Penggunaan Algoritma Boyer Moore untuk Memindai Berkas dari Virus

Fajar Nugroho - 13515060
Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung Jalan Ganesha 10 Bandung 40132
13515060@std.stei.itb.ac.id

***Abstract*—Algoritma Boyer Moore merupakan algoritma pencocokan pola yang cukup terkenal. Algoritma ini berfungsi mencocokkan suatu pola dengan sampel yang ada. Kemampuan algoritma ini dapat diterapkan pada pendeteksian berkas komputer yang terindikasi terinfeksi oleh virus komputer. Dengan mengetahui pola dari sebuah virus, anti virus dapat mendeteksi keberadaan sebuah virus pada berkas dengan melakukan pencocokan pola. Pada dasarnya pemindaian berkas oleh anti virus yaitu dengan membaca berkas dalam bentuk binary string kemudian mencocokkannya dengan database pola virus. Serangan cyber yang marak dewasa ini membuat masyarakat lebih berhati-hati dalam menggunakan komputer. Karena virus dapat menyerang komputer siapa saja dan penyebarannya sangat cepat melalui internet saat ini.**

Karena algoritma untuk pencocokan pola di dalam berkas banyak, maka hendaknya kita seharusnya mempertimbangkan kompleksitas algoritma yang ada dengan menganalisis kelebihan dan kekurangan dari algoritma pencarian string agar diperoleh efisiensi di dalam pencarian kata pada program aplikasi.

Keywords—Boyer Moore, Virus, Malware, Cyber, komputer, internet.

I. PENDAHULUAN

Dewasa ini sering terjadi kejahatan di dunia maya atau yang lebih dikenal dengan *Cyber Crime*. Hal tersebut bukan kejahatan musiman yang hanya terjadi beberapa saat namun dapat terjadi kapan saja dan di mana saja tanpa mengenal korban, waktu dan tempat karena pada dasarnya dunia maya – atau istilahnya cyber – merupakan jaringan global dan semua sektor membutuhkannya untuk menunjang keberjalanan sektor tersebut.

Baru-baru ini terjadi serangan ransomware yang menjangkiti komputer-komputer yang menjalankan Sistem Operasi Windows. Ransomware tersebut mengenkripsi berkas-berkas dan pembuatnya meminta tebusan sejumlah uang agar korban bisa mendapatkan berkasnya kembali. Hal tersebut belum terlalu berbahaya jika dibandingkan malware yang dapat mengambil data-data penting ataupun mengendalikan komputer tersebut secara diam-diam.

Serangan-serangan siber harus dihindari dan dicegah sedini mungkin karena saat ini semua sektor kehidupan di dunia bergantung pada teknologi komputer. Untuk mencegah serangan-serangan yang berupa penyusupan program-program berbahaya dapat menggunakan anti virus yang bekerja dengan mencocokkan isi sebuah berkas dengan sampel virus atau malware.

Saat ini, anti virus sudah banyak beredar di kalangan umum mulai dari anti virus lokal yang menangani virus-virus biasa sampai anti virus yang dapat menangani virus yang agresif. Dari sekian banyaknya anti virus ada yang bersifat freeware dan ada yang bersifat shareware. Perbedaan anti virus gratis dan berbayar hanya terletak pada fiturnya saja sedangkan kinerja dari anti virus tersebut masih tetap sama. Penggunaan anti virus merupakan hal yang paling efektif dalam mencegah penyebaran virus di dalam sistem komputer.

Melalui mata kuliah IF2211 Strategi Algoritma, saya akan melakukan penelitian tentang penggunaan algoritma pencocokan pola boyer moore yang memeriksa berkas apakah mengandung pola-pola tertentu yang mengindikasikan virus atau program berbahaya.

