Chapter 10

APPLYING TOPIC MODELING TO FORENSIC DATA

Alta de Waal, Jacobus Venter and Etienne Barnard

Abstract Most actionable evidence is identified during the analysis phase of digital forensic investigations. Currently, the analysis phase uses expression-based searches, which assume a good understanding of the evidence; but latent evidence cannot be found using such methods. Knowledge discovery and data mining (KDD) techniques can significantly enhance the analysis process. A promising KDD technique is topic modeling, which infers the underlying semantic context of text and summarizes the text using topics described by words. This paper investigates the application of topic modeling to forensic data and its ability to contribute to the analysis phase. Also, it highlights the challenges that forensic data poses to topic modeling algorithms and reports on the lessons learned from a case study.

Keywords: Digital investigation, analysis phase, evidence mining, topic modeling

1. Introduction

The four major phases in digital investigation are acquisition, examination, analysis and reporting [14]. The value of the information obtained in digital investigations has been questioned by several researchers [1, 11]. In particular, they argue that the analysis phase, where most of the actionable evidence is gathered, lacks sufficient definition and support in terms of principles, methods and tools [14, 17]. Knowledge discovery and data mining (KDD) has the potential to enhance the analysis phase [14, 17]. The use of KDD principles and tools in digital investigations is referred to as "evidence mining" [17].

Textual artifacts are important in many digital investigations [1, 11]. These "documents" include e-mails, reports, letters, notes, text messages, etc. In a typical case, the evidence set may contain thousands

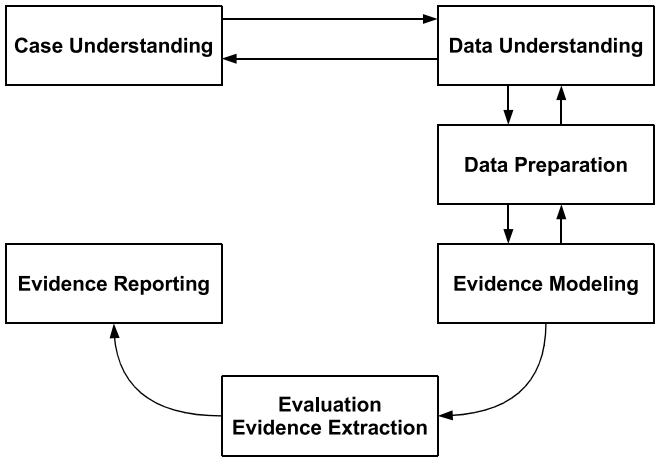

Figure 1. CRISP-EM process.

of documents. Often, a very small proportion of these documents are relevant and an even smaller proportion of the relevant documents may contain actionable evidence. Manually processing thousands of text documents to discover evidence is a difficult and time-consuming task.

Expression-based searches are often used to analyze digital data. Such searches require a good understanding of the evidence being sought. Furthermore, the retrieved information is not ranked (e.g., based on relevance to the case). Thus, latent evidence – evidence that exists but is not directly accessible to the investigator – will not be found. Evidence mining, on the other hand, uses KDD principles and techniques to uncover electronic artifacts that assist in developing crime scenarios [17]. These artifacts include known evidence as well as latent evidence.

CRISP-EM, a specialization of the CRISP-DM process [5], is intended to support evidence mining [17]. The work described in this paper falls within the scope of the data preparation phase of CRISP-EM (Figure 1). Data preparation covers all the activities involved in constructing a data set used for event reconstruction and modeling. Data set construction is a challenging task that involves a trade-off between selecting relevant data and losing vital information used for event reconstruction. A summary of the data would be extremely useful to an investigator; it would facilitate better understanding of the data content and assist in focusing the data preparation task on gathering relevant data.

Topic modeling is a powerful latent variable analysis technique that can help associate relevant documents by modeling the underlying (latent) topics in a collection of text documents. Additionally, it suggests prevalent themes within the text, thereby providing a useful summary of

the document collection. As a KDD technique, topic modeling has the potential to discover latent evidence that is often missed by expression-based searches. However, digital evidence is non-homogeneous in terms of format and content, which poses unique challenges to KDD techniques. This paper investigates the primary issues involved in applying topic modeling to forensic data. Also, it examines the utility of topic modeling in a real investigation.

