Log Classification using K-Means Clustering for Identify Internet User Behaviors

Muhammad Zulfadhilah
Department of Informatics
Politeknik Hasnur
Banjarmasin, Indonesia

Imam Riadi
Department of Information Systems
Ahmad Dahlan University
Yogyakarta, Indonesia

Yudi Prayudi
Department of Informatics
Universitas Islam Indonesia
Yogyakarta, Indonesia

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
Clustering, K-Means, Network, Log, Cyber-profiling

1. INTRODUCTION
The most important thing in the management of a network is to know the characteristics of the network users; the traffic analysis can help administrators to create policies and protect network security [1]. APJI survey that was conducted in 2014 shows the sequence of activities of Internet users in Indonesia are: users of social media, information search, chat, news, video, and email. These results indicate that the search for news and email usage are not included in a popular activity [2].

At this digital era, the use of social media produces more information than ever before. This suggests that big data requires data mining technology to address these challenges [3]. The clustering algorithm is one of the effective algorithms to analyze big data to describe the attributes available [4] . In general, the behavior of mobile phone users often makes browsing activity. This shows that the relationship between browsing activities and daily activities [5]. Cyber-profiling studies are the exploration of data to determine the characteristic of user activity when using the computer/ internet.

One of the algorithms that can be used in helping the cyber-profiling is a K-Means algorithm. This algorithm will perform the categorization of Internet users based on the number of visitors to a website. This will show what is frequently accessed by the user so they will know the behavior of their users activities on the internet. Profiling is the process of collecting data of individuals or groups who can produce something interesting, surprising and significant [6]. Cyber-profiling has been brought a good step for forensic computer science; it is based on the experience that has been done [7].

In accessing the internet, the locations will provide significant information to determine a person's behavior [8]. The use of internet services on a campus that can petrify the educational activities is sometimes also used for criminal or illegal activities. So to find out what can be accessed by Internet users in educational institutions, a cyber-profiling is needed.

2. CURRENT RESEARCH
In research conducted by [1] stated that cyber-profiling could assist administrators in determining the policies and the need for repairs to a network of user information. A company needs to collect and analyze customer data to identify the characteristics of its target customers [9]. Another study conducted by [10] states that the conclusion of cyber-profiling must use the deductive method. This is because if they only make inductive inference will occur extremely unreliable, and may cause misunderstanding in the analysis.

In the study conducted by [5] states that there is a correlation of Internet usage with daily activities. A survey by [2] showed that in 2014 there were 88 million Internet users in Indonesia. These results stated there are three main reasons people accessing the Internet, namely communication, a daily source and up to date. Based on the three main reasons that there are four main activities of Internet use, namely social media, find information, chat and search for the latest news.

In the study conducted by [11] states that to control the policy and network congestion, network operators are encouraged to design appropriate mechanisms in the supply of resources for consumers based on various categories of application used. Besides that, a study conducted by the [6] also mentioned that the presence of profiling could give a warning to the teenage users in sharing personal information when accessing the internet.

According to a survey conducted by [12] stated that there are nine categories of internet users' behavior, namely NetTerrorist, NetStreiver, NetAvoider, NetPublisher, Networker, NetCrawler, NetAdvocate, NetJungki and NetRookie.

3. BASIC THEORY

3.1 Data Mining

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The information can be used to increase revenue, cut costs or both, various data mining algorithms such as classification, clustering, the association used to extract information from the data potential [3].

Data mining has stages like in Figure 1 [3]. Data mining involves four tasks, namely Clustering, Classification, Regression and Association [13]. Data mining can detect useful knowledge from large datasets, such as pattern recognition, rules and trends [3].

3.2 K-Means

Clustering is a technique of grouping a set of objects in the same group (called clusters) that are more similar to each other than to those in the other groups (clusters). This is a major task in exploring data mining, and general techniques for the analysis of statistical data. Clustering is also used in various fields, including machine learning, pattern recognition, image analysis, information retrieval, and bio-informatics [14].

The main theory of the K-Means algorithm is a description of the K center point for each cluster. The selection of such centers should be exactly to their needs, because the selection affects the central point of the results obtained [15].

