Text Clustering for Digital Forensics Analysis

Sergio Decherchi[1], Simone Tacconi[2], Judith Redi[1], Fabio Sangiacomo[1], Alessio Leoncini[1] and Rodolfo Zunino[1]

[1]Dept. Biophysical and Electronic Engineering, University of Genoa,
16145 Genoa, Italy
{sergio.decherchi, judith.redi, fabio.sangiacomo,alessio.leoncini, rodolfo.zunino}@unige.it

[2]Servizio Polizia Postale e delle Comunicazioni
Ministero dell'Interno

***Abstract*:** In the last decades digital forensics has become a prominent activity in modern investigations. Seized digital devices can provide precious information and evidences about facts and/or individuals on which the investigational activity is performed. Due to the complexity of this inquiring activity and to the large amount of the data to be analyzed, the choice of appropriate digital tools to support the investigation represents a central concern. In this paper an effective digital text analysis strategy, relying on clustering-based text mining techniques, is introduced for investigational purposes. The proposed methodology is experimentally applied to the publicly available Enron dataset that well fits a plausible forensics analysis context.

***Keywords*:** text clustering, forensics analysis, digital investigation.

1. Introduction

Digital evidence, as defined as the information and data of investigative value that are stored on, received, or transmitted by a digital device [1], has become lately a crucial component in law enforcement agencies investigations. The relevance of this kind of evidence, collected when electronic data and devices are seized, is established by digital forensics analysts, which more and more often have to deal with massive amounts of data, still increasing with the capacity of mass storage devices.

In the investigative activity, two key aspects can be identified: the acquisition and retrieval of information extracted from digital devices, and the following data analysis [2-4], fundamental in depicting a clearer vision of the context of interest. The latter, in particular, is usually performed through a time-effort expensive human-based process: during this phase analysts are requested to carry on a heavy and complete study on the contents obtained from forensic acquisition, hence a selective strategy, aiming at identifying the most relevant data, is mandatory at least in a preliminary phase.

Textual information represents one of the core data sources that may contain significant information. The amount of available textual data is usually extremely large, in the order of thousands of texts. The analyst, in this context, encounters objective difficulties in data content analysis and in finding important investigational patterns. For this reason, the typical requirement that emerges is a semi-automated text content analysis tools.

In this paper, a two-steps investigative process is proposed, based on (1) textual information extraction and (2) textual data analysis via clustering-based text mining tools. Textual information extraction evolves in two phases, and aims at generating a collection of raw text file from information stored in digital devices. The first step involves well-known digital forensics techniques, designed for bit-stream acquisition and early analysis; the second step consists instead in textual information extraction from relevant files previously found.

The subsequent text mining process relies on powerful tools able to deal with large amounts of unstructured textual information [5,6]. Text mining has been proved to be able to profitably support intelligence and security activities in identifying, tracking, extracting, classifying and discovering patterns useful for building investigative scenarios. The general area of text-mining methods comprises various approaches [6]: detection/tracking tools to continuously monitor specific topics over time; document classifiers to label individual files and build up models for possible subjects of interest; clustering tools for documents processing and detection of relevant relations among those subjects.

This work addresses text clustering for forensics analysis based on a dynamic, adaptive clustering model to arrange unstructured documents into content-based homogeneous groups [7]. The approach is validated by using the publicly available Enron emails database [8] as experimental domain. The research presented here shows that the document clustering framework [7] can find consistent structures suitable for investigative issues that can considerably aid the analyst during the inquiry activity.

2. Forensic Acquisition and Early Analysis

Digital evidence data acquisition is a delicate process that articulates in several phases, each addressed at maximizing the amount of potentially useful information retrievable from the seized device.

According to well known best practices of digital forensics, the first step of data extraction is the acquisition of data from devices, performed by means of a bit-stream copy, i.e. a copy of every bit of data, including for example the file slack and the unallocated file space. It is important to consider that from the latter in particular deleted files and e-mails can be recovered. In digital forensic analysis, deleted files are fundamental elements, being potentially very interesting from the investigative point of view. Hence, deleted files recovery constitutes the second phase of the process.