2. Topic Modeling

Large collections of digital data are widely available and are growing at an incredible pace. Attempting to understand the meaning of the data is a difficult task and, in general, the first option is to perform expression (keyword) searches. However, the results of these searches do not adequately describe the meaning of the data collection, especially when the user has limited insight into the collection. A summary of the collection that encapsulates the main topics within the data would be very useful [12]. An example of a data collection is a text corpus of newspaper articles. For this corpus, a list of topics might include politics, sport, finance, culture and local news.

A text corpus is a collection of documents, each with an underlying semantic context. The semantic context refers to the intended meaning of a document and develops as the document is generated. For example, a newspaper article reports on a news event and, as the article is read, the reader becomes aware of the ideas the reporter intended to communicate. The "hidden" semantic context is represented by the words of a document. Topic modeling, which addresses the retrieval of semantic context from a text corpus, can be formalized as a statistical inference problem. Given a set of data (words), the latent semantic context from which it was generated can be inferred [7]. A topic is defined as a probability distribution over words. In statistical terms, a topic model is a latent variable model where the latent variables describe the topics [2].

Figure 2 presents an example involving two topics from a subset of the TREC AP corpus [8]. The ten words with the highest probabilities for each topic are presented along with their probabilities. These top-10 words describe the two topics. Topic A clearly has to do with financial markets whereas Topic B deals with a naval incident in Saudi Arabia.

The fundamental assumption in topic modeling is that the semantic context of a document is a mixture of topics [7]. A "bag-of-words" approach is commonly adopted for topic modeling, which means that a document is treated as a collection of words while ignoring the structure of the document. The output of the bag-of-words approach is a Word

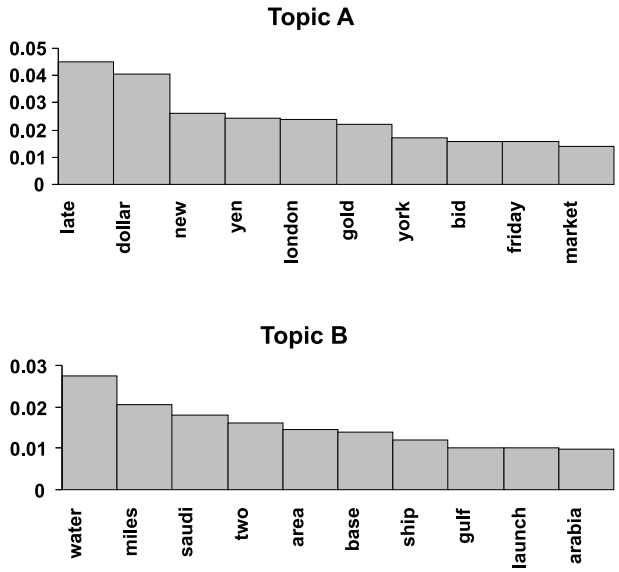

Figure 2. Word probability distributions for two topics (top 10 words).

$\times$ Document frequency matrix where cell_{ij} represents the frequency of word_i in document_j.

3. Topic Modeling Applied to Forensic Data

When applied to text data, topic modeling provides a summary of the documents by describing the latent topics in the data as illustrated in Figure 2. This leads to two useful outputs: a verbal summary of the topics and a visual representation of the document space.

3.1 Topic Modeling Process

Figure 3 illustrates the six-level process involved in applying topic modeling to the analysis of real forensic data. Each level represents a different data set. Level 1 represents the original forensic data set. Levels 2 through 4 represent data sets generated during data filtering. Data pre-processing produces a Word $\times$ Document matrix (Level 5), which is the input for topic modeling. The Level 6 data set represents the results of topic modeling.

3.2 Data Sets

The data sets produced during the topic modeling process can be described in parallel with the levels in the process graph in Figure 3.