The method of K-means algorithm as follows [16]:

1) Initialization: determine the value of K as the number of clusters desired.
2) Select the K data from the dataset as a centroid.
3) Allocate all data to the nearest centroid by a predetermined distance metric.
4) Recalculate centroid based on data that follows each cluster.
5) Repeat steps 3 and 4 until the convergence condition is reached (no data is moved).

K-Means algorithm is one of the popular clustering algorithm; it is also an unsupervised algorithm used in clustering. These algorithms select the centroid and compare it with data points based on the similarity of characteristics at any point, so that the establishment of clusters based on the distance of the data points to the centroid [17].

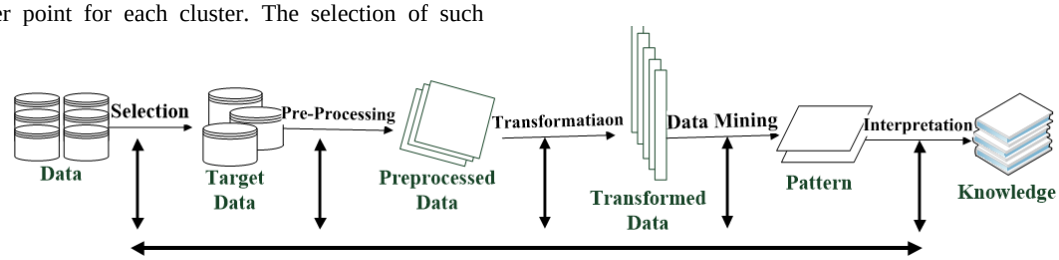

Figure 1 : Data Mining Process

3.3 Log

By definition, the log is a record of daily activities. While in the computer world is a log file that records computer activity. In digital forensics activities, log used as a support in the investigation process [16].

Table 1 is an example of data obtained.

Table 1 Example of Data

UNIX TIME	IP	PROTOKOL	WEBSITE
1460509501	192.168.	TCP_DENIED/407	armmf.adobe.com:443
1460509504	192.168.	TCP_DENIED/407	clients2.google.com:443
1460509510	192.168.	TCP_DENIED/407	clients2.google.com:443
1460509510	192.168.	TCP_DENIED/407	clients2.google.com:443
1460509510	192.168.	TCP_MISS/200	pixel.facebook.com:443
1460509524	192.168.	TCP_MISS/200	connect.facebook.net:443
1460509526	192.168.	TCP_MISS/200	syndication.twitter.com:443
1460509526	192.168.	TCP_MISS/200	d.kapanlaginetwork.com:443
1460509537	192.168.	TCP_MISS/200	static.xx.fbcdn.net:443
1460509539	192.168.	TCP_DENIED/407	tools.google.com:443
1460509539	192.168.	TCP_DENIED/407	tools.google.com:443
1460509542	192.168.	TCP_MISS/200	scontent-sin1-1.xx.fbcdn.net:443
1460509554	192.168.	TCP_DENIED/407	tools.google.com:443

3.4 Cyber-profiling

The use of data is fundamental to look at the relationship characteristics of users with mobile devices and applications as well as the type of access network [11]. Profiling is an individual or group information accumulated, stored and used for various purposes. One of which was to determine the usefulness of profiling activities of Internet users [7].

There are two types of profiling, which is deductive and inductive. Deductive profiling based on forensic evidence at the scene and the victim. While inductive profiling is an overview of the psychological regarding criminal behavior obtained from the tests and the cases that have been resolved [18].

Profiling on the Internet must be done by using inductive and deductive methods. This is done to prevent misunderstandings about the behavior of Internet users, because of the behavior on the Internet sometimes differ from the behavior in the real world [10].

Cyber-profiling is a result of the conclusion of the interests, characteristics, behavior, intentions and preferences of current user activity on the Internet [19]. Internet user profiles created to explain the background knowledge of the user [20].

4. RESEARCH METHOD

The data of Internet log activity obtained from educational institutions not only contain what is accessed by users, but other data packets on the network traffic activity also recorded. Therefore, the need for data cleansing process called pre-processing.

The steps of the K-Means algorithm:

1) Determine K as the number of clusters formed.

2) Generate K centroid (the center point of the cluster) beginning at random.
3) Calculate the distance of each object to each centroid of each cluster.
4) Allocate each object into the nearest centroid.
5) Iterating, then specifies the new centroid position.
6) Repeat step 3 if the new centroid is not the same.

