PERBANDINGAN KINERJA ALGORITMA ID3 DAN C4.5 DALAM KLASIFIKASI SPAM-MAIL

Sofi Defiyanti
Jurusan Sistem Informasi
Universitas Gunadarma

D. L. Crispina Pardede
Sistem Komputer
Universitas Gunadarma
pardede@staff.gunadarma.ac.id

Abstrak*-Klasifikasi spam mail digunakan untuk memisahkan spam-mail dari non spam mail (legitimate mail). Klasifikasi spam mail berguna untuk menghemat waktu dan biaya yang digunakan untuk menghapus spam mail dari inbox. Untuk itu diperlukan metode yang paling baik untuk melakukan klasifikasi spam mail. Algoritma decision tree merupakan salah satu metode yang untuk klasifikasi spam mail. Algoritma decision tree telah banyak mengalami pengembangan. Algoritma ID3 dan C4.5 adalah salah satu pengembangan dari algoritma decision tree. Penelitian ini membandingkan kinerja dari dua algoritma tersebut dalam melakukan klasifikasi spam mail. Pengukuran dilakukan menggunakan sekelompok data uji untuk mengetahui persentase precision, recall dan accuracy. Hasil pengukuran menunjukkan algoritma ID3 memiliki kinerja yang lebih baik dibandingkan algoritma C4.5.*
Kata Kunci : Decision Tree, ID3, C4.5, Klasifikasi, Spam-mail.

I. PENDAHULUAN

Spam messages membanjiri internet dengan mengirimkan salinan pesan-pesan yang sama untuk memaksa agar pesan-pesan tersebut sampai kepada pemakai yang tidak memilih untuk menerimanya. Akibatnya banyak pemakai yang merasa terganggu oleh banyaknya waktu yang dihabiskan untuk menghapus pesan spam, besarnya biaya yang harus dikeluarkan, dan besarnya *bandwidth* jaringan.

Untuk mengatasi hal ini, diperlukan suatu filter anti-spam dengan algoritma tertentu yang dapat memisahkan antara spam-mail dengan *non spam mail* (atau yang biasa disebut *ham* atau *legitimate mail*). Banyak algoritma *anti-spam filter* yang tersedia, diantaranya adalah algoritma *decision tree, naïve bayes, support vector machine (SVM), neural network* dan lain-lain.

Perbandingan kinerja antara algoritma *svm, neural network, naïve bayes*, dan *decision tree* yang memakai algoritma C4.5 yang dilakukan oleh Youn dan McLeod (2006) membuktikan bahwa *decision tree* dengan algoritma C4.5 lebih efisien dan paling sederhana jika dibandingkan dengan ketiga algoritma yang lain. Dengan kesederhanaannya, algoritma C4.5 memberikan hasil yang lebih baik untuk klasifikasi *spam-mail*.

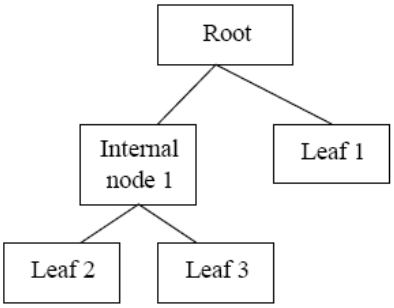
Dari penelitian lain yang dilakukan oleh Jyh-Jian Sheu (2008) diperoleh hasil bahwa metode ID3 dari *decision tree* merupakan metode yang paling baik jika dibandingkan dengan *naïve bayes* dan *k-nearest neighbors* (KNN). Dari penelitian tersebut diketahui bahwa ID3 mempunyai *precision* dan *accurancy* lebih baik dari pada *naive bayes* dan KNN.

Berdasarkan kedua penelitian tersebut, dapat dilihat bahwa kedua algoritma, ID3 dan C4.5 mempunyai kinerja yang baik dalam mengidentifikasi apakah suatu *email* adalah *spam* atau *non-spam*. Namun, belum diketahui algoritma mana diantara keduanya yang lebih unggul kinerjanya. Oleh karena itu kedua algoritma ini perlu dibandingkan.