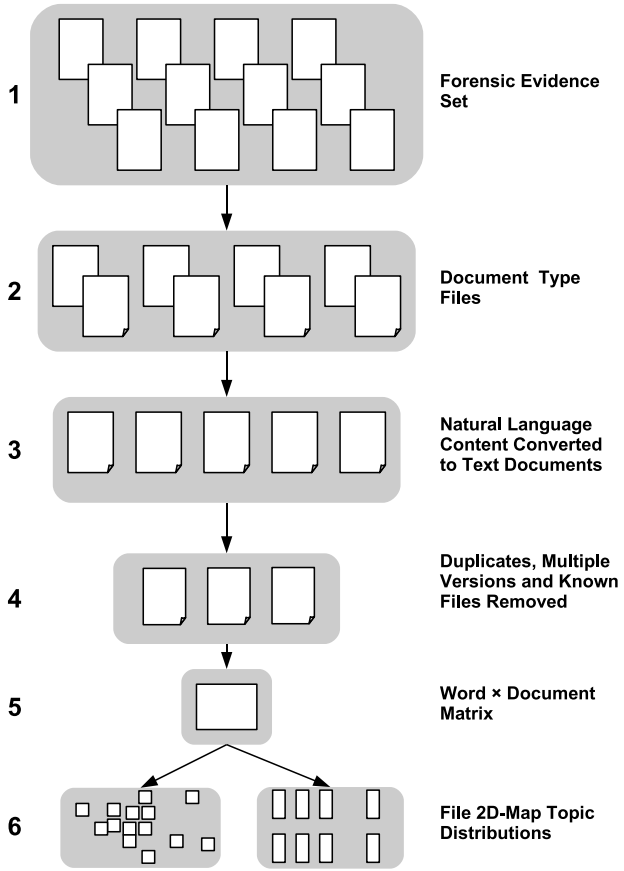

Figure 3. Topic modeling output and interpretation scheme.

- The text corpus (Level 1) was taken from a real investigation. It contained more than 100,000 entities such as documents, operating system files, deleted entities and page files.
- The data set and data type were selected according to CRISP-EM Task 3.1-A (Select Sites/Equipment/Device) and CRISP-EM Task 3.1-B (Select Types of Data to be Included). All the document files (`.doc`, `.txt`, `.pdf`, `.html` and `.rtf`) in the evidence set were extracted using FTK. The files were restricted to allocated or logical files. This data set (Level 2) contained 12,483 documents.
- The data set was reduced to documents with natural language content according to CRISP-EM Task 3.2-A (Reduce Data). After converting the documents to text files (CRISP-EM Task 3.5-A (Convert Data Formats)), the data set (Level 3) contained 1,661

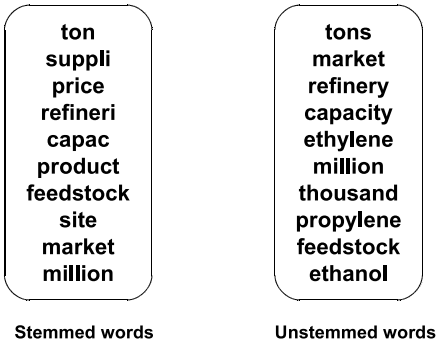

Figure 4. Topic comparison with and without stemming.

documents. Removing files such as keystroke logs, software documentation, multiple versions of the same documents and files with no text (CRISP-EM Task 3.2-A: (Reduce Data)) produced a data set of 837 files (Level 4).

3.3 Data Pre-Processing

Data pre-processing, which corresponds to CRISP-EM Task 3.3-D (Perform Text Processing), was programmed in Python. In this step, stop words (common words appearing frequently in the text), words occurring only once in the corpus, and numbers, special characters and words with two characters or less were removed. The result was a Word $\times$ Document matrix with approximately 11,000 words $\times$ 837 documents (Level 5). This matrix was the input for the topic modeling step.