Figure 2 is a flow of the algorithm K-Means:

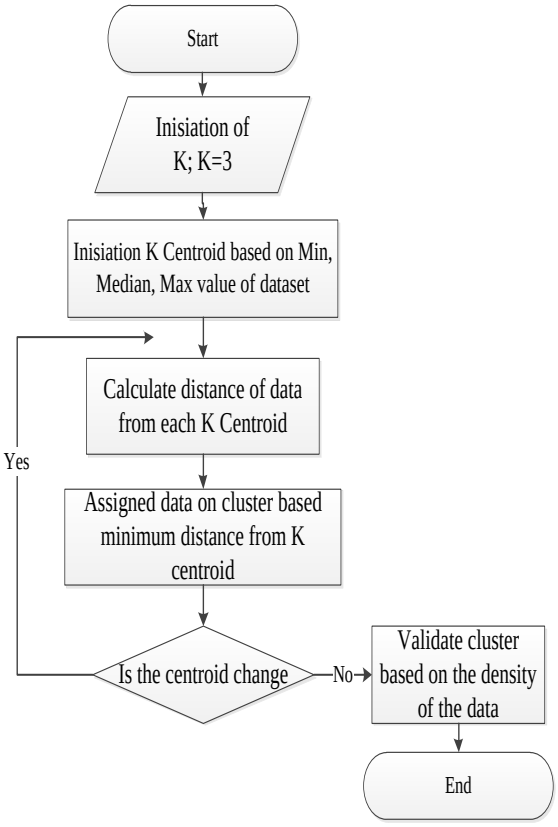

Figure 2 K-Means Algorithm

Data obtained from educational institutions as much as 320.773 records; this data still need to be processed before the clustering process. One of the steps before clustering is a pre-processing stage; this stage is to conduct a cleansing of data that is not needed in the research process.

The results of pre-processing shown that the data will be used as much as 1,638 records. The next stage after the pre-processing completes the process of clustering algorithms using K-Means. K-Means algorithm performed using SPSS and RapidMiner. The results of the cluster K-Means Algorithm Data will be analyzed to aid cyber-profiling process.

5. RESULT

5.1 Data Category

5.1.1 K-Means Clustering based on the Number of Visitor

This category of data will be in the cluster based on many visits to a website. Clustering performed using the K-Means algorithm in SPSS and RapidMiner. SPSS and RapidMiner are used to determine the cluster results obtained whether it is appropriate to proceed at this stage of cyber-profiling analysis.

K-Means algorithm implementation performed by the application SPSS and RapidMiner resulted in three clusters, namely low, medium and high. The first cluster is a cluster with low traffic levels have a total members of 1479 websites, the second cluster is a cluster with moderate traffic levels have a total members of 126 websites, and the last third cluster with high levels of traffic have a members of 33 websites.

Initialization of the initial cluster center in the clustering process can be seen in Table 2.

Table 2 Initialization Beginning of Cluster Center

	Initialization of Cluster Center		
	1	2	3
Number of Visitors	1	37	71

Initialize of initial values of the data in the cluster based on the highest value, the average and the smallest value. In this study there are eight iterations produced to get the right result. This initialization is performed by the application of SPSS and RapidMiner.

Iteration history in the clustering process can be seen in Table 3.

Table 3 Iteration History

Iteration	**Changes In Cluster Centers**		
	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

Table 3 shows that the need for 8 (eight) iterations to get the proper cluster. SPSS application states that the minimum distance between initial centers is 34. The result of the iteration process in determining the initial clustering center can be seen in Table 4.

Table 4 Final Result of Cluster Center

	Cluster Center		
	1	2	3
Number of Visitors	2	19	46

The results of clustering that has been done can be seen in Figure 3.

