A Term Distribution Visualization Approach to Digital Forensic String Search

Moses Schwartz and L.M. Liebrock

New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA
moses@nmt.edu, liebrock@cs.nmt.edu

Abstract. Digital forensic string search is vital to the forensic discovery process, but there has been little research on improving tools or methods for this task. We propose the use of term distribution visualizations to aid digital forensic string search tasks. Our visualization model enables an analyst to quickly identify relevant sections of a text and provides brushing and drilling-down capabilities to support analysis of large datasets. Initial user study results suggest that the visualizations are useful for information retrieval tasks, but further studies must be performed to obtain statistically significant results and to determine specific utility in digital forensic investigations.

Keywords: Term distribution visualizations, digital forensics, text string search.

1 Introduction

Digital forensic string search is a vital component of the forensic discovery process [2,3,12]. By searching through strings, an analyst may identify forensic artifacts residing in slack space, in deleted files and unallocated space, or in existing files without considering format details [2,12]. However, the state of the art method for identifying artifacts in these datasets is to use a conventional search tool such as Grep [7] and then rely on a human analyst to read through all of the identified hits [3,6]. This task is different from most other string search tasks in that the dataset is almost completely unstructured and the number of hits is extremely high [3]. There has been very little work on reducing the information retrieval overhead and information overload associated with this task [2,3]. Information visualization is one approach to addressing this problem [3].

We propose the use of term distribution visualizations (discussed in more detail in [16]) to ease the task of digital forensic string search. These visualizations, based on the TileBars method [11], show the frequency of a set of search terms throughout a document. This may act as a primary navigation aid for an analyst, allowing her to quickly identify sections of the dataset that may contain relevant information, and then present the text of identified sections. A Focus+Context mechanism provides support for large datasets by allowing the analyst to brush (or select) a large, potentially relevant section, and then drill

J.R. Goodall, G. Conti, and K.-L. Ma (Eds.): VizSec 2008, LNCS 5210, pp. 36–43, 2008.

down (or zoom) to a finer granularity version of the visualization. The visualization and Focus+Context model is demonstrated in Fig. 2 in Section 4.

Section 2 of this paper provides an overview of related work. Section 3 presents the term distribution visualizations and the Focus+Context model. We provide an example of the use of the visualizations in a simulated forensic case in Section 4, and describe our user study and initial results in Section 5. Plans for future work and conclusions are presented in Sections 6 and 7.

2 Related Work

We have found no work that directly applies visualization to digital forensic string search. There has been some work on using advanced search methods for digital forensics [2,3] and much on visualizations that might be applied to the problem [4,5,9,10,11,14,15,17,18,19,20], but none that explicitly discuss the application of visualizations to digital forensic string search.

In [2], Beebe and Dietrich discuss the need for improved digital forensic text string searching; their focus is clustering algorithms. In [3], Beebe and Clark use a clustering algorithm for digital forensic string search. Visualization is suggested in [3] to improve the digital forensic search process, but is not elaborated on.

Early work on visualizations of term distribution focused primarily on their use as relevance-feedback mechanisms for conventional search engines. TileBars, as presented in [11], compactly indicates relative document length, query term frequency, and query term distribution throughout a document (e.g., see Fig. 1(a)). In [13] and [14], TileBars are included as part of a set of visualizations to be used for improving World Wide Web search results. Relevance Curves are also included, which are similar to the histogram visualizations presented here. In [4], a TileBar-inspired term distribution visualization is placed in a scrollbar as an unintrusive and effective within-document search aid.

There has been considerable work on visualizations for text mining, e.g., [5], [18], and [20]. Text mining to identify relevant queries is an important aspect of information retrieval; however, only [5] has a facility to directly view portions of the text and these projects do not place much emphasis on within-document information retrieval. None of this work considers digital forensics.

