Penerapan K-Means *Clustering* untuk Klasifikasi Serangan *Cyber* pada Syslog *File*

I Wayan Ardiyasa
Institut Teknologi dan Bisnis STIKOM Bali
e-mail: ardi@stikom-bali.ac.id

Diajukan: 11 Mei 2020; Direvisi: 29 Juni 2020; Diterima: 10 Juli 2020

Abstrak

Cybercrime adalah aktivitas kejahatan yang menggunakan teknologi komputer sebagai sarana bertujuan untuk mendapatkan akses informasi dan data yang bersifat privasi sehingga menimbulkan kerugian. Kejahatan penggunaan teknologi informasi meningkat seiring tingginya pengguna teknologi dikarenakan akses informasi saat ini yang sangat mudah dan tidak pedulinya pengguna terhadap keamanan data maupun sistem bagi pihak penyedia maupun pengelola. Selain itu semakin mudahnya akses informasi pada suatu website yang terhubung dengan jaringan internet yang mengakibatkan meningkatnya kejahatan komputer seperti web defacing. Syslog merupakan sebuah protokol untuk system logging dan mencatat aktivitas pengguna dengan format file text pada suatu perangkat seperti perangkat komputer server. Permasalahan muncul ketika file syslog memiliki ribuan catatan aktivitas serangan, sehingga sangat sulit untuk mendapatkan informasi serangan secara cepat. Untuk itu diperlukan clustering untuk mengelompokkan jenis serangan pada syslog file. Jenis serangan yang dilakukan clustering yaitu SQL Injection, XSS Attack dan LFI Attack. Kebaruan dari penelitian ini adalah klasifikasi serangan cyber pada file syslog.log dengan menggunakan metode K-Means Clustering Untuk clustering serangan cyber pada file syslog.log dan pembobotan menggunakan metode TF-IDF untuk mendapatkan data numerik. Penelitian ini menghasilkan aplikasi analisis serangan cyber pada file syslog.log berbasis web untuk membantu pihak investigator digital forensic di dalam analisis dan mendapatkan informasi serangan cyber.

Kata kunci: *K-Means Clustering, Cyber, Syslog, Digital forensic.*

Abstract

Cybercrime is a criminal activity that uses computer technology as a means intended to gain access to information and data that constitutes privacy that causes harm. The crime of using information technology is increasing With the increase in technology users accessing information today which is very easy and does not care about the user's data or system security for providers as well as providers. In addition, easy access to information on web sites that are connected to the internet network related to computer problems such as web defacing. Syslog is a protocol for logging systems and records user activity in a text file format on a device such as a computer server device. The problem arises with compiling the syslog file to have a lot of records of attack activity, it is very difficult to get attack information quickly. This requires a cluster to classify types of attacks on the syslog file. Types of attacks carried out are SQL Injection, XSS Attack and LFI Attack. The novelty of this research is the collection of cyberattacks on the syslog.log file using the K-Means Clustering method for clustering cyberattacks on the syslog.log file and weighting using the TF-IDF method to obtain numerical data. This research makes a cyberattack analysis application in the web-based syslog.log file to help digital investigators in forensic analysis and obtain cyberattack information.

Keywords*: K-Means Clustering, Cyber, Syslog, Digital forensic.*

1. Pendahuluan

Teknologi sebagai perangkat yang dikembangkan untuk membantu pekerjaan manusia di dalam kehidupannya sehari-hari. Dengan penggunaan teknologi tepat guna seperti teknologi informasi dan komunikasi masyarakat sangat terbantu untuk melakukan komunikasi dan mencari informasi dengan mudah dan cepat. Semakin pesat perkembangan teknologi informasi dan komunikasi pengguna menjadi terbantu di dalam mengelola dan mengolah informasi yang bisa diakses secara luas karena seluruh komponen perangkat keras dan perangkat lunak terhubung menjadi satu ke dalam jaringan internet atau

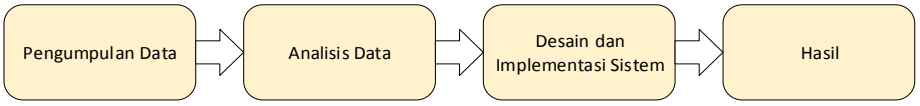
publik. Informasi yang berada pada suatu *website*, baik informasi yang terdapat pada halaman *interface* maupun administrator tidak secara otomatis dijamin aman dikarenakan adanya risiko terdapatnya celah pada lapisan keamanan yang memungkinkan disalahgunakan oleh pihak yang tidak berwewenang [1].