II. TINJAUAN PUSTAKA

Decision tree adalah salah satu metode klasifikasi yang paling popular karena mudah untuk diinterpretasi oleh manusia. Konsep dasar algoritma *Decision Tree* adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (*rule*).

Pembangunan *tree* dimulai dengan data pada simpul akar (*root node*) yang dilanjutkan dengan pemilihan sebuah *atribut*, formulasi sebuah *logical test* pada *atribut* tersebut dan pencabangan pada setiap hasil dari *test*. Langkah ini terus bergerak ke subset ke contoh yang memenuhi hasil dari simpul anak cabang (*internal node*) yang sesuai melalui proses rekursif pada setiap simpul anak cabang. Langkah-langkah tersebut diulangi hingga dahan-dahan dari *tree* memiliki contoh dari satu kelas tertentu. Gambar 1 memuat contoh dari sebuah *decision tree*. Beberapa model *decision tree* yang sudah dikembangkan antara lain adalah IDS, ID3, C4.5, CHAID dan CART.



Gambar 1. *Decision Tree*

A. Algoritma ID3

Algoritma ID3 atau *Iterative Dichotomiser* 3 (ID3) merupakan sebuah metode yang digunakan untuk membuat pohon keputusan. Algoritma pada metode ini menggunakan konsep dari entropi informasi. Secara ringkas, langkah kerja Algoritma ID3 dapat digambarkan sebagai berikut:

1. Penghitungan *Information Gain* dari setiap atribut dengan menggunakan

$$Gain(S,A) = Entropy(S) - \sum_{v \in nilai(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Dimana

$$Entropy(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

2. Pemilihan atribut yang memiliki nilai *information gain* terbesar,
3. Pembentukan simpul yang berisi atribut tersebut,
4. Ulangi proses perhitungan *information gain* akan terus dilaksanakan sampai semua data telah termasuk dalam kelas yang sama. Atribut yang telah dipilih tidak diikutkan lagi dalam perhitungan nilai *information gain.*

B. Algoritma C4.5

Algoritma C4.5 adalah pengembangan dari algoritma ID3. Oleh karena pengembangan tersebut algoritma C4.5 mempunyai prinsip dasar kerja yang sama dengan algoritma ID3. Hanya saja dalam algoritma C4.5 pemilihan atribut dilakukan dengan menggunakan *Gain Ratio* dengan rumus :

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)}$$

Atribut dengan nilai G*ain Ratio* tertinggi dipilih sebagai atribut test untuk simpul. Dengan *gain* adalah *information gain. SplitInfo* menyatakan *entropi* atau informasi potensial dengan rumus :

$$SplitInfo(S,A) = -\sum_{i=1}^{c} \frac{S_i}{S} \log_2 \frac{S_i}{S}$$

Dengan perbedaan dalam pemilihan atribut ini, C4.5 memiliki keunggulan dibandingkan ID3, yaitu dapat mengolah data numerik (kontinyu) dan kategori (diskret), dapat menangani nilai atribut yang hilang, dan menghasilkan aturan-aturan yang mudah diinterpretasikan.

C. Spam E-mail

Paul Graham (2002) mendefinisikan spam atau email sampah adalah sebagai *unsolicited automated email* (*e-mail* tidak diinginkan yang dikirimkan secara otomatis). Atau *Spam-mail* dapat didefinisikan sebagai "*unsolicited bulk e-mail*" yaitu *e-mail* yang dikirimkan kepada ribuan penerima (*recipient*). Jadi dapat ditarik kesimpulan bahwa spam adalah email yang tidak diinginkan yang dikirim secara otomatis kepada ribuan penerima. *Spam mail* biasanya dikirimkan oleh suatu perusahaan untuk mengiklankan suatu produk. Karena fasilitas *e-mail* yang murah dan kemudahan untuk mengirimkan pesan kepada sejumlah penerima, maka *spam mail* menjadi semakin merajalela. Pada survey yang dilakukan oleh Cranor dan La Macchia (1998), ditemukan bahwa 10% dari mail yang diterima oleh suatu perusahaan adalah *spam-mail*.

