Text Visualization Techniques: Taxonomy, Visual Survey, and Community Insights

Kostiantyn Kucher* Andreas Kerren†

ISOVIS Group, Department of Computer Science, Linnaeus University, Växjö, Sweden

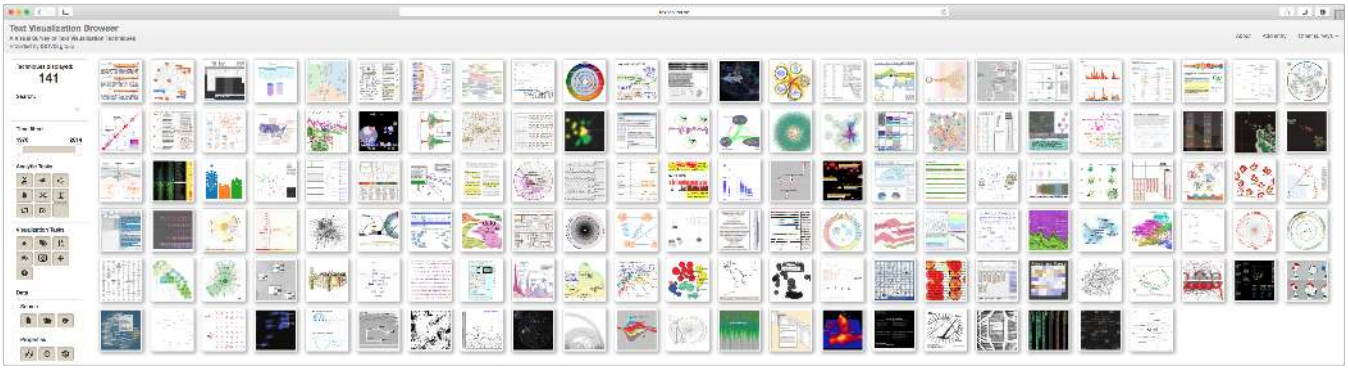

Figure 1: The web-based user interface of our visual survey called *Text Visualization Browser*. By using the interaction panel on the left hand side, researchers can look for specific visualization techniques and filter out entries with respect to a set of categories (cf. the taxonomy given in Sect. 3). Details for a selected entry are shown by clicking on a thumbnail image in the main view. The survey contains 141 categorized visualization techniques by January 19, 2015.

Abstract

Text visualization has become a growing and increasingly important subfield of information visualization. Thus, it is getting harder for researchers to look for related work with specific tasks or visual metaphors in mind. In this paper, we present an interactive visual survey of text visualization techniques that can be used for the purposes of search for related work, introduction to the subfield and gaining insight into research trends. We describe the taxonomy used for categorization of text visualization techniques and compare it to approaches employed in several other surveys. Finally, we present results of analyses performed on the entries data.

Keywords: Visualization, text visualization, survey, interaction, web-based systems, taxonomy, community analysis

1 Introduction

The interest for text visualization and visual text analytics has been increasing for the last ten years. The reasons for this development are manifold, but for sure the availability of large amounts of heterogeneous text data (caused by the popularity of online social media) and the adoption of text processing algorithms (e.g., for topic modeling) by the InfoVis and Visual Analytics communities are two possible explanations. Inspired by the TreeVis.net [27] and TimeVis [29] projects, we propose an interactive visual survey of text visualization techniques that can be used for getting an overview of the field, teaching purposes, and finding related work based on various categories defined in a survey taxonomy. Our web-based survey browser is available at:

`http://textvis.lnu.se/`

The term 'text visualization' is typically used for information visualization techniques that in some cases focus on raw textual data, in other cases on results of text mining algorithms. In the same way, they can be rather general or very specialized and dedicated to specific analytic tasks or application domains. This is the reason why we have decided to construct a taxonomy with numerous categories and subcategories that is exploited by the survey browser in order to facilitate the interactive exploration of the current set of entries. Our visual survey has been implemented as an interactive web page and includes 141 techniques at present originating from peer-reviewed work in InfoVis, Visual Analytics and other relevant research fields. After a short discussion on relevant surveys in the following, we highlight the taxonomy used by our survey browser, some implementation details, and the results of analyses conducted on the collected entries data. The present paper is based on our previous poster abstract [20].

