 ISBN: 978-602-60280-1-3

Pengelompokan Pengguna Internet Menggunakan Metode K-Means pada Data Log Akses Server

Adam Prayogo Kuncoro[1], Dwi Prasetyo Hutomo[2] , Muhammad Zulfadhilah[3]
[1,2] Program Studi Sistem Informasi – STMIK Amikom Purwokerto
[3] Program Studi Informatika – STIKES SARIMULIA Banjarmasin
Jl. Letjen Pol Sumarto Watumas Purwanegara Purwokerto, Banyumas
Telp : (0281) 623321, Fax : (0281) 623196, Email : amikom@amikompurwokerto.ac.id
adam@amikompurwokerto.ac.id[1], dwi.hutomo@amikompurwokerto.ac.id[2], muhammadfadhilah@stikessarimulia.acid[3]

Abstrak

Internet telah menjadi kebutuhan di masyarakat saat ini, informasi apapun dapat diakses di internet melalui browser. Namun, aktivitas ini bisa berdampak pada pengguna, salah satunya perubahan perilaku. Penelitian ini berfokus pada aktivitas pengguna internet berdasarkan jaringan data *log* pada institusi pendidikan. Data yang digunakan dalam penelitian ini dihasilkan dari pengamatan satu minggu dari salah satu universitas di Yogyakarta. Data *log* aktivitas jaringan adalah salah satu jenis data yang besar, sehingga diperlukan penggunaan data mining dengan algoritme K-Means sebagai solusi untuk mengetahui perilaku pengguna internet. Algoritme K-Means digunakan untuk *clustering* berdasarkan jumlah pengunjung. Jumlah *cluster* pengunjung dibagi menjadi tiga, yaitu rendah dengan 1479 jumlah data, sedang dengan 126 jumlah data, dan tinggi dengan 33 jumlah data. Kategorisasi juga dilakukan oleh waktu akses dan didasarkan pada isi *website* yang ada dalam data. Hasil penelitian ini adalah kelompok *website* yang sering dikunjungi dengan urutan: mesin pencari, media sosial, berita, dan informasi. Studi ini juga mengungkapkan bahwa *cyber-profiling* yang dilakukan sangat dipengaruhi oleh faktor lingkungan dan aktivitas sehari-hari.

Kata Kunci : Klasifikasi, Jaringan, K-Means, *Log*.

Abstract

The Internet has become a necessity in today's society, any information is accessible on the internet via web browser. However, these activities could have an impact on users, one of which changes in behavior. This study focuses on the activities of Internet users based on the log data network at an educational institution. The data used in this study resulted from one-week observation from one of the universities in Yogyakarta. Data log network activity is one type of big data, so it is needed to use of data mining with K-Means algorithm as a solution to determine the behavior of Internet users. The K-Means algorithm used for clustering based on the number of visitors. Cluster number of visitors divided into three, namely low with 1479 amount of data, medium with 126 amount of data, and high with 33 amount of data. Categorization also performed by the access time and is based on website content that exists in the data. It is to compare the results by the K-Means clustering algorithm. The results of the educational institution show that each of these clusters produces websites that are frequented by the sequence: website search, social media, news, and information. This study also revealed that the cyber-profiling had been done strongly influenced by environmental factors and daily activities.

Keywords: Classification, Network, K-Means, Log.

I. PENDAHULUAN

Yang terpenting dalam pengelolaan jaringan adalah mengetahui karakteristik pengguna jaringan, analisis lalu lintas dapat membantu administrator membuat kebijakan dan melindungi keamanan jaringan (Deliang, 2016). Survei APJI yang dilakukan pada tahun 2014 menunjukkan urutan aktivitas pengguna internet di Indonesia yaitu: pengguna media sosial, pencarian informasi, *chatting*, berita, video, dan email. Hasil ini menunjukkan bahwa pencarian berita dan email tidak termasuk dalam aktivitas populer (APJII, 2015).