Untuk mengatasi hal ini, diperlukan suatu filter anti-spam dengan algoritma tertentu yang dapat memisahkan antara spam-mail dengan *non spam mail* (atau yang biasa disebut *ham* atau *legitimate mail*).

Klasifikasi adalah salah satu metode dalam *data mining* yang dapat mengkasifikasikan email sebagai spam atau non-spam. Pengklasifikasian ini berdasarkan karakteristik dari spam (Lambert, 2003) yaitu :

1. Alamat pengirim yang tidak benar.
2. Pemalsuan header mail untuk menyembunyikan email sesungguhnya sehingga akan sulit menetapkan sebagi spam atau non-spam.
3. Identitas penerima tidak nyata.
4. Kamus alamat penyerang. Alamat email yang berada dalam 'To' memiliki variasi alamat email penerima.
5. Isi *subject* tidak berhubungan dengan isi email.
6. Isi email memiliki sifat keragu-raguan.
7. *Unsubscribe* tidak bekerja pada spam mail.
8. Mengandung script tersembunyi.

D. Pengukuran Kinerja

Untuk permasalahan dalam klasifikasi, pengukuran yang biasa digunakan adalah *precision, recall* dan *accuracy* [Jyh-Jian Sheu, 2008]. Karena spam merupakan *binary classification*, maka *precision*, *recall* dan *accuracy* dapat dihitung dengan cara seperti pada Tabel 1.

Tabel 1. Tabel Penilaian

	diidentifikasi sebagai *non-spam*	diidentifikasi sebagai *spam*
Non-spam	a	b
Spam	c	d

1. Precision

Precision adalah bagian data yang di ambil sesuai dengan informasi yang dibutuhkan. Rumus *precision* adalah :

$$\Pr ecision = \left(\frac{d}{b+d} \right) x100\%$$

Dalam klasifikasi binari, *precision* dapat disamakan dengan *positive predictive value* atau nilai prediktif yang positif.

2. *Recall*
   Recall adalah pengambilan data yang berhasil dilakukan terhadap bagian data yang relevan dengan *query*. Rumus *Recall* adalah :

$$\text{Re}\,call = \left(\frac{d}{c+d}\right) x100\%$$

   Dalam klasifikasi binari, *recall* disebut juga dengan *sensitivity*. Peluang munculnya data relevan yang diambil sesuai dengan query dapat dilihat dengan *recall.*

3. Accuracy
   Accuracy adalah persentase dari total e-mail yang benar diidentifikasi. Rumus *Accurasy* adalah :

$$Accuracy = \left(\frac{a+d}{totalemail}\right) x100\%$$

E. WEKA

WEKA adalah sebuah alat yang digunakan untuk membandingkan beberapa algoritma *machine learning* yang bisa diaplikasikan untuk permasalahan *data mining*. WEKA dikembangkan oleh University of Wakaito, New Zealand yang bersifat *open source.*

Penelitian yang dilakukan oleh Seongwook Youn dan Dennis Mcleod (2006) menggunakan WEKA sebagai alat bantu untuk membandingkan kinerja tiga algoritma yaitu *Neural Network, Support Vektor Mechine (SVM),Naïve Bayesian dan C4.5*. Keempat algoritma tersebut digunakan dalam kasus yang sama yaitu mengklasifikasikan email menjadi spam atau non-spam.

Beberapa kelebihan yang dimiliki WEKA antara lain mudah digunakan, berbasis GUI (*Graphical Interface User*) dan bisa digunakan untuk mengintegrasikan metode baru yang dibuat sendiri dengan beberapa ketentuan.