3.4 Experimental Setup

Early in experiments it became clear that forensic data poses unique challenges for topic modeling. A major challenge is the use of stemming, i.e., reducing derived words to their stems. For example, the words, "waiting," "waits" and "'waited," are reduced to their stem, "wait." The Porter stemming algorithm [15] in the Natural Language Toolkit of Python was used to perform stemming. Stemming was planned as a standard pre-processing task, but the stemmed words hampered the intelligibility and interpretation of topic distributions.

We ran two experiments. The first applied stemming to words. The second used inflections and derived versions of words without stemming.

Figure 4 presents the results obtained with and without stemming. It is important to understand the influence that stemming has on the interpretation of results. If stemming hampers an investigator from grasping

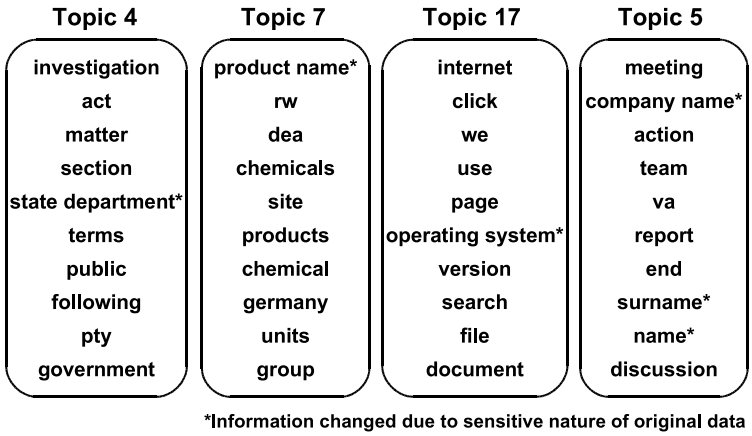

Figure 5. Sample topics modeled from forensic data.

the gist of a topic because he/she is unable to see the original unstemmed word, then it is more appropriate to develop topics without stemming. This is despite the fact that not using stemming increases the dimensionality of the problem.

Several topic models are available, each with different assumptions about the distribution of topics [7]. The Latent Dirichlet Allocation (LDA) model assumes that the set of topics has a Dirichlet distribution. It produces a more reasonable mixture of topics compared with earlier approaches that do not use explicit models [2].

Our experiments used LDA as the topic model. For simplicity, the number of topics was fixed at 20. In the future, the LDA model will be extended by defining the number of topics as a random variable; this will permit the model to infer the natural number of topics inherent in the text corpus. The Matlab Topic Modeling Toolbox [6] was used to perform LDA topic modeling.

3.5 Experimental Results

The output of topic modeling is a Word × Topic matrix and a Topic × Document matrix, which correspond to the data set at Level 6 (Figure 3).

- **Word × Topic Matrix:** Each column of this matrix represents a topic as a probability distribution over words. The top-10 words (words with the highest probabilities) provide a good description of a topic. Listing the top-10 words for each topic provides a summary of the document collection. Figure 5 presents sample topics modeled from forensic data. Topic 17 deals with computer use and

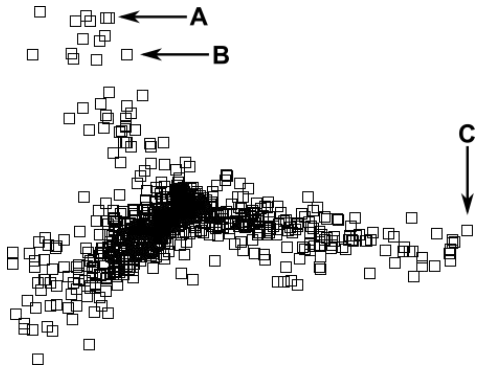

Figure 6. Visualization of documents in a 2D map.

Internet access/search. Topic 5 relates to company meetings that were attended by a specific individual.

