Evaluasi Sistem Pendeteksi Intrusi Berbasis Anomali dengan N-Gram dan

Incremental Learning

I Made Agus Adi Wirawan, Royyana Muslim Ijtihadie, dan Baskoro Adi Pratomo  
Jurusan Teknik Informatika, Fakultas Teknologi Informasi  
Institut Teknologi Sepuluh Nopember (ITS)  
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia  
*e-mail*: {roy, baskoro}@if.its.ac.id

***Abstrak*—*Keberadaan teknologi informasi yang terus berkembang dengan pesat menjadikan kebutuhan akan penggunaannya semakin hari semakin meningkat. Transaksi data melalui internet telah menjadi kebutuhan wajib hampir dari semua perangkat lunak yang ada saat ini. Perangkat lunak seperti media social, colud server, online game, aplikasi layanan pemerintah, aplikasi pengontrol suatu tempat secara remote, dsb. Tentu dengan berbagai macam penggunaan internet tersebut dibutuhkan metode untuk mengamankan jaringannya.***

***Sistem pendeteksi intrusi atau yang pada umumnya disebut IDS (Intrusion Detection System) merupakan solusi untuk mengamankan suatu jaringan. Sistem ini nantinya bertugas untuk menentukan apakah suatu paket merupakan bentuk serangan atau paket biasa sesuai dengan kondisi tertentu. Saat ini telah banyak dikembangkan aplikasi IDS (Intrusion Detection System), namun sebagian besar yang dikembangkan berbasis signature atau menggunakan rule, dan sebagaian kecil menggunakan anomali. Anomali adalah suatu metode untuk mencari penyimpangan dalam sebuah data.***

***Pada aplikasi ini konsep IDS yang diterapkan adalah IDS berbasis anomali dimana analisis datanya pada infromasi paket data yang dikirimkan. Pada tugas akhir ini menggunakan dua metode, yaitu metode n-gram yang digunakan untuk mengitung distribusi byte karakter pada paket data sedangkan metode mahalanonis distance digunakan untuk menghitung jarak antara paket data normal dan paket data yang berupa intrusi.***

***Metode mahalanobis distance dapat membedakan paket data yang normal dan paket data yang berupa intrusi dengan menghitung rata-rata dan standar deviasi dari paket data.***

*Kata Kunci*—*N-Gram*, Mahalanobis Distance, Incremental Learning.

# PENDAHULUAN

S

EMAKIN pesatnya perkembangan teknologi informasi memudahkan orang-orang untuk saling tukar menukar data baik melalui internet maupun intranet. Tentunya dengan mudahnya berbagi data itulah sangat memungkinkan terjadinya serangan terhadap data tersebut terutama melalui jaringan komputer. Sistem pendeteksi intrusi atau yang pada umumnya disebut IDS (*Intrusion Detection System*) [1] merupakan senjata utama untuk mengamankan suatu jaringan dimana sistem ini nantinya bertugas untuk mengidentifikasi dan mencatat apakah suatu paket data tersebut merupakan bentuk serangan atau paket data bisa.

Saat ini telah banyak dikembangkan aplikasi IDS (*Intrusion Detection System*), namun sebagian besar yang dikembangkan berbasis *signature* atau menggunakan *rule*, dan sebagian kecil menggunakan *anomaly*. *Anomaly* pada dasarnya adalah mencari data yang menyimpang dari sekumpulan data normal. IDS yang berbasis pada *anomaly* [2] bersifat lebih fleksibel, karena dapat mengenali pola serangan baru tanpa harus memperbaharui basis data pola serangan. IDS yang berbasis pada anomali memiliki sebuah kecerdasan buatan yang mampu mendeteksi dan mengenali sebuah serangan. IDS yang berbasis anomali menggabungkan metode analisis dan statistik untuk mengenali penyimpangan tersebut. Kelemahan dari metode ini adalah kemungkinan salah identifikasi pada data yang diolah.

