Cyber Profiling using Log Analysis and K-Means Clustering

By Imam Riadi

Cyber Profiling using Log Analysis and K-Means Clustering

A Case Study Higher Education in Indonesia

Muhammad Zulfadhilah
Departement of Informatics
Politeknik Hasnur
Banjarmasin, Indonesia

Yudi Prayudi
Departement of Informatics
Universitas Islam Indonesia
Yogyakarta, Indonesia

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

***Abstract*—The Activities of Internet users are increasing from year to year and has had an impact on the behavior of the users themselves. Assessment of user behavior is often only based on interaction across the Internet without knowing any others activities. The log activity can be used as another way to study the behavior of the user. The Log Internet activity is one of the types of big data so that the use of data mining with K-Means technique can be used as a solution for the analysis of user behavior. This study has been carried out the process of clustering using K-Means algorithm is divided into three clusters, namely high, medium, and low. The results of the higher education institution show that each of these clusters produces websites that are frequented by the sequence: website search engine, social media, news, and information. This study also showed that the cyber profiling had been done strongly influenced by environmental factors and daily activities.**

***Keywords*—Clustering; K-Means; Log; Network; Cyber Profiling**

I. Introduction

The increasing number of applications, hardware (device), and an Internet connection has affected the behavior of its users. In this case, APJI has been reported that in 2014 the order of the activities of Internet users in Indonesia is: users of social networks (social media), information search, chat (messaging), news search, video, email as a user internet activity in order of popularity. The data also indicate that the search for news and email usage is not a popular activity [1].

In general, cyber profiling studies is the exploration of data to determine what user activity at the time of internet access. One method that can be used to support the profiling process is a K-Means algorithm. Through these algorithms, the data can be grouped by the number of websites visited. This grouping aims to see what the user frequently accesses websites.

The data of internet users access at an institution can be categorized as a large data type so that the analysis can be done with data mining. In this case, the cluster algorithm as one of data mining techniques can be used to find groups (clusters) of a useful object, which the used are depends on the purpose of data analysis [2]. Clustering analysis is one of the most useful methods for the acquisition of knowledge and is used to find clusters that are a fundamental and important pattern for the distribution of the data itself [3].

Profiling is the process of collecting data from individuals and groups which can produce something interesting, surprising and significant, correlations that by using a machine that has good strength calculations to detect such data, while we as humans cannot [4]. Meanwhile, cyber profiling brings a good step in forensic computer science, based on the experience that has been achieved in the process of handling that have been made [5].

Educational institutions are one of the most likely group to conduct Internet activities. User behavior in educational institutions is also necessary to know the characteristics of user profiling and access to what is being done. Among the Indonesian has not been any research related to this issue, that's way cyber profiling would be very useful to know the behavior of Internet users in higher education in Indonesia.

Internet usage in higher education should be utilized by the user to support the educational process, but sometimes the facts obtained they used the Internet for the purpose outside of education, even less so there is an indication of such a user on educational institutions leading to cyber-crime. For that, we need to know more whether the use of the Internet in education is in line with the scope of activity in the education process activities.

II. Current Research

A survey by APJI [1] showed that Internet users in Indonesia in 2014 reached 88 million. The survey was stated that there are three main reasons people use the Internet, namely access to social facilities/communications (72%), daily source (65%), and follow the development of the world (51%). The main reasons of internet access are practiced through four main activities, namely the use of social media (87%), searching for information (69%), instant messaging (60%) and search for the latest news (60%).

Research related to profiling, among others, performed by[6]. In these studies, [6] used machine learning to help the process of profiling to assist the experts in analyzing the crime.

Another study conducted by [4] and profiling results obtained knowledge of the risks of children and adolescents in accessing the internet. Based on these studies, [4] provide recommendations for caution in the use of personal data because the data will be accumulated and stored is likely to be used by parties who are not responsible.

In the study conducted by [7], the results of profiling can know the habits of Internet users and help network

administrators to improve the quality, security, and policy in the Internet network based on user behavior.

Meanwhile, [8] have also been doing profiling of Facebook users using the inductive method. However, the study also revealed that cyber profiling still has to use the deductive method because the cyber profiling process still requires additional data from the user completely. It to support the existence of differences in the behavior of individuals, because inductive generalizations extremely unreliable, and may cause misunderstanding in the analysis.