III. ANALISIS DAN PEMBAHASAN

A. Data yang Digunakan

Data uji yang digunakan dalam penelitian ini bersumber pada database spam-mail yang diperoleh dari UCI *Machine Learning Repository* http://www.ics.uci.edu/~mlearn/MLRepository.html. Database terdiri dari koleksi e-mail dari bulan Juni sampai Juli 1999. Database terdiri dari total 4601 *e-mail*, dimana 1813 (39.4%) adalah *spam* dan 2788 (60.6%) adalah *non-spam*. Koleksi *spam-email* berasal dari HP *e-mail* dan *spam-email* individu. Koleksi *non-spam email* berasal dari *e-mail* kantor dan *e-mail* perseorangan.

Setiap *e-mail* telah dianalisa dan terdapat 58 atribut (57 atribut input dan 1 atribut target atau kelas) yang menjelaskan tentang *spam-email.* Rincian dari atribut tersebut adalah :

1. 48 atribut bertipe *continuous* [0,100] yang beranggotakan kata. Kata yang dimaksud antara lain :

Make	*address*	*all*	*3d*
Our	*Over*	*Remove*	*Internet*
Order	*mail*	*Receive*	*Will*
People	*Report*	*Addresses*	*Free*
Business	*Email*	*You*	*Credit*
Your	*Font*	*000*	*Money*
Hp	*Hpl*	*George*	*650*
Lab	*Labs*	*telnet*	*857*
Data	*415*	*85*	*Technology*
1999	*Parts*	*Pm*	*Direct*
Cs	*Meeting*	*Original*	*Project*
Re	*Edu*	*Table*	*Conference*

   Dengan persentase:

$$\frac{\text{Jumlah Kata Yang Muncul Dalam E – mail}}{\text{Total Keseluruhan Kata Dalam E – mail}} \text{x100\%}$$

2. 6 atribut bertipe *continuous* [0,100] yang beranggotakan karakter berikut:
   ";" "(" "[" "!" "$" "#".
   Dengan persentasi :

$$\frac{\text{Jumlah Karakter Yang Muncul Dalam E – mail}}{\text{Total Keseluruhan Karakter Dalam E – mail}} \text{x100\%}$$

3. 1 atribut bertipe *continous real* [1,…] yang berisi nilai rata-rata deret huruf kapital yang tidak bisa dipecahkan.
4. 1 atribut bertipe *continous real* [1,…] yang berisi nilai terpanjang deret huruf kapital yang tidak bisa dipecahkan
5. 1 atribut bertipe *continous real* [1,…] yang berisi nilai jumlah deret huruf kapital yang tidak bisa dipecahkan

B. Transformasi data

Data yang didapat dari dataset bertipe numerik, sedangkan pengujian ini memerlukan data tipe kategori. Teknik yang digunakan untuk mengubah data numerik menjadi data kategori adalah teknik *distribusi frekuensi*. Data tipe numerik ini dikelompokkan ke dalam empat grup, yaitu 1, 2, 3, dan 4. Di mana 1 untuk rendah dan 4 untuk tinggi maka nilai 2 dan 3 berada di antara keduanya.

IV. Pengukuran Kinerja Algoritma

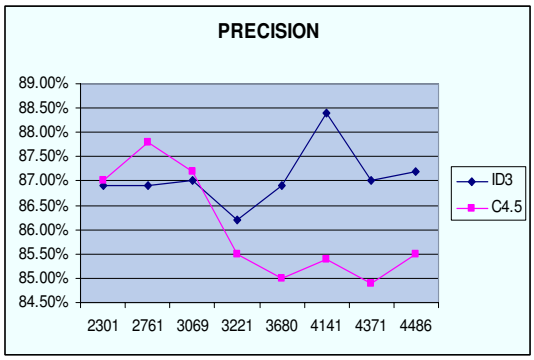
Proses penambangan data dilakukan dengan bantuan perangkat lunak data mining, yaitu WEKA. Algoritma yang diujikan adalah algoritma ID3 dan C4.5 yang berada pada modul *classify.* Pengukuran kinerja dilihat dari *spam precision*, *spam recall* dan *accuracy* .