Pada era digital ini, penggunaan media sosial menghasilkan lebih banyak informasi dari pada sebelumnya. Ini menunjukkan bahwa data yang besar membutuhkan teknologi data *mining* untuk mengatasi tantangan ini (Gole, 2015). Algoritme *clustering* adalah salah satu algoritme yang efektif untuk menganalisis data besar dan menggambarkan atribut yang ada (Dong, 2015). Secara umum, perilaku pengguna ponsel sering menjelajah internet. Dan penelitian ini menunjukkan bahwa terdapat hubungan antara aktivitas menjelajah internet dengan aktivitas sehari-hari (Gao, 2015). Studi pembuatan *cyber-profiling* adalah eksplorasi data untuk mengetahui karakteristik aktivitas pengguna saat menggunakan komputer / internet.

Salah satu algoritme yang bisa digunakan dalam membantu *cyber-profiling* adalah algoritme K-Means. Algoritme ini akan melakukan pengkategorian pengguna internet berdasarkan jumlah kunjungan ke sebuah website. Ini akan menunjukkan apa yang sering diakses oleh pengguna sehingga mereka akan mengetahui perilaku aktivitas penggunanya di internet. *Profiling* adalah proses pengumpulan data individu atau kelompok yang bisa menghasilkan sesuatu

yang menarik, mengejutkan dan signifikan (Irvine, 2010). *Cyber-profiling* telah membawa langkah bagus untuk ilmu komputer forensik. Hal ini didasarkan pada pengalaman yang telah dilakukan (Berg, 2013).

Dalam mengakses internet, lokasi akan memberikan informasi penting untuk menentukan perilaku seseorang (Zhou, 2016). Penggunaan layanan internet di kampus yang bisa membatu kegiatan edukasi terkadang juga digunakan untuk kegiatan kriminal atau ilegal. Jadi untuk mencari tahu apa yang bisa diakses oleh pengguna internet di institusi pendidikan, diperlukan pembuatan *cyber profiling*.

II. METODE PENELITIAN

Dalam penelitian yang dilakukan oleh (Deliang, 2016) disebutkan bahwa cyber-profiling dapat membantu administrator dalam menentukan kebijakan dan kebutuhan perbaikan informasi pengguna. Perusahaan perlu mengumpulkan dan menganalisis data pelanggan untuk mengidentifikasi karakteristik target pelanggannya (Liao, 2015). Studi lain yang dilakukan oleh (Yu, 2013) menyatakan bahwa kesimpulan dari cyber-profiling harus menggunakan metode deduktif. Hal ini karena jika mereka hanya membuat kesimpulan dengan metode induktif sangat tidak dapat diandalkan, dan dapat menyebabkan kesalahpahaman dalam analisis.

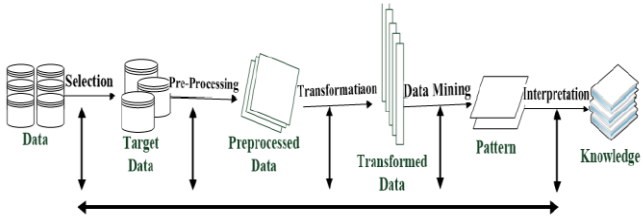
Dalam penelitian yang dilakukan oleh (Gao, 2015) disebutkan bahwa ada korelasi penggunaan internet dengan aktivitas sehari-hari. Sebuah survei menunjukkan bahwa pada tahun 2014 ada 88 juta pengguna internet di Indonesia. Hasil ini menyatakan ada tiga alasan utama orang mengakses internet, yaitu komunikasi, sumber informasi dan sehari-hari dan mengikuti perkembangan informasi terbaru. Berdasarkan tiga alasan utama ada empat aktivitas utama penggunaan internet, yaitu media sosial, mencari informasi, chatting dan mencari berita terbaru.