Another study conducted by [9] to use Twitter using ontology-based modeling OWL (Web Ontology Language), it is known that cyber profiling can be used to determine user interest based on URLs that have been shared via Twitter. The use of ontology also applied by [10], and the study revealed that cyber profiling using these methods could facilitate in providing information to the user when performing a search on a website.

III. Basic Theory

A. Data Mining

Data mining is an iterative and interactive process to find a new pattern or model valid, useful and understandable in a very large database. Data mining provides the search for patterns or trends that are desirable in a large database to help make decisions in the future. This pattern is recognized by a particular device that can provide a useful analysis and insightful data that can then be studied more carefully. The results of these patterns may be used in devices other decision support [2]. Data mining has stages like in Figure 1.

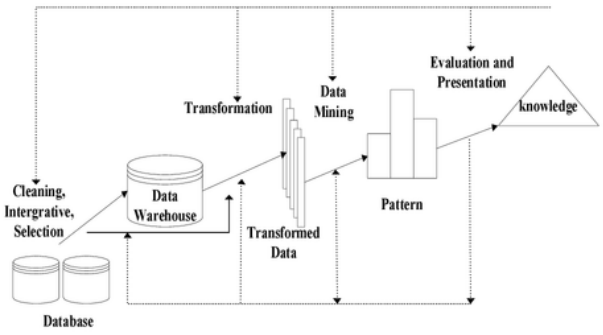

Fig. 1. Data Mining Process [10]

Data mining involves four tasks [11]:

1) Clustering – It is the task of finding a group and structure the data in some way or the "similar", without using known structures in the data.

2) Classification – It is the task of generalizing known structure to apply to new data. For example, an email program to attempt to classify an email as legitimate email or as spam.

3) Regression – Attempts to find a function which models the data with the least error.

4) Association rule learning – search the relationship between variables. For example, a supermarket might gather data on customer habits. Association rule learning can help supermarkets to determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

B. K-Means

Clustering is used to create a group (cluster) of the data so that it can easily find the necessary data. Clustering is a classification of similar objects into several different groups, it is usually applied in the analysis of statistical data which can be utilized in various fields, for example, machine learning, data mining, pattern recognition, image analysis and bioinformatics [11].

Clustering including supervised learning types. There are four types of clustering algorithms that have been compared based on performance, such as K-Means, hierarchical clustering, self-organization map (SOM) and expectation maximization (EM Clustering). Based on these test results can be concluded that the k-means algorithm performance and EM better than a hierarchical clustering algorithm. In general, partitioning algorithms such as K-Means and EM highly recommended for use in large-size data. This is different from a hierarchical clustering algorithm that has good performance when they are used in small size data [12].

The method of K-means algorithm as follows [13]:

1) Determine the number of clusters k as in shape. To determine the number of clusters K was done with some consideration as theoretical and conceptual considerations that may be proposed to determine how many clusters.

2) Generate K centroid (the center point of the cluster) beginning at random. Determination of initial centroid done at random from objects provided as K cluster, then to calculate the i cluster centroid next, use the following formula:

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad ; i=1,2,..,n \qquad (1)$$

v : cluster centroid xi : the object to-i

n : the number of objects to be members of the cluster

3) Calculate the distance of each object to each centroid of each cluster. To calculate the distance between the object with the centroid author using Euclidian Distance.

$$d(x,y) = \|x-y\| = \sqrt{\sum_{j=1}^{n}(x_i - y_i)^2} \qquad (2)$$

xi= object x to-i n= the number of object

yi= object y to-i

4) Allocate each object into the nearest centroid. To perform the allocation of objects into each cluster during the iteration can generally be done in two ways, with a hard K-means, where it is explicitly every object is declared as a

member of the cluster by measuring the distance of the proximity of nature towards the center point of the cluster, another way to do with fuzzy C-Means.

5) Do iteration, then specify a new centroid position using equation (1).

6) Repeat step 3 if the new centroid position is not the same.

