Timeline Visualization of Keywords

Wynand Van Staden

Chapter 13

TIMELINE VISUALIZATION OF KEYWORDS

Wynand van Staden

Abstract Visualizations of communications between actors are typically presented as actor interactions or as plots of the dates and times when the communications occurred. These visualizations are valuable to forensic analysts; however, they do not provide an understanding of the general flow of the discussed topics, which are identified by keywords or keyphrases. The ability to view the content of a corpus as a timeline of discussion topics can provide clues to when certain topics became more prevalent in the discussion, when topics disappeared from the discussion and which topics are outliers in the corpus. This, in turn, may help discover related topics and times that can be used as clues in further analyses. The goal is to provide a forensic analyst with assistance in systematically reviewing data, eliminating the need to manually examine large amounts of communications.

This chapter focuses on the timeline-based visualization of keywords in a text corpus. The proposed technique employs automated keyword extraction and clustering to produce a visual summary of topics recorded from the content of an email corpus. Topics are regarded as keywords and are placed on a timeline for visual inspection. Links are placed between topics as the timeline progresses. Placing topics on a timeline makes it easier to discover patterns of communication about specific topics instead of merely focusing on general discussion patterns. The technique complements existing visualization techniques by enabling a forensic analyst to concentrate on the most interesting portions of a corpus.

Keywords: Email corpus, topic extraction, timeline visualization

1. Introduction

Data visualization in social network analysis is typically organized around communication patterns – who knows whom and how the ac-

tors interact. During an investigation, social network analysis is used with great effect to determine which actors might be colluding, where to search for potential material evidence and which actors should be interviewed for additional data. When a large corpus has to be examined, the visualization of interactions and communication patterns is an important part of the analysis process.

As discussed in the next section on related work, considerable research has concentrated on data visualization. However, in nearly all cases, visualization has focused on the quantitative aspects of the data – for example, plotting email messages as two-dimensional data where one dimension corresponds to the email user and the other dimension corresponds to the time when the email was sent. Such visualization may directly assist in discovering evidence.

Taking a cue from information cartography, this chapter proposes a technique for visualizing topical communication on a timeline by extracting keywords (representative of topics) from a corpus. This provides a means for a forensic analyst to follow the progression of topics (and related topics) based on a particular time window. The idea of providing a timeline visualization of topics during analysis comes from the work of Shahaf et al. [20]. However, the proposed approach is different because, as a first-step technique that assists a forensic analyst in an investigation, it does not hide information and important clues from the analyst. Specifically, it enables a forensic analyst to identify the various topics in a corpus and how they flow through time, providing valuable assistance in understanding actor interactions and supporting focused explorations of a corpus. The timeline-based visualization of keywords is tested on the well-known Enron email corpus.

2. Related Work

Digital forensics is becoming a mature discipline [13] with standardized processes and commercial tools. Email forensics has developed into a separate area within the discipline and comes with its own challenges [8].

Several researchers have investigated visualization as a means to discover digital evidence. Schrenk and Poisel [18] survey visualization techniques used in digital forensics. Olsson and Boldt [12] have developed a visualization tool that provides timelines for event reconstruction from files and file content. Fei et al. [3] have employed self-organizing maps to discover anomalies in data that can help identify sources of evidence.

Devendran et al. [2] have conducted a comparative study of five popular open-source email forensic tools; their study indicates that visu-

alization is restricted to standard content inspection. However, tools that support email visualization are becoming more prevalent. For example, Stadlinger and Dewald [22] have developed a tool that depicts email communications between different accounts. The tool provides histograms of email volumes per hour and per day, and the most active users. It also creates a link graph that highlights the flow of email from accounts to other accounts (i.e., outbound communication patterns between accounts).

Haggerty et al. [5] report that most investigative tools for email visualization support quantitative data analyses. Their triage system makes use of link analysis and tag clouds to visualize interactions and actor relationships. The tag clouds highlight important words and concepts shared by actors. However, their approach focuses the attention of a forensic analyst on searching for evidence instead of appreciating the patterns and events that are latent in the data. Understanding the nascent patterns and events assists the analyst in developing complex and concrete ideas about where to search for evidence.