Sistem kerja intrusi ini pada dasarnya dikirimkan lewat jaringan dengan paket-paket data yang sama dengan paket data normal. Dengan banyaknya paket data yang masuk kedalam sebuah host, tentunya host ini harus bisa mengenali paket data, apakah paket data tersbut terdapat paket data yang berupa intrusi atau tidak. Hal tersebut dapat dikenali dengan cara mengelompokkan data berdasarkan beberapa hal yang membedakan antara paket data normal dengan paket data yang berupa intrusi.

Maka untuk membedakan hal tersebut diperlukan sebuah sistem deteksi intrusi dimana nantinya sistem deteksi intrusi tersebut menggunakan gabungan metode analisis dan statistik yang berfungsi mengenali perbedaan paket data normal maupun paket data berupa intrusi. Selain itu, sistem deteksi intrusi yang dapat mempelajari paket data normal yang baru sebagai data *training*.

Untuk dapat menghitung jarak mahalanobis dari paket data, diperlukan metode yang dapat merubah informasi paket data menjadi nilai yang dapat dihitung. Metode n-gram dapat digunakan untuk mebuat model yang sederhana dan cepat untuk dihitung khususnya menghitung distribusi karakter pada suatu paket data. metode n-gram merupakan metode yang paling efesien dan efektif dalam membuat model dari suatu paket data.

Metode mahalanobis *distance* berguna untuk membedakan paket-paket data berdasarkan anomali yang terjadi. Untuk dapat mempelajari paket data normal yang baru menggunakan metode *incremental learning*, dimana metode ini nantinya memperbaharui rata-rata dan standar deviasi dari model paket data yang ada pada data *training*.

# DESAIN DAN PERANCANGAN

## Alur Kerja Sistem Secara Umum

Secara umum sistem yang dibangun terdiri dari tiga proses utama, yaitu proses *training* data set, proses *sniffing* dan proses identifikasi serangan. Alur kerja sistem secara umum dapat dilihat pada Gambar 1.

Cara kerja sistem yang lebih detail yaitu, membaca *file* data set, membuat model data *training*, menyimpan model data *training*, menangkap paket, memproses paket dan membandingkan jarak mahalanobis paket dengan model serta pengambilan keputusan terhadap hasil perbandingan. Membuat model data *training* adalah proses dimana membaca *file* paket data dan ditampung pada *array of object.* *File* yang dapat dibaca hanya *file* yang berekstensi *\*.cap*, \*.*pcap*, *\*.tcpdump* dengan memanfaatkan *library* Jpcap [2]. Proses selanjutnya adalah menangkap *packet* dari *network* interface dengan bantuan library Jpcap lalu menyimpannya pada *array of object.* Selanjutnya adalah proses membandingkan jarak mahalanobis *packet* dengan model data. *Packet* yang dihitung hanya *packet* yang memiliki port tujuan yang kurang dari 1024. Setelah terdapat *packet* yang memenuhi syarat tersebut akan dilakukan proses perhitungan, mulai dari perhitungan rata-rata dan standar deviasi dari *packet*. Dan selanjutnya proses pemanggilan fungsi Mahalanobis *Distance* untuk menghitung jarak mahalanobis *packet* dengan model data *training*. Setelah mendapatkan jarak mahalanobis, lalu dibandingkan dengan nilai *threshold* yang sudah ditentukan sebelumnya. Nilai *threshold* setiap port memiliki besaran yang berbeda. Jika jarak mahalanobis *packet* melebihi nilai *threshold,* maka *packet* tersebut dapat dikategorikan paket yang tidak normal.

Proses diatas diulangi sampai aplikasi dihentikan oleh pengguna dan menulis hasil keputusan terhadap proses perbandingan ke sebuah *file log*.



Gambar 1 Diagram alir kerja sistem secara umum

## N-Gram

Pada dasarnya, model N-Gram [3] adalah model probabilistik yang awalnya dirancang oleh ahli matematika dari Rusia pada awal abad ke-20 dan kemudian dikembangkan untuk memprediksi *item* berikutnya dalam urutan *item.* Item bisa berupa huruf / karakter, kata, atau yang lain sesuai dengan aplikasi. Salah satunya, model *n-gram* yang berbasis kata digunakan untuk memprediksi kata berikutnya dalam urutan kata tertentu. Dalam arti bahwa sebuah *n-gram* hanyalah sebuah wadah kumpulan kata dengan masing-masing memiliki panjang *n* kata. Sebagai contoh, sebuah *n-gram* ukuran 1 disebut sebagai *unigram;* ukuran 2 sebagai *bigram*; ukuran 3 sebagai *trigram*, dan seterusnya.