II. DASAR TEORI

1. Virus

Virus komputer adalah program komputer yang dapat menggandakan dan menyisipkan dirinya ke berkas ataupun ke program komputer yang sedang berjalan dan bersifat merusak. Umumnya virus komputer dibuat untuk tujuan yang tidak baik, antara lain merusak sistem komputer, memanipulasi data-data di dalam komputer, mencuri informasi dan data di dalam komputer dan lain sebagainya. Virus komputer terdiri bermacam-macam jenis sesuai cara kerjanya dan tujuan virus

tersebut diciptakan. Jenis-jenis virus yang umum adalah *trojan horse*, *backdoor*, dan *worm.*

2. *Trojan Horse*

Trojan horse adalah virus komputer yang berbentuk program dan jika dieksekusi akan mengirimkan data dan informasi kepada pembuat virus.

3. *Backdoor*

Backdoor biasanya berbentuk program yang jika dieksekusi maka pembuat backdoor dapat mengendalikan komputer yang terinfeksi secara *remote* (jarak jauh melalui koneksi internet) dan korban tidak menyadarinya.

4. *Worm*

Worm bekerja dengan cara menduplikasikan dirinya pada media penyimpanan sehingga penyimpanan akan penuh dengan worm tersebut dan memperlambat kinerja komputer.

5. Anti virus

Anti virus adalah program yang digunakan untuk menangkal terjadinya infeksi virus dalam sebuah sistem komputer. Anti virus bekerja dengan memindai berkas-berkas dan mencocokkan isi berkas dengan database pola virus yang dimilikinya. Selain menggunakan teknik pemindaian anti virus juga menggunakan metode lain untuk mendeteksi keberadaan virus, yaitu dengan static heuristic dan integrity check namun keakuratan kedua metode tersebut jauh lebih kecil dibandingkan dengan metode pemindaian yang secara langsung mengetahui isi dari berkas apakah mengandung kode virus atau tidak.

6. Algoritma Boyer Moore

Algoritma Boyer Moore mempunyai empat konsep dasar di dalam proses pencarian string, yaitu :

a. Preprocessing

b. Right-to-left-scan

c. Bad-character-rule

d. Good-suffix-rul

Algoritma Boyer-Moore diperkenalkan oleh Bob Boyer dan J.S. Moore pada tahun 1977. Algoritma ini merupakan algoritma pencocokan string yang cukup terkenal dan lebih efisien dibandingkan algoritma pencocokan pola KMP dan brute force. Pada algoritma ini pencocokan kata dimulai dari karakter terakhir kata kunci menuju karakter awalnya. Jika terjadi perbedaan antara karakter terakhir kata kunci dengan kata yang dicocokkan maka karakter-karakter dalam potongan kata yang dicocokkan tadi akan diperiksa satu per satu. Hal ini dimaksudkan untuk mendeteksi apakah ada karakter dalam potongan kata tersebut yang sama dengan karakter yang ada pada kata kunci. Apabila terdapat kesamaan, maka kata kunci akan digeser sedemikian rupa sehingga posisi karakter yang sama terletak sejajar, dan kemudian dilakukan kembali pencocokan karakter terakhir dari kata kunci. Sebaliknya jika tidak terdapat kesamaan karakter, maka seluruh karakter kata kunci akan bergeser ke kanan sebanyak m karakter, di mana m adalah panjang karakter dari kata kunci.

Algoritma Pattern Matching Booyer-Moore ini berbasis pada 2 metode yaitu :

a. The Looking-Glass Technique

The Looking-Glass Technique melakukan perbandingkan suatu karakter akhir pada kata w dengan suatu karakter pada teks s. Jika karakter tersebut sama maka jendela karakter akan berjalan mundur pada kedua string dan mengecek kembali kedua karakter. Mencari Suatu kecocokan String pada Teks dengan pola yang akan dicari dengan cara memindahkan atau menggesernya sampai Teks string selesai.