Frau et al. [4] have developed a tool that depicts email as glyphs whose size and color change based on their locations in the email folder hierarchy and their overall size. Email messages are presented as a scatter-plot on a timeline.

Viegas et al. [24] have created a forensic tool that provides an overview of topics discussed between users and their contacts. However, their tool is not designed for email visualization and does not link topics over time.

Nordbo [11] has developed a visualization tool that considers user interactivity to discover digital evidence. The tool provides email timeline views per user, activity histograms summarized per day, week and overall, frequencies of messages sent and received, and popular communication times. Several other visualization techniques and tools have been developed and interested readers are referred to the work of Joorabi [7], Appan et al. [1], and Sudarsky and Hjelsvold [23].

Very little research has focused on topics and timelines as a means for visualization. One exception is the work of Shahaf et al. [20], which introduces the concept of information maps ("metro maps") that convey the knowledge and evolution of stories in curated news articles. The maps are generated based on properties and constraints – coherence, coverage and connectivity. A highly-coherent map provides storylines in which each point in the story relates to the previous and next "stops." High coverage ensures that a storyline provides as much information as possible about a story and promotes diversity (i.e., the storyline provides as much information about a particular topic as possible). Connectivity ensures that links between different aspects of a story are provided,

meaning that the connections between different aspects of the story are present (as a reader might expect). The concept of metro maps has been extended to the visualization of academic (research) papers [19].

The work of Shahaf et al. [20] has motivated the timeline visualization of topics described in this chapter. However, the notions of coverage and diversity are difficult to apply to an email corpus during an investigation because the objective of a forensic analyst is not to acquire new knowledge. Instead, the analyst is interested in discovering material evidence and may not care how well an email message covers a particular topic, just that the topic is present in the message. The considerations of coverage and diversity in the case of an email corpus and topic visualization are left for future research.

3. Proposed Technique

The proposed technique incorporates three processes: (i) data acquisition; (ii) topic extraction and preprocessing; and (iii) visualization:

- **Data Acquisition:** The data acquisition process involves data preparation, extraction and storage in the appropriate formats and locations. This requires the email messages to be parsed in various formats, such as UNIX mbox [6] and Microsoft PST/OST. Data may be stored in a normalized relational database, in a NoSQL database that handles large data volumes more effectively and scalably, or in a container format. The principle is that it should be easy to query the data.

- **Topic Extraction and Preprocessing:** Topic extraction and preprocessing involve the following steps:

  - Automated extraction of keywords using established techniques such as word co-location analysis and named entity recognition.
  - Normalization of extracted keywords, which includes automated spelling correction. This step can be difficult because most spelling corrections are curated: the user is present and can provide guidance to the spelling corrector. Extracting keywords and making automatic corrections require assumptions to be made about the correct spellings of words, which could form a vernacular that is unique to the entity. This issue is discussed by Samanta and Chaudhuri [17].
  - Generation of common lexicographic rendering indexes related to normalization. The lexicographic renderings of keywords may differ slightly in the corpus due to the writing

habits of individuals (e.g., misspellings of words and uncommon renderings of company names). These minor variations are united to present a consistent view of keywords.

- **Visualization:** Visualization involves the following steps:
  - Acceptance of a search query.
  - Finding related or similar keywords.
  - Clustering topics and email messages based on keywords.
  - Rendering topics and keywords on a timeline.

The proposed technique is implemented using a lightweight SQL relational database system to store the data. Email messages are stored in one table. Another table contains the keywords. A bridging table is used to present the many-to-many relationships between email messages and keywords. A final table contains the normalized keywords and their one-to-many relationships.

The reference implementation was evaluated using the well-known Enron dataset. However, timeline comparisons of reported events and corpus events were not performed.

The next two sections describe the topic extraction and preprocessing phase and the visualization phase in detail.

