Forensic Analysis using Text Clustering in the Age of Large Volume Data: A Review

Bandar Almaslukh
Department of Computer Science, College of Computer Engineering and Sciences
Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

***Abstract*—Exploring digital devices in order to generate digital evidence related to an incident being investigated is essential in modern digital investigation. The emergence of text clustering methods plays an important role in developing effective digital forensics techniques. However, the issue of increasing the number of text sources and the volume of digital devices seized for analysis has been raised significantly over the years. Many studies indicated that this issue should be resolved urgently. In this paper, a comprehensive review of digital forensic analysis using text-clustering methods is presented, investigating the challenges of large volume data on digital forensic techniques. Moreover, a meaningful classification and comparison of the text clustering methods that have been frequently used for forensic analysis are provided. The major challenges with solutions and future research directions are also highlighted to open the door for researchers in the area of digital forensics in the age of large volume data.**

Keywords*—Digital investigation; forensic analysis; text clustering

I. Introduction

Digital Forensic Investigation (DFI) is the process of exploring digital devices in order to generate digital evidence related to an incident being investigated [1]. The six steps of the Digital Forensic Investigation (DFI) process as stated by DFRWS (Digital Forensic Research Workshop) illustrated in Fig. 1. First, the identification phase where all the components, devices, and data related to the incident are determined. After that, the preservation phase is conducted by avoiding any activities that can damage the collected digital information.

The next step is collecting the digital information that could be related to the incident under investigation, named the collection phase. Then, the examination phase is used for in-depth systematic search of evidence related to the incident being investigated. In the analysis phase, the investigator derives a conclusion for the evidence collected in the examination phase. Finally, the findings are summarized and presented to the court of law in the presentation phase.

However, over several years, the issue of digital investigation in large volume data has been raised increasingly. Many studies indicated that this issue should be addressed to find efficient solutions. For example, in [2] authors state that the coming digital forensic crisis is the growing size of storage devices since the tasks of collecting and analyzing and presenting a terabyte of data in a short report is more challenges. In addition, the ever growing in storage number and capacity with lack of adequate automated analyzing techniques are considered as one of the main current challenges in digital forensics filed [3-6]. In [7] the challenges posed to the digital forensics by the problem of big data are discussed.

The problem of big data can lead to wrong decision-making, falling to find evidence or loss of life in dangerous cases [7]. Specifically, the task of examination and analysis become more challenges in the age of big data since the current forensics tools cannot cope with large volumes of data. The limitation of these tools is designed for a relatively small volume of data (up to 1 Terabyte).

However, it is common in the age of big data that the volume of data that need to be analyzed can extend from a number of terabytes up to a couple of petabytes. To cope with the large volumes of data researches have used clustering algorithms as an alternative approach to speed up the examination and analysis phases. Since a great deal of the stored data is linguistics in nature (textual) [8], specifically the text clustering techniques have been utilized.

Text clustering is thematically assigning the text documents to separate groups where documents in the same group are more similar than other groups. Clustering methods are usually used for data analysis in which there is no prior or little knowledge about the data [9]. This is specifically the case in numerous applications of computer digital forensics addressed in this work.

In this paper, we conduct a comprehensive review for the state of the art research works that utilizing text clustering in the digital forensics investigation process. The main objective of the paper is to provide the reader with the recent text clustering techniques used in the context of digital forensics and benchmarking these techniques. In addition, the ideas of where the research might go next for the researchers who are interesting in this filed are provided.

The rest of the paper is structured as follow: Section II states the literature review of text clustering algorithms. Problem statement and motivation are presented in Section III, as well as, the summary of related works is shown in Section IV. Section V gives the applicability of clustering techniques in the literature that are utilized for large textual data. Conclusion is drawn in Section VI. Finally, future trends are presented in Section VII.

