Interactive Hierarchical Quote Extraction for Content Insights

Johannes Knittel[1], Steffen Koch[1] and Thomas Ertl[1]

[1]Institute for Visualization and Interactive Systems, University of Stuttgart

Abstract

This work presents a new approach to visually summarize large micro-document collections such as tweets. We extract frequent patterns of phrases as shortened quotes to present analysts an overview of popular snippets and statements, enabling more specific insights into large text collections compared to keyword-based visualizations. In our hierarchical structure, each quote can be the starting point to extract more fine-grained patterns on a subset of sentences that match the parent pattern. We show that our approach is scalable by applying it to millions of tweets.

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics;*** • ***Computing methodologies*** → *Information extraction;*

1. Introduction

Extracting popular opinions and statements from social media is of vital interest for many stakeholders, including journalists investigating developing stories, and agencies monitoring brand exposure. Unfortunately, dealing with large amounts of micro-documents such as tweets is challenging in several ways, particularly regarding information extraction tasks [ICDV15]. Approaches to summarize documents often do not adapt well to very short texts, and advanced methods may run into scalability issues. The variety of spelling and lack of grammar further complicates the analysis. While trending hashtags or keywords can give analysts an overview of popular topics, they do not convey more concrete concepts and statements. Conversely, displaying a selection of frequently shared and liked documents ignores the vast majority of content.

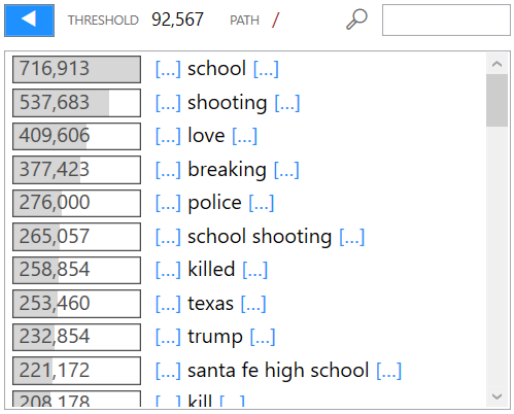

Figure 1: *Initial top layer results after analyzing 10m tweets from May 18, 2018. Analysts can double-click on items to further extract quotes on a subset of tweets matching the clicked pattern.*

To overcome these obstacles, several concepts have been proposed to visualize unstructured text content while preserving the linguistic structure in some way. However, there are a number of shortcomings concerning previous work. Tree-based approaches visualizing how sentences continue [WV08] [CL10] require the analyst to provide a starting pattern and typically do not scale well for many items and variants, because every possible pathway is considered. Hu et al. [HWS17] employ node-link diagrams to improve scalability. However, they state that their static visualization may imply a pattern that does not occur in the underlying data set, a caveat also shared by Van Ham et al. [VWV09]. Furthermore, these approaches make it difficult to quantify the popularity of patterns.

We introduce a new concept to visualize aggregated concepts that is inspired by the way how quotes are often shortened to convey the main idea. We automatically extract such quotes at varying levels of detail to aggregate a range of different micro-documents that talk about similar things, but are phrased slightly different. We assume that many text items share multiple chunks, but are possibly scattered throughout the sentence. Connecting these chunks supports analysts to make sense of the content. Analysts can iteratively dive into patterns to extract more detailed quotes regarding themes of interest, down to the level of individual posts.

2. Quote Extraction

Each quote is a sequence of subsequences, e.g., '*... santa fe ... shooting ...*' represents all texts containing the two subsequences in that order with arbitrary content in-between. Extracting such pat-

DOI: 10.2312/eurp.20191134

terns is non-trivial because of the sheer number of theoretically possible variations. Thus, we first extract all *single* subsequences that meet an adjustable *threshold* and are not entirely comprised of stop words. To reduce the number of redundant variants, we drop patterns that match longer patterns, but occur similarly often. For instance, *'... how old are ...'* is removed, because it matches the equally frequent pattern *'... how old are you ...'* in that data set.

For each resulting item, we then try to find patterns with an additional subsequence (chunk), e.g. *'... school ... shooting ...'* if *'... school ... '* was the initial subsequence. We iteratively repeat this process with increasing numbers of subsequences, until no new pattern is found meeting the threshold.

We use collected word and word pair statistics of the processed set to only count promising patterns that have a high chance of making the cut. We set the default threshold at one percent of the total number of documents for top level results, and at five percent of the number of child items for lower-level results.

3. System Design

The system analyzes the loaded data set and presents a list of patterns as depicted in Figure 1, ordered by popularity to quickly guide the attention to the most frequent quotes. Here, the top level results of 10 million tweets published on May 18, 2018 are displayed, the day on which a school shooting occurred in Texas. Each row shows to the left the number of tweets matching the respective pattern.

Analysts can double-click on any item to retrieve quotes concerning only the data that matches the clicked parent pattern. This parent pattern is then stuck to the top of the list with a distinct color. The related, more fine-grained quotes underneath are aligned with the parent, and the words matching the parent pattern are highlighted with the corresponding color. This way, the path of the analyst is visualized on their way down the tree, it is always obvious which parts of the quote contain new information, and the variety of content between the anchor chunks becomes evident.