C. Log

Log (record keeping) is a file that records events in the computer program. Meanwhile, according to the definition of the log is a record of daily activities. Activities that are recorded directly called the transaction log. The log file can be used as a support in the process of cyber forensics to obtain digital evidence during the investigation stage [14].

The cleaning process must precede analysis of log data or preprocessing. Preprocessing is performed to remove duplication of data, check the data inconsistency, and correct errors in the data, such as print errors (typography) [15].

In Table 1 is an example of data on educational institutions

TABLE I. EXAMPLES OF DATA

Waktu		IP		Protokol		Website
1460509257	166960	192.168.15.	0	TCP_MISS/200	10018	a.tribalfusion.com:443
1460509207	115146	192.168.15.	0	TCP_MISS/200	5665	a248.e.akamai.net:443
1460624909	115780	192.168.15.	0	TCP_MISS/200	5945	accounts.google.com:443
1460509823	1425613	192.168.15.	0	TCP_MISS/200	84404	accounts.google.com:443
1460510173	115941	192.168.15.	0	TCP_MISS/200	3551	accounts.google.com:443
1460510343	115456	192.168.15.	0	TCP_MISS/200	3343	accounts.google.com:443
1460510463	116206	192.168.15.	0	TCP_MISS/200	3247	accounts.google.com:443
1460508537	5476	192.168.15.	0	TCP_MISS/200	5607	ad.turn.com:443
1460677224	115980	192.168.15.	0	TCP_MISS/200	21069	addons.cdn.mozilla.net:44
1460625101	121110	192.168.15.	0	TCP_MISS/200	21037	addons.cdn.mozilla.net:44

D. Cyber Profiling

The idea of cyber profiling is derived from criminal profiles, which provide information on the investigation division to classify the types of criminals who were at the crime scene. Profiling is more specifically based on what is known and not known about the criminal [8].

Profiling is information about an individual or group of individuals that are accumulated, stored, and used for various purposes, such as by monitoring their behavior through their internet activity [4].

Difficulties in implementing cyber profiling is on the diversity of user data and behavior when online is sometimes different from actual behavior. Given the privilege in personal behavior, inductive generalizations can be very reliable but can also lead to a misunderstanding of behavior analysis. Therefore the cyber-profiling process is via a combination of deductive and inductive methods [8].

For investigation, the cyber-profiling process gives a good, contributing to the field of forensic computer science. Cyber Profiling is one of the efforts made by the investigator, to know the alleged offenders through the analysis of data patterns that include aspects of technology, investigation, psychology, and sociology.

Cyber Profiling process can be directed to the benefit of:

- Identification of users of computers that have been used previously.
- Mapping the subject of family, social life, work, or network-based organizations, including those for whom he/she worked.
- Provision of information about the user regarding his ability, level of threat, and how vulnerable to threats
- Identify the suspected abuser

In a broader scope of cyber profiling can provide support information in a case, such as counterintelligence and counterterrorism [5].

The process of profiling against criminals often also known as cyber-criminal profiling criminal investigation or analysis. Criminal profiles generated in the form of data on personal traits, tendencies, habits, and geographic-demographic characteristics of the offender (for example: age, gender, socio-economic status, education, origin place of residence). Preparation of criminal profiling will relate to the analysis of physical evidence found at the crime scene, the process of extracting the understanding of the victim (victimology), looking for a modus operandi (whether the crime scene planned or unplanned), and the process of tracing the perpetrators were deliberately left out (signature) [16].

The new approach to cyber profiling is to use clustering techniques to classify the Web-based content through data user preferences. This preference can be interpreted as an initial grouping of the data so that the resulting cluster will show user profiles [17].

User profiling can be seen as the conclusion of the interests of users, intentions, characteristics, behavior and preferences [9]. User profiles are created for a description of the background knowledge of the user. User profile represents a concept model which is owned by the user when searching for information web [18].

IV. RESEARCH METHODS

To determine cyber-profiling of the higher educational institutions, so in this study the sample data is a log of Internet activities from one educational institution. Log data do not only contain any websites accessed by the user, but also includes packets received and sent over the network traffic. Data obtained containing the activities of network traffic for five days and produce data as much as 320.773 records.