Pada pembangkitan karakter, *N-gram* terdiri dari *substring* sepanjang *n* karakter dari sejumlah *string* dalam definisi lain *n-gram* adalah potongan sejumlah *n* karakter dari sebuah *string.* Metode *n-gram* ini digunakan untuk mengambil potongan-potongan karakter huruf sejumlah *n* dari sebuah kata secara kontinuitas dibaca dari teks sumber sehingga akhir dari dokumen. Sebagai contoh : kata “TEXT” dapat diuraikan ke dalam beberapa *n-gram* berikut :

*uni-gram* :T, E, X, T

*bi-gram* : TE, EX, XT

*tri-gram* : TEX, EXT

*quad-gram* : TEXT, EXT\_

dan seterusnya.

Sedangkan pada pembangkit kata, metode *n-gram* ini digunakan untuk mengambil potongan kata sejmlah *n* dari sebuah rangkaian kata (kalimat, paragraf, bacaan) yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen. Sebagai contoh : kalimat “saya dapat melihat cahaya itu.” Dapat diuraikan ke dalam beberapa *n-gram*  berikut :

*uni-gram* :saya, dapat , melihat, cahaya, itu

*bi-gram* : saya dapat, dapat melihat,

melihat cahaya, cahaya itu

*tri-gram* : saya dapat melihat, dapat melihat

cahaya, melihat cahaya itu\_

dan seterusnya.

Salah satu keunggulan menggunakan *n-gram* dan bukan suatu kata utuh secara keseluruhan adlah bahwa *n-gram* tidak terlalu sensitif terhadap kesalahan penulisan yang terdapat pada suatu dokumen.

## Simplified Mahalanobis Distance

Mahalanobis distance [3] adalah sebuah metode statistika untuk menghitung jarak antara titik P dan distribusi D. Prinsip Mahalanobis *Distance* adalah mengitung jarak di ruang multidimensional antara sebuah pengamatan dengan pusat dari semua pengamatan. Pada Tugas Akhir ini Mahalanobis *Distance* digunakan untuk menghitung jarak antara distribusi *byte* karakter dari payload baru terhadap model yang ada pada data *training*. Semakin jauh jaraknya, semakin besar kemungkinan payload ini tidak normal.

Mahalanobis *distance* dari sebuah payload baru dapat dihitung jika sistem sudah mempunyai data *training*. Selanjutnya menghitung rata-rata dan standar deviasi dari model yang ada pada data *training*. Untuk menghitung rata-rata dari model yang ada pada data *training* dapat dilihat pada persamaan 1. Sedangkan untuk menghitung standar deviasi dari model yang ada pada data *training* dapat dilihat pada persamaan 2. Setelah selesai menghitung rata-rata dan standar deviasi dari model yang ada pada data *training* baru dapat menghitung jarak mahalanobis dari payload baru dengan menggunakan persamaan 1. Format data kasar yang ada pada Mahalanobis *disatance* dapat dilihat pada Tabel 1.

(1)

dimana,

= jarak mahalanobis

= variable ke-i dari *payload* baru

= rata-rata variable ke-i dari model data *training*

= standar deviasi variable ke-i dari model data *training*

= *smoothing factor*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Variabel (karakteristik) | | | | | | |
| Object |  |  | … |  | … |  |  |
| 1 | . | . | … | . | … | . | . |
| 2 | . | . | … | . | … | . | . |
| 3 | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| K |  |  | … |  | … |  |  |
| . | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| N |  |  | … |  | … |  |  |
| Average |  |  | … |  | … |  |  |
| Standar deviation |  |  | … |  | … |  |  |

Tabel 1 Format data kasar didalam Mahalanobis Distance

Persamaan untuk mencari rata-rata, yaitu:

(2)

dimana,

= rata-rata variabel ke-i

= jumlah object model

= nilai variabel ke-i

Persamaan untuk mencari nilai standar deviasi, yaitu:

(3)

dimana,

= standar deviasi variabel ke-i

= nilai dari variabel ke-i

= rata-rata variabel ke-i

= jumlah object model

## Incremental Learning

*Incremental Learning* merupakan proses untuk memperbaharui nilai rata-rata dan standar deviasi dari model yang ada pada data *training* ketika menambahkan payload baru. Proses ini diperlukan untuk meningkatkan akurasi dari setiap model ketika ditambah data sampel baru.