Ada informasi yang sifatnya terbuka dan ada pula informasi yang bersifat *private*. Namun, di satu sisi sistem yang menyimpan informasi tersebut banyak yang memiliki *vulnerability*, namun banyak juga sistem yang memiliki tingkat keamanan yang tinggi. Dengan adanya kelemahan pada sistem, akan mengakibatkan adanya peluang bagi *attacker* untuk melakukan tindakan *cybercrime* maka diperlukan sistem pengamanan baik dari sisi aplikasi maupun komputer *server*.

Server merupakan perangkat utama dalam sebuah sistem komunikasi jaringan yang berfungsi sebagai penyedia layanan atau *service* dalam sebuah jaringan. Sebagai penyedia layanan, *service* sistem yang berjalan pada *server* harus mampu berjalan secara *real-time*. Di dalam *monitoring service running* pada komputer *server* diperlukan pencatatan dalam bentuk *log service* secara *real-time* untuk mencatat aktivitas *service* yang berjalan pada *server*. *Log service* adalah catatan atau riwayat akses ke sebuah sistem *service* yang dijalankan oleh sebuah *server* [2]. Syslog adalah standar untuk meneruskan pesan *log* di jaringan IP. Istilah untuk Syslog sering digunakan untuk protokol Syslog yang sebenarnya, serta aplikasi atau pustaka yang mengirim pesan Syslog [3]. Komputer server yang terhubung ke dalam jaringan publik sangat rentan terhadap serangan *cyber*, apabila itu terjadi maka diperlukan langkah investigasi untuk mendapatkan informasi sumber serangan pada *file* syslog.log. Aplikasi berbasis web dengan menggunakan metode K-means *Clustering* untuk mengelompokkan jenis serangan pada *file* syslog.log dan menggunakan metode TF-IDF untuk konversi data serangan menjadi data *numeric*. Kebaruan dari penelitian ini adalah klasifikasi serangan *cyber* pada *file* syslog.log secara komputerisasi menggunakan metode K-Means *Clustering*, sehingga mampu mengelompokkan jenis serangan sesuai dengan jenis serangan yang ada dan menghasilkan suatu informasi serangan *cyber*. Tujuan dari penelitian ini adalah untuk membantu *investigator digital forensic* di dalam melakukan investigasi jenis serangan *cyber* pada *file* syslog.log berbasis web.

2. Metode Penelitian

Metode penelitian menggunakan empat tahap di dalam penelitian ini untuk mendapatkan hasil analisis dari *file* syslog.log. Adapun metode penelitian yang digunakan adalah sebagai berikut:



Gambar 1. Metode penelitian.

2.1. Pengumpulan Data

Pada tahap pengumpulan data menggunakan jenis data yaitu data sekunder. Data sekunder yang dimaksud adalah data yang didapatkan dari sumber yang sudah ada dalam hal ini adalah *file* syslog.log pada *web server* yang mencatat aktivitas serangan pada *web server*. Sumber data yang digunakan didapatkan dari situs https://www.indonesianbacktrack.or.id/forum/thread-5506.html?highlight=httpd di mana data yang diberikan adalah data *file* syslog.log dari *web server*.

2.2. Analisis Data

Pada tahap analisis data, *file* syslog.log tersebut berisi informasi tentang IP *Address*, aktivitas pengguna, tanggal, waktu akses, serta *method* yang digunakan dengan format *text*. Untuk bisa diaplikasikan ke dalam metode K-Means *Clustering* diperlukan teknik analisis awal untuk mencari hubungan antara frase/kalimat jenis serangan pada *file* syslog.log serta melakukan pembobotan menjadi data *numeric* dengan metode TF-IDF. Algoritma K-Means adalah metode partisi terkenal untuk *clustering*. Metode pengelompokan K-Means, mengelompokkan data berdasarkan kedekatannya satu sama lain sesuai dengan jarak *Euclidean*. Dibutuhkan *kY* sebagai parameter *input* dan memartisi sekumpulan objek *n* dari kluster *kY*. Nilai rata-rata objek diambil sebagai kesamaan parameter untuk membentuk kelompok. Klaster atau pusat rata-rata dibentuk oleh pemilihan acak dari objek *kY*. Dengan membandingkan sebagian besar kesamaan objek lain yang ditugaskan ke *cluster*. Untuk setiap vektor data, algoritma ini menghitung jarak antara vektor data dan setiap *centroid* kluster menggunakan persamaan [4]. Setelah didapatkan informasi dari *file* syslog.log tahap berikutnya dilakukan tahap desain dan implementasi aplikasi.