Visualizations of term distribution have been used for more general trend analysis, as in [10], [15], [17], and [20]. The interaction paradigm in [17] is very similar to the one presented here. It visualizes arbitrary time-series textual data in a histogram format, based on user-supplied queries. However, its primary applications are for information technology tasks such as auditing system logs, its visualizations are relatively coarse, and it is not intended to be used as a general-purpose information retrieval tool.

The most relevant work explicitly visualizes text for information retrieval, e.g., [8], [9], and [19]. [19] visualizes term distribution in a histogram to support information retrieval from speech archives; [19] is relatively specialized, does not explicitly support very large datasets, and uses a relatively coarse visualization. [9] and an accompanying case study, [8], present ProfileSkim, a tool to

visualize a document and provide a user interface for browsing and information retrieval. ProfileSkim uses a language modeling approach to relevance profiling and visualizes a document as a sequential histogram of relevance scores based on user-supplied queries. ProfileSkim's visualization does not provide granular information on the distribution of each search term, nor support searching very large datasets. Although much of this related work could be applied to digital forensics, none explore the potential applications.

3 Visualization Techniques

TileBars and histograms, in conjunction with a Focus+Context model, comprise a Query-Browse (QB) information retrieval model [1,21]. Both visualizations support variable-granularity term distributions, which may be calculated using either a sliding window or discrete blocks of text. In this paper, we show only visualizations calculated with discrete blocks, exclude color TileBars [16] and present a new visualization variant (filled-line histograms). The visualizations shown in the following section correspond directly to those used in our user study, as discussed in Section 5. All example images in this section have been generated from a simulated digital forensic scenario, as discussed in Section 4.

3.1 TileBars

As in the original TileBars [11], the TileBar visualizations in this work are matrices of tiles. Along the horizontal axis, each block represents a block of text. The darkness the block indicates the number of occurrences of a search term in the block. Fig. 1(a) shows an example of the results from our TileBar implementation. Term distribution appears to be obvious and intuitive in this visualization. However, with large numbers of terms this visualization may become harder to interpret and less intuitive. Quantifying this effect is a subject for future work.

3.2 Histograms

Histograms [16] are an extension to the original TileBars [11] visualization concept, and very similar to Relevance Curves [14]. Here distributions are plotted on a graph as a sequential histogram. This supports identification of frequency as the height of a peak, as well as overlap by overlap of the distribution graphs.

Fig. 1(b) shows a greyscale histogram. Overlapping areas appear darker, so distribution overlap is very apparent and intuitive, but there is no indicator of which terms are overlapping or where each term occurs.

Fig. 1(c) shows a color histogram, where the lighter color (than the legend) is used to permit color mixing. Where overlaps occur, the colors are mixed based on how many terms are in the block of text. In this case, term-specific information is readily available and distribution overlap is intuitive. However, interpreting color blending as distribution overlap requires additional cognitive effort.

