International Workshop on Visual Analytics (2012)
K. Matkovic and G. Santucci (Editors)

Visual Text Analytics using Semantic Networks and Interactive 3D Visualization

Philipp Drieger [1,2]

[1]University Eichstätt-Ingolstadt [2]noumentalia.de - digital arts

Abstract

Facing the growth of textual information, the analysis of unstructured text data remains a challenge for visual analytics. Most text visualizations are based on models that use word frequencies for text vectorization and representation. As semantics reveal from word relations, we propose an integrated visual text analytics approach that utilizes semantic networks in an interactive 3D workspace for exploration and analysis. Semantic networks act as an intermediary structure for data modeling and interactive visualization to support the visual analytics process. Focussing on the integration of text analysis and visualization, this paper describes our system design and its preliminary implementation. Discussing typical usage scenarios and a practical field-test, we present a strategy for text exploration and analysis to illustrate the usage of our system in an exemplary use case.

Categories and Subject Descriptors (according to ACM CCS): Computer Graphics [I.3.6]: Methodology and Techniques – Interaction techniques—Document and Text Processing [I.7.0]: General—Information Interfaces and presentation [H.5.2]: User interfaces—

1. Introduction

Visual Analytics propose the tight integration of *automated data analysis* and *interactive visualization* for exploration and analytical reasoning [TC05]. According to [KMS*08] a typical visual analytics process enables analysts to iteratively refine their insights by interacting with the visualization and the data analysis model. To leverage human visual abilities for knowledge building, user interfaces and interaction styles have to be optimized for a intuitive visual communication with the system [KAF*08]. Additionally, analysts must be able to manipulate the visual analytics process according to their analytical intuitions [WTP*95]. Since large amounts of data are available as unstructured text, *visual text analytics* remain an urgent challenge [RKPW08], especially regarding semantics [KMS*08]. The crucial point for visual text analytics can be identified in the integration of *text visualization* and *automated analysis of text* [RKPW08] that includes text mining techniques [FS07, BK10].

Considering the challenge of semantics in visual analytics, we propose a concept for an integrated visual text analytics system that utilizes semantic networks for extendable data modeling and 3D visualization for improved spatial exploration. The analyst directly interacts with the network by manipulation, annotation and reorganization, thereby synthesizing meta data that can be used as a feedback in the analytic process.

2. Related work

A typical problem for visual text analytics is the analysis of document collections to gain overview, discover unexpected patterns by visual exploration (e.g. SPIRE [WTP*95]) or analyze documents to support hypothesis building (e.g. Jigsaw [SGL08]). Most word frequency based approaches rely on tokenization, vectorization, dimensionality reduction, spatialization and labeling to provide representations that encode conceptual similarity through spatial proximity [RKPW08]. Conceptual entities can be identified and connected collaboratively to create concept maps (e.g. VizCept [CYM*10]). In order to model semantic contexts and concepts automatically, relation extraction techniques (e.g. AutoMap [DC04]) rely on information retrieval that can be enhanced with natural language processing [FS07]. Our approach is based on semantic networks to directly model semantic information using extracted word relations. Statisti-

DOI: 10.2312/PE/EuroVAST/EuroVA12/043-047

cal features like word frequencies are integrated as attributes in the graph. With this, we can build on methods and metrics of network analysis [Bra05, New06] that may be adopted for semantic link analysis and exploration [FS07].

According to [CCP09], text visualizations may be divided into two groups: *Synoptic visualizations* summarize document contents for overview like DocuBurst [CCP09], TextArc [Pal02] or tag clouds. Visualizations that focus on *pattern recognition* try to reveal repetitions [Wat02] or features [DZMG*07] while contexts are mainly visualized in tree structures [LPP*06]. Although each of these types offer unique approaches in text visualization, a structurally appropriate representation for text – semantic networks – is still underrepresented as a medium for visual text analytics. As networks offer obvious advantages for the visual exploration of contextually related information [Ber83], our work focusses on the exploration and analysis of semantics that reveal from semantic network visualization.