*e-mail: kostiantyn.kucher@lnu.se

†e-mail: andreas.kerren@lnu.se

2 Related Surveys

There are a number of survey papers in the literature that focus on text visualization or its specific subproblems. Šilić and Bašić [30] classify about 30 text visualization methods with regard to data source, underlying text representation and processing method, temporal aspects, and supported user interactions. Alencar et al. [2] describe roughly 30 techniques by means of data source, underlying text representation, visual metaphor, layout, and supported user tasks. Gan et al. [15] discuss approx. 40 techniques with regard to data source, user tasks, visual representation, and supported interactions. Nualart-Vilaplana et al. [24] categorize about 50 techniques on the basis of data source, underlying text structure and corresponding processing method, support for temporal aspect (as well other special data properties), data domain, and visual metaphor. The recent work of Wanner et al. [31] on event detection in texts classifies approx. 50 visualization approaches with regard to data source, text processing methods, event detection methods, visualization representations, and tasks. Table 1 provides an overview of all these surveys and the taxonomies they used.

Finally, the aforementioned visual survey projects use dimensionality, visualization metaphor and alignment to classify tree-oriented techniques [27]; and data properties, temporal properties, visual representation to classify time-oriented techniques [29].

In Proceedings of the 8th IEEE Pacific Visualization Symposium (PacificVis '15), visualization notes track, pages 117-121, Hangzhou, China, 2015. IEEE Computer Society Press.

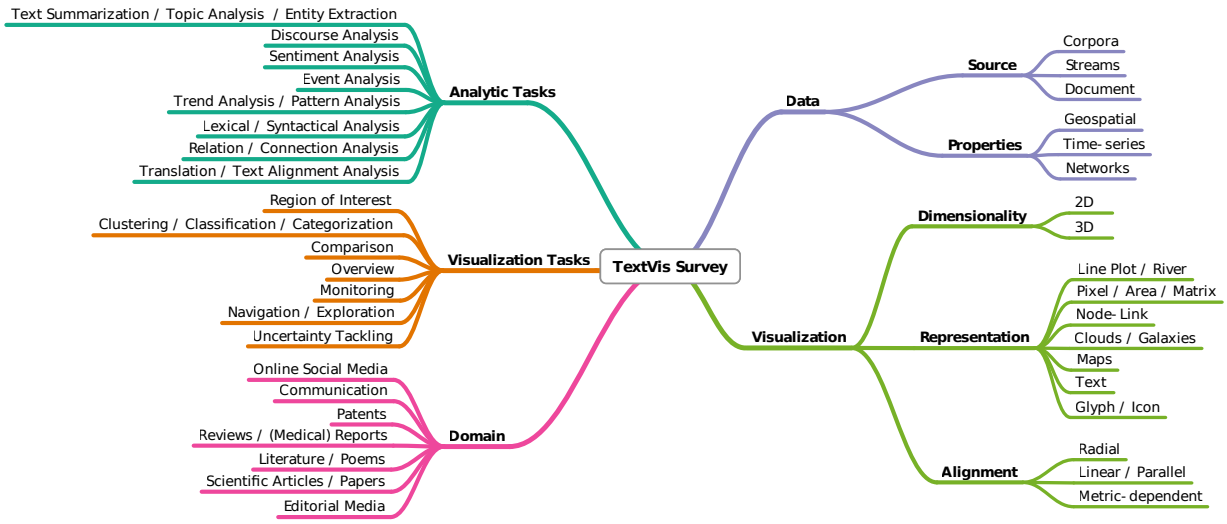

Figure 2: The taxonomy of text visualization techniques used in our visual survey. We focus on the description on the left hand side of the figure in this paper and only briefly summarize the right side in Subsect. 3.4 which should be self-evident for the visualization community.

Table 1: The comparison of text visualization taxonomies. Supported categories are marked by '+', partial support denoted by '(+)'.