Dalam penelitian yang dilakukan oleh (Cheng, 2015) menyatakan bahwa untuk mengendalikan kebijakan dan padatnya jaringan, operator didorong untuk merancang mekanisme yang sesuai dalam penyediaan sumber daya bagi konsumen berdasarkan berbagai kategori aplikasi yang digunakan. Selain itu, sebuah studi yang dilakukan oleh (Irvine, 2010) juga menyebutkan bahwa kesadaran akan tindakan profiling bisa memberi peringatan kepada pengguna remaja dalam berbagi informasi pribadi saat mengakses internet.

Menurut sebuah survei yang dilakukan oleh (Shekhawat, 2014) disebutkan bahwa ada sembilan kategori perilaku pengguna internet, yaitu NetTerrorist, NetStreiver, NetAvoider, NetPublisher, Networker, NetCrawler, NetAdvocate, NetJungki dan NetRookie.

A. Data Mining

Data *mining* adalah proses menganalisis data dari perspektif yang berbeda dan meringkasnya menjadi informasi yang berguna. Informasi tersebut dapat digunakan untuk meningkatkan pendapatan, memotong biaya atau keduanya, berbagai algoritme data *mining* seperti klasifikasi, *clustering*, asosiasi yang digunakan untuk mengekstrak informasi dari potensi data (Gole, 2015). Data *mining* memiliki tahapan seperti pada Gambar 1 (Gole, 2015). Data mining melibatkan empat tugas, yaitu *Clustering*, *Classification*, *Regression* dan *Association* (Kumar, 2011). Data *mining* dapat mendeteksi pengetahuan yang berguna dari kumpulan data yang besar, seperti pengenalan pola, aturan dan tren (Gole, 2015).



Gambar 1. Proses *data mining*

B. K-Means

Clustering adalah teknik mengelompokkan sejumlah objek dalam kelompok yang sama (disebut *cluster*) yang lebih mirip satu sama lain daripada kelompok lain (*cluster*). Ini adalah tugas utama dalam mengeksplorasi data *mining*, dan teknik umum untuk analisis data statistik. *Clustering* juga digunakan di berbagai bidang, termasuk pembelajaran mesin, pengenalan pola, analisis citra, pencarian informasi, dan bio-informatika (Luan, 2015).

Teori utama algoritme K-Means adalah deskripsi titik pusat K untuk setiap *cluster*. Pemilihan pusat-pusat tersebut harus sesuai dengan kebutuhan mereka, karena pemilihan tersebut mempengaruhi titik pusat dari hasil yang diperoleh (Azar, 2016).

Metode algoritme K-means sebagai berikut (Riadi, 2016):

1) Inisialisasi: tentukan nilai K sesuai jumlah *cluster* yang diinginkan.

2) Pilih data K dari dataset sebagai *centroid*.

3) Alokasikan semua data ke *centroid* terdekat dengan metrik jarak yang telah ditentukan.

4) Hitung ulang *centroid* berdasarkan data yang mengikuti setiap cluster.

5) Ulangi langkah 3 dan 4 sampai kondisi konvergensi tercapai (tidak ada data yang dipindahkan).

Algoritme K-Means adalah salah satu algoritme *clustering* yang populer; Ini juga merupakan algoritme tanpa pengawasan yang digunakan dalam pengelompokan. Algoritme ini memilih *centroid* dan membandingkannya dengan titik data berdasarkan kemiripan karakteristik pada setiap titik, sehingga pembentukan *cluster* berdasarkan jarak titik data ke *centroid* (Rahmani, 2014).

C. *Log*

Menurut definisi, *log* adalah catatan kegiatan sehari-hari. Sedangkan di dunia komputer *log* adalah file yang mencatat aktivitas komputer. Dalam kegiatan forensik digital, log digunakan sebagai pendukung dalam proses investigasi (Riadi, 2016).