Considering semantic network representations, visualization tools for graphs and networks are also related to our approach. Although there are a lot of elaborated tools for network analysis [INS12], only few are closely related to visual analytics criteria as stated in [vLKS*10]. Integrated graph analytical approaches like GraphDice [BCD*10] show the importance of a seamless integration of automated data analysis and visualization. Most graph visualization tools require data preprocessing (e.g. Gephi [BHJ10]), what can be unhandy for analysts who want to gain fast insights. Thus, we still see demand in the research on dedicated visual text analytic systems that tightly integrate adjustable automated text analysis with interactive visualization. Our approach contributes to that goal proposing the concept of a lightweight, but performant and flexible system design.

3. System Design

Facing the challenge of visual text analytics, we built our system upon the visual analytics process described by [KMS*08] who postulates the seamless integration of automated data analysis and visual data exploration using interactive user interfaces. To achieve a highly responsive system that allows for direct interaction with data model and visualization, we identified four key objectives that can be stated as design principles for our approach on an integrated visual text analytics system and its implementation on desktop-sized workstations.

Maximize flexibility of the data structure (section 3.1) with an expandable graph-based data model to remain scalable and flexible for automated data analysis and user attributed creation and manipulation of meta data.

Minimize user interface complexity (section 3.2) by simple, visually supported operation commands for selection and manipulation in synchronized views.

Maximize visual interactivity (section 3.3 and 3.4) using a real-time responsive 3D environment to provide an immersive workspace that supports spatial orientation and preserves complex mental models of semantic networks.

Maximize parallel data processing in general to achieve fast automated data analysis and fluid visualization by implementing a hybrid CPU/GPU model to achieve real-time performance for fluid workflows.

3.1. Data model and meta data

In contrast to frequency based document vectorization we propose to model unstructured text with a generalized semantic network model [Sow91] that is created after relation extraction similar to [BMZ02, DC04]. We tokenize different words that can optionally be grouped by paragraphs and sentences, filtered by a stoplist or being merged by applying stemming or language-dependent thesauri [FS07]. Using windowing techniques, we build word relations based on k-next-neighborhood models with user-adjustable k. The resulting network is an undirected, weighted graph where attributed nodes represent words and edges stand for extracted word relations. Word frequencies adjust node-based weighting and optionally also refine edge-based weights. For faster data processing we partition the text input from one or many documents and apply network retrieval in parallel. Relevant network statistics like centrality measures are computed concurrently using hybrid CPU/GPU processing.

We choose a graph-based data model to provide flexibility and scalability for additional information that can be attributed to nodes or edges as *meta data*. In our concept meta data is created by the analyst when interacting with the semantic network. Meta data includes *implicit interacting information* like the manipulation history and *explicit synthesized information* that is created by the analyst. With this, the analyst can explicitly annotate and manipulate the semantic network by creating new nodes and edges or reorganize nodes in groups and abstraction layers. In contrast to ontology-driven systems that support automated reasoning, we adopt analyst annotations as lightweight "folksonomies" [Mik05] in the sense of "dynamic ontologies" [Sow06] that are embedded as meta data into the semantic network. As analysts mostly rely on highly domain-specific knowledge that may even be domain-variant in interdisciplinary teams, we can't assume that they come with a suitable ontology that already fits their analytic goals. Thus, our data model comprises the retrieved semantic network and arbitrary meta data that has been created by the analyst.

3.2. User Interface

Figure 1 describes the user interface which provides different data representations in synchronized views to maintain coherency of user interactions and operations.