2.3. Desain dan Implementasi Aplikasi

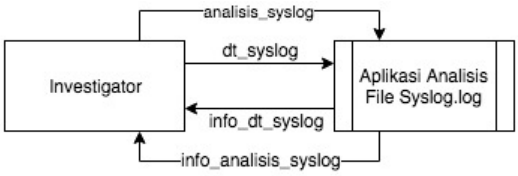
Pada tahap desain dan implementasi aplikasi merupakan tahap dibangunnya aplikasi untuk analisis Syslog berbasis web. Pada perancangan aplikasi menggunakan DFD (*Data Flow Diagram*) serta menggunakan bahasa pemrograman PHP untuk tahap implementasi.

2.4. Hasil

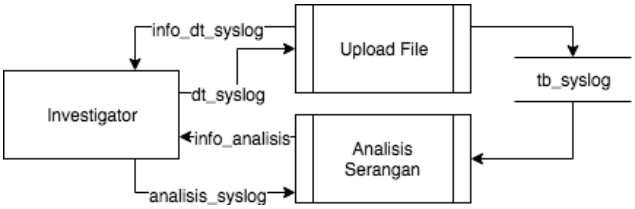
Pada tahap hasil, menampilkan informasi dari jenis serangan yang ada di dalam *file* syslog.log, di mana jenis serangan tersebut diklasifikasikan menggunakan metode K-Means *Clustering* sehingga informasi jenis serangan bisa diketahui.

3. Hasil dan Pembahasan

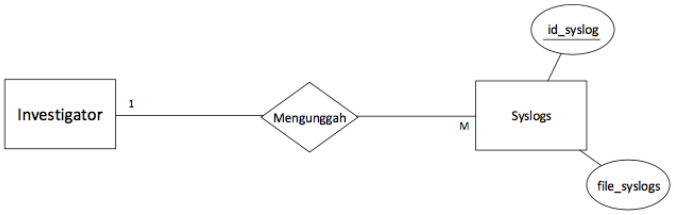
Untuk mendapatkan informasi jenis serangan *cyber* pada *file* syslog.log diperlukan analisis secara manual dan sangat membutuhkan waktu yang lama. Untuk mengatasi hal tersebut, diaplikasikan metode K-Means *Clustering* untuk mengelompokkan jenis serangan ke dalam aplikasi berbasis *web*. Di dalam perancangan aplikasi berbasis web menggunakan *Data Flow Diagram* dan perancangan basis data menggunakan ERD (*Entity Realtionship Diagram*). Berikut hasil rancangan aplikasinya:



Gambar 2. Diagram konteks aplikasi analisis *file* Syslog.



Gambar 3. DFD Level 0.



Gambar 4. Konseptual *database*.

Untuk mengelompokkan jenis serangan pada *file* syslog.log menggunakan metode K-Means *Clustering* dan TF-IDF. Tahap pertama mencari frekuensi atau TF (*Term Frequency*) dari kemunculan suatu serangan pada *file* syslog.log, yang kedua dilakukan pencarian hubungan ketersediaan istilah dalam hal ini adalah jenis serangan yaitu SQL *Injection*, XSS *Attack*, LFI *(Local File Inclusion)* ketiga jenis serangan tersebut memiliki istilah-istilah atau parameter serangan untuk mencari hubungan yang cocok untuk dibobotkan sehingga menghasilkan data *numeric* yang digunakan pada proses *clustering*. Untuk mendapatkan data *numeric* pada *file* syslog.log dilakukan konversi menggunakan metode TF-IDF sebagai berikut:

Tabel 1. Tabel konversi dengan TF-IDF.