Category / Taxonomy	Šilić and Bašić [30]	Alencar et al. [2]	Gan et al. [15]	Nualart-Vilaplana et al. [24]	Wanner et al. [31]	**Our proposed taxonomy**
Analytic Tasks		+	+		+	+
Visualization Tasks	+	+	+		+	+
Data Domain				+		+
Data Source	+	+	+	+	+	+
Data Properties (temporal, etc.)	+			+	+	+
Visual Dimensionality		+	+	+	+	+
Visual Representation (metaphor)		+	+	+	+	+
Visual Alignment (layout)		+	+	+	+	+
Underlying Data Representation	+	+		+		
Data Processing Methods	+			+	+	(+)

3 Survey Taxonomy

We have arranged a taxonomy (cf. Fig. 2) with multiple categories and items in order to classify the techniques with fine granularity. The presented taxonomy is the result of refinements occurring while categorizing entries for the survey, i.e., the choice of concrete category items is motivated by the underlying data. While we cannot claim that our classification is absolutely definite (numerous techniques have been ambiguous, especially in case of hybrid approaches), we have tried to base the choice of category items for particular entries on the description and claims of the original author(s). For example, certain techniques could be easily applied to domains other than originally described, but we do not reflect that in our choice of category items for those techniques. On the other hand, some papers mentioned specific domains only for the sake of giving examples, though the corresponding techniques were not tailored for those domains. In such cases, we have not assigned entries to those domain items. In the remainder of this section, we introduce the categories and items comprising our taxonomy. Due to space limitations, we only briefly discuss those categories that should be familiar to the visualization community in Subsect. 3.4.

3.1 Analytic Tasks

This category describes high-level analytic tasks that are facilitated by corresponding techniques: these items are critical to the main analysis goals that users expect to achieve when employing a text visualization technique.

Text Summarization / Topic Analysis / Entity Extraction We have decided to combine entity extraction/recognition with topic analysis/modeling in a single category item, since visualization techniques treat entity names simply as topics in most cases encountered by us.

Discourse Analysis This item concerns the linguistic analysis of the flow of text or conversation transcript.

Sentiment Analysis We have used this item for techniques related to the analysis of sentiment, opinion, and affection.

Event Analysis While event analysis and visualization is in fact a separate subfield, some of the corresponding techniques deal with the extraction of events from the text data or involve visualization of text in some different manner.

Trend Analysis / Pattern Analysis This item denotes the tasks of both automated trend analysis and manual investigation directed at discovering patterns in the textual data.

Lexical / Syntactical Analysis We have included this item to represent various linguistic tasks, for instance, analysis of lexemes and sentences in poems.

Relation / Connection Analysis This item is dedicated to comparison of data items, including the analysis of explicit relationships exposed by visualizations.

Translation / Text Alignment Analysis We use this item for corpus linguistics tasks, for instance.

3.2 Visualization Tasks

This category describes lower-level representation and interaction tasks that are supported by the text visualization techniques. In comparison to analytic tasks, we have included more instrumental items here, for example, clustering could be used in various visualizations as merely an auxiliary feature.

Region of Interest This task denotes the automatic highlighting/suggestion of data items/regions that could be of interest to the user for more detailed investigation.

Clustering / Classification / Categorization Here, we combine several tasks related to (semi-)automatic tagging or grouping of data elements.

Comparison This item denotes the comparison of several entities facilitated by the visualization technique, e.g., laying out several objects side by side or marking discrepancy (also cf. the survey paper [17]).

Overview We use a very general notion of "overview" for this item, including both techniques that provide "the big picture" by displaying a significant portion of the data set as well as techniques which use special aggregated representations to provide overview while reducing the visual complexity.

Monitoring This task is related to visualization techniques that are designed to alert users to the changes in the data.

Navigation / Exploration We use this item for techniques that facilitate the process of navigating around the data set, while possibly switching the visual representations or underlying data types.

Uncertainty Tackling This task—which is currently not very prominent in the present techniques—is generally related to techniques that handle and/or visualize uncertainty in source or processed data as well as uncertainty in computations.

3.3 Domain

This category describes the dedicated text domains a technique was developed for.

Online Social Media Twitter, Facebook, blogs, forums, etc.

Communication We include email, instant messaging logs, or snail mail letters into this item.

Patents Official patents for detailed disclosure of inventions.

Reviews / (Medical) Reports This item denotes user reviews, medical report data alongside reviews and reports from other sources.

Literature / Poems Various artistic, historical and documentary texts.

Scientific Articles / Papers Scientific texts of various genres and fields.