4. Topic Extraction and Preprocessing

The topic extraction and preprocessing phase involves harvesting keywords (keyphrases) from a corpus and preparing the extracted topics for querying. Three types of models may be employed for keyword extraction: (i) statistical models; (ii) supervised models; and (iii) unsupervised models [21]. Each model type has its own advantages and disadvantages for use in different domains [21]. However, a system used by a forensic analyst should provide as much information as possible with little configuration overhead. Specifically, it should provide a starting point for the analyst.

In the case of the reference implementation, an unsupervised model was considered that would be relatively fast and would not require the number of topics to be predetermined (e.g., latent Dirichlet allocation). For this reason, the rapid automated keyword extraction (RAKE) model [15] was selected. Note that the idea was to create a reference implementation that would permit the replacement of one model with another to adapt topic extraction to a particular domain. However, it should be clear that any other model could be used after sufficient testing in a given domain.

Rapid automated keyword extraction considers stop words as boundaries between potential keywords and scores the individual words based on co-occurrence. The highest scoring candidates are used as keywords for the text. The choice of the stop-word list can play a role in the coherence of the keywords that are harvested. Standard stop-word lists were selected for the reference implementation. However, domain-specific stop words may yield additional benefits [9]; this topic is the subject of future research.

The standard stop-word list provided by NLTK [10] introduced too much noise during initial testing. Therefore, the SMART stop-word list [16] was chosen.

The second part of preprocessing involves getting the keywords ready for searching. Stemming was deemed to be an appropriate preprocessing step for information retrieval.

Stemming maps a known word to a common form. Stemming words that have similar lexical and semantic representations produces a common lexicographic representation that can be used to perform string comparisons more easily. This enables an analyst to enter variations of the words to compare against the corpus. For example, stemming "activities" and "activity" maps both words to the same common lexicographic representation.

This work opted for the well-known stemming technique of Porter [14], which maps "activities" and "activity" to "activ." Thus, a forensic analyst could use the search term "activities," but still obtain all the email messages that contain the term "activity."

In order to facilitate searches based on keywords, all the words in the extracted topics were stemmed and an index was created on the actual topics to which the words referred. Words in the search terms were also stemmed and matches were performed on the stemmed words.

A common representation index was created to account for slight variations of keywords such as "securities exchange commission" and "security exchange commission." The common representation index employed a maximum likelihood estimator to produce a consistent visual rendering of such topics. The estimator mapped keywords (using stemming) to the most common representation found in the list of keywords. In this case:

$$c(k) = \text{argmax}_x \delta(x, k)$$

where $\delta(x, k)$ is the number of times x appears as a topic related to k. This approach ensures that minor variations in topic spellings produce a single consistent representation during visualization.

5. Visualization

The visualization of topics is presented after a search query (term) is issued. The stemming of the search terms works in the same way as the stemming of the keywords obtained via the rapid automated keyword extraction technique. Each keyword is simply stemmed and the stemmed version is used when finding candidate keywords.

5.1 Finding and Ranking

Finding related topics is similar to finding related documents in information retrieval. Several techniques can be used to find related documents, the most common being TF-IDF (term frequency/inverse document frequency). In basic information retrieval, searching and matching are variations of the bag-of-words approach.

Upon conducting sampling and statistical analyses of the words in a collection of documents, it is possible to determine how different the documents are from each other. In such cases, the search term is considered to be a document and the searching system simply finds all the documents that are similar to the search term document. These search techniques are extremely powerful for large corpora with large documents. However, since the approach presented here preprocesses the keywords, the search essentially matches words in the short search term against words in the short keyword/topic list. This makes the matching technique simple and fast.

The matching technique employs a similarity score based on the Jaccard index J, which is defined as:

$$J(x, y) = \frac{|x \cap y|}{|x| + |y|}$$

where x and y are the words being compared.

Using the stemmed index created in the preprocessing stage, all the candidate keywords are found and then ranked according to the Jaccard distance measure. Algorithm 1 specifies the details of the search.

The search results return the top $n = 30$ exact keyword matches in order to reduce the amount of noise that can bleed into the listed keywords and reduce the information presented in the visualization. Note that the parameter n can be adjusted to de-clutter the results.