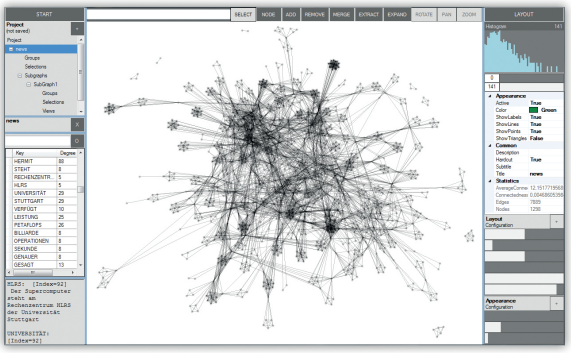

Figure 1: *Screenshot of the user interface. (Top) Menu, global search bar and operation buttons. (Left) Project tree for data organization; sortable table of nodes; source information drill-down for active selection. (Center) Interactive 3D visualization of the semantic network. (Right) Statistical view including a histogram with degree threshold filter; adjustable node properties and statistics; sliders for adjusting layout parameters and visual appearance interactively.*

3.3. 3D Visualization

The central viewing component involves a real-time responsive 3D visualization (OpenGL) of the semantic network for spatial exploration and user interaction. The network is drawn with selectable point and line primitives. We provide shading for triangles in the graph to improve spatial appearance for presentation (see figure 2). All drawing modes allow transparency adjustments and custom node coloring. The network layout is computed continuously (OpenCL) using a spring-embedding algorithm [DB99] with adjustable parameters for node padding and strength of forces. Due to the similarity of this layout algorithm with multidimensional scaling [IMO09], the layout of the weighted network is used to code semantic relatedness by spatial proximity. Meta data is differently symbol-coded to distinguish from the underlying network.

Common concerns about 3D visualization address occlusion, perspective distortion and navigation issues [Mun00, FS07]. To reduce these concerns, we optionally provide functions to diminish them. The layout can be flattened dynamically to fit into a plane. Node and label positions can be readjusted to prevent occlusion. We implemented support for a 3D navigation input device to greatly improve spatial navigation. Despite those concerns we deem the 3D visualization of semantic networks appropriate for four major reasons. 1. According to [WM08], three dimensional space offers better options for perceiving more complex graph structures. 2. Natural human perception in three dimensions allows to build and preserve a mental model [FR06]. Semantic structures can be examined like a crystal or sculpture that may act as a metaphor for spatial orientation. 3. To explore larger structures and their interrelations [WFPD97], spatial visualizations provide sufficient virtual space that can be exploited with level-of-detail (LOD) and subdivision mechanisms like k-d trees [BHJ10]. 4. In contrast to 2D layouts, 3D space offers better options to integrate explicit meta data in the model (e.g. in planar abstraction layers or spherical hulls) and to provide more advanced layout options (see section 6).

3.4. Interaction Techniques

Referring to [vLKS*10], we describe techniques for interacting with the data model and the visualization. All **visual interactions** are synchronized in all relevant views. Selected nodes are highlighted in a primary color, adjacent nodes in a secondary color while the rest of the network appears transparent to improve visual clarity [MCH*09]. Manipulation operations such as deleting, adding or grouping nodes are synchronized and affect the network layout accordingly. For example, when exploring a semantic network's robustness by temporary node removal, the layout shows the separated clusters falling apart. The exploration of larger networks is eased by using subgraphs that can be constructed from a selection and expanded according to the underlying data model. For the exploration of a graph structure it may also be helpful to steer the network layout [vLKS*10]. We introduce functions to pin nodes and move them to rearrange the network manually. Pinned nodes are excluded from layout changes, but still affect the layout of their adjacencies. Newly inserted nodes on abstraction layers are initially treated as pinned nodes.

The concept of semantic interaction can improve sensemaking by translating user interactions directly into parametric adjustments and model refinements [EFN11]. In the context of semantic networks we propose six relevant operations for **data model interaction**: *Deleting nodes* may indicate stopwords that can be assigned to the stoplist of the data model. *Merging nodes* to aggregated meta nodes can indicate synonymy or concept similarity in a semantic text analysis model. *Adding nodes* can indicate synthesized meta information like annotations. *Inserting edges* may indicate the (re)connection of concepts and allow the (re)organization of data and meta data. *Merging edges* to hyperedges can indicate relational groups that are represented as meta nodes. *Deleting edges* enables the elimination of connections by user decision. By employing those interaction techniques, analysts are able to refine the data analysis model and integrate domain-specific knowledge as explicit meta data into the semantic network. This meta data can be used again for second order analytics or to evaluate collaborative analytic processes [HA07].