Jenis Serangan			
No	**SQL *Injection***		**Hasil**
1.	select	IDF = log (2/1) = 0,3010 W = 1 * 0,3010	0,3010
2.	union	IDF = log (2/1) = 0,3010 W = 6 * 0,3010	1,8061
No.	**XSS *Attack***		**Hasil**
1.	alert	IDF = log (2/1) = 0,3010 W = 3 * 0,3010	0,9030
2.	script	IDF = log (2/1) = 0,3010 W = 2 * 0,3010	0,6020
No.	**LFI (*Local File Inclusion*)**		**Hasil**
1.	etc	IDF = log (2/1) = 0,3010 W = 8 * 0,3010	2,4082
2.	passwd	IDF = log (2/1) = 0,3010 W = 9 * 0,3010	2,7092

Tabel 2. Hasil konversi dengan TF-IDF.

No.	select	etc	passwd	union	script	alert
1	0.3010	2.4082	2.7092	1.8061	0.6020	0.9030

Setelah ditentukan tiap kata dalam bentuk *numeric* dengan rumus TF-IDF adalah sebagai berikut:

$$DF_j = log\ (D / df_j) \quad (1)$$

di mana:
D : jumlah dokumen
Df : jumlah dokumen yang mengandung term (t_j)
Selanjutnya untuk menghitung bobot (*w*) digunakan formula sebagai berikut:

$$W_{ij} = TF_{ij} \times IDF_j \quad (2)$$

di mana:
W_{ij} : adalah bobot term (t_j) terhadap dokumen (d_i)
TF_{ij} : jumlah kemunculan term (t_j) dalam dokumen (d_i)

Maka proses selanjutnya adalah pengelompokan data menggunakan K-Means. Untuk melakukan pengelompokan data dengan K-Means penulis menentukan terlebih dahulu:

1. Tentukan jumlah *cluster*.
2. Ambil sembarang data sebanyak jumlah *cluster* secara acak sebagai pusat klaster (*centroid*).
3. Hitung jarak antar pusat dengan *cluster* dengan rumus:

$$D_{(i,j)} = \sqrt{(X_{1i} - X_{1j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (3)$$

Keterangan:
D(i,j)= Jarak data ke i ke pusat *cluster j*
Xki = Data ke-*i* pada atribut *k*
Xkj = titik pusat ke-*j* pada atribut *k*

4. Hitung kembali pusat *cluster* dengan keanggotaan yang baru. Jika pusat *cluster* tidak berubah maka proses *cluster* telah selesai. Jika belum ulangi langkah nomor 4 sampai pusat *cluster* tidak berubah lagi [5].

Untuk mendapatkan informasi jenis serangan pada *file* syslog.log, dilakukan *cluster* dengan menggunakan nilai *K* = 3, dari 3 *cluster* tersebut setelah dilakukan perhitungan dengan rumus K-Means maka hasilnya adalah *cluster* pertama terdapat 12 sampel, *cluster* kedua 2 sampel, dan *cluster* ketiga 15 sampel dalam satu dokumen pada *file* syslog.log di mana *cluster* pertama adalah SQL *Injection*, *cluster* kedua adalah XSS *Attack*, dan *cluster* ketiga adalah LFI *(Local File Inclusion*). Dari hasil perhitungan tersebut dengan K-Means maka data tersebut dapat ditampilkan pada halaman web. Berikut hasil dari proses K-Means *Clustering* pada Tabel 3.

Tabel 3. Hasil proses dengan K-Means *Clustering*.

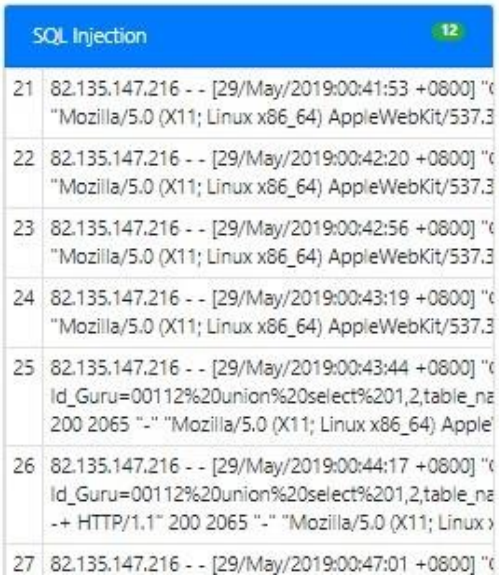
Cluster	Banyaknya Serangan	Kategori
Cluster 1	12	SQL *Injection*
Cluster 2	2	XSS *Attack*
Cluster 3	15	LFI *(Local File Inclusion)*