Two major strategies can be applied for deleted files

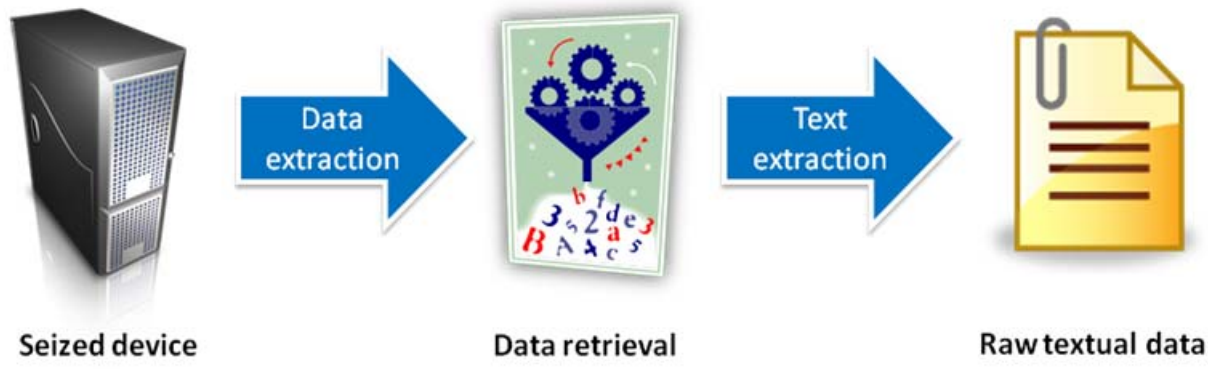

Figure 1. The two-steps data extraction process. First data is acquired from seized devices as a bit-stream. Then, an advanced analysis process individuates textual data and collects them.

recovery: a metadata-based approach and an application-based one [9]. The first method relies on metadata of deleted files: the related entry record of involved parent directories is recovered provided that such metadata still exist. In case that metadata were reallocated to a new file or had been wiped, an application-based strategy is needed. In this case, chunks of data are searched for signatures that correspond to the header (start) and/or the footer (end) of known file types. This task is generally performed on the unallocated space of the file system, and allows also recovering files that are not pointed by any metadata, provided that their clusters were contiguously allocated. In this phase, obviously, one also extracts current files, i.e. files that are logically stored in the media.

The third phase, applied both to current files and to recovered deleted files, is devoted to file type identification and classification. This goal is not achievable by means of file extensions examination, since users can easily change them, but requires the analysis of headers and footers, applying to each file the same methodology of data carving.

All the above phases are implemented in widely spread forensic analysis tool, both in commercial ('Guidance Encase' [17], 'Access Data Forensic Toolkit (FTK)' [18] etc.) and open-source ('Sleuthkit / Authopsy' [19] etc.).

3. Textual Information Extraction

Once all the potentially relevant digital information has been collected, an action devoted to text extraction from files belonging to significant categories is needed. At this stage, the analyst may have availability of both documental and non-documental files. With regard to documental files, in the simplest case, textual information can be found in a plain form, i.e. in raw text files. More often, textual information is present in a latent form (e-mail database, documents produced by Microsoft Word or Adobe Acrobat, web pages etc.). Documental files that are not purely textual need to be converted to pure text, since text miners most likely work on raw text files.

Concerning non-documental files, it is possible to extract the external existing metadata within the related entry record of the parent directories. Indeed, each one of these files has a set of textual metadata (name, path, MAC times etc.) maintained by the file system. Moreover, some types of non documental files could have internal textual metadata stored inside the file itself by software applications (author in a Microsoft Word document, exif-data in images etc.). In these cases, textual information has to be extracted by developing appropriate procedures, not discussed in the following.

At this point, a collection of raw text files is ready to be further processed by the text mining tool. The extraction of all textual information is not a trivial task: as for the authors' best knowledge, no tool able to automatically perform such activity currently exists. However, some software tools implement specific functions which are useful in this context. In particular, 'Text Mining Tool' [20] is a program for extraction of text from files of diverse types: Portable Document Format (*.pdf),, Microsoft Word Document (*.doc), Rich Text Format (*.rtf), *.chm and hypertext (*.html) files. It also offers the console tool 'minetext' for automation of text conversion. Additionally, Metadata Extraction Tool [21], developed by the National Library of New Zealand, automatically extracts metadata from a range of file formats like PDF documents, multimedia files and Microsoft office documents.

4. Text Clustering

Thanks to the method presented in sections 2 and 3, unstructured information can be collected from seized devices and converted into text. To extract actual knowledge from these data, though, some more advanced analysis is required. The analyst in this phase has to deal with a huge amount of unlabeled text, and manual check of all the information can prove sometimes prohibitive.

Text mining [22-26],[40] provides an effective, automatic platform to support the analysis of digital textual evidences, which is a key issue for homeland security [10,11].

Clustering algorithms [27-39] can be applied to text mining to allow the automatic recognition of some sort of structure in the analyzed set of documents. In particular, clustering is designed to discover groups in the set of documents such that the documents within a group are more similar to one another than to documents of other groups. The core idea is to provide the analyst with clusters including documents semantically related, as a starting point for determining investigation paths.