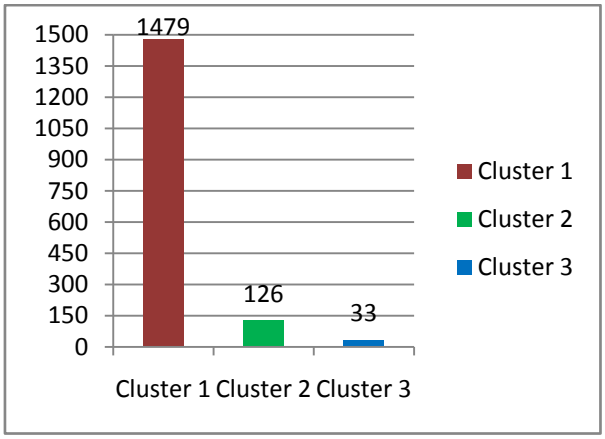

Figure 3 Result of Clustering

The results of clustering will be explained as follows:

- Cluster 1: this cluster is a cluster with the highest number of members, namely 1479 websites. The first cluster is a cluster with the level of user traffic slightly, ranging from 1-10 visits per website. This cluster has members mostly a website advertising.
- Cluster 2: websites which were included in this cluster as many as 126, with the number of clusters it then entered in the intermediate category because it has a higher value than the average value generated in the process of clustering ranged at 11-13 visits per website. This cluster contains more information and news sites.
- Cluster 3: this cluster has the fewest members, which is only 33 websites. However, this cluster has the highest traffic levels compared to other clusters. Values in this cluster are at 34-64 visits per website. This cluster contains more search engine and social media websites.

5.1.2 Category Based Access Time

In this section, the categorization is based on Internet access time by the user. This categorization is divided into four times of access, namely the "morning-afternoon", "afternoon-evening", "evening-night" and "night-morning".

In the category of "morning-afternoon", there are 1204 websites accessed by users with a total of as many users as 5150. In this category, the search site is one of the most visited by users, followed by video streaming sites that also have a lot of visitors in this category. This category has the most visitors in comparison to other categories because of in the morning-afternoon is a productive time for users to do activities related to the Internet.

In the category of "afternoon-evening", there are 962 websites that are accessed by 3,000 users. In this category, the search engine is one of the most visited websites, followed sequentially by streaming video sites, social media, and e-mail sites. This category is the second largest category; and the access time is still in productive hours.

In the category of "evening-night", there are 143 websites that are accessed by 157 users. In this category, the search engines are sites frequently accessed by users, but the number of visitors slightly. This is caused by the productive time is over, so that the productivity to access the Internet also declined.

In the category of time "night-morning", there are 27 websites that are recorded on the network data traffic with very minimal use. This is caused by the use of the Internet which is very rare. In this category, access to streaming video sites there is only one user, it indicates that there are those who use the Internet at these times.

The results of category based on access time can be seen in Figure 4.

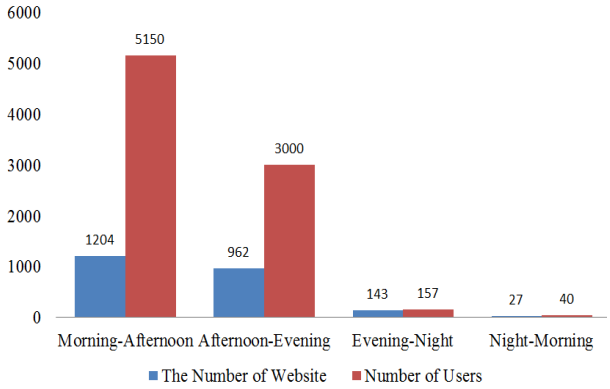

Figure 4 Result of Category Based on Access Time

Based on the categories of access time, the website that has the most visitors is the search engine, this is indicated in each time category contained search engine became the most accessible website.

5.1.3 Category Based Website Content

This section will explain the categories based on website content contained in the research data, the categorization of the website is taken from various sources on the Internet. Based on the 1638 websites obtained, there are 22 types of websites that can be categorized.

In Figure 5 is a type of website that successfully categorized.

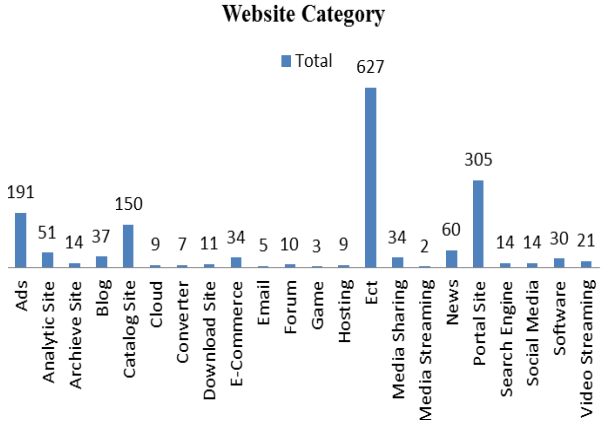

Figure 5 Website Category

These results are used to determine the categories of websites that are frequently accessed by Internet users in the educational institutions, so as to assist in concluding on cyber-profiling process.