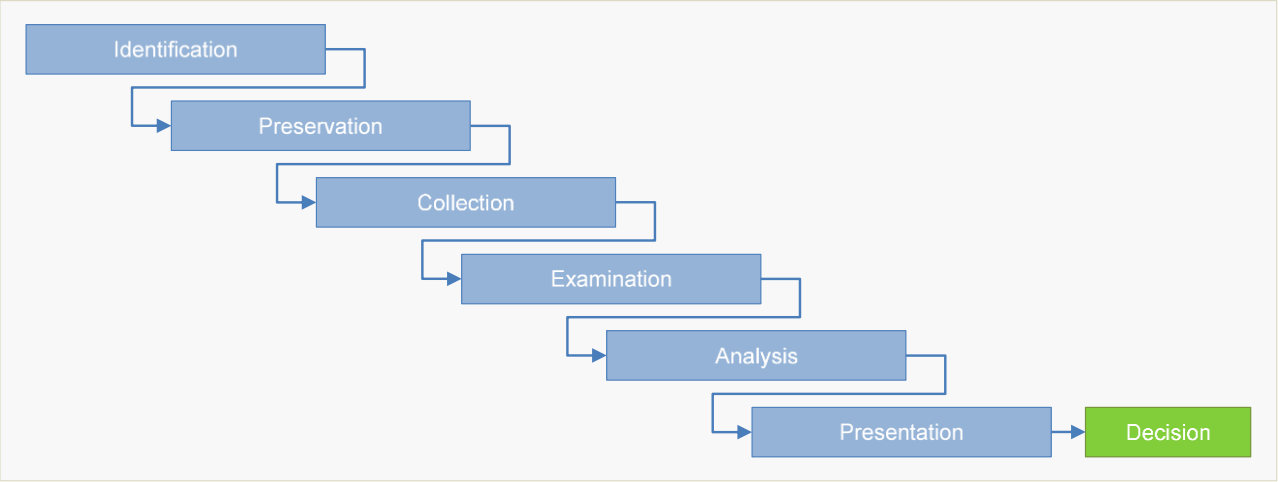

Fig. 1. The Digital Forensic Investigation (DFI) Process as Stated by the Digital Forensic Research Workshop (DFRWS).

II. Literature Review

Clustering algorithms have been studied since 1990; the literature on this subject is enormous. However, only few researchers have utilized the text clustering algorithms in the digital forensics context. The main algorithms used are SSOM (scalable self-organizing map), k-means, kernel k-means, hierarchical clustering, and Latent Dirichlet Allocation (LDA). Essentially, we can classify the related works to fifth classes based on the clustering techniques used, as follow:

A. Scalable Self-Organizing Map (SSOM)

A Kohonen self-organizing map (SOM) [18] is a type of unsupervised learning based on neural network approaches. This network consists of two layers: an input layer and an output layer (normally two-dimensional map). The input layer considered as a distributed layer, where the number of nodes in this layer as many as the input object features. After that, the network is trained to transform high-dimensional objects (input layer) to two-dimensional objects (output layer). However, in [19], scalable self-organizing map (SSOM) takes advantage of sparsity in the text document feature vector to improve the computational complexity of the SOM.

In the context of digital forensics analysis, the conceptual documents clustering using SSOM have utilized. For example, in [10], the SSOM clustering algorithm was used to cluster the search hits retrieved by the computer forensics tools. The software named Grouper was developed to evaluate the proposed method. Despite the required computer processing time, the proposed method reduced the examiner analytical time by around 80%. The limitation of this research is that testing the proposed method on a relatively small dataset (40 gigabytes) which is not the case in the age of big data.

B. Subject-based Clustering

The study offered in [11] proposed a novel subject-based semantic document clustering approach. This approach clustering the documents stored on seized devices according to the subjects provided by the investigator. The main intuition behind this algorithm is to extend semantically the subjects provided by the investigator using WorldNet [20] and a list of forensic specific synonyms. To model the proposed algorithm, a new subject vector space model (SVSM) was formulated. This model based on the vector space model (VSM) [21] and topic-based vector space model (TVSM) for information retrieval [22]. SVSM is an n-dimensional space model. Each dimension in the space characterizes one subject, where each dimension is orthogonal from each other (independent from each other). Terms and documents representation in SVSM is similar to their representation in TVSM.