Dalam sistem *spam filtering*, sebuah *email spam* yang salah identifikasi memiliki masalah yang tidak terlalu serius dibandingkan dengan *email non-spam* yang salah identifikasi. Dengan kata lain, salah identifikasi *email non-spam* lebih beresiko dibandingkan salah identifikasi *email spam*, maka *precision* harus besar dan *recall* juga harus besar.

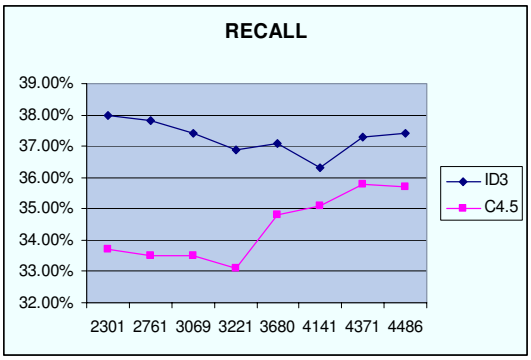
A. Pengukuran Kinerja Berdasarkan Jumlah Data

Pengukuran dilakukan berdasarkan jumlah data yang dibagi ke dalam delapan kelompok pengujian (Raghavan, 2006), yaitu kelompok yang memuat 2301 (50%), 2761 (60%), 3069 (66.7%), 3221 (70%), 3680 (80%), 4141 (90%), 4371 (95%), dan 4486 (97.5%) data. Hasil pengukuran menunjukkan algoritma ID3 mencapai nilai *precision* tertinggi pada jumlah data 4141 dengan nilai *precision* 88,4% (Gambar 2). Sedangkan algoritma C4.5 mencapai nilai *precision* tertinggi pada jumlah data 2761 (87,8%). Secara keseluruhan, algoritma ID3 menunjukkan nilai *precision* lebih tinggi dari pada algoritma C4.5, meskipun pada jumlah data 2301, 2761, dan 3069 algoritma C4.5 memiliki nilai *precision* yang lebih tinggi dibandingkan dengan algoritma ID3.

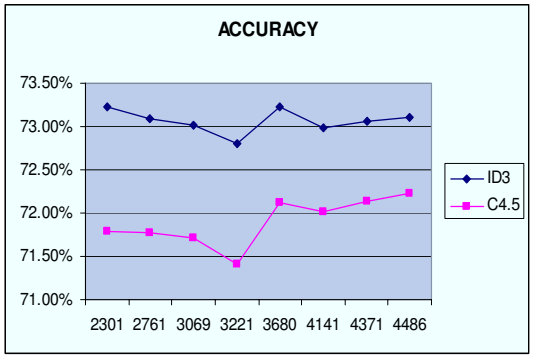
Algoritma ID3 mencapai nilai *recall* tertinggi pada jumlah data 2301 dengan nilai *recall* 38%. Sedangkan pada saat jumlah data 4141 algoritma ID3 mencapai titik terendah yaitu sebesar 36.30%. Algoritma C4.5 mencapai nilai *recall* tertinggi pada jumlah data 4486 dengan nilai *recall* 35,7%. Sedangkan pada jumlah data 3221 nilai *recall* pada algoritma C4.5 adalah yang paling rendah yaitu sebesar 33.10%. Secara keseluruhan, algoritma ID3 menunjukkan nilai *recall* lebih tinggi dibandingkan algoritma C4.5. (Gambar 3).



Gambar 2. Grafik *Precision* Berdasarkan Jumlah Data



Gambar 3. Grafik *Recall* Berdasarkan Jumlah Data



Gambar 4. Grafik Accuracy Berdasarkan Jumlah Data

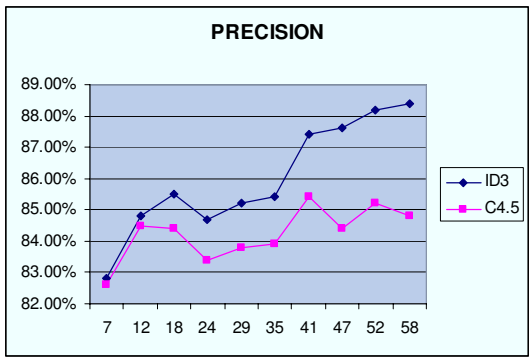
Gambar 4. menunjukkan bahwa algoritma ID3 mencapai nilai *accuracy* tertinggi pada jumlah data 2301 dan 3221 dengan nilai *accuracy* 73.23%. Sedangkan algoritma C4.5 mencapai nilai *accuracy* tertinggi pada jumlah data 4486 dari jumlah data dengan nilai *accuracy* 72.23%. Maka dapat disimpulkan bahwa nilai *accuracy* algoritma ID3 lebih baik dari pada algoritma C4.5.