- **Topic × Document Matrix:** Each column of this matrix represents a mixture of topics for a document. The mixture of topics describes the semantic context or gist of the document [7]. Documents with similar topic distributions are closely related in terms of semantic context. This "relatedness" of documents can be visualized in a 2D map, which presents the symmetrized Kullback-Leibler divergence [10] between each pair of topic distributions. (The Kullback-Leibler divergence measures the difference between two probability distributions.) Classical multidimensional scaling is used to visualize all pairwise document distances in the 2D map. Figure 6 shows a 2D visualization of the forensic document collection, where each block represents a document. Documents A and B are closely related based on their mixtures of topics (semantic context). On the other hand, Documents A and C differ significantly in terms of their semantic context. Thus, if Document A is relevant to the case at hand, the investigation should focus on Document B rather than Document C. A similar 2D map can be generated for topics to convey the relatedness between topics. In general, if a topic is identified as being relevant to a case, other topics can be prioritized for investigative purposes based on their proximity to the original topic in the 2D map.

4. Forensic Benefits

Topic modeling can assist digital forensic analysts and investigators in several ways. In large cases, with multiple data sets from multiple sites, performing topic modeling on natural language data can provide

analysts and investigators with valuable information about the semantic context of the data. A summary of the natural language data also enables investigators to prioritize the data to be analyzed. A 2D map helps identify closely related documents that would not typically be identified via keyword searches. The map also assists in expanding the set of relevant documents. Moreover, the topics can be used to augment the existing keyword set. When an existing keyword is a top-10 word for a topic, the other words defining the topic can be included in the keyword set. Note that such an expansion of the keyword set is based on the actual characteristics of the forensic data, not on prior knowledge of the case.

5. Lessons Learned

Topic modeling is a promising technique because it reduces the quantity of data to be reviewed by human analysts and suggests prevalent themes within a set of documents to be analyzed. Although much research remains to be done on algorithm development and performance evaluation, our work has shown that even off-the-shelf algorithms can function very well. One issue that deserves attention is the design of performance metrics that reflect modeling goals. This is a significant challenge for standard applications of topic models [16], more so for digital forensic applications. The metrics should reflect the requirements of the forensic environment (e.g., intelligibility to human analysts and salience of detected topics).

Our study identified several other practical matters.

- Many documents have multiple versions. Treating these versions as independent documents increases the computational overhead and skews the results (topics). On the other hand, attempting to detect the different versions of each document is a difficult problem. For example, it is not clear how to deal with two documents that have a small overlap or how to merge different versions of documents without losing relevant information.

- Named entities (e.g., person names, locations and organizations) have high evidence potential, but need to be treated with care. We recommend that named entities be recognized [9] and removed from documents temporarily (to exclude them from data pre-processing tasks such as stemming and removal of stop words). Newman, *et al.* [13] have combined topic models and named entity recognizers to jointly analyze named entities and topics. This enables topics to be used to relate entities, which provides a wealth

of information on people, organizations and locations mentioned in the text corpus.

- Documents written in different languages may be present in a corpus. Such documents should be treated separately for several reasons, e.g., investigators may not be proficient in all the languages, data pre-processing tasks such as stemming and spell checking are language-dependent, and existing algorithms cannot perform topic modeling across languages. An automated system (see, e.g., [3]) may be used to separate documents written in different languages.

- Stemming reduces the number of parameters in a corpus and consolidates semantically-related words. Also, it increases the number of occurrences of individual words in a corpus, which leads to better modeling. However, as discussed earlier, using stemming on forensic data may hamper the understanding of topic distributions. It may, therefore, be advisable to revert to the original words when presenting topics to an investigator.

- "Known files" (e.g., `readme.txt` and other help files, license agreements, etc.) must be removed from a corpus to reduce the amount of spurious data presented to the analyst. This can be done very efficiently by screening known documents using hash values.

- Spelling mistakes add parameters to the model and give rise to incorrect word statistics (the count for one word is assigned to multiple variants). However, it is difficult to automate spell checking in a reliable manner, especially in an informal context where important neologisms and jargon could be transcribed incorrectly. It may be preferable to have low precision as opposed to correcting spelling mistakes in an incorrect manner. This matter deserves further investigation.