However, the limitation of this research work is that the produced clusters based on the investigator experience to provide the appropriate subjects. In addition, it seems unlikely that the expert investigator can know all the criminals' events occurred in seized devices. To overcome this limitation, the work in [12] applies the same subject-based clustering algorithms in [11], with the help of subject suggestion that provided to the investigator. The provided subject suggestion improves the task of an investigator in terms of accuracy and speed. However, the subject-based semantic document clustering method in [11] assigns each document to the appropriate cluster, and all documents that do not belong to any subject are grouped in a separate cluster (named "general"). The problem is that the documents in the "general" cluster are belonging to different subjects. Therefore, the research work in [13] solves this problem by clustering the documents in the "general" cluster using bisection k-means algorithm.

C. Kernel K-mean

The "k-means" term was used first time by James Macqueen in 1967 [23]. It has been using intensively in the field of data mining as an unsupervised learning technique. K-mean grouping the set of observation $\mathcal{D} = \{D_i; i = 1, .., N\}$ into a set of K clusters, where $K \leq N$. The N observations are grouped in K disjoint clusters where the intra-cluster similarity between observation more than inter-cluster. The limitation of conventional k-means, it cannot detect the non- linear separable clusters accurately. Therefore, the kernel version of k-means was proposed in [24], to detect the non-linear separable clusters. The intuition of kernel k-means is to map all observation, D_i to another space using one of the will know kernel function such as the sigmoid kernel. By mapping to the new space, the observation clusters will be linearly separable.

However, the research work in [17] utilized the kernel-based version of k-means to clustering the documents stored on suspect devices. It adopts the method in [25] to measure the similarity between documents, where the Euclidian distance and term base on stylistic information used. Enron dataset used to measure experimentally the performance of the proposed work. The weakness of this research is that evaluating the proposal using the email dataset only which not reflect the real world cases where the textual data from different resources.

D. LDA (Latent Dirichlet Allocation)

LDA is a generative probabilistic model of a discrete data collection such as text corpus [26]. Essentially, it represents the documents as random mixtures over latent topics, and each topic is considered as a distribution over words. LDA has been considering as one of the best text modeling approaches, which automatically discover hidden topics from document corpus.

However, Authors in [14] showed how the topic modeling approaches could be applied to the forensic data. Specifically, the LDA clustering algorithm used to facilitate the examination and analyses phases in DFI process. In addition, the challenges posed by digital forensic data to the topic modeling algorithms mentioned. They conclude that topic modeling is beneficial for realizing the semantic of text documents in forensic data as well as summarizing the content of the documents.

In addition, a profound comparison between keyword-based search techniques and LDA is accomplished in [7]. The comparison was conducted on Real Data Corpus (RDC), which was collected from 2400 disks belonging to a real user. They conclude that LDA topic analysis should not be considered as a replacement of keyword-based search techniques, but it offers some benefits. The first benefits are relaxing the condition to match the keyword with the exact word appeared in the document. In addition, facilities the documents browsing by grouping documents based on the topic.