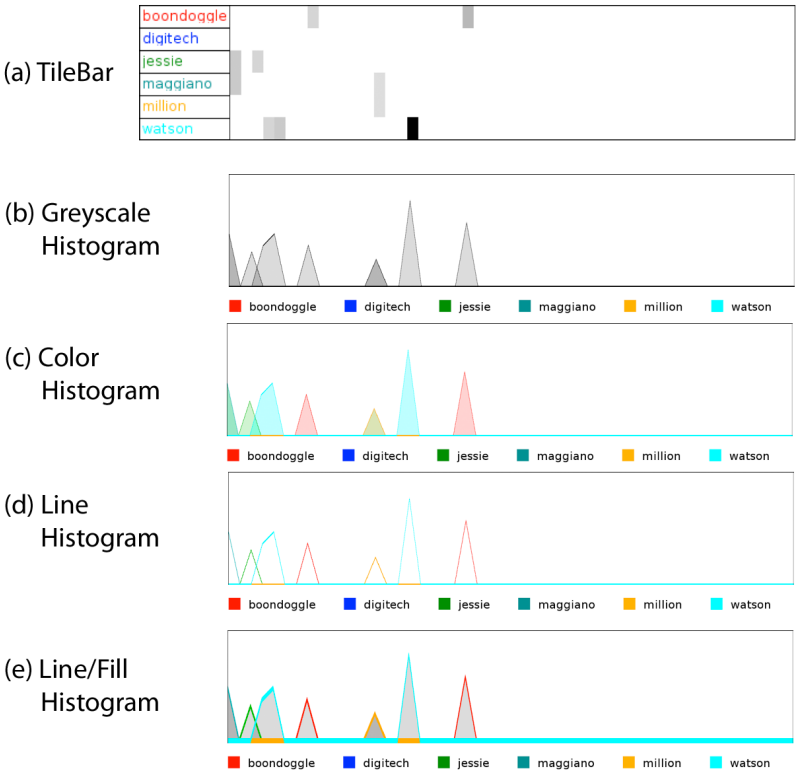

Fig. 1. All visualization variants used in the user study. (a) shows a TileBar, and (b) through (e) show histogram variants. All visualizations were generated with a simulated digital forensic case, discussed in Section 4, and the search terms "Boondoggle," "Digitech," "Jessie," "Maggiano," "Million," and "Watson."

Fig. 1(d) and Fig. 1(e) show two variants of color histograms—line histograms, which present the same information but do not use color blending, and filled-line histograms, which represent overlap by dark fill underneath a color line. In line histograms, there is no need to lighten the color for mixing, so outline colors more closely match the legend. In the filled-line histograms, the fill is done in grey and terms' occurrences are outlined in the legend color. This clarifies to some extent the concentrations and the set of terms in any overlap area.

3.3 Focus+Context

The Focus+Context model allows a user to brush an area of interest within a TileBar or histogram and drill down to visualize the dataset with finer granularity. The previous visualization remains visible to indicate relative location within the overall dataset (see Fig. 2).

4 Digital Forensic Analysis

To illustrate the use of our visualization model for digital forensic string search, we apply it to a digital forensics training module developed by Sandia National Laboratories. In this exercise, we are presented with a seized hard drive image and must perform a digital forensic string search on unallocated and slack space on the drive image to find artifacts. In the fictional scenario, Roberta Hutchins has been accused of attempting to sell trade secrets for Digitech's Project Boondoggle to an individual named Jessie. Interviews have revealed that Roberta planned to meet Jessie at the restaurant Maggiano's, at which point she would be given 1.5 million dollars.

We preprocess the dataset by running Grep, with a list of search terms, on the extracted strings. Next, we use our visualization tool. Fig. 2 shows the reduced dataset in the visualization utility with relevant sections brushed and drilled down to a particular artifact supporting the case against Roberta.

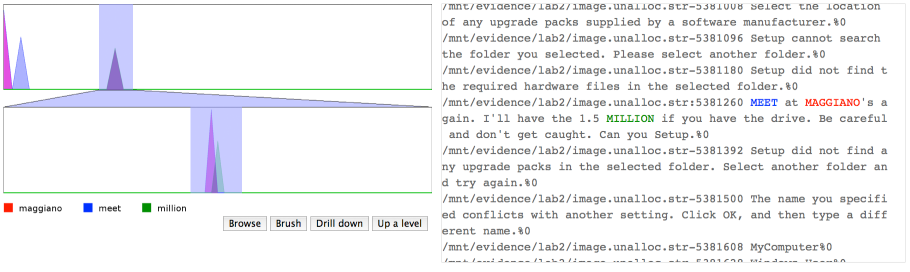

Fig. 2. The visualization tool applied to a notional forensic case. The left pane shows the dataset visualized with the queries "maggiano," "meet," and "million." The section that appears to contain all terms (as indicated by color blending) is brushed, and the visualization drilled down. The drilled section may then be easily browsed. The right pane shows the text of the selected area.