Editorial Media Text data from organizations (newspapers, etc.) as well as pre-moderated websites (e.g., Wikipedia).

3.4 Data and Visualization

We have decided to even categorize the techniques with regard to both data source and special data properties (if any supported). Here, we list the categories with some references to prominent examples. Data *sources* include the following self-evident items: **Document** [33], **Corpora** [25], and **Streams** [19]. The special data *properties* include **Geospatial** [11], **Time-series** [14], and **Networks** [6].

Finally, to categorize the techniques with regard to the used visualization approach, our taxonomy uses three subcategories. While visual *dimensionality* does not require additional description, we list the others. *Representation* includes the following items: **Line Plot / River** [9, 18], **Pixel / Area / Matrix** [13, 7, 4], **Node-Link** [32], **Clouds / Galaxies** [1, 3], **Maps** [34], **Text** [26], and **Glyph / Icon** [28, 10]. *Alignment*, i.e., layout, includes **Radial** [35], **Linear / Parallel** [8], and **Metric-dependent** [22].

4 Interactive Browser

We have implemented our visual survey as an interactive HTML/JavaScript page that merely requires a modern web browser for access, see Fig. 1 for a screenshot. The survey browser has a main view with a collection of thumbnails (ordered by time) that represent the individual visualization techniques as well as filter controls that comprise text search field, publication year range slider, and category radio buttons. Since the included technique entries may be assigned with arbitrary sets of category tags, and the filtering is based on logical OR operation, the interface contains additional category filters for "Other" entries to support precise filtering, e.g., to display only entries that are not associated with any domain.

After clicking on an entry's thumbnail image, the corresponding details are displayed in a dialog box. Here, a slightly larger thumbnail, a complete list of assigned category tags, a bibliographical reference, a URL (optional), and a link to a BibTeX file (if available) are displayed, see Fig. 3.

Figure 3: Details of a survey entry.

We have also provided an additional form for authors who wish to add a new entry to our survey. The form generates a JSON entry [12] that can be sent to us via email to prevent direct-manipulation of the survey browser content. Finally, we visualize some basic statistics about the current entry set in the "About" dialog. Since the techniques can be assigned with multiple category tags, the sets of corresponding techniques overlap for sibling categories—therefore, we currently use simple bar charts for showing the statistics (as depicted in Fig. 4).

5 Discussion and Analysis

Our decision to design a rather extensive taxonomy was motivated by the need for fine-grained technique search or filtering as well as for the comparison of entries. We have compared our resulting taxonomy to the ones described in the related text visualization surveys (cf. Table 1). In order to match the taxonomies, we have mapped the categories used by the other surveys into several fine-grained categories proposed in our taxonomy. We have not included the category "event detection methods" into the comparison, since it is used only by a single, more specialized survey. As displayed in the table, our proposed taxonomy includes most of the categories except for two: we believe that the underlying data representation (e.g., bag-of-words vs. language model [30] or whole text vs. partial text [24]) is more relevant to the underlying computational methods than to observable visualization techniques. And the same naturally holds for data processing methods (e.g., the specification of involved MDS methods [2]) that are partially covered by other categories in our taxonomy, for instance, the analytic task of topic analysis implies the usage of corresponding computational methods. However, we do not negate the possibility of extending our taxonomy as a part of future work.

Using the data collected for the survey, we have been able to analyze the general state of the text visualization field, to compare the usage of various analysis and visualization techniques (with regard to our taxonomy), and to analyze the information about researchers in this field. According to our current set of entries,

Figure 4: The basic statistics for technique categories which are displayed as bar charts in the "About" dialog on demand.

the trend for rapid increase of text visualization techniques started around 2007. With regard to category statistics (cf. Fig. 4), there is an obvious interest for tasks related to topic modeling (56% of all entries). The majority of the techniques support corpora as data sources (70% of all entries), and a lot of them support time-dependent data (43% of all entries). Another result—which is probably expected—is that only less than 4% of all entries use 3D 3-dimensional visual representations.

We have also taken a look at the authorship statistics for the current data set. The top five authors with regard to number of techniques are Daniel A. Keim (17 entries), Shixia Liu (12 entries), Christian Rohrdantz (9 entries), Daniela Oelke (7 entries), and Huamin Qu (7 entries).