4. Usage scenarios and practical testing

We developed our preliminary implementation by steadily testing it in different usage scenarios to improve performance and handling. To test our system with different

types of text data, we examined shorter texts like news or wikipedia articles which yield suitable semantic networks. Complete books or larger document collections resulted in very dense and complex graphs that are hard to study without using filtering or subgraph strategies. Usage scenarios are mainly focussed on the analysis of semantic networks to make sense about semantic contexts that are encoded in word relations. For practical testing we are actively cooperating with a group of business consulting analysts who field-test our system in social media analysis. One of their goals is the identification of relevant topics in unstructured text data whilst taking relations to other topics into account. Another goal is the analysis of semantic contexts at a given point of interest to obtain qualitatively differentiated sentiments or valuations. The resulting visualizations can easily be integrated in presentations to communicate the analyzed results. As detailed case studies are work in progress, we present a frequently used exploration strategy to illustrate the usage of our system for a typical use case.

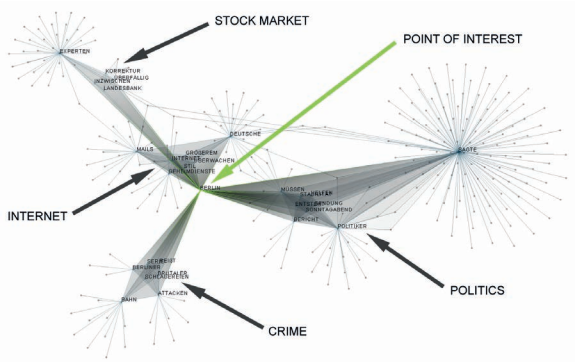

Figure 2: *Exemplary use case for the exploration of semantic contexts using a subgraph at a given point of interest (green highlighted node) by expanding adjacent nodes. The underlying semantic network has been retrieved from ten short news articles. In this example semantic contexts reveal current topics (marked by black arrows).*

5. Strategy for text exploration and analysis

As [KMS*08] suggested for the visual analytics process to analyze first (1) - show the important (2) - zoom, filter and analyze further (3) - details on demand (4), text exploration strategies may also benefit from this guideline considering link and network analysis [Bra05, FS07]. A first analysis (1) concerns the preselection of text sources that are relevant for the analytic goals. The retrieved semantic network was initially filtered by a degree threshold to show important nodes (hubs) which gave an abstract overview and can be used as a starting point for subgraph exploration (2). By gradually decreasing the degree threshold, the network can be analyzed from top down, showing finer structures (e.g. clusters) that can be zoomed and examined more closely (3). Depending on the analytic strategy, subgraphs can be extracted and analyzed further (3) to explore semantic contexts. Further details are displayed on demand (4) as a drill-down to the source text which helps to confirm or reject hypothesis. The data model can iteratively be refined by removing and merging nodes to obtain clearer networks. Figure 2 shows the result of this strategy after being applied on a collection of news articles to provide an exemplary use case.

6. Conclusions and future work

We presented the concept of an integrated visual text analytics approach that relies on the 3D visualization of semantic networks for the exploration and analysis of unstructured text data. In contrast to large-scale text analytic systems we aimed at the exploration of contextual semantic information using a fast and lightweight system. Although our implementation still can't be applied to larger document collections, on-going case studies and field-tests in social media analysis strengthened our approach. As we operate on unstructured text data, our system can handle a variety of input data that may be analyzed as a semantic network and thus remain flexible for various applications.

In the near future we will work on our implementation to improve automated data analysis and visualization staying focussed on parallel approaches like [IMO09]. Network retrieval can still be refined with more advanced text mining techniques [FS07] to obtain clearer semantic networks. For handling larger text collections we consider to store the retrieved networks in a database to query subgraphs on demand. As our visualization greatly depends on the graph layout, we are to improve layout performance [FT07] and quality by using more advanced 3D layout techniques. We are also researching on alternative layouts that include network dynamics and preserve textual linearity. For multiple document analysis we suggest to position each retrieved network in a base plane by applying document classification techniques (e.g. SOM [KKL*00] or MDS [RKPW08]). With this, networks can be arranged in a spatial landscape, so that analysts can draw links between them and annotate meta information. Annotations may be organized spherically around each network in different radii or in abstraction layers relative to the base plane of the represented document collection. The usability of spatial network layouts in a 3D workspace has to be examined in future user studies that can also provide insights in typical user tasks to further improve functionalities and handling. By continuing field-testing we will be able to elaborate more detailed case studies that also focus on methodical and domain-specific issues. The evaluation of meta data may set the stage for a development towards a collaborative 3D environment for visual text analytics.

