutionary Computation: Komputasi Berbasis "Evolusi" dan "Genetika"* , Bandung: Informatika.

[3] O'Neill M., Ryan Conor, 2003, *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*, Kluwer Academic Publishers.

[4] Khorsi, Ahmed, 2007, *An Overview of Content-Based Spam Filtering Techniques,* Department of Computer Science, Djillali Liabes University, Bel Abbes, 22000, Algeria

[5] Mertz D, 2002, *Spam filtering techniques. Six approaches to eliminating unwanted e-mail*, www.ibm.com/developerworks/linux/library/l-spamf.html. Diakses tanggal 8 Oktober 2009.

[6] Bortzmeyer, Stephane, 2006, *Anti-spam filtering techniques,* AFNIC, ITU

[7] Cohen, W.W, 1996, *Learning Rules that Classify E-Mail*. Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access, Stanford, California.

[8] Apte, C. and F. Damerau, 1994, *Automated Learning of Decision Rules for Text Categorization*. ACM Transactions on Information Systems, 12(3):233–251.

[9] Cranor, L.F. and B.A. LaMacchia, 1998, *Spam!* Communications of ACM, 41(8):74–83.

[10] Graham, Paul, 2002, *A Plan for Spam*, http://www.paulgraham.com/spam.html. Diakses tanggal 18 November 2009.

[11] Akira Hara, Tomohisa Yamaguchi, Takumi Ichimura, and Tetsuyuki Takahama, *Multi-chromosomal Grammatical Evolution*, Fourth International Workshop on Computational Intelligence & Applications. IEEE SMC Hiroshima Chapter, Hiroshima University, Japan, December 10 & 11, 2008