This simple digital forensics example shows that these visualizations can be effective in focusing attention very quickly to the area of the data set that is most likely of interest. This tool is even more effective in complex data sets where search terms appear in many locations and the user must find where certain terms appear in proximity to others to quickly find relevant evidence.

5 Usability Study

We performed a small pilot user study. Here we describe the study, show some preliminary results, and draw what conclusions we can from the preliminary data.

5.1 Study Design

A pilot study was conducted on five subjects from New Mexico Institute of Mining and Technology. Subjects were senior undergraduate students, graduate students, or faculty in the Department of Computer Science. All subjects had prior exposure to this research, but none had previously used the interface.

The study used eight electronic documents and was administered through a web interface. Lewis Carroll's "Through the Looking Glass" was used for training and 8000-line excerpts from the United States Federal Register were used for testing. Unique information was identified in each excerpt by randomly identifying a 400-line section and then selecting specific facts within that section. From these "answers" we created questions and multiple-choice quizzes to determine whether the subject found the correct information.

The study used five visualization variants and two web-based versions of Grep. The visualization variants included: TileBars, greyscale histograms, color histograms, line histograms, and filled-line histograms (see Fig. 1). The Grep variants showed either all occurrences of all search terms as generated by:

```
grep -C2 -aif terms_file target_file
```

or all overlapping sections as generated by:

```
grep -C15 -i term0 target_file | ... | grep -C15 -i termN
```

Subjects filled out a survey before beginning the study, after each trial, and at the end of the study. The first survey gathered basic demographic information. The other surveys elicited qualitative feedback.

Before the actual trials, subjects were given as much time as they wanted to familiarize themselves with the interface. Training used Lewis Carrol's "Through the Looking Glass" with the Grep interface, TileBars, and color histograms.

Subjects were presented with each document sequentially, with one of the described tools. The order of the documents was fixed, but the search aids (visualization or Grep) were counterbalanced to minimize carryover effects. In each trial, the time until a subject answered the quiz (which presumably coincides with finding desired information) was measured. Answer correctness was also recorded.

5.2 Usability Study Results

Fig. 3 shows the mean time to complete an information retrieval task for each search aid and the time to complete the information retrieval task for each file.

5.3 Usability Study Analysis

Since the pilot study results involve a very small sample, we do not dwell on analysis, as any claims would be suspect. However, many of the visualizations appear to perform comparably to Grep with all hits shown—this is a very positive early result. To see why, consider how long these searches would have taken with no search aid at all! Grep that only shows overlapping occurrences of terms seems to have been the most effective search aid, but excluding so much data may render the technique unsuitable for digital forensic purposes.

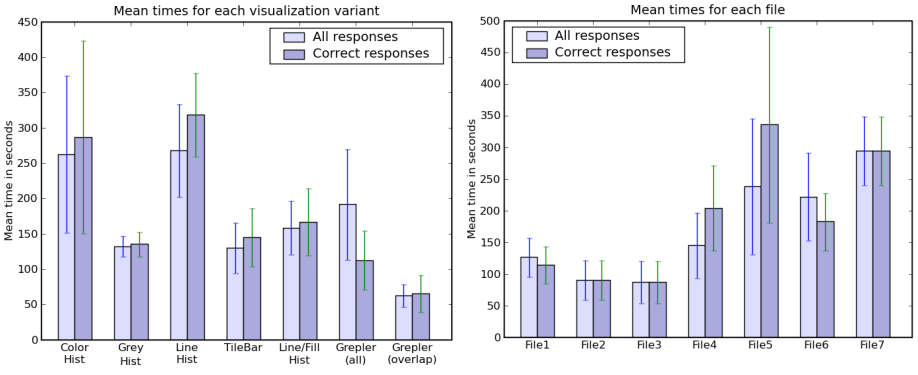

Fig. 3. Mean time to complete an information retrieval task for each search aid and mean time to complete the task for each file. Error bars show standard error.