The document clustering problem can be defined as follows. Once the set of documents of interest $\mathcal{D} = \{D_1, \ldots, D_n\}$ has been defined, a similarity measure (or distance metric), and a

Figure 2. Extraction of knowledge from raw textual data. Pure text is first processed to obtain groups of documents with similar content. The following analysis allows the final knowledge extraction.

partitioning criterion, which is usually implemented by a cost function, are selected. In the case of flat clustering, the desired number of clusters, Z has to be set as well. The goal is to compute a membership function $\phi : \mathcal{D} \rightarrow \{1, \ldots, Z\}$ such that ϕ minimizes the partitioning cost with respect to the similarities among documents. Conversely, hierarchical clustering doesn't envision any cardinality definition, and applies a series of nested partitioning tasks which eventually yield a hierarchy of clusters.

4.1 Knowledge base representation

Every text mining framework should always be supported by an information extraction (IE) model [12,13] which is designed to pre-process digital text documents and to organize the information according to a given structure that can be directly interpreted by a machine learning system. Thus, a document D is eventually reduced to a sequence of terms and is represented as a vector, which lies in a space spanned by the dictionary (or vocabulary) $\mathcal{T} = \{t_j; j = 1,.., n_T\}$. The dictionary collects all terms used to represent every document D, and can be assembled empirically by gathering the terms that occur at least once in the document collection $\mathcal{D}$; by this representation one loses the original relative ordering of terms within each document. Different models [9,10] can be used to retrieve index terms and to generate the vector that represents a document D. However, the vector space model [14] is the most widely used method for document clustering. In [7] an augmented vector space model is proposed; this augmented space is characterized in the subsequent section.

In the following, $\mathcal{D} = \{D_u; u = 1,..,n_D\}$ will denote the corpus, holding the collection of documents to be clustered. The set $\mathcal{T} = \{t_j; j = 1,.., n_T\}$ will denote the vocabulary, which is the collection of terms that occur at least once in $\mathcal{D}$ after the pre-processing steps of each document $D \in \mathcal{D}$ (e.g., stop-words removal, stemming [12]).

4.2 Clustering framework

The clustering strategy is mainly based on two aspects: the notion of distance between documents and the involved clustering algorithm.

Following the approach adopted in [7], a distance consisting in a weighted Euclidean distance and a term based on stylistic information was used. Having defined a weight α, the Euclidean term $\Delta^{(f)}$ and the stylistic term $\Delta^{(s)}$, then the distance between D_u and D_v can be worked out as:

$$\Delta(D_u,D_v) = \alpha \cdot \Delta^{(f)}(D_u,D_v) + (1-\alpha)\cdot \Delta^{(s)}(D_u,D_v) \qquad (1)$$

To support the proposed document distance measure, a document D is here represented by a pair of vectors, $\mathbf{v}'$ and $\mathbf{v}''$. Vector $\mathbf{v}'(D)$ addresses the content description of a document D; it can be viewed as the conventional n_T-dimensional vector that associates each term $t \in \mathcal{T}$ with the normalized frequency, tf, of that term in the document D. Therefore, the k-th element of the vector $\mathbf{v}'(D_u)$ is defined as:

$$v'_{k,u} = tf_{k,u} \Big/ \sum_{l=1}^{n_T} tf_{l,u} \qquad (2)$$

where $tf_{k,u}$ is the frequency of the k-th term in document D_u. Thus, $\mathbf{v}'$ represents a document by a classical vector model, and uses term frequencies to set the weights associated to each element.

The stylistic information $\mathbf{v}''$ involved in (1) aims at exploiting the structural properties of a document, D. These properties are represented by a set of probability distributions associated with the terms in the vocabulary. Each term $t \in \mathcal{T}$ occurring in D_u is associated with a distribution function that gives the spatial probability density function (*pdf*) of t in D_u. Such a distribution, $p_{t,u}(s)$, is generated under the hypothesis that, when detecting the k-th occurrence of a term t at the normalized position $s_k \in [0,1]$ in the text, the spatial *pdf* of the term can be approximated by a Gaussian distribution centered around s_k. To derive a formal expression of the *pdf*, assume that the u-th document, D_u, holds n_O occurrences of terms after simplifications; if a term occurs more than once, each occurrence is counted individually when computing n_O, which can be viewed therefore as a measure of the length of the document. The spatial *pdf* can be defined as:

$$p_{t,u}(s) = \frac{1}{A}\sum_{k=1}^{n_O} G(s_k,\lambda) = \frac{1}{A}\sum_{k=1}^{n_O}\frac{1}{\sqrt{2\pi}\lambda}\exp\left[-\frac{(s-s_k)^2}{\lambda^2}\right] \qquad (3)$$

where A and λ are normalization terms. In practical situations, one uses a discrete approximation of (3). First, the document D is segmented evenly into S sections. Then, an S-dimensional vector is generated for each term $t \in \mathcal{T}$; each element of that vector estimates the probability that the term t occurs in the corresponding section of the document. As a

result, $\mathbf{v}''(D)$ is an array of n_T vectors having dimension S.