E. Benchmarking different Clustering Algorithms

The research works under this section was proposed to benchmark diverse clustering algorithms for forensic analysis. For instance, in [15] authors have proposed an approach that utilizes document clustering algorithms for the forensic analysis of seized digital devices. They realized that the majority of the related works in digital forensics assumes that the number of clusters is prior known, but in reality, the number of clusters varies from one case to the other. Thus, the relative validity index such as Silhouette [27] has used to estimate the number of clusters from the data automatically. The proposed approach is demonstrated by carrying out an extensive comparative study of the six well-known clustering algorithms (k-means, k-medoids, hierarchical clustering (Single/Average/complete link) and cluster ensembles (CSPA)), with a different mixture of parameters. In order to make the comparison more realistic, these algorithms applied to five real-world investigation cases seized by the Brazilian Federal Police Department. They conclude that the hierarchal algorithms (Average Link, complete Link) produce the best result in term of accuracy and stability. In addition, it had been shown how the hierarchal algorithms could be used to facilitate the digital examination task. However, we believe that hierarchal clustering techniques are not some good choices when the datasets are very large since they have a high computation cost. Another interesting research to benchmark some of the clustering algorithms in the context of digital forensic analysis proposed in [16]. In order to analyze the text string search output, k-means, SOM, LDA followed by k-means, and LDA followed by SOM clustering algorithms were used and evaluated. It realized that LDA follows by k-means accomplished the best performance; also, k-means and SOM achieve a better performance when they combined with LDA. Unfortunately, the poor evaluation was achieved since only small size synthetic data is used (up to 10 gigabytes), which is not the case in real-world data. To improve the performance of document clustering algorithms for criminal news, the authors in [28] proposed to use a Memetic Algorithm Feature Selection (MAFS) method with k-means and Spherical k-means (Spk) clustering algorithms. They achieved in somehow good clustering labels. In [29], the authors used a hierarchical clustering algorithm to distribute the chunks of a certain file type from memory into their corresponding files. This method does not need any information about the number of clusters. However, the dataset size used in this work is very small. Moreover, the hierarchical clustering algorithm is more sensitive to noise and outliers, as well as it is difficult to handle different sized clusters and clusters with convex shapes.

III. SUMMARY OF THE LITERATURES

In order to provide the reader with the main characteristics of the related work, a summary is provided in this section. The main features used to designate the related work are the size of the datasets used, including the semantic between words and identifying the cluster label. The size of datasets is considered as an important feature since it reflects the scalability in term of time complexity, and considering semantic between words provide a more accurate result. In addition, good cluster labeling helps the investigator to identify the semantic content of the clusters. Based on text clustering techniques that are used for digital forensics, the related works are classified into five classes as shown in Table I.

TABLE I. RELATED WORKS SUMMARY

Clustering Algorithms		Paper (Publication Year)	Is semantic between words is included	Good Cluster Labeling	Achieved Accuracy (%)	Dataset Size
SSOM (Scalable Self-Organizing Map)		[10] (2011)	No	No	70.9	Small (up to 40 gigabytes)
Subject-based Clustering (Actually this in not unsupervised clustering techniques since the clustering here based on the subjects provided by investigator)		[11] (2013)	Yes	Yes *	72	Small (up to 3893 documents)
		[12] (2014)	Yes	Yes *	80	Very Small (up to 100 documents)
		[13] (2014)	Yes	Yes *	65	Small (up to 3893 documents)
Kernel k-means		[2] (2009)	No	No	-	Small (up to 3331 documents)
LDA (Latent Dirichlet Allocation) (topic modeling)		[7] (2014)	Yes	Yes	-	Relatively Large (up to 98529 documents)
		[14] (2008)	Yes	Yes	-	Very Small (up to 837 documents)
Benchmarking Different Clustering Algorithms	k-means, k-medoids, hierarchical clustering (Single/Average/complete link) and cluster ensembles (CSPA)	[15] (2013)	No	No	91	Very Small (up to 131 documents)
	K-mean, SOM, LDA + K-mean and LDA+ SOM	[16] (2014)	Yes	Yes	67	Small (up to 40 gigabytes)
Memetic Algorithm Feature Selection with k-means and Spherical k-means (Spk)		[28] (2018)	No	No	-	Small (up to 4195 documents)
Hierarchical clustering algorithm		[29] (2018)	No	No	84.9	Very Small (20 files)
	Yes *: means that the cluster labeling based on the subjects provided by investigator.					