D. *Cyber-Profiling*

Penggunaan data sangat mendasar untuk melihat karakteristik hubungan pengguna dengan perangkat dan aplikasi seluler serta jenis akses (Cheng, 2015). *Profiling* adalah informasi individu atau kelompok yang terakumulasi, disimpan dan digunakan untuk berbagai keperluan. Salah satunya adalah untuk mengetahui kegunaan kegiatan pembuatan profil pengguna internet (Berg, 2013).

Ada dua jenis profil, yaitu deduktif dan induktif. Deduktif profil berdasarkan bukti forensik di tempat kejadian dan korban. Sementara profil induktif adalah analisa psikologis mengenai perilaku kriminal yang diperoleh dari tes dan kasus yang telah diselesaikan (Atkins, 2015).

Profil di Internet harus dilakukan dengan menggunakan metode induktif dan deduktif. Hal ini dilakukan untuk mencegah kesalahpahaman tentang perilaku pengguna internet, karena tingkah laku di Internet terkadang berbeda dengan tingkah laku di dunia nyata (Yu, 2013).

Cyber-profiling adalah hasil dari kesimpulan dari minat, karakteristik, perilaku, niat dan preferensi aktivitas pengguna saat ini di Internet (Mayo, 2013). Profil pengguna internet dibuat untuk menjelaskan latar belakang pengetahuan pengguna (Shobana, 2014).

E. *Metode Riset*

Data aktivitas log internet yang diperoleh dari sebuah gateway tidak hanya berisi apa yang diakses oleh pengguna, namun paket data lain pada aktivitas lalu lintas jaringan juga tercatat. Oleh karena itu, diperlukan proses pembersihan data yang disebut *pre-processing*.

Langkah-langkah algoritme K-Means:

1) Tentukan K sebagai jumlah *cluster* yang terbentuk.

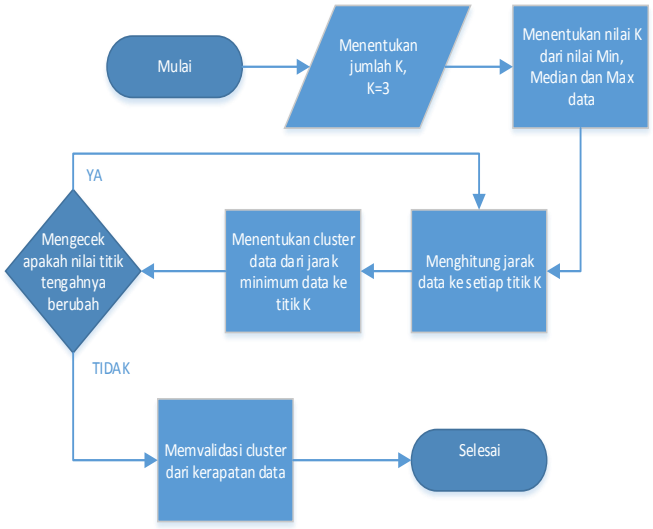
2) Menghasilkan K *centroid* (titik pusat *cluster*) dimulai secara acak.

3) Hitung jarak setiap objek ke masing-masing *centroid* setiap *cluster*.

4) Alokasikan setiap objek ke *centroid* terdekat.

5) Iterasi, kemudian tentukan posisi *centroid* baru.

6) Ulangi langkah 3 jika *centroid* baru tidak sama.



Gambar 2. Aliran algoritme K-Means

Data yang diperoleh dari gateway sebanyak 320.773 buah, data ini masih perlu diproses sebelum proses pengelompokan. Salah satu langkah sebelum pengelompokan adalah *pre-processing*. Tahap ini adalah melakukan pembersihan data yang tidak diperlukan dalam proses penelitian. Hasil *pre-processing* menunjukkan bahwa data akan digunakan sebanyak 1.638 catatan. Tahap selanjutnya setelah *pre-processing* menyelesaikan proses pengelompokan algoritme menggunakan K-Means. Algoritme K-Means dilakukan dengan menggunakan SPSS dan RapidMiner. Data hasil *clustering* dengan algoritme K-Means akan dianalisis untuk membantu proses *cyber-profiling*.