Vectors $\mathbf{v}'$ and $\mathbf{v}''$ support the computation of the frequency-based distance, $\Delta^{(f)}$, and of the stylistic distance, $\Delta^{(s)}$, respectively. Strictly speaking (1) is not a metric space because does not guarantee the triangular inequality, for this reason equation (1) can be more properly considered a similarity measure. This distance measure has been employed in the well known Kernel K-Means [7] clustering algorithm.

The conventional k-means paradigm supports an unsupervised grouping process [15], which partitions the set of samples, $\mathcal{D} = \{D_u; u= 1,..,n_D\}$, into a set of Z clusters, C_j ($j = 1,\ldots, Z$). In practice, one defines a "membership vector," which indexes the partitioning of input patterns over the K clusters as: $m_u = j \Leftrightarrow D_u \in C_j$, otherwise $m_u = 0$; $u = 1,\ldots, n_D$. It is also useful to define a "membership function" $\delta_{uj}(D_u,C_j)$, that defines the membership of the u-th document to the j-th cluster: $\delta_{uj} =1$ if $m_u = j$, and 0 otherwise. Hence, the number of members of a cluster is expressed as

$$N_j = \sum_{u=1}^{n_D} \delta_{uj} \; ; \quad j = 1,\ldots, Z \tag{4}$$

and the cluster centroid is given by:

$$\mathbf{w}_j = \frac{1}{N_j} \sum_{u=1}^{n_D} \mathbf{x}_u \delta_{uj} \; ; \quad j = 1,\ldots, Z \tag{5}$$

where $\mathbf{x}_u$ is any vector-based representation of document D_u. The kernel based version of the algorithm is based on the assumption that a function, Φ, can map any element, D, into a corresponding position, $\Phi(D)$, in a possibly infinite dimensional Hilbert space. In the new mapped space clustering centers become:

$$\Psi_j = \frac{1}{N_j} \sum_{u=1}^{n_D} \Phi_u \delta_{uj} \; ; \quad j = 1,\ldots, Z \tag{6}$$

According to [7] this data mapping allows different salient features able to ease the clustering procedure.

The ultimate result of the clustering process is the membership vector, **m**, which determines prototype positions (6) even though they cannot be stated explicitly. As per [7], for a document, D_u, the distance in the Hilbert space from the mapped image, Φ_u, to the cluster Ψ_j as per (6) can be worked out as:

$$d(\Phi_u, \Psi_j) = \left\| \Phi_u - \frac{1}{N_j} \sum_{v=1}^{n_D} \Phi_v \right\|^2 =$$

$$= 1 + \frac{1}{(N_j)^2} \sum_{m,v=1}^{n_D} \delta_{mj}\delta_{vj} \Phi_m \cdot \Phi_v - \frac{2}{N_j} \sum_{v=1}^{n_D} \delta_{vj} \Phi_u \cdot \Phi_v \tag{7}$$

By using expression (7), one can identify the closest prototype to the image of each input pattern, and assign sample memberships accordingly.

5. Forensic Analysis on Enron Dataset

In this study, for simulating an investigational context, Enron emails dataset [8] was used. This choice was guided by the exigency of a publicly available dataset that well simulates an investigative context. The previously explained data extraction process was not performed because all data is already available in archival form; despite this fact, the overall proposed framework is still valid in the general case where explicit data recovery on the seized device is needed.

The Enron email dataset [8] provides a reference corpus to test text-mining techniques that address investigational applications [2-4]. The Enron mail corpus was posted originally on Internet by the Federal Energy Regulatory Commission (FERC) during its investigation on the Enron case. FERC collected a total of 619,449 emails from 158 Enron employees, mainly senior managers. The original set was suffering from document integrity problems; hence an updated version was later released by SRI International for the CALO project [16]. Eventually, William Cohen from Carnegie Mellon University published the cleaned dataset online [8] for researchers in March 2004. Other processed versions of the Enron corpus had been made available on the web, but were not considered in the present work because the CMU version made it possible fair comparison of the obtained results with respect to established, reference corpora in the literature.

Of the 158 authors (i.e. employees), the emails of five of them were randomly selected: *White S., Smith M., Solberg G., Ybarbo P. and Steffes J* (table 1).

Each message in the dataset includes: the email addresses of the sender and receiver, date and time, the subject, the body and text. When applying the clustering algorithm, only the 'body' sections of the emails were used, and sender/receiver, date/time information were discarded. Experiments were aimed to test the effectiveness of the approach on two different issues: information retrieval and authorship.