The time taken for each file varied considerably, indicating that the information retrieval difficulty was not uniform; this casts some doubt on the utility of our results. However, with more subjects the counterbalancing will largely negate these effects.

6 Future Work

We have identified numerous avenues of future work. Maximizing efficacy of the visualizations is an obvious extension and will be greatly aided by performing further work on the user study. Extensions to make the visualization more applicable to digital forensics, such as providing support for collaborative analysis and showing file boundaries in the visualization will be explored.

7 Conclusions

Visualization for digital forensic string search is a virtually untouched field of research. Our visualization model appears to be effective as an aid for digital forensic string search, but needs full validation through further user studies, as well as further specialization for digital forensics.

Acknowledgements. This work was supported in part by NSF Grant #0313885.

References

1. Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press / Addison-Wesley (1999)
2. Beebe, N., Dietrich, G.: A New Process Model for Text String Searching. Springer, Norwell (2007)
3. Beebe, N.L., Clark, J.G.: Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. In: Digital Investigation, September 2007, vol. 4(suppl. 1) (2007)

4. Byrd, D.: A scrollbar-based visualization for document navigation. In: Proceedings of the Fourth ACM International Conference on Digital Libraries (1999)
5. Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., Plaisant, C.: Discovering interesting usage patterns in text collections: integrating text mining with visualization. In: CIKM 2007: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 213–222. ACM Press, New York (2007)
6. Forte, D.: The importance of text searches in digital forensics. In: Network Security, April 2004, pp. 13–15 (2004)
7. Free Software Foundation. Tool: GNU Grep
8. Harper, D., Koychev, I., Sun, Y., Pirie, I.: Within-document retrieval: A user-centred evaluation of relevance profiling. In: Information Retrieval, vol. 7, pp. 265–290 (2004)
9. Harper, D.J., Coulthard, S., Yixing, S.: A language modelling approach to relevance profiling for document browsing. In: JCDL 2002: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital Libraries, New York, NY, USA (2002)
10. Havre, S., Hetzler, E., Whitney, P., Nowell, L.: ThemeRiver: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics 8(1), 9–20 (2002)
11. Hearst, M.A.: Tilebars: visualization of term distribution information in full text information access. In: CHI 1995: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, pp. 59–66. ACM Press/Addison-Wesley Publishing Co (1995)
12. Mandia, K., Prosise, C., Pepe, M.: Incident Response & Computer Forensics. McGraw-Hill/Osborne, California (2003)
13. Mann, T., Reiterer, H.: Case study: A combined visualization approach for www-search results. In: Proc. IEEE Information Visualization 1999, pp. 59–62 (1999)
14. Mann, T.M.: Visualization of WWW-search results. In: DEXA Workshop, pp. 264–268 (1999)
15. Mao, Y., Dillon, J.V., Lebanon, G.: Sequential document visualization. In: IEEE Transactions on Visualization and Computer Graphics, November/December 2007, vol. 13(6), pp. 1208–1215 (2007)
16. Schwartz, M., Hash, C., Liebrock, L.: Term distribution visualizations with a focus+context model. Technical report, New Mexico Institute of Mining and Technology (2008), `http://cs.nmt.edu/~liebrock/papers/SchwartzHashLiebrock.pdf`
17. Splunk, Inc. Application: Splunk
18. Paley, W.B.: TextArc: Showing word frequency and distribution in text. Poster presented at IEEE Symposium on Information Visualization (2002)
19. Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F.C.N., Singhal, A.: SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In: Research and Development in Information Retrieval, pp. 26–33 (1999)
20. Wong, P.C., Cowley, W., Foote, H., Jurrus, E., Thomas, J.: Visualizing sequential patterns for text mining. In: INFOVIS 2000: Proceedings of the IEEE Symposium on Information Vizualization 2000, p. 105 (2000)
21. Zhang, J.: Visualization for Information Retrieval, 1st edn. Springer, Heidelberg (2007)