The final step in the process is to retrieve all the email messages in the corpus that contain the listed keywords. These candidate email messages are used to construct the timeline.

Algorithm 1: Search algorithm.

```
Input: T: Search term
Result: R: Ranked list of keywords
S ← StemmedWordList(T);
foreach s in S do
    C ← C ∪ KeywordLookup(s);
end
foreach c in C do
    R ← R ∪ (c,J(c,T))
end
return R;
```

5.2 Clustering

The results must be clustered prior to visualization. The choice of clustering method impacts the eventual summary and display of data. However, in the case of the reference implementation, it was decided to employ as much information as possible in the clustering. Topics were clustered per day and subsequently by merging related keywords common to email messages on the same day. The resulting cluster contained common keywords in a collection of email messages on a particular day.

To illustrate the clustering technique, assume that the search term "accounting irregularities" yields several email messages on a single day. This gives rise to one of two scenarios: (i) several email messages on the given day, where all the messages contain the same keyword used in the search; or (ii) several email messages on the given day, where some messages contain only the keyword related to the search term and some messages contain the keyword as well as additional keywords.

In the first scenario, clustering emails containing the same keyword would de-clutter a visual rendering of the timeline. In the second scenario, clustering email messages with the same keyword would also declutter the display. However, these keywords are not assimilated into multi-keyword clusters because the assimilation could obscure unique email clusters.

5.3 Rendering

The timeline is rendered by displaying topic clusters using the topics in the email messages per day. Each topic cluster is then linked to the cluster corresponding to a following day based entirely on the topic. Linking is performed using the following rules:

- For each topic on a given day, find the first future occurrence of the same topic and create a link to the topic.

- For each topic on a given day, find the first future occurrence of the same topic that is present in a multi-topic cluster and create a link to the topic.

- For each topic in a multi-topic cluster, find the first future occurrence of the topic and create a link to the topic.

6. Results

This section presents the visualization renderings for searches using two keyphrases: (i) "accounting irregularities;" and (ii) "securities exchange commission." The renderings illustrate the utility of timeline visualization of keywords/keyphrases. Presenting visual renderings on paper is always difficult and the renderings have been reduced in size for presentation purposes.

In the case of the keyphrase "accounting irregularities," some of the top search terms returned – "accounting irregularities," "accounting irregularities disclosed," "accounting irregularities leaked," "creative accounting" – were found in nearly 400 email messages.

Figure 1 shows the visualization corresponding to the keyphrase "accounting irregularities." The timeline clearly shows two interesting periods during which several clusters of email messages discussed accounting irregularities – between January 14 and 18, as well as between January 29 and March 3. Each period contains a flurry of email messages that discuss the same topics. A merge/split on the topics reveals that the phrase "creative accounting" appears in the second period. However, the keyphrase meanders throughout the timeline. The appearances of "creative accounting" and "accounting problem" yield a potential area of interest because it appears that these topics were being discussed after the initial shock of the Enron exposé.

In the case of the keyphrase "securities exchange commission," some of the top search terms returned – "securities exchange commission inquiry," "exchange commission," "exchange commission opens" – were found in nearly 1,700 email messages

Figure 2 shows the visualization corresponding to the keyphrase "securities exchange commission." Since Enron had regular dealings with the U.S. Securities and Exchange Commission (SEC), the keyphrase search does not provide a significant amount of information – the timeline is riddled with references on an almost daily basis. Moreover, since the corpus was collected around the time of the investigation by the U.S. Securities and Exchange Commission, many email messages would be expected to include the keyphrase "securities exchange commission." The point is that a search using this particular keyphrase delivers very few outliers.