3.1. Implementasi Aplikasi Analisis *File* Syslog

Aplikasi analisis serangan *cyber* pada *file* syslog.log menggunakan metode K-Means *Clustering* untuk klasifikasi serangan *cyber* berbasis web dibangun menggunakan bahasa pemrograman PHP serta manajemen *database* menggunakan *PHPMyAdmin*. Aplikasi ini bertujuan untuk membantu pengguna khususnya *investigator* di dalam melakukan analisis serangan *cyber* pada *file* syslog.log. Berikut hasil implementasi aplikasi berbasis web:

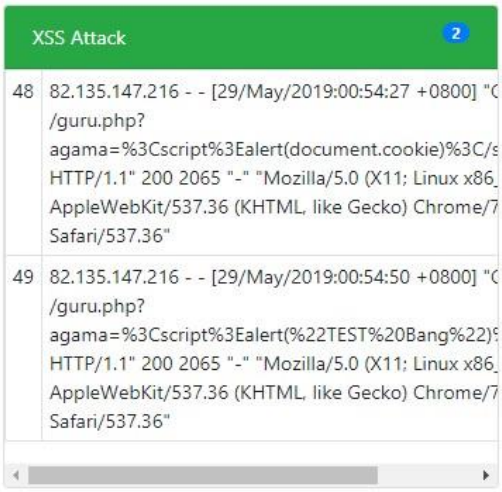
Gambar 5. Antarmuka aplikasi berbasis *web*.

Pada Gambar 2 merupakan tampilan antarmuka aplikasi yang terdapat 3 menu yaitu *collection*, *analysis*, dan *report*. Pada menu *collection* merupakan menu untuk isian *profile* pengguna dan *upload file* syslog.log. Pada menu *analysis* merupakan menu untuk melakukan analisis serangan *cyber* pada *file* syslog.log untuk mendapatkan informasi jenis serangan dan klasifikasi jenis serangan dan menu *report* merupakan menu untuk *print out* hasil klasifikasi serangan *cyber*.



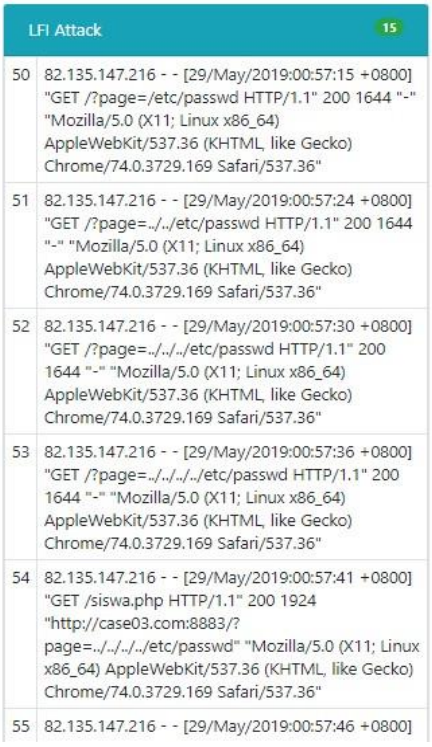
Gambar 6. Analisis SQL *injection*.

Pada gambar 3 merupakan hasil pengelompokan jenis serangan *cyber* SQL *Injection*. Teknik pengelompokan menggunakan metode K-Means *Clustering* menggunakan frase *select* dan *union* dan didapatkan informasi serangan SQL *Injection* sebanyak 12 *cluster* dengan serangan SQL *Injection.*



Gambar 7. Analisis XSS *attack*

Pada Gambar 4 didapatkan hasil *clustering* jenis serangan *cyber* XSS *Attack*. Teknik pengelompokan menggunakan metode K-Means *Clustering* menggunakan frase *alert* dan *script* dan didapatkan informasi serangan XSS *Attack* sebanyak 2 *cluster* dengan serangan XSS *Attack.*



Gambar 8. Analisis LFI (*Local File Inclusion*).