III. HASIL DAN PEMBAHASAN

A. *Kategori Data*

Kategori data ini akan di-*cluster* berdasarkan banyaknya kunjungan ke situs web. *Clustering* dilakukan dengan menggunakan algoritme K-Means di SPSS dan *Rapid Miner*. SPSS dan *Rapid Miner* digunakan untuk menentukan hasil *cluster* yang didapat apakah itu tepat untuk melanjutkan tahap analisis dari *cyber-profiling* ini.

Implementasi algoritme K-Means dilakukan oleh Aplikasi SPSS dan *Rapid Miner* menghasilkan tiga *cluster*, yaitu rendah, sedang dan tinggi. *Cluster* pertama adalah *cluster* dengan tingkat lalu lintas yang rendah memiliki total anggota 1.479 situs *web*, *cluster* kedua adalah *cluster* dengan tingkat lalu lintas sedang memiliki total anggota 126 situs *web*, dan yang terakhir *cluster* dengan tingkat lalu lintas yang tinggi memiliki anggota 33 situs *web*.

Inisialisasi pusat *cluster* awal dalam *clustering*. Proses dapat dilihat pada Tabel 1.

Tabel 1. Inisialisasi awal pusat *cluster*

	Inisialisasi Pusat *Cluster*		
	1	2	3
Jumlah pengakses	1	37	71

Inisialisasi nilai awal data dalam cluster berdasarkan nilai tertinggi, rata-rata dan nilai terkecil. Di dalam Studi ada delapan iterasi yang dihasilkan untuk mendapatkan hasil yang tepat. Inisialisasi ini dilakukan dengan aplikasi SPSS dan *Rapid Miner*.

Proses iterasi dalam proses pengelompokan dapat dilihat pada Tabel 2.

Tabel 2. Rekap proses iterasi

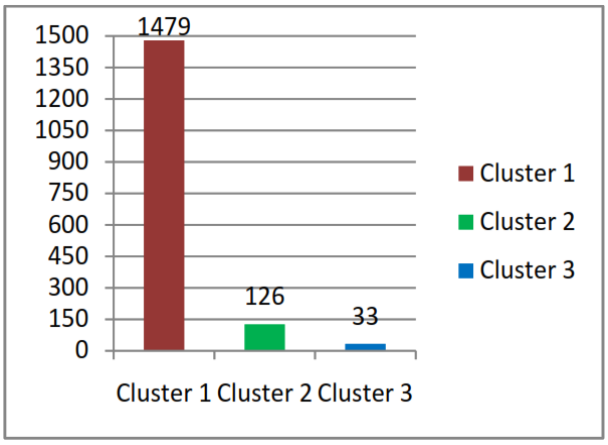
Iterasi	Perubahan di pusat *cluster*		
	1	2	3
1	1,522	6,620	10,429
2	0,150	3,805	4,857
3	0,147	3,173	4,000
4	0.158	2,332	2,194
5	0,060	1,221	1,727
6	0,067	1,109	1,262
7	0,000	0,113	0,410
8	0,000	0,000	0,000

Tabel 2 menunjukkan bahwa dibutuhkan 8 (delapan) iterasi untuk mendapatkan *cluster* yang tepat Aplikasi SPSS menyatakan bahwa minimum jarak antara pusat awal adalah 34. Hasil dari proses iterasi dalam menentukan pusat pengelompokan awal dapat dilihat pada Tabel 3.

Tabel 3. Hasil akhir pusat *cluster*

	Inisialisasi Pusat *Cluster*		
	1	2	3
Jumlah pengakses	1	37	71

Hasil *clustering* yang telah dilakukan dapat dilihat pada Gambar 3.