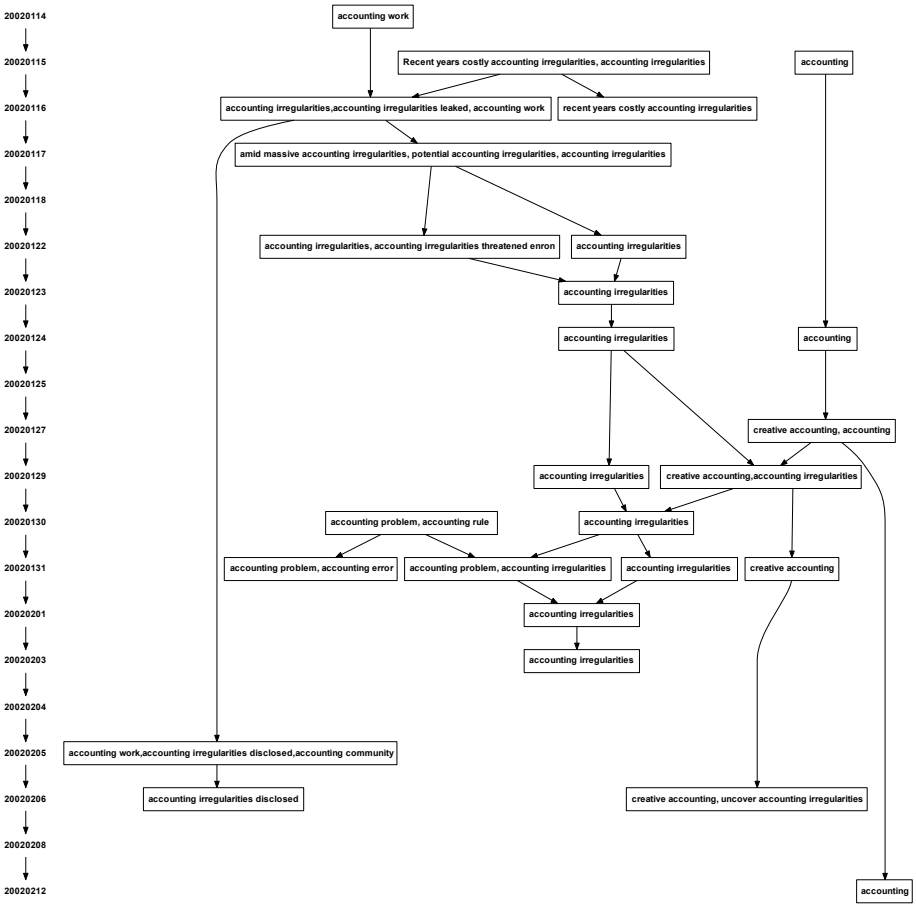

Figure 1. Visualization corresponding to "accounting irregularities."

7. Conclusions

Most digital forensic investigations engage typical social network visualization approaches that depict person-person communications on a timeline and person-person links. Some approaches even provide keyword listings per day or word clouds. However, these visualizations do not provide an understanding of the general flow of the discussed topics, which are identified by keywords or keyphrases.

The proposed timeline-based visualization of keywords draws on the concept of metro maps of science [19]. It leverages automated keyword extraction and clustering to produce a visual summary of topics in an email corpus. Topics are regarded as keywords and are placed on a timeline for visual inspection; links are then placed between topics as

Figure 2. Visualization corresponding to "securities exchange commission."

the timeline progresses. Placing topics on a timeline makes it easier for forensic analysts to discover patterns of communication about specific topics instead of manually analyzing general discussion patterns. Also,

the technique complements existing visualization techniques by enabling forensic analysts to concentrate on the most interesting portions of a corpus, including zooming in on specific times and specific communications.

Future research will attempt to develop an interactive system that will address the problems associated with paper-based visualizations of dense data, enabling forensic analysts to explore corpora more efficiently and effectively. Research efforts will also focus on understanding topical conversations in a corpus by incorporating news events. For example, in the Enron case, the timelines could be correlated with reports of arrests and other relevant events that could enhance human understanding of the case. Finally, research will investigate the choice of stop-word lists and perform rigorous tests on the use of rapid automated keyword extraction on email messages.

Acknowledgement

This research was supported by the South African National Research Foundation under Grant No. 114848.

References

[1] P. Appan, H. Sundaram and B. Tseng, Summarization and visualization of communication patterns in a large-scale social network, *Proceedings of the Tenth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 371–379, 2006.