Name	Number of Emails
White S.	3272
Smith M.	1642
Solberg G.	1081
Ybarbo P	1291
Steffes J	3331

Table 1. Names and corresponding number of Emails

In the first case one simulates the inquiry activity on the seized device under investigation. In this context the interest is in analyzing the content of the emails of each author; for this reason an independent clustering process is applied to the emails of each author.
In the second case, all emails are kept together and clustered; in this experiment one is interested in the ability of the framework in recovering the correct authors from the email corpus.

5.1 Information Retrieval

Collected emails (see tab.1) were processed separately, thus obtaining five different scenarios for each employee. The underlying hypothesis was that email contents can also be characterized by the role the mailbox owner played within the company. The performed experiments used a number of 10 clusters: this choice was guided by the practical demand of obtaining a limited number of informative groups. Tables from 2 to 6 report on the results obtained by these experiments. Each table shows the terms that characterize

Cluster	Most Frequent and Relevant Words
1	employee, business, hotel, Houston, company
2	pipeline, social, database, report, link, data
3	ECT, EnronXg
4	coal, oil, gas, nuke, west, test, happy, business
5	Yahoo, compubank, NGCorp, Dynegi, night, plan
6	shank, trade
7	travel, hotel, continent, airport, flight, Sheraton
8	Questar, Paso, price, gas
9	schedule, London, server, sun, contact, report
10	trip, weekend, plan, ski

Table 2. Smith results

Cluster	Most Frequent and Relevant Words
1	Paso, iso, empow, ub, meet
2	schedule, detected, California, ISO, parsing
3	ub, employee, EPE, benefit, contact, ubsq
4	schedule, EPMI, NCPA, sell, buy, peak, energy
5	dbcaps97, data, failure, database
6	trade, pwr, impact, London
7	awarded, California, ISO, westdesk, Portland
8	error, pasting, admin, SQL, attempted
9	failure, failed, required, intervention, crawl
10	employee, price, ub, trade, energy

Table 3. Solberg results

Cluster	Most Frequent and Relevant Words
1	FERT, RTO, EPSA, NERC
2	market, FERC, Edison, contract, credit, order, RTO
3	FERC, report, approve, task, imag, attach
4	market, ee, meet, november, october
5	California, protect, attach, testimony, Washington
6	stock, billion, financial, market, trade, investor
7	market, credit, ee, energy, util
8	attach, gov, energy, sce
9	affair, meet, report, market
10	gov, meet, november, imbal, pge, usbr

Table 4. Steffes results

Cluster	Most Frequent and Relevant Words
1	meet, chairperson, Oslo, invit, standard, smoke
2	confidential, attach, power, internet, copy
3	West, ECT, meet, gas
4	gopusa, power, report, risk, inform, managment
5	webster, listserv, subscribe, htm, blank, merriam
6	report, erv, asp, EFCT, power, hide
7	ECT, Rhonda, John, David, Joe, Smith, Michae,l Mike
8	power
9	mvc, jpg, attach, meet, power, energy, Canada
10	calendard, standard, Monica, vacation, migration

Table 5. White results

Cluster	Most Frequent and Relevant Words
1	report, status, week, mmbtu, price, lng, lpg, capacity
2	tomdd, attach, ship, ect, master, document
3	London, power, report, impact, gas, rate, market, contact
4	dpc, transwestern, pipeline, plan
5	inmarsat, galleon, eta, telex, master, bar, fax, sea, wind
6	rate, lng, price, agreement, contract, meet
7	report, Houston, Dubai, dial, domest, lng, passcode
8	power, Dabhol, India, dpc, mseb, govern, Maharashtra
9	cargo, winter, gallon, price, eco, gas
10	arctic, cargo, methan

Table 6. YBarbo results

each of the clusters provided by the clustering framework for each employee. For each cluster, the most descriptive words between the twenty most frequent words of the cluster are listed. Reported terms actually include the following peculiar abbreviations: "ECT" stands for Enron Capital & Trade Resources, "HPL" stands for Houston Pipeline Company, "EPMI" stands for Enron Power Marketing Inc, "MMBTU" stands for Million British Thermal Units, "dynegi" stands for Dynegy Inc, a large owner and operator of power plants and a player in the natural gas liquids and coal business, which in 2001 made an unsuccessful takeover bid for Enron. From the investigational point of view, some interesting aspects emerges: in *Smith* results there is an interesting cluster (cluster 10) in which the context seems not strictly adherent to the workplace usual terms. Hence, analyzing that bunch of email may allow the acquisition of sensible private life information potentially useful for investigation.

Smith emails does not underline a particular trend: it could be interesting, from the investigative point of view, to analyze more in depth cluster 4 in which business and personal life words mix up.

Analyzing *Solberg* emails, no particular activity seems to rise from the clustering results. However, it is curious to observe how his emails are full of server errors.