Dari pengukuran kinerja kedua algoritma yang telah dilakukan berdasarkan jumlah data, maka dapat disimpulkan algoritma ID3 memiliki kinerja yang lebih baik dibandingkan algoritma C4.5.

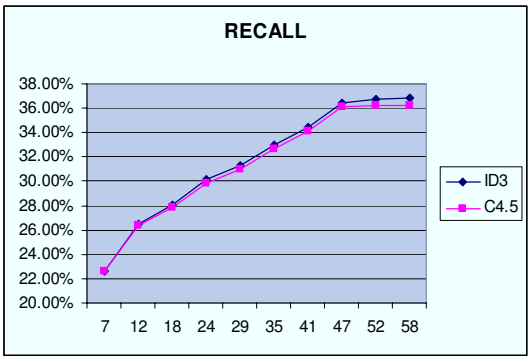
B. Pengukuran Kinerja Berdasarkan Jumlah Atribut

Selain pengukuran kinerja berdasarkan jumlah data, pengukuran kinerja juga dilakukan dengan jumlah atribut (*feature size*) dengan pemilihan atribut (*feature selection*) menggunakan x^2 *statistic* (CHI). Pemilihan atribut diambil dari nilai *chi* terbesar ke *chi* terkecil dengan jumlah persentase 10% sampai 100% dari jumlah atribut yang ada (Feng Tan , 2007). Jumlah atribut yang diperoleh adalah 7, 12, 18, 24, 29, 35, 41, 47, 52, dan 58 atribut.

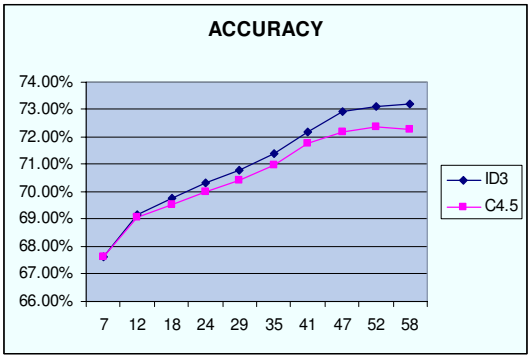
Nilai *precision* tertinggi (88.40%) dicapai oleh algoritma ID3 pada jumlah atribut 58 dari jumlah atribut yang ada. Algoritma C4.5 mencapai nilai *precision* tertinggi pada jumlah atribut 41 dari jumlah atribut dengan nilai *precision* 85.40% (Gambar 5). Secara keseluruhan, algoritma ID3 selalu berada di atas nilai *precision* algoritma C4.5.



Gambar 5. Grafik *Precision* Berdasarkan Jumlah Atribut



Gambar 6. Grafik *Recall* Berdasarkan Jumlah Atribut



Gambar 7. Grafik *Accuracy* Berdasarkan Jumlah Atribut

Algoritma ID3 mencapai nilai *recall* tertinggi pada jumlah atribut 58 dengan nilai *recall* 36.80%. Sedangkan algoritma C4.5 mencapai nilai *recall* tertinggi pada jumlah atribut 52 dan 58 dengan nilai *recall* 36.20%. Secara keseluruhan, algoritma ID3 menunjukkan nilai *recall* lebih tinggi dari pada algoritma C4.5 (Gambar 6).