5.2 Analysis Result

In this study, the log data of networks obtained from educational institutions. The data obtained was performed 3 categorizations, which is categorized by a number of visits to the website, based on the time of internet access by the user and based on the content of the website.

The categorization of data is divided into three categories: low, medium and high. The process of categorization performed by using the K-Means algorithm implemented by SPSS and RapidMiner. The clustering results obtained from the implementation of the K-Means algorithm showed that the use of the Internet for an educational institution to access the search engine, information websites and social media websites. This study is slightly different from the results of a survey conducted [2] which states that the use of the Internet is in this order: networks (social media), information search, chat (messaging), news search, video and email.

Categories of data based on the content website show that the Portal Site category is one category that has a website that is accessed by many users. This is because many websites related to information, but the average of every website in the category of this Portal Site has a low level of traffic, which is included in the first cluster when viewed based on the results of K-Means algorithm. For the category Etc., the website contained in this category are also very much, but websites that have categorized are still cannot be categorized more specifically because of the limitations of the data source, so the website Etc. is assumed as a pop-up that often appears on websites that accessible by Internet users.

Based on the results of the category of the access time, the morning-afternoon and afternoon-evening are the time that has the most visitors. This is because in these times Internet users in educational institutions do activities that require Internet access. In this category, the search engine on a website frequently accessed by users, followed by social media and streaming video website. While in the late evening-night visited websites have started a little bit, this is because productivity has ended. At the time of access night-morning, Internet usage was minimal despite the persistence of the users who use the Internet access during this period.

The result of the profiling process that has been done shows that the search engine and content information is frequently accessed by the user. This indicates that the daily activities in the real world and the environment in which access the Internet affect the activity of the user. This is according to research conducted by [10] which states that demographic factors affect the activity on the Internet. The results of the study also showed that Internet users in educational institutions belong to the category Networker and NetJunki; this is according to research conducted by [12].

The results of this research have been able to meet the definition of cyber-profiling for providing information on the use of Internet-based daily activities. The results of this study may be used by network administrators to enhance security, policy and network quality, this is according to research conducted by [1], [7] and [21].

6. CONCLUSION

Analyzes network traffic log data using an algorithm K-Means for the profiling process shows results in line with expectations of research, because it has a good degree of accuracy. K-Means algorithm produces three categories of traffic to the website, namely high, medium and low. The results are also similar to the results based on the categorization of data access time and based on the content of the website's content.

The results of this study also showed that the websites that have high levels of traffic in the sequence are searching website, information and social media. Results from cyber-profiling in this study are Internet users entered in character Networker and NetJungki, based on these characteristics indicate that cyber-profiling has been done strongly influenced by environmental factors and daily activities.

This study has limitations on the source data in the profiling process. To get the maximum results in the process of profiling, data obtained should contain about activity on the computer has been used. In future research, the cyber-profiling process should use the data obtained from the computer activities that have been used and also data from the computer user. It is expected to get the results of the analysis of cyber-profiling better.

7. REFERENCES

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

[2] APJII, "Indonesian Internet User Profile 2014," 2015.

[3] 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.

[4] 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.

[5] 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.

[6] J. J. Irvine, "Digital Forensic Analysis & Cyber-profiling," no. 703, pp. 1–32, 2010.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] F. Gharehchopogh, N. Jabbari, and Z. Azar, "Evaluation

of Fuzzy K-Means And K-Means Clustering Algorithms In Intrusion Detection Systems," *Int. J. Sci. …*, vol. 1, no. 11, 2012.

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

[17] 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.

[18] R. Shaw and A. S. Atkins, "Conceptual Analysis of Cybercrime Events in Profiling Business Attacks."

[19] 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.

[20] P. Jayakumar and P.Shobana, "Creating Ontology Based User Profile for Searching Web Information," no. 978, 2014.

[21] T. Bakhshi and B. Ghita, "Traffic Profiling : Evaluating Stability in Multi-Device User Environments," 2016.