The initial top level results in Figure 1 reveal that more than a quarter of a million tweets contained *school shooting*. In this case, the analyst wants to find out more about the shooting and double clicks on that item, then on *'... texas ... school shooting ...'* and finally on *'... killed ... texas ... school shooting ...'*. The resulting visualization is shown in Figure 2 with several (partial) statements about killed students, including different reports on the number of casualties and their quantitative relation. If the analyst double-clicks the same pattern again, the current subcollection is processed again with an even lower threshold, e.g., 15 instead of 314.

The top level results suggest popular themes to facilitate exploring unknown collections. In addition, analysts can also search for specific terms that are not in the list using the search box. Then, popular quotes are extracted regarding all posts matching the query. Figure 3 shows an example in which the analyst searched for `london` and gets results about a stabbing, among others.

4. Discussion and Conclusion

Keyword-based approaches to summarize document collections without preserving the word order can densely compress big data sets, but at a cost of less interpretable results. Instead, our approach to extract possibly shortened quotes enables analysts to *read* the results, supporting sense-making tasks. While it is slightly more verbose than showing just keywords, the system is still capable of aggregating large collections by omitting chunks of words in-between. Importantly, only patterns that actually occur in the data set are visualized, and the popularity of each pattern is exactly quantified. In addition, analysts do not have to provide a starting pattern, enabling the exploration of unknown collections.

The hierarchical approach not only reduces the needed processing time, it also supports the interactive analysis of big data sets, because every selection step drastically reduces the number of relevant items. However, the analyst can influence the depth of the hierarchy. Setting a lower threshold shows more detailed quotes already at the top level, which flattens the hierarchy.

With roughly 500 million tweets each day, high scalability is important. Our implementation can handle millions of sentences. The initial processing of one million tweets takes about 20 s, and diving into one of the popular top-level results takes about 5 s.

In the future, we want to experiment with new design spaces to reduce visual redundancies and to improve the global awareness which part of the data set the analyst currently explores. Additionally, we want to refine our algorithms to speed up the process even more and further improve the quality of the aggregation.

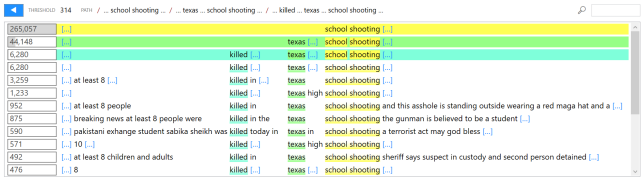

Figure 2: *Extracted quotes from 10m tweets published on May 18, 2018 matching the selected pattern '... killed ... texas ... school shooting ...'. The number of matches is shown on the left, and the highlighting indicates which parent pattern the part belongs to.*

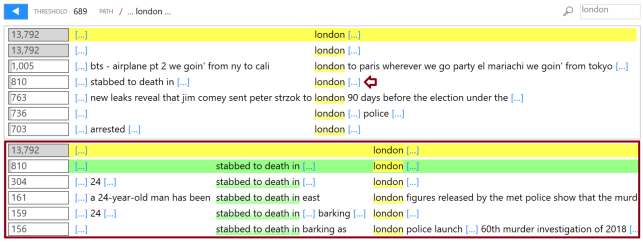

Figure 3: *Upper part: extracted quotes after an analyst searched for 'london'. Lower part: (partial) results after double-clicking on entry 'stabbed to death in [...] london'.*

Acknowledgments

This work was funded by the DFG ER 272/13-1 project Visual Analytics of Online Streaming Text (VAOST).

References

[CL10] CULY C., LYDING V.: Double tree: An advanced KWIC visualization for expert users. In *Proceedings of the International Conference on Information Visualisation* (2010). doi:10.1109/IV.2010.24. 1

[HWS17] HU M., WONGSUPHASAWAT K., STASKO J.: Visualizing Social Media Content with SentenTree. *IEEE Transactions on Visualization and Computer Graphics* (2017). doi:10.1109/TVCG.2016.2598590. 1

[ICDV15] IMRAN M., CASTILLO C., DIAZ F., VIEWEG S.: Processing Social Media Messages in Mass Emergency: A Survey. *ACM Comput. Surv. 47*, 4 (2015), 67:1—-67:38. URL: http://doi.acm.org/10.1145/2771588, doi:10.1145/2771588. 1

[VWV09] VAN HAM F., WATTENBERG M., VIÉGAS F. B.: Mapping text with phrase nets. In *IEEE Transactions on Visualization and Computer Graphics* (2009). doi:10.1109/TVCG.2009.165. 1

[WV08] WATTENBERG M., VIÉGAS F. B.: The word tree, an interactive visual concordance. In *IEEE Transactions on Visualization and Computer Graphics* (2008). doi:10.1109/TVCG.2008.172. 1