b. The Character-Jump Technique

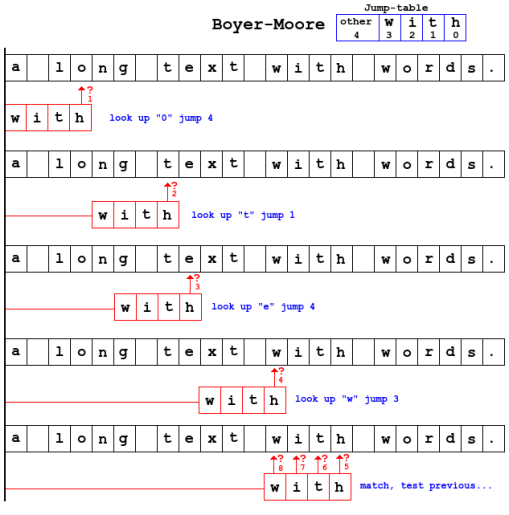
Character-jump Technique melakukan suatu aksi ketika perbandingan antara dua karakter yang berbeda. Ada dua aksi yang tergantung pada teks s dan kata w yang dimiliki; jika p yaitu karakter pada s yang sedang diproses yang tidak cocok maka ada dua kemungkinan aksi. Mencari karakter yang sesuai dan cara penggeseran sebuah karakter perbandingan terakhir.

Dalam algoritma boyer moore ada beberapa kasus yaitu :

Kasus 1 : Jika P mengandung karakter T[i] maka geser P sehingga T[i] = P[j].

Kasus 2 : Jika P mengandung karakter T[i] namun tidak memungkinkan menggeser P maka atur i menjadi i+1 dan j kembali ke P paling akhir.

Kasus 3 : Jika kasus 1 dan 2 tidak terpenuhi maka geser P sehingga P[0] sejajar dengan T[i+1].



III. STUDI KASUS

A. *Pemindaian Berkas yang Mengandung Kode "VIRUS"*

Langkah pertama dalam pemindaian ini yaitu membuka berkas dan membaca semua isinya sebagai binary dan mengubahnya ke bentuk string dengan enkoding utf-8 sehingga akan terdapat beberapa bagian yang tidak sempurna dibaca menjadi string dan terbaca sebagai kode heksadesimal diawali dengan *escape character '\x'*. Pengubahan kode biner menjadi string memudahkan dalam pencocokan pola dan tidak perlu membongkar berkas menjadi bentuk lain yang bias merusak isi berkas. Selain itu, Karena semua berkas belum tentu berbentuk sama, misal berkas media, program atau teks namun semua berkas tersebut disimpan dalam biner di dalam media penyimpanan sehingga kita bias mengubahnya menjadi string. Di bawah ini merupakan contoh berkas uji yang berbentuk teks biasa yang nantinya akan dibaca oleh program sebagai binary file dan di dekode menjadi string :

```
.ELF.......................4..
........4. ...(........4...4.
..4... ... ..............T...T
...T..........................
..............................
........ ...8................
.............................h
...h...h...D...D..........P.td
0...0...0...\...\..........Q.t
d............................R.
td............................/
lib/ld-linux.so.2.............
GNU........... ..............
.GNU..h94.EU..^e3.......J......
..........#. ........)....K...
VIRUS........................
..............&.............
.2..............K.......... .
..9...............+...(........
....................libc.so.6._
IO_stdin_used.gets.fflush.puts.
```

Dengan menggunakan algoritma boyer moore didapatkan bahwa berkas diatas mengandung kode sesuai sampel yang digunakan. Dalam program anti virus yang asli kode tersebut tidak hanya satu tetapi bisa banyak dan berada di lokasi yang terpisah pisah. Dibawah ini merupakan contoh program yang dibuat untuk mendeteksi keberadaan virus dengan memanfaatkan algoritma boyer moore dan hasil yang menandakan berkas "sample" terinfeksi virus.

```
with open(fileName, mode='rb') as file:
    fileContent = file.read()
text = fileContent.decode('utf-8', 'backslashreplace')
pattern = "VIRUS"
show_match(text, pattern)
```

Kode program untuk mencari kode "VIRUS"