IV. PROBLEM STATEMENT AND MOTIVATION

In recent years, a major challenge to digital forensic examination and analysis phases is the ever growing in the number and volume of digital devices seized by the digital forensic agencies for investigation. This is a consequence of the ongoing development of storage capacity and computing technologies, as well as the number of devices seized per case has increased. In addition, the number of backlogs of seized devices waiting for analysis (regularly many months to years) has increased rapidly.

However, in order to reduce the overall examination time, many digital forensic tools such as FTK[1], Encase[2], etc. have been developed. The main techniques utilized in these tools are keyword search, regular expression search, and approximate matching search. These tools are designed to accomplish 100% query recall to retrieve all relevant documents, regardless of the extremely high proportion of non-relevant documents retrieved (very low precision). The limitations of these techniques are applied against the entire stored data (e.g., email document, internet history, instant message, word documents PDF files, etc.) without prior knowledge about the similarity amongst the documents. In addition, they are limited to the background knowledge about the case as well as the used search terms from the investigator's personal experience. Thus, the search hit of these techniques suffers from a large number of false negative and false positive. Therefore, the examiners still have to analyze the data manually in order to find potential evidence. However, this process is time-consuming, exceed the expert examiner ability and prone to human error.

Indeed, these challenges have led many researchers [10,11,12,13,14,7,15,16,17] to intentionally use different approaches such as machine learning and data mining in digital forensics for semi-automatic data analysis, in particular algorithms for document clustering. The clustering algorithms normally utilized for exploratory data analysis, when there is no prior knowledge about the data. This is exactly the case in the majority of digital investigation cases. However, the main idea behind document clustering algorithms is to group the objects from different clusters where the similarity between these objects within a cluster is more than the similarity between the objects in different clusters. Therefore, the examiner can perform preliminary analysis by investigating the representative documents of each produced cluster; making the task of examining the entire documents is avoided. Moreover, the investigator has the ability to prioritize the analysis of each cluster based on the relationship strength with the case under investigation.

The encouraging results of text clustering techniques in many fields are motivated the researchers to discover the usability of these techniques as a substitute approach to finding evidence in digital forensics filed. This encourages us to conduct a comprehensive review of the research works that addressing the problem of analyzing digital textual data in digital forensics using document-clustering techniques. However, this work is considered as a starting point for the researchers who interested in improving the accuracy and speed up the analyzing of large-scale textual data in digital forensics.

[1] http://accessdata.com/products/computer-forensics/ftk
[2] https://www.guidancesoftware.com/

V. Applicability of Clustering Techniques utilized in the Literature for Large Textual Data

Analyzing large volumes of text data in the digital forensics fields, we develop a set of criteria to evaluate these techniques as shown in Table II. These criteria are time complexity, tackling high dimensionality, number of input parameters. Time complexity is very important since we deal with a large volume of data. Tackling high dimensionality is very important criteria since the data type is text, which is high dimensional data. In addition, the number of input parameters is very important because a large number of parameters might reduce the cluster quality. Finally, for large datasets, the main strength and weakness of each technique are presented in Table II.

TABLE II. Proposed Evaluation Criteria

Text Mining Method	Technique	Time Complexity	Tackling High Dimensionality	Number of Input Parameters	Strength	Weakness
Traditional Clustering Methods	k-means	$O(NKd)$	No	1 (number of clusters)	Scalable	only discovering cluster with spherical shape
	Kernel k-means	If sampling is used $O(NmK + Nmd)$ Otherwise $O(N^2K + N^2d)$	Yes	1 (number of clusters)	Detect the non-linear separable clusters	Not Scalable
	Hieratical Clustering	$O(N^2)$	No	0	Can provide clusters at different levels of granularity	Not Scalable
	SOM	$O(N^2)$	Yes	5	It makes similarities between data easier to be observed and interpreted.	Similar objects could be split to more than one cluster
Topic Modeling	LDA	$O(NWKd)$ [24]	Yes	1 (number of topics)	Topics in corpus is identified clearly	It is hard to know interpretable topics when LDA is working and difficult when the design is not balanced.
Clustering Based on Information Retrieval Model	Subject-based Clustering	$O(N)$	Yes	Many set of words each set represents one subject	Assume that user has prior knowledge about data	Based on subjects provided by user

N: Number of document in the corpus K: Number of clusters D: Number of iteration W: Number of words in the document.