Steffes results seem extremely interesting instead. Cluster 6 underlines a compact group of emails in which important financial aspects of Enron group are discussed. In particular some key words as F.E.R.C. (Federal Energy Regulatory Commission) are particularly expressive.

From *White* analyzing the most occurring terms one can observe that cluster 2 contains several time the word "*confidential*" making this group interesting and worth of further analysis. Cluster 7 exhibits a strange content; it is characterized completely by names of people. This could indicate that these emails may concern private life.

YBarbo emails have no particular features. The only aspect that can be understood is that his position is tightly linked to international affairs.

Table 7 summarizes the most significant words detected for each author.

Author	Words
Smith	Dynegi, ECT, London, Server
Solberg	dbcaps97, EPMI, NCPA, ISO California
Steffes	stock, billion, financial, market, trade, investor, F.E.R.C, California, Washington
White	ECT, Rhonda, John, David, Joe, Smith, Michae,l Mike, confidential, ECT, Jpeg, Canada
YBarbo	power, Dabhol, India, London, mmbtu, Houston, Dubai

Table 7. Authors most significant words

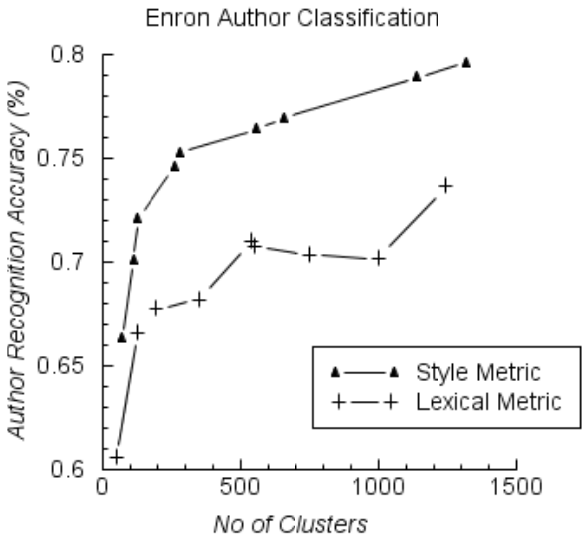

Figure 3. Authorship recognition rate with varying metric and number of clusters

5.2 Authorship on email corpus

In this second group of experiments one is concerned with recovering the correct author by analyzing the email content. To this aim, all collected emails were kept together in order to form a heterogeneous dataset; the obtained size was 10609. Accuracy on author recognition is computed with varying number of clusters and using two different configurations of the metric (1); α was chosen with values 0.9 and 0.2, giving low and high importance, respectively, to the stylistic element of the metric. Table 8,9 and figure 3 illustrate the obtained results.

This group of experiments showed that results were considerably influenced by the stylistic metric; it proved very effective in improving accuracy performance, confirming previous results in [7]. Figure 3 shows that an acceptable accuracy level (70%) can be reached using about 100 clusters, being two orders of magnitude less than the total number of emails.

#Clusters	#Errors	Accuracy %
52	4182	60.58%
129	3546	66.57%
198	3427	67.69%
354	3378	68.15%
539	3079	70.97%
554	3109	70.69%
753	3155	70.26%
1002	3166	70.15%
1245	2794	73.66%

Table 8. Authorship results for lexical-based distance (α=0.9)

#Clusters	#Errors	Accuracy %
70	3576	66.29%
117	3179	70.03%
130	2962	72.08%
267	2696	74.58%
283	2627	75.23%
559	2505	76.38%
657	2455	76.85%
1139	2241	78.87%
1319	2167	79.57%

Table 9. Authorship results for stylistic-based distance (α=0.2)

From the operational point of view one can assert that both for authorship and content analysis, the proposed framework can be considered a valuable aid to the process of the inquiring activity. In particular, beside the fact that a human operator is still necessary, the automation level on the investigational process can be consistently speeded up.

6. Conclusions

The presented work gives an overview on the possibilities offered by textual clustering when applied to Digital Forensics analysis. The current research used the text mining clustering-based framework developed in [7],[40].

In order to assess the effectiveness of the clustering engine in such a context, the Enron email dataset has been employed: this real world dataset well renders an investigative context. Obtained results in both information retrieval and authorship confirm the suitability of the approach.

An analyst, using the proposed tool, can exploit the obtained clusters in order to get useful investigative information; in particular the tool proves effective when one has to cope with a notable amount of data, when a human operator cannot manually proceed to inspection.

Future research could inquiry on the possibility of endowing in the cluster calibration procedure semantic information. Another line of research could consist in using supervised learning tools to categorize data on already defined categories for investigative purposes.