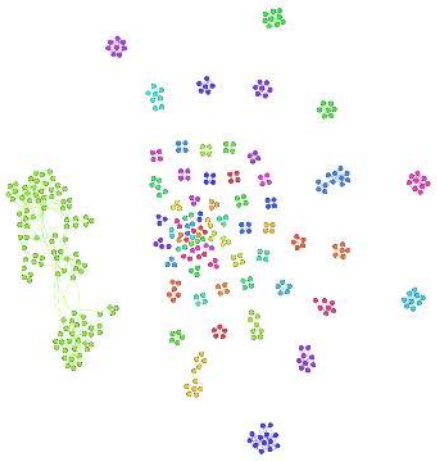

Figure 5: Co-authorship network for current survey entries visualized in Gephi with the ForceAtlas layout. Note the big connected component on the left hand side containing 106 author nodes.

After extracting the co-authorship network (406 nodes, 956 edges), we have analyzed it with Gephi [5]. As seen in Fig. 5, the majority of author nodes are included into isolated connected components of small sizes (less than 10 nodes) while there is a big connected component with 106 nodes present in the graph. The two major clusters in that component represent the research groups from the University of Konstanz and Microsoft Research Asia with Daniel A. Keim and Shixia Liu as cluster center nodes. This fact is quite interesting when we set this in relation to the geolocation statistics of visits to our interactive survey browser (according to Google Analytics) that list United States, China, United Kingdom, Germany and France as the Top 5 user locations.

We have also analyzed several network centralities [23] in this co-authorship network, for instance, *closeness*: the largest closeness value is shared by Frank van Ham and Jonathan Feinberg. However, we were mostly interested in the *betweenness* centrality, because betweenness in co-authorship networks has the largest effect on the research impact as shown by Li et al. [21]. Shixia Liu and Daniel A. Keim happen to have the 1st and the 2nd largest betweenness values in the graph, respectively. While these two researchers have no direct collaboration with regard to our data set, they both have collaborated with Dongning Luo and Jing Yang who both share the 3rd largest betweenness value.

A GMap [16] was generated to facilitate the exploration of the current co-authorship graph. The resulting map is available online[1]. By the middle of January 2015, the interactive survey browser has been visited by approximately 10,300 users from 108 countries (according to Google Analytics). We have already received quite positive feedback from various researchers interested in text visualization as well as a good number of entry submissions from technique authors.

6 Conclusions and Future Work

In this paper, we presented a visual survey of text visualization techniques. The main contributions of this work are the proposed taxonomy, the interactive web-based browser that currently includes 141 techniques, and the insights on the current state of the text visualization field gained by analyzing the current survey entries. In the future, we plan to continue including new entries, refine/extend the taxonomy, create additional visualizations based on the data provided by the entries, and support automatic suggestions based on user behavior.

Acknowledgements

The authors wish to thank Hans-Jörg Schulz and Christian Tominski from the University of Rostock, Germany, for the inspiration and access to their source codes as well as their valuable comments and input. The authors also wish to thank the anonymous reviewers for their feedback. This work was partly funded by the framework grant "The Digitized Society – Past, Present, and Future" from the Swedish Research Council (Vetenskapsrådet) [grant number 2012-5659].

[1] http://gmap.cs.arizona.edu/map/3280/

REFERENCES

[1] C. Albrecht-Buehler, B. Watson, and D. Shamma. Visualizing live text streams using motion and temporal pooling. *Computer Graphics and Applications, IEEE*, 25(3):52–59, May 2005.

[2] A. B. Alencar, M. C. F. de Oliveira, and F. V. Paulovich. Seeing Beyond Reading: A Survey on Visual Text Analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):476–492, 2012.

[3] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The InfoSky visual explorer: Exploiting hierarchical structure and document similarities. *Information Visualization*, 1(3-4):166–181, 2002.

[4] D. Angus, A. Smith, and J. Wiles. Conceptual recurrence plots: Revealing patterns in human discourse. *Visualization and Computer Graphics, IEEE Transactions on*, 18(6):988–997, June 2012.

[5] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an Open Source Software for Exploring and Manipulating Networks. In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM '09)*, pages 361–362, 2009.

[6] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. FacetAtlas: Multifaceted visualization for rich text corpora. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1172–1181, Nov 2010.