Untuk menghitung Mahalanobis *distance* versi *Incremental Learning* diperlukan rata-ratadan standar deviasi dari masing-masing karakter ASCII untuk setiap sampel baru yang dihitung. Untuk menghitung rata-rata dari sebuah karakter dapat dilihat pada persamaan 2. Selanjutnya agar dapat memperbaharui nilai rata-rata dari model yang ada pada data *training*, diperlukan jumlah sampel yang telah dihitung sebelumnya [6]. Untuk menghitung nilai rata-rata yang baru dapat dilihat pada persamaan 4.

Sedangkan untuk mengitung standar deviasi yang baru diperlukan rata-rata dari pada model sebelumnya. Untuk menghitung standar deviasi yang baru dapat dilihat pada persamaan 5.

Persamaan untuk menghitung rata-rata baru dari model yang diamati, yaitu:

(4)

dimana,

=rata-rata baru

= nilai dari variabel yang baru

=jumlah sampel sebelumnya

Persamaan untuk menghitung standar deviasi baru dari model yang diamati, yaitu:

(5)

dimana,

= standar deviasi variabel ke-i

= nilai dari variabel ke-i

= nilai dari variabel yang baru

= jumlah object model

# UJI COBA DAN EVALUASI

## Skenario Pengujian

Pengujian yang dilakukan dibagi kedalam dua bagian, yakni:

1. Pengujian akurasi tanpa menambahkan proses *incremental learning*; dan
2. Pengujian akurasi dengan menambahkan proses *incremental learning*.

Parameter pengujian yang digunakan adalah besarnya ukuran *window*. *Window size* yang dimasksud adalah jumlah paket data yang ditangkap, jika jumlah paket data yang ditangkap sudah memenuhi *window size* selanjutnya paket data tersebut diproses untuk mengetahui jarak mahalanobis setiap paket. Data uji yang digunakan adalah *file* paket data hasil tangkapan paket data pada jaringan eksternal DARPA [4] pada minggu ke-4 yang berjumlah 5 *file* paket data.

(6)

## Data Pengujian

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **Ukuran *window*** | **Port TCP** | | | | **Port UDP** |
| 1 | 10000 | 21 | 23 | 25 | 80 | 53 |
| 2 | 15000 | 21 | 23 | 25 | 80 | 53 |
| 3 | 20000 | 21 | 23 | 25 | 80 | 53 |

Tabel 2 Data uji

## Hasil Pengujian

### Pengujian akurasi tanpa menambahkan proses incremental learning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | ***Window size*** | **Jumlah connection** | **Jumlah**  **Paket normal** | **Jumlah**  **Paket serangan** |
| 1 | 10000 | 5328 | 5308 | 20 |
| 2 | 20000 | 8515 | 8478 | 37 |

Tabel 3 Hasil Uji Data Testing minggu ke-5 tanpa proses incrmental learning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | ***Window size*** | **Kelas** | | | |
| **A** | **B** | **C** | **D** |
| 1 | 10000 | 6 | 355 | 14 | 4935 |
| 2 | 20000 | 15 | 552 | 22 | 7926 |

Tabel 4 Confussion matrix uji coba 1

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Jenis Penilaian** | **Nilai** | **Persentase** |
| 1 | Akurasi (AC) | 0.9307 | 93.07% |
| 2 | *True positive rate* (TP) | 0.0166 | 1.66% |
| 3 | *False negative rate* (FN) | 0.9834 | 98.34% |
| 4 | *False positive rate* (FP) | 0.0028 | 0.28% |
| 5 | *True negative rate* (TN) | 0.9972 | 99.72% |
| 6 | Presisi (P) | 0.3 | 30.0% |