In the early stages of research data collection, then do preprocessing that the data did not meet the criteria can be eliminated. Preliminary data obtained from 320.773 into a 1.638 record with the results of preprocessing. Furthermore, the mechanism of clustering using K-Means algorithm running on Rapid Miner and SPSS applications. The cluster data is then analyzed to make the process of profiling against internet users.

Figure 2 is a flow of the application of K-Means algorithm in the profiling process.

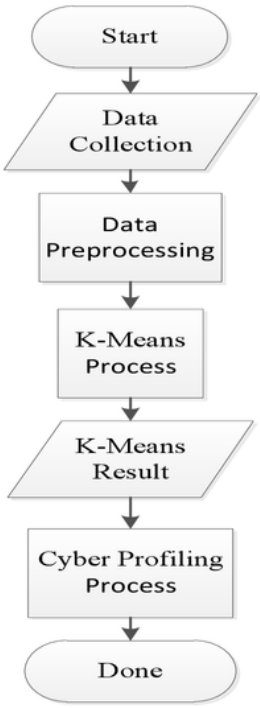

Fig. 2. Flow Research

Figure 3 is a flow of the algorithm K-Means:

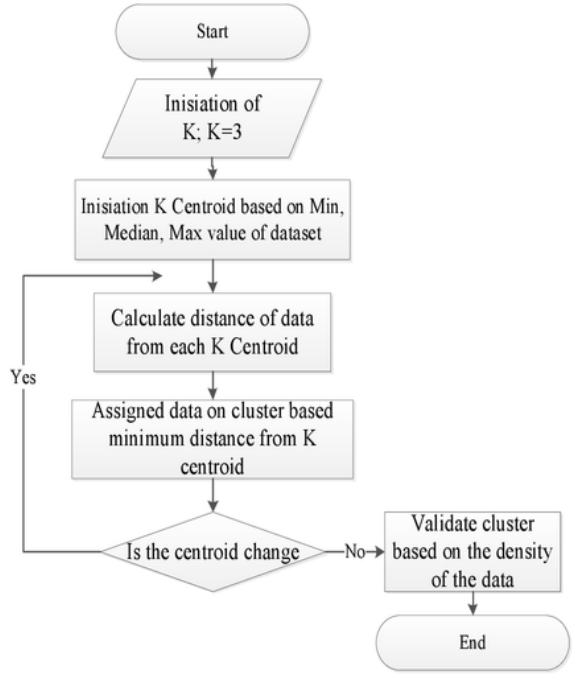

Fig. 3. K-Means Algorithm

V. Result

A. K-Means Clustering

Implementation of the K-Means algorithm, the result obtained is a level of visits to the website. The visit is divided into three groups: low, medium, and high.

Clustering by Rapid Miner and SPSS application indicates that the output produced has the same cluster of data. Based on the results of the cluster, it appears there are three clusters whose value is different, even on the first cluster value reached in 1479 (90.30%), the second is worth 126 (7.70%), and the third is worth 33 (2.00%). Those values represent the number of websites that have been divided in each cluster. Clustering results have shown this process has been running as expected to research.

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

TABLE II. Initialization Beginning Cluster Center

	Cluster		
	1	2	3
Number of Visitors	1	37	71

An initial value is determined based on the data that have the highest value, the median value, and the smallest value. Those values are at the center of the initial cluster that will be followed in the process of K-Means.

The new centroid calculations will continue to do (iteration) until the discovery of iterations where centroid result is the same as the results of the previous centroid. In this study, there were eight iterations to determine the exact outcome of the 1638 cluster object. The iteration process can be seen in Table 3.

TABLE III. Clustering Process

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

The result of the iteration process in determining the initial clustering center can be seen in Table 4.

TABLE IV. Final Result of Cluster Center

	Cluster		
	1	2	3
Number of Visitors	2	19	46

The results of clustering details will be explained as follows:

- Cluster-1. On the results of clustering that has been done, the first cluster has as much data as 1467 records. The first cluster has the most members, but this cluster has a value which is below the overall average of the

data studied. In the first cluster has a data value in the range of 1-10, because in this cluster of existing data has a low level of traffic. Thus, cluster unity categorized on the website that has the least traffic from another cluster.