Acknowledgements: We would like to thank Andreas Harrer for all helpful discussions and our cooperation partners from Tourismuszukunft for field-testing and feedback.

References

[BCD*10] BEZERIANOS A., CHEVALIER F., DRAGICEVIC P., ELMQVIST N., FEKETE J.-D.: Graphdice: A system for exploring multivariate social networks. In *Proceedings of Eurographics (EuroVis)* (2010). 2

[Ber83] BERTIN J.: *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, 1983. 2

[BHJ10] BASTIAN M., HEYMANN S., JACOMY M.: Using computer games techniques for improving graph visualization efficiency. In *Poster Abstracts at Eurographics / IEEE-VGTC Symposium on Visualization* (2010). 2, 3

[BK10] BERRY M. W., KOGAN J. (Eds.): *Text Mining: Applications and Theory*. Wiley, 2010. 1

[BMZ02] BATAGELJ V., MRVARY A., ZAVERŠNIK M.: Network analysis of texts. In *T. Erjavec, J. Gros (Eds.), Proc. of the 5th International Multi-Conf. Information Society - Language Technologies* (2002), pp. 143–148. 2

[Bra05] BRANDES U. (Ed.): *Network analysis: methodological foundations*. Lecture notes in computer science; 3418. Springer, Berlin, 2005. 2, 4

[CCP09] COLLINS C., CARPENDALE S., PENN G.: Docuburst: Visualizing document content using language structure. *Eurographics/ IEEE-VGTC Symposium on Visualization* (2009). 2

[CYM*10] CHUNG H., YANG S., MASSJOUNI N., ANDREWS C., KANNA R., NORTH C.: Vizcept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis. In *IEEE VAST* (2010), IEEE, pp. 107–114. 1

[DB99] DI BATTISTA G. (Ed.): *Graph drawing : algorithms for the visualization of graphs*. Prentice Hall, NJ, 1999. 3

[DC04] DIESNER J., CARLEY K. M.: *AutoMap1.2 - Extract, analyze, represent, and compare mental models from texts*. Tech. rep., Carnegie Mellon University School of Computer Science ISRI - CASOS, 2004. 1, 2

[DZMG*07] DON A., ZHELEVA E., M. GREGORY S. T., AUVIL L., CLEMENT T., SHNEIDERMAN B., PLAISANT C.: Discovering interesting usage patterns in text collections: Integrating text mining with visualization. *Proc. of the Conf. on Information and Knowledge Management* (2007). 2

[EFN11] ENDERT A., FIAUX P., NORTH C.: Unifying the sensemaking loop with semantic interaction. In *Workshop on Interactive Visual Text Analytics for Decision-Making at the IEEE VisWeek* (2011). 3

[FR06] FREIRE M., RODRÍGUEZ P.: Preserving the mental map in interactive graph interfaces. In *Proceedings of the working conference on advanced visual interfaces* (New York, 2006), ACM, pp. pp. 270–273. 3

[FS07] FELDMAN R., SANGER J.: *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007. 1, 2, 3, 4

[FT07] FRISHMAN Y., TAL A.: Multi-level graph layout on the gpu. *IEEE Transactions on Visualization and Computer Graphics* (2007). 4

[HA07] HEER J., AGRAWALA M.: Design considerations for collaborative visual analytics. In *IEEE Visual Analytics Science & Technology (VAST)* (2007), pp. 171–178. 3

[IMO09] INGRAM S., MUNZNER T., OLANO M.: Glimmer: Multilevel mds on the gpu. *IEEE Trans. Visualization and Computer Graphics* (2009), pp. 249–261. 3, 4