Tabel 5 Hasil penilaian percobaan 1

### Pengujian akurasi dengan menambahkan proses incremental learning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | ***Window size*** | **Jumlah connection** | **Jumlah**  **Paket normal** | **Jumlah**  **Paket serangan** |
| 1 | 10000 | 5328 | 1577 | 3751 |
| 2 | 20000 | 8515 | 6711 | 1804 |

Tabel 6 Hasil Uji Data Testing minggu ke-5 dengan proses incremental learning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | ***Window size*** | **Kelas** | | | |
| **A** | **B** | **C** | **D** |
| 1 | 10000 | 187 | 174 | 3564 | 1403 |
| 2 | 20000 | 314 | 253 | 6397 | 1551 |

Tabel 7 Confussion matrix uji coba 2

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Jenis Penilaian** | **Nilai** | **Persentase** |
| 1 | Akurasi (AC) | 0.2984 | 29.84% |
| 2 | *True positive rate* (TP) | 0.518 | 51.8% |
| 3 | *False negative rate* (FN) | 0.482 | 48.2% |
| 4 | *False positive rate* (FP) | 0.7175 | 71.75% |
| 5 | *True negative rate* (TN) | 0.2825 | 28.25% |
| 6 | Presisi (P) | 0.0499 | 4.99% |

Tabel 8 Hasil penilaian percobaan 2

# KESIMPULAN

## Kesimpulan

Dalam proses pengerjaan Tugas Akhir yang melalui tahap perancangan, implementasi, serta uji coba, didapatkan kesimpulan sebagai berikut :

1. Sistem deteksi intrusi yang dibuat dapat membaca data set dari DARPA IDS tahun 1999 dengan bantuan *library* Jpcap.
2. Sistem deteksi intrusi yang dibuat dapat menangkap paket data dari *network interface* secara *real-time* dengan bantuan *library* Jpcap.
3. Sistem deteksi intrusi yang dibuat dapat menerapkan metode *n-gram* dalam membuat model sebuah paket data.
4. Metode Mahalanobis *Distance* dapat mengklasifikasikan antara paket data normal dan paket data yang berupa intrusi, namun tidak berlaku untuk protokol HTTP.
5. Sistem yang dibuat untuk pendeteksi intrusi menggunakan metode Mahalanobis *Distance* tanpa proses *incremental learning* dapat mendeteksi intrusi dengan persentase kebenaran sekitar 93%, namun dengan tambahan proses *incremental learning* hanya dapat mendeteksi intrusi dengan persentase kebenaran sekitar 20%.

## Saran

Adapun saran-saran yang diberikan untuk pengembangan sistem ini selanjutnya adalah karena membedakan paket data normal dengan paket data serangan menggunakan metode mahalanobis *distance* dengan proses *incremental learning* kurang akurat dibandingkan tanpa proses *incremental leraning*. Hal ini dikarenakan dengan menambahkan proses *incremental learning,* rata-rata dan standar deviasi pada model diperbaharui tetapi *threshold* yang digunakan untuk mendeteksi intrusi tidak diperbaharui, sehingga *threshold* yang ada tidak akurat untuk mendeteksi intrusi. Perlu ada implementasi metode lain sehingga dapat membantu meningkatkan keakuratan pendeteksian intrusi.

DAFTAR PUSTAKA

|  |  |
| --- | --- |
| [1] | SANS Institute, “Understanding Intrusion Detection System,” *SANS Institute Reading Room,* pp. 1-9, 2001. |
| [2] | K. Fuji, “a Java library for capturing and sending network packets,” Jpcap, 15 May 2007. [Online]. Available: http://jpcap.gitspot.com/. [Diakses 23 May 2016]. |
| [3] | A. Hanafi, “Pengenalan Bahasa Suku Bangsa Indonesia Berbasis Teks Menggunakan Metode N-gram. IT TELKOM,” 2009. |
| [4] | “Mahalanobis distance,” [Online]. Available: https://en.wikipedia.org/wiki/Mahalanobis\_distance. [Diakses 22 June 2016]. |
| [5] | D. E. Knuth, “The Art of Computer Programming,” *Fundamental Algorithms. Addison Wesley,* vol. 1, 1973. |