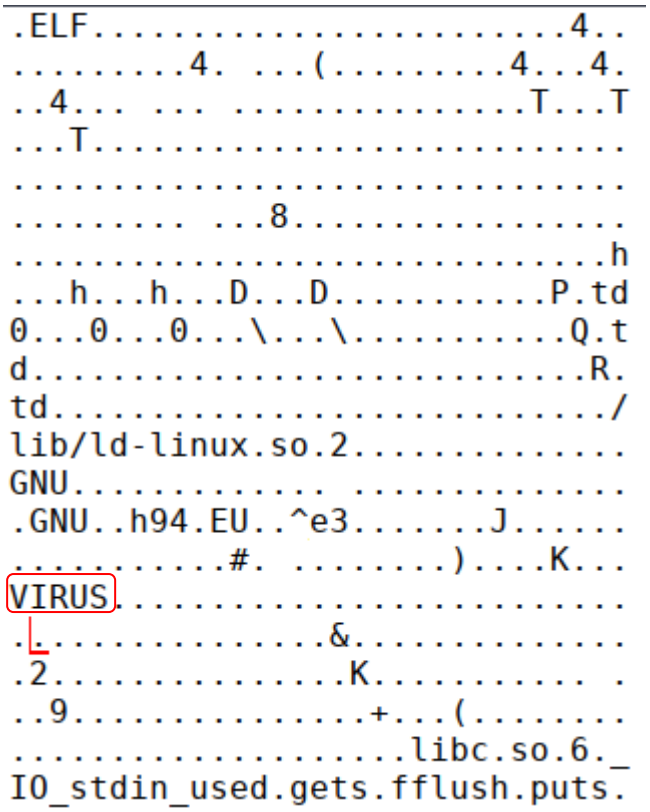
```
fajar@ArchF  ~/stima  python antivirus.py sample

Berkas terinfeksi virus

fajar@ArchF  ~/stima 
```

Hasil Eksekusi dari program yang menemukan kode "VIRUS" pada berkas bernama "sample"

Dari hasil uji di atas dapat disimpulkan bahwa berkas "sample" mengandung kode "VIRUS". Sejauh ini program hanya dapat memeriksa string dengan rentang 'a-z', 'A-Z' dan '0-9' saja. Namun pada program anti virus yang beredar di kalangan umum saat ini sudah lebih canggih sehingga dapat mencari sebuah pola virus yang jauh lebih rumit. Selain dengan pencocokan pola anti virus juga menggunakan metode heuristik yang sangat membantu anti virus memindai berkas lebih cepat karena melakukan pendekatan lain yang meminimalkan ruang pemindaian dengan mengamati ukuran berkas, perilaku ketika dijalankan dan lain sebagainya.

B. *Pemindaian Berkas yang Mengandung Kode "SAFE"*

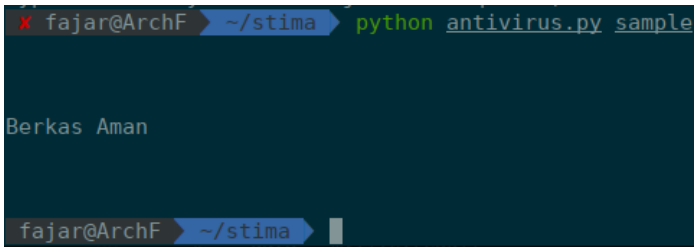
Langkah-langkah yang digunakan masih sama yaitu membuka berkas yang akan dipindai sebagai *binary* kemudian melakukan decoding dan mengubahnya menjadi string, lalu menggunakan algoritma boyer moore mnecari kode SAFE dalam berkas tersebut. Bila ditemukan adanya kode "SAFE" dalam berkas maka program akan menampilkan hasil yang menandakan bahwa berkas terinfeksi, jika tidak maka sebaliknya., program akan menampilkan hasil bahwa berkas yang dipinda aman dari infeksi virus.



Isi berkas yang digunakan masih sama tidak diubah sama sekali. Dan berikut potongan kode program yang akan memindai berkas.

```
with open(fileName, mode='rb') as file:
    fileContent = file.read()
text = fileContent.decode('utf-8', 'backslashreplace')
pattern = "TROJAN"
show_match(text, pattern)
```

Kode program untuk mencari kode “SAFE”



Hasil Eksekusi dari program yang tidak menemukan kode “SAFE” pada berkas bernama “sample”

Kode program yang digunakan masih sama hanya mengubah pola “VIRUS” menjadi “SAFE”. Dan pada hasil eksekusi dinyatakan bahwa berkas tersebut aman.