[INS12] INSNA: International network for social network analysis. http://www.insna.org/, 2012. 2

[KAF*08] KEIM D., ANDRIENKO G., FEKETE J.-D., GÖRG C., KOHLHAMMER J., MELANÇON G.: Visual analytics: Definition, process, and challenges. *Information Visualization: Human-Centered Issues and Perspectives* (2008), pp. 154–175. 1

[KKL*00] KOHONEN T., KASKI S., LAGUS K., SALOJARVI J., PAATERO V., SAARELA A.: Self organization of a massive document collection. *IEEE Transactions on Neural Networks 11*, 3 (2000), pp. 574–585. 4

[KMS*08] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., THOMAS J., ZIEGLER H.: Visual analytics: Scope and challenges. In *Visual Data Mining*, Simoff S. J., Böhlen M. H., Mazeika A., (Eds.). Springer, Berlin, 2008, pp. 76–90. 1, 2, 4

[LPP*06] LEE B., PARR C., PLAISANT C., BEDERSON B., VEKSLER V., GRAY W., KOTFILA C.: Treeplus: Interactive exploration of networks with enhanced tree layouts. tvcg, 2006. 2

[MCH*09] MOSCOVICH T., CHEVALIER F., HENRY N., PIETRIGA E., DANIEL FEKETE J.: Topology-aware navigation in large networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM, 2009). 3

[Mik05] MIKA P.: Ontologies are us: A unified model of social networks and semantics. In *International Semantic Web Conference* (2005), pp. 522–536. 2

[Mun00] MUNZNER T.: *Interactive visualization of large graphs and networks*. Tech. rep., Stanford University, 2000. 3

[New06] NEWMAN M. (Ed.): *The structure and dynamics of networks*. Princeton Univ. Press, 2006. 2

[Pal02] PALEY W. B.: Textarc: Showing word frequency and distribution in text. In *Proc. of the IEEE Symp. on Information Visualization* (2002). 2

[RKPW08] RISCH J., KAO A., POTEET S. R., WU Y. J.: Text visualization for visual text analytics. In *Visual Data Mining*, Simoff S. J., Böhlen M. H., Mazeika A., (Eds.). Springer, Berlin, 2008, pp. 154–171. 1, 4

[SGL08] STASKO J., GÖRG C., LIU Z.: Jigsaw: supporting investigative analysis through interactive visualization. In *Information Visualization* (2008), vol. 7. 1

[Sow91] SOWA J. F.: *Principles of Semantic Networks*. Morgan Kaufmann, 1991. 2

[Sow06] SOWA J. F.: A dynamic theory of ontology. In *Proceedings of the conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference* (Amsterdam, 2006), IOS Press, pp. 204–213. 2

[TC05] THOMAS J. J., COOK K.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics. National Visualization and Analytics.* IEEE Press, Los Alamitos, 2005. 1

[vLKS*10] VON LANDESBERGER T., KUIJPER A., SCHRECK T., KOHLHAMMER J., VAN WIJK J., FEKETE J.-D., FELLNER D.: Visual analysis of large graphs. In *Proceedings of EuroGraphics: State of the Art Report* (2010). 2, 3

[Wat02] WATTENBERG M.: Arc diagrams: Visualizing structure in strings. In *IEEE Symposium on Information Visualization* (2002). 2

[WFPD97] WARE C., FRANCK G., PARKHI M., DUDLEY T.: Layout for visualizing large software structures in 3d. In *Proceedings of VISUAL '97* (1997), Springer, pp. 215–223. 3

[WM08] WARE C., MITCHELL P.: Visualizing graphs in three dimensions. *ACM Trans. Appl. Percept. 5*, 1 (2008), 2:1–2:15. 3

[WTP*95] WISE J. A., THOMAS J. J., PENNOCK K., LANTRIP D., POTTIER M., SCHUR A., CROW V.: Visualizing the nonvisual: Spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Information Visualization Symposium InfoViz* (1995). 1