[2] V. Devendran, H. Shahriar and V. Clincy, A comparative study of email forensic tools, *Journal of Information Security*, vol. 6(2), pp. 111–117, 2015.

[3] B. Fei, J. Eloff, H. Venter and M. Olivier, Exploring forensic data with self-organizing maps, in *Advances in Digital Forensics*, M. Pollitt and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 113–123, 2005.

[4] S. Frau, J. Roberts and N. Boukhelifa, Dynamic coordinated email visualization, *Proceedings of the Thirteenth International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, pp. 187–193, 2005.

[5] J. Haggerty, S. Haggerty and M. Taylor, Forensic triage of email network narratives through visualization, *Information Management and Computer Security*, vol. 22(4), pp. 358–370, 2014.

[6] E. Hall, The application/mbox Media Type, RFC 4155 (`data tracker.ietf.org/doc/rfc4155`), 2005.

[7] M. Joorabchi, EmailTime: Visualization and Analysis of Email Dataset, Master's Thesis, School of Interactive Art and Technology, Simon Fraser University, Burnaby, Canada, 2010.

[8] H. Lalla and S. Flowerday, Towards a standardized digital forensic process, *Proceedings of the Information Security South Africa Conference*, 2010.

[9] M. Makrehchi and M. Kamel, Extracting domain-specific stop words for text classifiers, *Intelligent Data Analysis*, vol. 21(1), pp. 39–62, 2017.

[10] NLTK Project, Natural Language Toolkit (`www.nltk.org`), 2019.

[11] A. Nordbo, Data Visualization for Discovery of Digital Evidence in Email, Master's Thesis, Department of Computer Science and Media Technology, Gjovik University College, Gjovik, Norway, 2014.

[12] J. Olsson and M. Boldt, Computer forensic timeline visualization tool, *Digital Investigation*, vol. 6(S), pp. S78–S87, 2009.

[13] G. Palmer, A Road Map for Digital Forensic Research, DFRWS Technical Report, Technical Report DTR-T001-01 Final, Air Force Research Laboratory, Rome, New York, 2001.

[14] M. Porter, An algorithm for suffix stripping, in *Readings in Information Retrieval*, K. Sparck-Jones and P. Willet (Eds.), Morgan Kaufmann, San Francisco, California, pp. 313–316, 1997.

[15] S. Rose, D. Engel, N. Cramer and W. Cowley, Automatic keyword extraction from individual documents, in *Text Mining: Applications and Theory*, M. Berry and J. Kogan (Eds.), John Wiley and Sons, Hoboken, New Jersey, pp. 1–20, 2010.

[16] G. Salton, *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Upper Saddle River, New Jersey, 1971.

[17] P. Samanta and B. Chaudhuri, A simple real-word error detection and correction using local word bigram and trigram, *Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing*, pp. 211–220, 2013.

[18] G. Schrenk and R. Poisel, A discussion of visualization techniques for the analysis of digital evidence, *Proceedings of the Sixth International Conference on Availability, Reliability and Security*, pp. 758–763, 2011.

[19] D. Shahaf, C. Guestrin and E. Horvitz, Metro maps of science, *Proceedings of the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1122–1130, 2012.

[20] D. Shahaf, C. Guestrin and E. Horvitz, Trains of thought: Generating information maps, *Proceedings of the Twenty-First International Conference on World Wide Web*, pp. 899–908, 2012.

[21] S. Siddiqi and A. Sharan, Keyword and keyphrase extraction techniques: A literature review, *International Journal of Computer Applications*, vol. 109(2), pp. 18–23, 2015.

[22] J. Stadlinger and A. Dewald, A forensic email analysis tool using dynamic visualization, *Journal of Digital Forensics, Security and Law*, vol. 12(1), article no. 6, 2017.

[23] S. Sudarsky and R. Hjelsvold, Visualizing electronic mail, *Proceedings of the Sixth International Conference on Information Visualization*, pp. 3–9, 2002.

[24] F. Viegas, S. Golder and J. Donath, Visualizing email content: Portraying relationships from conversational histories, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 979–988, 2006.