- Cluster-2. In the second cluster, members who entered at this cluster of some 126 records. The value of the results of the second cluster is in the range 11-31. This value indicates that the members of the second cluster have a medium level visits, because it has a higher value than the average value generated by clustering. Thus, the second cluster of clusters categorized as having moderate traffic levels.
- Cluster-3. On the results of the third cluster, cluster members who sign on as many as 33 records. The results of this third cluster have the fewest number of members in comparison with other clusters, but the members of this cluster have the highest value of the data that has been generated. The value in this cluster is in the 34-63 range, pointing to a result that the third cluster has a value far above average. Thus, the third cluster is categorized as a cluster that has the highest traffic levels.

The Results of clustering that has been done can be seen in Figure 4.

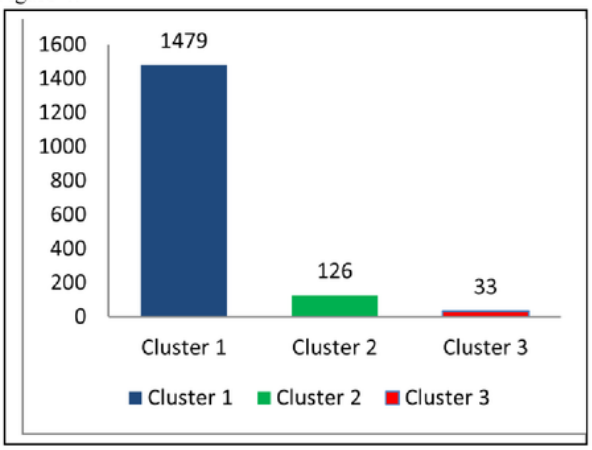

Fig. 4. The Result of Clustering

B. Analysis Results

In this study, the algorithm K-Means clustering has been implemented to perform in line with expectations. In the early stages of primary data obtained containing information about the websites accessed by users via the internet. In addition to the data contained informative website also contains data that updates to the operating system, the update of the web browser, and website advertising that usually appears as a pop-up.

Based on the results of the K-Means as shown in Figure 4 indicate that each cluster obtained having a different number of significant cluster members.

In the first cluster have shown low levels of traffic, but has some websites most. Data on the first cluster contains most of the advertising media website that coincided with a visit to a website activity. Meanwhile, in the second cluster that has moderate traffic levels, the data indicate a cluster member news sites that are in this cluster.

On the results of the third cluster is a group of websites with the highest traffic levels, but has the least number of websites. Data in this cluster shows that social media is a website with traffic levels were relatively high. Other data from the third cluster shows that Internet users access website search engine more frequently than from other websites including social media websites.

Although the research by [19] mention that the K-Means algorithm has shortcomings in central initialization beginning, in this study there were no such deficiencies. Clustering by two applications, Rapid Miner and SPSS produce the same data, this indicates that the determination of the value of the initial cluster center of these two applications has the same initial value.

Based on the results of clustering has been obtained that Internet users in educational institutions often make access to website search engine for information related to their field. This study also obtained results, that the social media website and streaming video sites accessed more frequently than the information and news website. These results caused by the delivery of information in the digital age now entering the realm of social media compared to other information websites.

In this study, social media included in the website frequently accessed by the user. This is according to research conducted by [1] [9] which states that social media becomes one of the surfing activity frequently accessed.

Cyber profiling process that has been done shows that the search for information more frequently accessed by users coming from educational institutions. This indicates that environmental factors and daily activities affect on what is accessed by the user. These results refer to the study [8] which states that the process of cyber profiling to predict based on the demographic information has a high degree of accuracy.

Based on the above data profiling results to higher educational institutions indicate that the use of the Internet has been used to support the educational process. The source of the data obtained showed no user activity in the area of higher education that leads to cybercrime.

In this case was supposed to complete the profiling results as mentioned by [5], the source of data should contain log activity on a computer that had been used, but in this research, the data has only been in the form of log data network traffic.

In this study cyber profiling can only provide information about the Internet activities performed by users, but for the threat of crime and profiling based on data from computers that have been used as mentioned by [5] are still not exist.

The results of this research have been able to meet the definition of cyber profiling, because it provides information about the user based on the current activity connected with the internet. The results of this study can be used by network administrator to improve the quality of services, policies, and security as mentioned by [7].