Pada Gambar 5 didapatkan hasil *clustering* jenis serangan *cyber* LFI *(Local File Inclusion)*. Teknik pengelompokan menggunakan metode K-Means *Clustering* menggunakan frase *etc* dan *passwd* dan didapatkan informasi serangan LFI *(Local File Inclusion)* sebanyak 15 *cluster* dengan serangan LFI *(Local File Inclusion)*.

Hasil dari penelitian ini sebagai solusi untuk membantu pengguna khususnya pihak *investigator* di dalam melakukan analisis barang bukti berupa *file* syslog.log yang bertujuan untuk mendapatkan informasi serangan *cyber* dari 3 jenis serangan *cyber* yaitu SQL *Injection,* XSS *Attack*, dan LFI *(Local File Inclusion).* Aplikasi analisis serangan *file* syslog.log berbasis *web* ini, menjadikan *tools* analisis berbasis grafis *user interface* menjadi lebih cepat dan *user friendly.*

3.2. Pengujian

Pengujian dengan metode *blackbox* ini dilakukan dengan cara mencoba fungsionalitas dari sistem tersebut agar sesuai yang diharapkan, berikut hasil dari pengujian menggunakan metode *blackbox.*

Tabel 4. Hasil pengujian dengan *blackbox testing*.

No.	Skenario Pengujian	Hasil	Kesimpulan
1.	Mengunggah *file* Syslog pada kolom *upload file*, lalu klik tombol *submit*	Memunculkan pesan '*File* sudah berhasil di-*upload*!'	Valid
2.	Mengosongkan kolom *upload file*, lalu klik tombol *submit*	Memunculkan pesan 'gagal'	Valid
3.	Mengunggah *file* Syslog pada kolom *upload file* lebih dari 2MB, lalu klik tombol *submit*	Memunculkan pesan '*File* terlalu besar!'	Valid
4.	Mengunggah *file* Syslog yang sama pada kolom *upload file*, lalu klik tombol *submit*	Memunculkan pesan '*File* sudah ada!'	Valid
5.	Mengunggah *file* Syslog pada kolom *uploadfile* dengan tipe *file* yang berbeda, lalu klik tombol *submit*	Memunculkan pesan 'Maaf! Tipe *file* tidak sesuai'	Valid
6.	Menghapus salah satu *file* Syslog, lalu klik ikon tombol hapus	Memunculkan pesan 'Maaf! Tipe *file* tidak sesuai'	Valid

4. Kesimpulan

Telah dihasilkannya aplikasi analisis *file* syslog.log dengan menggunakan metode K-Means *Clustering* untuk klasifikasi serangan-serangan *cyber* dan metode TF-IDF untuk melakukan konversi dari data *text* menjadi data *numeric*. Hasil pengujian yang dilakukan dengan menggunakan *blackbox testing* didapatkan hasil pengujian yang sesuai. Tujuan dari aplikasi ini yaitu untuk membantu pihak pengguna khususnya *Investigator Digital Forensic* di dalam mencari informasi jenis serangan *cyber* pada *file* syslog.log.

Daftar Pustaka

[1] A. M. Elu, "Rancang Bangun Aplikasi Pendeteksian Vulnerability Structured Query Language (Sql) Injection untuk Keamanan Website," *J. Teknol. Inf.*, vol. VII, no. 1, pp. 111–124, 2013.

[2] N. K. Ariasih and D. P. Hostiadi, "Monitoring Log Service Pada Server Berbasis Web," *Semin. Nas. Inform.*, pp. 190–194, 2014.

[3] Braun, Uri, Y. Zaslavasky, and Y. Teitz, "Syslog Parser," US11/875,955, 2007.

[4] S. Shah and M. Singh, "Comparison of a time efficient modified k-mean algorithm with K-mean and K-medoid algorithm," *Proc. - Int. Conf. Commun. Syst. Netw. Technol. CSNT 2012*, pp. 435–437, 2012.

[5] E. Yulian, "Text Mining dengan K-Means Clustering pada Tema LGBT dalam Arsip Tweet Masyarakat Kota Bandung," *J. Mat. "MANTIK,"* vol. 4, no. 1, pp. 53–58, 2018.