VI. Conclusion

The literature survey identified the potential future works remain in relation to forensic analysis using text clustering in the age of large volume data; for instance, validating text clustering on real world and large scale data, investigating the automatic approach for cluster labeling and bilingual clustering, etc. The related works are classified to fifth categories base on clustering techniques. The categories are SSOM, Kernel k-means and subject-based clustering, LDA, and benchmarking different clustering algorithms. In the last categories, more than one clustering techniques are compared in the context of digital forensics. However, the applicability of the clustering techniques utilized in the literature to analyze a large volume of text data is investigated.

VII. Future Trends

In this section, we stated several promising venues for future works. It is important to note that some of these future works are addressed partially in the literature and others not addressed at all. However, we can summarize the promising spots for the future works as follows:

1) Since the majority of the related works are validated on a relatively small dataset (up to 1 terabyte) or synthetic data, it is important to investigate the applicability of these methods on real-world as well as large datasets (from a number of terabytes up to a couple of petabytes).

2) Investigating the applicability of other clustering techniques, such as density-based clustering and bisection k-means.

3) Exploring the automatic approaches for cluster labeling to facilitate the analyzer task by identifying the semantic content of clusters.

4) In some country, it is common that seized devices have a document from two different languages such as English and Arabic. Therefore, the task of bilingual clustering needs to be investigated in the context of digital forensics analysis.

5) The majority of the related works represent document features as a bag of words that do not represent semantic relations between words. Therefore, it is beneficial to integrating the WorldNet [20] ontology with clustering

algorithms to enhance document-clustering quality. The WorldNet uses to find the semantic relation between words. In addition, it is useful to use the state-of-the-art deep learning technique called words embedding. Word embedding is considered as one of the most robust representations of document vocabulary. It is capable of capturing context of the words, semantic and syntactic similarity with other words in a document [30-32].

6) Parallel and distributed processing methods might be used to speed up the analysis of digital forensic data.

REFERENCES

[1] Montasari, Reza, Richard Hill, Victoria Carpenter, and Amin Hosseinian-Far. "The Standardised Digital Forensic Investigation Process Model (SDFIPM)." In Blockchain and Clinical Trial, pp. 169-209. Springer, Cham, 2019.

[2] Garfinkel, Simson, et al. "Bringing science to digital forensics with standardized forensic corpora." digital investigation 6 (2009): S2-S11.

[3] Raghavan, Sriram. "Digital forensic research: current state of the art." CSI Transactions on ICT 1.1 (2013): 91-114.

[4] Kazem, Craig. “Extracting Hidden Topics from Texts using LDA Model” A.I. Optify (2014): http://www.aioptify.com/lda.php.

[5] W. Anwar, I. S. Bajwa, M. A. Choudhary, and S. Ramzan, "An Empirical Study on Forensic Analysis of Urdu Text Using LDA-Based Authorship Attribution," IEEE Access, vol. 7, pp. 3224-3234, 2019.

[6] I. Al-Jadir, K. W. Wong, C. C. Fung, and H. Xie, "Enhancing digital forensic analysis using memetic algorithm feature selection method for document clustering," in 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018, pp. 3673-3678.

[7] Noel, George E., and Gilbert L. Peterson. "Applicability of Latent Dirichlet Allocation to multi-disk search." Digital Investigation 11.1 (2014): 43-56.

[8] Beebe, Nicole Lang, and Jan Guynes Clark. "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results." Digital investigation 4 (2007): 49-54.