IV. Pembahasan

A. *Antivirus dengan algoritma boyer moore*

Pada dasanya anti virus menggunakan beberapa algoritma sekaligus untuk mendeteksi keberadaan virus. Selain pencocokan pola yang digunakan untuk memindai sebuah berkas ada metode lain yang digunakan yaitu static heuristic dan *integrity check.*

a. *Statik Heuristik*

Bila teknik pemindaian digunakan untuk melakukan deteksi terhadap malware yang sudah dikenali karakteristiknya, maka teknik Heuristic umumnya digunakan untuk malware yang belum dikenali karakteristiknya. Teknik ini tidak melakukan pencarian *signature* malware, tapi akan berusaha membuat *signature* baru. Jadi teknik ini berusaha menduplikasi cara seorang analis malware dalam mengenali malware dari source codenya. Umumnya teknik ini akan mencari pada kode sumber apakah terdapat aktifitas atau fungsi yang mencurigakan. Contohnya adalah adanya fungsi untuk menggandakan dirinya (replikasi), atau fungsi untuk mengelabui anti virus (*obfuscation*) dll. Teknik ini umumnya memiliki tingkat akurasi yang rendah dan banyak menghasilkan *false positive* (alarm palsu). Biasanya teknik ini akan mempunyai list banyak kriteria karakteristik malware yang diberi bobot nilai tertentu, kemudian bila kriteria tersebut ditemukan maka nanti nilai tersebut akan dijumlahkan. Bila hasil perhitungan tadi mencapai batas nilai tertentu maka antivirus akan memberikan alarm adanya dugaan malware. Teknik ini banyak menggunakan teori dari sistem pakar, *neural network* dan *data mining.*

b. *Integrity Check*

Malware yang menginfeksi sebuah berkas, umumnya akan melakukan modifikasi pada berkas. Sehingga bila terjadi perubahan pada sebuah berkas tanpa ada otorisasi yang jelas, maka aktifitas ini dicurigai sebagai adanya malware. Teknik ini umumnya menggunakan *checksum*, jadi antivirus akan melakukan *checksum* terhadap berkas, kemudian *checksum* ini akan di input ke dalam database. Bila antivirus melakukan pemindaian maka *checksum* terbaru akan dibandingkan dengan *checksum* yang ada di database. Bila terjadi perubahan maka antivirus akan memberikan alarm. Teknik ini memiliki kelemahan yaitu memberikan hasil akurasi yang kurang baik.

Penggunaan algoritma sudah cukup bagus untuk berkas yang berukuran kecil hingga sedang, namun dengan metode ini sangat lambat jika memindai berkas dengan ukuran sampai Gigabytes, sedangkan antivirus saat ini sudah sangat baik dengan kemampuan yang mumpuni tapi menggunakan

sumberdaya yang relative sedikit sehingga tidak membebani kinerja komputer.

B. Kompleksitas waktu algoritma boyer moore.

Algoritma Boyer Moore mempunyai kompleksitas waktu rata-rata O(m/n), di mana, m adalah jumlah karakter dalam teks dan n adalah jumlah karakter dalam pola. Sebagai perbandingan, algoritma brute force dalam pencocokan pola memiliki kompleksitas waktu rata-ratanya adalah O(m+n) dan algoritma pencocokan pola Knuth-Morris-Pratt kompleksitas waktunya adalah O(m+n). Jika dibandingkan, algoritma booyer moore memiliki kompleksitas waktu lebih baik dibanding algoritma brute force dan KMP.

C. Prosedur Pemindaian Berkas dalam Program

Implementasi anti virus sederhana ini bias menggunakan bahasa pemrograman apa saja. Prosedur pembacaan berkas dari kode biner menjadi string sampai pencocokan pola dengan algoritma Boyer Moore adalah sebagai berikut :

1. Membuka berkas dan membaca kode binernya
2. Mendekode kode biner yang telah dibaca menjadi string dengan sistem *utf-8*.
3. Melakukan pencocokan pola dengan algoritma boyer moore.
4. Menampilkan hasil pemindaian, jika fungsi bmMatch menghasilkan nilai -1 maka menampilkan hasil bahwa berkas aman, dan jika fungsi menghasilkan nilai lebih besar dari 0 maka menampilkan hasil bahwa bekas terinfeksi.