Gambar 3. Hasil *clustering* data

Hasil *clustering* akan dijelaskan sebagai berikut:

• *Cluster* 1: *cluster* ini adalah *cluster* dengan jumlah anggota tertinggi, yaitu 1479 *website*. *Cluster* pertama adalah *cluster* dengan tingkat lalu lintas pengguna sedikit, mulai dari 1-10 kunjungan per *website*. *Cluster* ini sebagian besar beranggota sebuah iklan *website.*

• *Cluster* 2: situs *web* yang termasuk dalam *cluster* ini sebanyak 126 buah, *website* yang masuk *cluster* ini karena memiliki nilai lebih tinggi dari pada nilai rata-rata yang dihasilkan dalam proses pengelompokan berkisar pada 11-13 kunjungan per situs web *cluster* ini berisi lebih banyak informasi dan situs berita.

• *Cluster* 3: *cluster* ini memiliki anggota paling sedikit, yang hanya 33 situs *web*. Namun, *cluster* ini memiliki tingkat lalu lintas tertinggi dibanding *cluster* lainnya. Nilai dalam kelompok ini adalah 34-64 kunjungan per situs *web*. *Cluster* ini berisi lebih banyak mesin pencari dan situs media sosial.

2. Hasil Analisa

Dalam penelitian ini, data *log* jaringan diperoleh dari *gateway* sebauh institusi pendidikan. Pengkategorian data dibagi menjadi tiga kategori: rendah, sedang dan tinggi. Proses pengkategorian dilakukan dengan menggunakan algoritme K-Means yang diimplementasikan menggunakan SPSS dan *Rapid Miner*. Hasil pengelompokan yang diperoleh dari implementasi algoritme K-Means menunjukkan bahwa penggunaan internet untuk institusi pendidikan untuk mengakses mesin pencari, situs informasi dan situs media sosial. Penelitian ini sedikit berbeda dengan hasil survei yang dilakukan (APJII, 2015) yang menyatakan bahwa penggunaan internet ada dalam urutan ini: jaringan (social media), pencarian informasi, *chatting* (*messaging*), pencarian berita, video dan email.

IV. KESIMPULAN

Kesimpulan Analisis data lalu lintas jaringan dengan menggunakan algoritme K-Means untuk proses *profiling* menunjukkan hasil yang sesuai dengan harapan penelitian, karena memiliki tingkat akurasi yang baik. Algoritme K-Means menghasilkan tiga kategori lalu lintas ke situs web, yaitu tinggi, sedang dan rendah. Hasilnya juga sama dengan hasil berdasarkan kategorisasi waktu akses data dan berdasarkan isi konten website.

Hasil penelitian ini juga menunjukkan bahwa situs-situs yang memiliki tingkat lalu lintas tinggi dalam bereksperimen adalah mencari situs web, informasi dan media sosial. Hasil dari *cyber-profiling* dalam penelitian ini adalah pengguna internet masuk karakter Networker dan NetJungki, berdasarkan karakteristik ini menunjukkan bahwa *cyber-profiling* telah dilakukan sangat dipengaruhi oleh faktor lingkungan dan aktivitas sehari-hari.

Penelitian ini memiliki keterbatasan data sumber dalam proses pembuatan profil. Untuk mendapatkan hasil maksimal dalam proses pembuatan profil, data yang didapat harus berisi tentang aktivitas pada komputer yang telah digunakan. Dalam penelitian selanjutnya, proses pembuatan profil cyber harus menggunakan data yang diperoleh dari aktivitas komputer yang telah digunakan dan juga data dari pengguna komputer. Diharapkan untuk mendapatkan hasil analisis profil yang lebih baik.

DAFTAR PUSTAKA

C. Deliang, "A Comparative Study on User Characteristics of Fixed and Wireless Network Based on DHCP," pp. 0–3, 2016.

(APJII, 2015) APJII, "Indonesian Internet User Profile 2014," 2015.