[9] B. S. Everitt, S. Landau, and M. Leese. "In book: Cluster Analysis, London: Arnold." (2001): 128.

[10] Beebe, Nicole Lang, et al. "Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies." Decision Support Systems 51.4 (2011): 732-744.

[11] Dagher, Gaby G., and Benjamin CM Fung. "Subject-based semantic document clustering for digital forensic investigations." Data & Knowledge Engineering 86 (2013): 224-241.

[12] Mascarnes, Sweedle, and Joanne Gomes. "Subject based Clustering for Digital Forensic Investigation with Subject Suggestion." International Journal of Computer Applications 102.11 (2014).

[13] Thilagavathi, G., and J. Anitha. "Document Clustering in Forensic Investigation by Hybrid Approach." International Journal of Computer Applications 91.3 (2014).

[14] de Waal, Alta, Jacobus Venter, and Etienne Barnard. "Applying topic modeling to forensic data." IFIP International Conference on Digital Forensics. Springer US, 2008.

[15] da Cruz Nassif, Luís Filipe, and Eduardo Raul Hruschka. "Document clustering for forensic analysis: An approach for improving computer inspection." IEEE transactions on information forensics and security 8.1 (2013): 46-54.

[16] Beebe, Nicole L., and Lishu Liu. "Clustering digital forensic string search output." Digital Investigation 11.4 (2014): 314-322.

[17] Decherchi, Sergio, et al. "Text clustering for digital forensics analysis." Computational Intelligence in Security for Information Systems. Springer Berlin Heidelberg, 2009. 29-36.

[18] Kohonen, Teuvo. "The self-organizing map." Proceedings of the IEEE 78.9 (1990): 1464- 1480.

[19] Roussinov, Dmitri G., and Hsinchun Chen. "A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation." Communication and Cognition in Artificial Intelligence Journal(1998).

[20] Miller, George A. "WorldNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

[21] Salton, Gerard. "Automatic text processing: The transformation, analysis, and retrieval of." Reading: Addison-Wesley (1989).

[22] Becker, Jörg, and Dominik Kuropka. "Topic-based vector space model." Proceedings of the 6th international conference on business information systems. 2003.

[23] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 14. 1967.

[24] Girolami, Mark. "Mercer kernel-based clustering in feature space." IEEE Transactions on Neural Networks 13.3 (2002): 780-784.

[25] Decherchi, Sergio, et al. "Hypermetric k-means clustering for content-based document management." Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS’08. Springer Berlin Heidelberg, 2009.

[26] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichle t allocation." Journal of machine learning research 3.Jan (2003): 993-1022.

[27] Kaufman, Leonard, and Peter J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. Vol. 344. John Wiley & Sons, 2009.

[28] Al-Jadir, Ibraheem, Kok Wai Wong, Chun Che Fung, and Hong Xie. "Enhancing digital forensic analysis using memetic algorithm feature selection method for document clustering." In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3673-3678. IEEE, 2018.

[29] Al-Sharif, Ziad A., Attaa Y. Al-Khalee, Mohammed I. Al-Saleh, and Mahmoud Al-Ayyoub. "Carving and Clustering Files in Ram for Memory Forensics." (2018).

[30] Duan, Tiehang, Qi Lou, Sargur N. Srihari, and Xiaohui Xie. "Sequential embedding induced text clustering, a non-parametric bayesian approach." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 68-80. Springer, Cham, 2019.

[31] Wu, Songze, Huaping Zhang, Chengcheng Xu, and Tao Guo. "Text Clustering on Short Message by Using Deep Semantic Representation." In Advances in Computer Communication and Computational Sciences, pp. 133-145. Springer, Singapore, 2019.

[32] Ihm, Sun-Young, Ji-Hye Lee, and Young-Ho Park. "Skip-gram-KR: Korean Word Embedding for Semantic Clustering." IEEE Access (2019).