Pendeteksian virus dengan pemindaian menghasilkan hasil yang cukup akurat dibanding metode heuristic atau integrity check, namun keakuratan tersebut harus dibayar mahal bila berkas yang dipindai berukuran sangat besar. Karena pada dasarnya satu karakter pada string berukuran 1 byte dan jika ukurannya sangat besar maka string yang akan dipindai bias terdiri triliunan karakter.

```
file=openFile('namaberkas','binary')
text = decode(file, 'utf-8')

pattern = "VIRUS"
result = bmMatch(text, pattern)

if (result = -1)
      print('berkas aman')
else
      print('berkas terinfeksi
virus')

closeFile(file)
```

Kode di atas merupakan pseudo code dari antivirus sederhana yang mendeteksi virus dengan memindai berkas menggunakan algoritma pencocokan pola boyer moore. Kode di atas merupakan bagian program utama yang merupakan pembacaan berkas dan pemindaian berkas dengan fungsi bmMatch.

V. KESIMPULAN

Program anti virus menggunakan algoritma pencocokan pola dalam memindai berkas. Pemindaian dengan algoritma tersebut terbukti lebih akurat dibanding dengan dengan metode heuristik ataupun integrity check. Algoritma boyer moore yang memiliki kompleksitas waktu lebih baik dibandingkan algoritma brute force dan KMP. Penggunaan algoritma boyer moore dalam program anti virus cukup bagus sehingga nantinya dapat dikembangkan lebih jauh lagi menjadi anti virus yang lebih baik dalam meminda virus-virus yang berbahaya di dalam sistem komputer. Dan dengan penambahan metode heuristic dan integrity check dapat mempercepat kinerja dan mengurangi sumber daya yang digunakan oleh anti virus.

VI. UCAPAN TERIMA KASIH

1. Allah Swt. yang telah memberikan nikmat kehidupan, sehat dan sempat dalam menuntut ilmu di Kampus Institut Teknologi Bandung.
2. Orang Tua yang telah memberi dukungan baik moral maupun materi agar selalu bersemangat menuntut ilmu.
3. Bapak Dr. Rinaldi Munir, S.T, M.T selaku dosen pengampu kuliah Strategi Algoritma dan telah

memberikan tugas ini agar saya lebih tentang ilmu yang telah beliau ajarkan.

4. Pihak-pihak yang telah mendukung saya dalam menyusun makalah ini.

VII. GLOSARIUM

Signature sebuah ciri khas dari virus berupa struktur berkas,ukuran berkas dan lain sebagainya.

Checksum Sebuah teknik untuk mendeteksi apakah sebuah data berubah pada saat transmisi.

Heuristik pemecahan masalah dengan mengamati pola-pola tertentu dalam masalah.

Neural Network jaringan dari sekelompok unit pemroses kecil yang dimodelkan berdasarkan sistem saraf manusia**.**

Data Mining ekstraksi pola yang menarik dari data dalam jumlah besar

String Kumpulan dari karakter-karakter

Enkode proses konversi informasi dari suatu sumber (objek) menjadi data

Dekode proses konversi data menjadi informasi.

UTF-8 standar internasional dalam pengkodean karakter menjadi bilangan biner.

REFERENCES

[1] https://www.it-jurnal.com/pengertian-dan-jenis-jenis-virus-pada-komputer/ dikunjungi pada Kamis, 18 Mei 2017 pukul 23:13 WIB.

[2] https://andynor.net/blog/430/ dikunjungi pada Jumat, 19Mei 2017 pukul 00:06 WIB.

[3] https://gist.github.com/dbrgn/1154006 dikunjungi pada Jumat, 19 Mei 2017 pukul 03.04 WIB.

[4] http://julismail.staff.telkomuniversity.ac.id/teknik-deteksi-malware/ dikunjungi pada Jumat, 19 Mei 2017 pukul 10.30 WIB.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 19 Mei 2017

ttd

Fajar Nugroho - 13515060