(Gole, 2015) S. Gole, "A survey of Big Data in social media using data mining techniques," *2015 IEEE Int. Conf. Adv. Comput. Commun. Syst.*, pp. 5–10, 2015.

(Dong, 2015) J. He, A. Wei, Y. Yang, and W. Dong, "Research on Degree of Video Completion of Internet Videos with Clustering Algorithms," pp. 89–95, 2015.

(Gao, 2015) J. Yan, Y. Qiao, J. Yang, and S. Gao, "Mining Individual Mobile User Behavior on Location and Interests," *2015 IEEE Int. Conf. Data Min. Work.*, pp. 1262–1269, 2015.

(Irvine, 2010) J. J. Irvine, "Digital Forensic Analysis & Cyber-profiling," no. 703, pp. 1–32, 2010.

(Berg, 2013) D. B. van den Berg, P. dr. A. de Vries, P. dr. S. van der Hof, M. Kakaris, and A. Theocharidis, "Online Identities , Profiling and Cyber Bullying," no. March, 2013.

(Zhou, 2016) C. Zhou, H. Jiang, Y. Chen, L. Wu, and S. Yi, "User Interest Acquisition by Adding Home and Work Related Contexts on Mobile Big Data Analysis," no. Bdsta, pp. 0–5, 2016.

(Liao, 2015) C. H. Liao, Y. H. Lei, K. Y. Liou, J. S. Lin, and H. F. Yeh, "Using Big Data for Profiling Heavy Users in Top Video Apps," *Proc. - 2015 IEEE Int. Congr. Big Data, BigData Congr. 2015*, pp. 381–385, 2015.

(Yu, 2013) S. Yu, "Behavioral Evidence Analysis on Facebook: a Test of Cyber-Profiling," *Defendologija*, vol. 16, no. 33, pp. 19–30, 2013.

(Cheng, 2015) J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet," *IEEE Trans. Emerg. Top. Comput.*, vol. 3, no. 1, pp. 95–106, 2015.

(Shekhawat, 2014) P. Shekhawat, "Netizens Buying Online Most Attracted to Digital Advertising," *http://www.markplusinsight.com/article/detail/34/netizens-buying-online-most-attracted-to-digital-advertising*, 2014.

(Kumar, 2011) A. Chauhan, G. Mishra, and G. Kumar, "Survey on Data Mining Techniques in Intrusion Detection," vol. 2, no. 7, pp. 2–5, 2011.

(Luan, 2015) L. Xue and W. Luan, "Improved K-means Algorithm in User Behavior Analysis," *2015 Ninth Int. Conf. Front. Comput. Sci. Technol.*, pp. 339–342, 2015.

(Azar, 2016) F. Gharehchopogh, N. Jabbari, and Z. Azar, "Evaluation *International Journal of Computer Applications (0975 – 8887) Volume 154 – No.3, November 2016* 39

(Riadi, 2016) A. Iswardani and I. Riadi, "Denial Of Service Log Analysis Using Density K-Mans Method," vol. 83, no. 2, pp. 299–302, 2016.

(Rahmani, 2014) Md. Khalid Imam Rahmani; Naina Pal; Kamiya Arora, "Clustering of Image Data Using K-Means and Fuzzy," *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 7, pp. 160–163, 2014.

(Atkins, 2015) R. Shaw and A. S. Atkins, "Conceptual Analysis of Cybercrime Events in Profiling Business Attacks."

(Mayo, 2013) P. Peña, R. del Hoyo, J. Vea-Murguía, C. González, and S. Mayo, "Collective knowledge ontology user profiling for twitter: Automatic user profiling," *Proc. - 2013 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2013*, vol. 1, pp. 439–444, 2013.

(Shobana, 2014) P. Jayakumar and P.Shobana, "Creating Ontology Based User Profile for Searching Web Information," no. 978, 2014.