[7] C. Collins, S. Carpendale, and G. Penn. DocuBurst: Visualizing document content using language structure. *Computer Graphics Forum*, 28(3):1039–1046, 2009.

[8] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST '09)*, pages 91–98, 2009.

[9] W. Cui, H. Qu, H. Zhou, W. Zhang, and S. Skiena. Watch the story unfold with TextWheel: Visualization of large-scale news streams. *ACM Trans. Intell. Syst. Technol.*, 3(2):20:1–20:17, Feb. 2012.

[10] P. DeCamp, A. Frid-Jimenez, J. Guiness, and D. Roy. Gist icons: Seeing meaning in large bodies of literature. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '05)*, 2005.

[11] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. LeadLine: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST '12)*, pages 93–102, Oct 2012.

[12] ECMA-404. JSON (JavaScript Object Notation). Available at `http://www.json.org`. Accessed: 2014-12-05.

[13] S. Eick, J. Steffen, and J. Sumner, E.E. Seesoft – a tool for visualizing line oriented software statistics. *Software Engineering, IEEE Transactions on*, 18(11):957–968, Nov 1992.

[14] E. Freeman and D. Gelernter. Lifestreams: A storage model for personal data. *SIGMOD Rec.*, 25(1):80–86, Mar. 1996.

[15] Q. Gan, M. Zhu, M. Li, T. Liang, Y. Cao, and B. Zhou. Document Visualization: An Overview of Current Research. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(1):19–36, 2014.

[16] E. Gansner, Y. Hu, and S. Kobourov. GMap: Visualizing Graphs and Clusters as Maps. In *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis '10)*, pages 201–208, 2010.

[17] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[18] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, Jan 2002.

[19] M. Krstajić, F. Mansmann, A. Stoffel, M. Atkinson, and D. A. Keim. Processing online news streams for large-scale semantic analysis. In *Proceedings of the 26th International IEEE Conference on Data Engineering Workshops (ICDEW '10)*, pages 215–220, March 2010.

[20] K. Kucher and A. Kerren. Text visualization browser: A visual survey of text visualization techniques. In *Poster Abstracts of IEEE VIS 2014*, 2014.

[21] E. Y. Li, C. H. Liao, and H. R. Yen. Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9):1515–1530, 2013.

[22] X. Lin. Visualization for the document space. In *Proceedings of the IEEE Conference on Visualization (Visualization '92)*, pages 274–281, Oct 1992.

[23] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[24] J. Nualart-Vilaplana, M. Pérez-Montoro, and M. Whitelaw. How We Draw Texts: A Review of Approaches to Text Visualization and Exploration. *El profesional de la información*, 23(3):221–235, 2014.

[25] E. Rennison. Galaxy of News: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology (UIST '94)*, pages 3–12, New York, NY, USA, 1994. ACM.

[26] G. G. Robertson and J. D. Mackinlay. The document lens. In *Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology (UIST '93)*, pages 101–108, New York, NY, USA, 1993. ACM.

[27] H.-J. Schulz. TreeVis.net: A Tree Visualization Reference. *Computer Graphics and Applications, IEEE*, 31(6):11–15, Nov 2011.

[28] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. A. Keim, and O. Deussen. Document cards: A top trumps visualization for documents. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1145–1152, Nov 2009.

[29] C. Tominski and W. Aigner. The TimeVis Browser. Available at `http://survey.timeviz.net/`. Accessed: 2014-06-18.

[30] A. Šilić and B. D. Bašić. Visualization of Text Streams: A Survey. In R. Setchi, I. Jordanov, R. Howlett, and L. Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6277 of *Lecture Notes in Computer Science*, pages 31–43. Springer Berlin Heidelberg, 2010.

[31] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. A. Keim. State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams. *Computer Graphics Forum*, 33(3), 2014.

[32] M. Wattenberg. Arc diagrams: Visualizing structure in strings. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '02)*, pages 110–116, 2002.

[33] M. Wattenberg and F. B. Viégas. The Word Tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1221–1228, Nov 2008.

[34] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '95)*, pages 51–58, Oct 1995.

[35] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. OpinionSeer: Interactive visualization of hotel customer feedback. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1109–1118, Nov 2010.