- It is standard practice in topic modeling to remove words that occur only once in a corpus. This usually leads to the removal of approximately 5% of the vocabulary of a corpus. However, when this practice was applied to the forensic data set, approximately 50% of the vocabulary was removed, suggesting that valuable information was discarded in the process. A better way for dealing with unique words is needed for topic modeling to be successfully applied to forensic corpora.

- Text corpora used for topic modeling are typically homogeneous (e.g., news articles, conference proceedings and book chapters). Forensic corpora, on the other hand, are generally mixtures of

documents, reports, letters, email bodies and faxes. It is important to modify topic modeling approaches to better handle non-homogeneous data, e.g., by avoiding the bias towards longer documents inherent in the statistical models used by current approaches.

6. Conclusions

This paper has reported on a case study of topic modeling applied to forensic data very early in an actual investigation. No evidence was discovered in this investigation, but the analysis indicates that, with certain refinements, topic modeling can be very useful for discovering the semantic context of text documents in a forensic corpus and for summarizing document content. Future research will investigate the role of metadata in forensic corpora and the application of topic modeling on corpora from different types of cases. Also, topic modeling algorithms will be augmented to address the temporal characteristics of data and the evolution of topics and changes in their importance [12, 18].

References

[1] N. Beebe and J. Clark, Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results, *Digital Investigation*, vol. 4S, pp. S49–S54, 2007.

[2] D. Blei, A. Ng and M. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] G. Botha, V. Zimu and E. Barnard, Text-based language identification for the South African languages, *Proceedings of the Seventeenth Annual Symposium of the Pattern Recognition Association of South Africa*, 2006.

[4] E. Casey, *Digital Evidence and Computer Crime*, Academic Press, London, United Kingdom, 2000.

[5] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartzrysler, C. Shearer and R. Wirth, CRISP-DM 1.0: Step-by-Step Data Mining Guide, The CRISP-DM Consortium, SPSS, Chicago, Illinois (www.crisp-dm.org/CRISPWP-0800.pdf), 1999.

[6] T. Griffiths and M. Steyvers, Finding scientific topics, *Proceedings of the National Academy of Sciences*, vol. 101(1), pp 5228–5235, 2004.

[7] T. Griffiths, M. Steyvers and J. Tenenbaum, Topics in semantic representation, *Psychological Review*, vol. 114(2), pp. 211–244, 2007.

[8] D. Harman, Overview of the first text retrieval conference, *Proceedings of the First Text Retrieval Conference*, pp. 1–20, 1992.

[9] A. Louis, A. de Waal and J. Venter, Named entity recognition in a South African context, *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, pp. 170–179, 2006.

[10] D. Mackay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, United Kingdom, 2003.

[11] C. McCue, *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*, Butterworth-Heinemann, Burlington, Massachusetts, 2007.

[12] Q. Mei and C. Zhai, Discovering evolutionary theme patterns from text: An exploration of temporal text mining, *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 198–207, 2005.

[13] D. Newman, C. Chemudugunta, P. Smyth and M. Steyvers, Analyzing entities and topics in news articles using statistical topic models, *Proceedings of the Intelligence and Security Informatics Conference*, pp. 93–104, 2006.

[14] M. Pollitt and A. Whitledge, Exploring big haystacks: Data mining and knowledge management, in *Advances in Digital Forensics II*, M. Olivier and S. Shenoi (Eds.), Springer, New York, pp. 67–76, 2006.

[15] M. Porter, An algorithm for suffix stripping, *Program*, vol. 13(3), pp. 130–137, 1980.

[16] L. Rigouste, O. Cappe and F. Yvon, Inference and evaluation of the multinomial mixture model for text clustering, *Information Processing and Management*, vol. 43(5), pp 1260–1280, 2007.

[17] J. Venter, A. de Waal and N. Willers, Specializing CRISP-DM for evidence mining, in *Advances in Digital Forensics III*, P. Craiger and S. Shenoi (Eds.), Springer, New York, pp. 303–315, 2007.

[18] X. Wang and A. McCallum, Topics over time: A non-Markov continuous-time model of topical trends, *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 424–433, 2006.