Nilai *accuracy* tertinggi dicapai oleh algoritma ID3 pada jumlah atribut 58 dari dengan nilai *accuracy* 73.20%. Sedangkan algoritma C4.5 mencapai nilai *accuracy* tertinggi pada jumlah atribut 52 dari jumlah atribut dengan nilai *accuracy* 72.38%. Maka dapat disimpulkan algoritma ID3 memiliki kinerja yang lebih baik dibandingkan algoritma C4.5 (Gambar 7).

Dari pengukuran kinerja kedua algoritma yang telah dilakukan berdasarkan jumlah atribut secara keseluruhan algoritma ID3 memiliki kinerja yang lebih baik dibandingkan algoritma C4.5.

V. PENUTUP

Dari pengukuran kinerja kedua algoritma yang telah dilakukan berdasarkan jumlah data, dapat disimpulkan algoritma ID3 memiliki kinerja (*precision, recall,* dan accuracy) yang lebih baik dibandingkan algoritma C4.5. Pengukuran kinerja kedua algoritma yang dilakukan berdasarkan jumlah atribut menunjukkan algoritma ID3 memiliki kinerja (*precision, recall,* dan accuracy) yang lebih baik dibandingkan dengan algoritma C4.5. Secara keseluruhan, dari percobaan yang telah dilakukan dapat disimpulkan bahwa ID3 mempunyai kinerja yang lebih baik dibandingkan algoritma C4.5

Pengukuran kinerja sebuah algoritma *data mining* dapat dilakukan berdasarkan beberapa kriteria antar lain seperti keakuratan prediksi, kecepatan/efisiensi, kehandalan, skalabilitas dan interpretabilitas. Penelitian ini menggunakan satu kriteria yaitu berdasarkan keakuratan prediksi. Dengan demikian penelitian lain dengan menggunakan kriteria lain dapat dilakukan.

DAFTAR PUSTAKA

1. Anonim, *A Data Mining Glossary,* http://www.thearling.com/index.htm tanggal 28 Agustus 2008.
2. *What is data mining - A Word Definition From the Webopedia Computer Dictionary*, http://www.webopedia.com/TERM/D/, tanggal 28 Agustus 2008.
3. Garcia, Flavio, Jaap-Henk Hoepman, dan Jeroen van Nieuwenhuizen. *Spam Filtering Analysis*. http://citeseer.ist.psu.edu/644060.html, 2004
4. Lambert, Anselm, *Analysis of Spam*, A dissertation in Computer Science at University of Dublin, 2003.
5. Raghavan, Ratheesh, *Study Of The Relationship Of Training Set Size To Error Rate In Yet Another Decision Tree And Random Forest Algorithms,* A Thesis In Computer Science at Texas Tech University, 2006.
6. Riza, Ramadan, *Penerapan Pohon Untuk Klasifikasi Dokumen Teks Berbahasa Inggris,* Program Studi Teknik Informatika Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung, 2006
7. S. Youn, D. Mcleod, *A Comparative Study for Email Classification*. Proceedings of International Joint Conferences on Computer, Information, System Sciences and Engineering, Bridgeport, CT, (2006).
8. Sheu, Jyh-Jian. *An Efficient Two-phase Spam Filtering Methode Based on E-mails categorization.* International Journal of Network Security, Vol. 8, No. 3, PP.334-343, Taiwan, May 2008
9. Tan, Feng, *Improving Feature Selection Techniques For Machine Learning,* Georgia Stage University, 2007
10. Teuku, Hasbullah, *Spam Filtering MenggunakanJaringan Syaraf Tiruan,* Tugas Akhir Jurusan Teknik Informatika Sekolah Tinggi Teknologi Telkom, 2005.
11. UCI *Machine Learning Repository* http://www.ics.uci.edu/~mlearn/MLRepository.html (diakses tanggal 17 september 2008)
12. Yudho Giri Sucahyo, *Data mining – Menggali Informasi Yang terpendam*, http://ikc.cbn.net.id/populer/yudho/yudho-datamining.zip, tanggal 07 Juni 2008.
13. http://lecturer.eepis-its.edu/~tessy/lecturenotes/datamining/ (diakses tanggal 20 Agustus 2008)