References

[1] U.S. Department of Justice, Electronic Crime Scene Investigation: A Guide for First Responders, I Edition, NCJ 219941, 2008, http://www.ncjrs.gov/pdffiles1/nij/219941.pdf

[2] Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: "Crime data mining: a general framework and some examples". *IEEE Trans. Computer*. 37, 50--56 (2004)

[3] Seifert, J. W.: Data Mining and Homeland Security: An Overview. CRS Report RL31798, www.epic.org/privacy/fusion/crs-dataminingrpt.pdf (2007)

[4] Mena, J.: *Investigative Data Mining for Security and Criminal Detection*. Butterworth-Heinemann (2003)

[5] Sullivan, D.: *Document warehousing and text mining.* John Wiley and Sons (2001)

[6] Fan, W., Wallace, L., Rich, S., Zhang, Z.: "Tapping the power of text mining". *Comm. of the ACM*. 49, 76--82 (2006)

[7] Decherchi, S., Gastaldo, P., Redi, J., Zunino, R.: Hypermetric k-means clustering for content-based document management, *First Workshop on Computational Intelligence in Security for Information Systems*, Genova. (2008)

[8] The Enron Email Dataset, http://www-2.cs.cmu.edu/~enron/

[9] Carrier, B., *File System Forensic Analysis*, Addison Wesley, 2005

[10] Popp, R., Armour, T., Senator, T., Numrych, K.: "Countering terrorism through information technology." *Comm. of the ACM*. 47, 36--43 (2004)

[11] Zanasi, A. (eds.): *Text Mining and its Applications to Intelligence*, CRM and KM. 2nd edition , WIT Press (2007).

[12] Manning, C. D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)

[13] Baeza-Yates, R., Ribiero-Neto, B.: *Modern Information Retrieval*. ACM Press (1999).

[14] Salton, G., Wong, A., Yang, L.S.: "A vector space model for information retrieval". *Journal Amer. Soc. Inform. Sci*. 18, 613—620 (1975)

[15] Linde, Y., Buzo, A., Gray, R.M.: "An algorithm for vector quantizer design". *IEEE Trans. Commun. COM-*28, 84--95 (1980).

[16] R. Bekkerman, A. McCallum, and G. Huang, "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora." *CIIR Technical Report IR-418* 2004, http://www.cs.umass.edu/~ronb/papers/email.pdf

[17] Guidance Encase: http://www.guidancesoftware.com/computer-forensics-ediscovery-software-digital-evidence.htm

[18] Access Data Forensic ToolKit: http://www.accessdata.com/forensictoolkit.html

[19] Sleuth Kit & Authopsy: http://www.sleuthkit.org/

[20] Text Mining Tool: http://text-mining-tool.com/

[21] Metadata Extraction Tool: http://meta-extractor.sourceforge.net/

[22] D. Hand, H. Mannila, P. Smyth. *Principles of Data Mining,* MIT Press, Cambridge, MA

[23] M. W. Berry, M. Castellanos. *Survey of Text Mining II*, Springer, 2008.

[24] A. Hotho, A. Nürnberger and G. Paaß. "A brief survey of text mining," LDV Forum - *GLDV Journal for Computational Linguistics and Language Technology*, 20, pp. 19-62, 2005

[25] F. Shahnaz, M. W. Berry , V. Paul Pauca, R. J. Plemmons. "Document clustering using nonnegative matrix factorization", *Information Processing and Management*, 42, pp. 373–386, 2006

[26] M.W. Berry, S.T. Dumais, G. W. O'Brien. "Using linear algebra for intelligent information retrieval", *SIAM Review*, 37, pp. 573–595, 1995

[27] J. Shawe-Taylor, N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.
[28] I. Dhillon, Y. Guan and B. Kulis. "A unified view of Kernel kmeans, Spectral Clustering and Normalized Cuts". *UTCS Technical Report #TR-04-25*, University of Texas at Austin, Department of Computer Sciences, Austin, TX 78712, 2005
[29] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*, Prentice Hall, 1988
[30] S. Ridella, S. Rovetta, and R. Zunino. "Plastic algorithm for adaptive vector quantization", *Neural Computing and Applications*, 7, pp. 37-51, 1998
[31] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2nd edition, 2000
[32] D. Cai, X. He, and J. Han. "Document Clustering Using Locality Preserving Indexing", IEEE Transaction on knowledge and data engineering, 17, pp. 1624-1637, 2005
[33] L. Jing, M. K. Ng, and J. Zhexue Huang. "An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data", *IEEE Transactions on knowledge and data engineering*, 19, pp. 1026-1041, 2007
[34] Y.-J. Horng, S.-M. Chen, Y.-C. Chang, and C.-H. Lee. "A New Method for Fuzzy Information Retrieval Based on Fuzzy Hierarchical Clustering and Fuzzy Inference Techniques", IEEE Transactions on fuzzy systems, 13, pp. 216-228, 2005
[35] W. –C. Tjhi, L. Chen. "A heuristic-based fuzzy coclustering algorithm for categorization of highdimensional data", *Fuzzy Sets and Systems*, 159, pp. 371–389, 2008
[36] K. M. Hammouda and M. S. Kamel. "Efficient phrasebased document indexing for web document clustering", *IEEE Transactions on Knowledge and Data Engineering*, 16, pp. 1279–1296, 2004.
[37] H. Chim, X. Deng. "Efficient Phrase-based Document Similarity for Clustering", *IEEE Transaction on Knowledge and Data Engineering*, 20, pp. 1217-1229, 2008.
[38] O. Zamir and O. Etzioni. "Grouper: A Dynamic Clustering Interface to Web Search Results", *Computer Networks*, 31, pp. 1361-1374, 1999.
[39] M. Girolami. "Mercer kernel based clustering in featurespace", *IEEE Trans. Neural Networks*, 13, pp. 2780-2784, 2002.
[40] S. Decherchi, P. Gastaldo, J. Redi and R. Zunino. "A Text Clustering Framework for Information Retrieval". *Journal of Information Assurance and Security*, Special Issue Cisis 2008.