VI. CONCLUSION

The results of log analysis datasets using the K-Means algorithm to cyber profiling process show that the algorithm has to group activity based on the data of internet users visited the website. This grouping is divided into three, namely the visit low, medium, and high.

In this study, the K-Means algorithm is used as an algorithm for the cyber profiling process. K-Means algorithm being used is in line with expectations from this study, because it has a simple algorithmic process with a good degree of accuracy. But the K-Means algorithm has disadvantages, namely the process of making an initial value initial random center. This can lead to differences in the results of the cluster.

The results of this study indicate that Internet users in higher educational institutions are more accessible to website for searching information. The results also show that social media has a high-level visit after website search engine.

This study has limitations in the source of data for the profiling process. For the perfection of the profiling, the process should contain the data of any computer activities. Therefore, further research is expected to perform better cyber profiling with the more complete data source.

REFERENCES

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

[2] Fajar Astuti Hermawati, Data Mining. Yogyakarta: CV. Andi Offset, 2013.

[3] H. Chunchun, L. Nianxue, Y. Xiaohong, and S. Wenzhong, "Traffic Flow Data Mining and Evaluation Based on Fuzzy Clustering Techniques," vol. 13, no. 4, pp. 344–349, 2011.

[4] 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.

[5] J. J. Irvine, "Digital Forensic Analysis & Cyber Profiling," no. 703, pp. 1–32, 2010.

[6] A. S. N. Chakravarthy, "Analysis of cyber-criminal profiling and cyber-attacks : A comprehensive study," no. September, 2014.

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

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

[9] P. Peña, R. Hoyo, J. Vea-murguía, C. González, and S. Mayo, "Collective Knowledge Ontology User Profiling for Twitter Automatic User Profiling," pp. 439–444, 2013.

[10] C. J. Lei Xu, J. Wang, J. Yuan, and Y. Ren, "Information Security in Big Data : Privacy and Data Mining," pp. 1149–1176, 2014.

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

[12] I. Riadi, J. E. Istiyanto, and Su. S. Saleh, "Internet Forensics Framework Based-on Internet Forensics Framework Based-on Clustering," no. January, 2013.

[13] N. S. Ediyanto, Muhlasah Novitasari Mara, "Characteristics classification by Method K-Means Cluster Analysis," Bul. Ilm., vol. 02, no. 2, pp. 133–136, 2013.

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

[15] Universitas Sumatera Utara, "Decision Tree," Repos. II.pdf, 2012.

[16] Margaretha, "Criminal Profiling dan Psychological Autopsy," http//psikologiforensik.com/2013/04/22/criminal-profiling-dan-psychological- Autops., 2015.

[17] P. B. Costa, S. Oliveira, and L. Nunes, "Profiling Web Users Preferences with Text Mining," pp. 1–4, 2013.

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

[19] S. Andayani, "Formation of clusters in Knowledge Discovery in Databases by Algorithm K-Means," 2007.

Cyber Profiling using Log Analysis and K-Means Clustering

ORIGINALITY REPORT

2%

SIMILARITY INDEX

PRIMARY SOURCES

1	the.uncanny.valley.en.ddds.info Internet	20 words — < 1%
2	Shan Feng, , Ruifang Liu, Qinlong Wang, and Ruisheng Shi. "Word distributed representation based text clustering", 2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems, 2014. Crossref	10 words — < 1%
3	Jing Hua. "Localized Feature Selection for Clustering and its Application in Image Grouping", Multimedia and Expo 2007 IEEE International Conference on, 07/2007 Crossref	9 words — < 1%
4	Bui, Quang Vu Sayadi, Karim Bui, Marc. "A multi-criteria document clustering method based on topic modeling and pseudoclosure function.(Repo", Informatica, June 2016 Issue Publications	9 words — < 1%
5	regpro.mechatronik.uni-linz.ac.at Internet	8 words — < 1%
6	www.stochastik.uni-hannover.de Internet	8 words — < 1%
7	ijircce.com Internet	8 words — < 1%
8	Advances in Intelligent Systems and Computing, 2016. Crossref	7 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	OFF		