Authors' Biographies

Sergio Decherchi (born Genoa, Italy, 1983) obtained the "Laurea" degree summa cum laude in Electronic Engineering in 2007 from Genoa University, Italy. Since 2005 he started collaborating with the Department of Biophysical and Electronics Engineering of Genoa University, where he is pursuing a PhD in Electronic Engineering and Computer Science on Machine Learning. His main research areas include: theoretical aspects of Machine Learning, large scale learning algorithms development, semi-supervised learning, dedicated hardware for learning machines and Text Mining.

Simone Tacconi (born in San Benedetto del Tronto, Italy, 1973) obtained the M.S. in Electronic Engineering (2000) from the University of Ancona and the Ph.D. in Artificial Intelligent Systems from the Polytechnic University of Marche (2004). He is currently Technical Director of Computer Forensics Unit at italian Postal and Communications Police. His main scientific interests include computer security, cryptography and digital forensics. In these fields he is co-author of several papers in International Journals and Conferences.

Fabio Sangiacomo (born Genoa, Italy, 1985) obtained the "Laurea" degree summa cum laude in Electronic Engineering in 2009 from Genoa University, Italy. He is beginning a PhD in Electronic Engineering, Information Technology, Robotics and Telecommunications at the Department of Biophysical and Electronics Engineering of University of Genoa. His main research areas include: machine learning, semantic methods for Text Mining, network security.

Alessio Leoncini (born Genoa, Italy, 1985) obtained the "Laurea" degree in Electronic Engineering in 2009 from University of Genoa, Italy. Since 2009 he started collaborating with the Department of Biophysical and Electronics Engineering, University of Genoa. He is beginning a PhD in Space Science and Engineering about Machine Learning and Data Mining.
His main research areas include: Text Mining with semantic elaborations, SVMs, practical aspects of neural networks, dedicated hardware for machine learning, network security.

Judith Redi obtained the "Laurea" degree in Electronic Engineering in 2006 from Genoa University, Italy. Her main research areas include the study of cryptographic algorithms and number theory, and the implementation of neural networks models for the objective assessment of visual quality. The latter subject is being developed in conjunction with Philips Research Labs, Eindhoven (NL). Since 2007 she joined SEALab as a PhD student in Space Engineering Sciences. She received a grant from Philips Research Labs, Eindhoven, for a five-month visiting period at Philips Labs to enhance research on visual quality assessment.

Rodolfo Zunino (born Genoa, Italy, 1961) obtained the "Laurea" degree cum laude in Electronic Engineering from Genoa University in 1985. From 1986 to 1995 he was a research consultant with the Department of Biophysical and Electronic Engineering (DIBE) of Genoa University.
He is currently with DIBE as an Associate Professor, teaching Electronics for Embedded Systems and Electronics for Security. His main scientific interests include intelligent systems for Computer Security, network security and Critical Infrastructure Protection, embedded electronic systems for neural networks, efficient models for data representation and learning, massive-scale text-mining and text-clustering methods, and advanced techniques for multimedia data processing.
Rodolfo Zunino coauthored more than 170 scientific papers in International Journals and Conferences; he has been the Co-Chairman of the two Editions of the International Workshop on Computational Intelligence for Security in Information Systems (CISIS'08 and CISIS'09).
Since 2001 he is contributing as Associate Editor of the IEEE Transactions on Neural Networks, and has participated in the Scientific Committees of several International Events (ICANN'02, ICANN'09, IWPAAMS2004, IWPAAMS2005, Applied Computing 2006). Rodolfo Zunino is a Senior Member of IEEE (CIS – Computational Intelligence Society).