SimBaTex: Similarity-based Text Exploration

D. Witschard[1], I. Jusufi[1], and A. Kerren[1,2]

[1]Department of Computer Science and Media Technology, Linnaeus University, Sweden
[2]Department of Science and Technology, Linköping University, Sweden

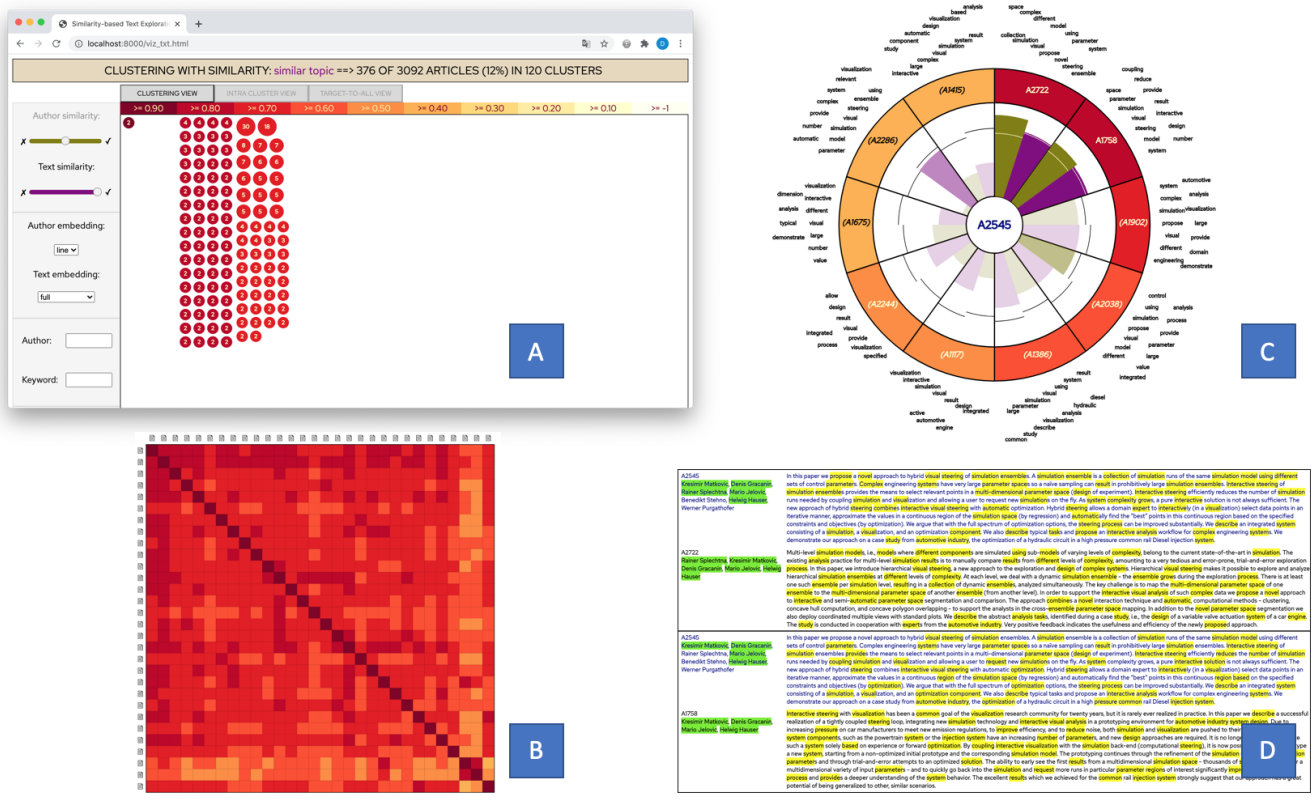

Figure 1: *An overview of the SimBaTex tool. In the Clustering View [A], the clustering result of the current similarity criteria is displayed. In the Intra Cluster View [B], the pairwise similarity scores can be assessed. The Target-to-all View [C and D] shows the Top 10 matches of a selected article for detailed comparison. The aim of the visualization is to let the user reveal latent similarity patterns of a set of documents and use them as a basis for efficient interactive search and exploration.*

Abstract

Natural language processing in combination with visualization can provide efficient ways to discover latent patterns of similarity which can be useful for exploring large sets of text documents. In this poster abstract, we describe the ongoing work on a visual analytics application, called SimBaTex, which is based on embedding technology, dynamic specification of similarity criteria, and a novel approach for similarity-based clustering. The goal of SimBaTex is to provide search-and-explore functionality to enable the user to identify items of interest in a large set of text documents by interactive assessment of both high-level similarity patterns and pairwise similarity of chosen texts.

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics;*** • ***Information systems*** → ***Content analysis and feature selection;***

DOI: 10.2312/evp.20211067

1. Introduction

Research results are often published in the form of text reports, and for many fields the publication rate makes it challenging for a single practitioner to get an overview of the vast source of textual information. Since close reading is time consuming, and time typically is a limiting factor, there is a high demand for supporting the navigation through large document sets and identifying information relevant for the task at hand [JFCS15]. Natural language processing (NLP) in combination with visualization can provide efficient ways to handle such challenges [BG19, KK15, FHKM17, LTW*18, JKLZ14, KKLS17], and a similarity-based approach [GF13] can be useful to answer key questions like "Are there any groupings of similar documents within the set?" or "Are there any documents within the set that are similar to this specific document?".

In this poster abstract, we describe the ongoing work on a visual analytics application, called SimBaTex, which aims to combine NLP approaches, similarity calculations and visualization methods to provide the user with efficient and interactive search-and-explore functionalities for large sets of scientific text documents. Furthermore, the tool implements a dynamic specification of the similarity criteria (i.e., different aspects can be included, excluded or even negated depending on the current level of interest), so that different types of similarity criteria can be interactively applied in order to discover latent patterns of similarity, which in turn could be the basis for further exploration. The data set that we use for our application is the IEEE VIS data set which contains information of articles published at the IEEE VIS conferences [IHK*17]. From this data, we have extracted about 3,000 articles published during the period 1990-2018. In our software implementation, we use the text from the article abstracts and also the associated co-author network.

2. Architecture

The basis of our application is cosine similarity calculations on embedding vectors obtained from the underlying data. We use similarity score thresholds to classify article pairs as similar or dissimilar (i.e., pairs with scores above or equal to the threshold are classified as similar; and pairs with scores below the threshold are classified as dissimilar). The article abstracts are embedded with the Universal Sentence Encoder (USE) [CYK*18, ML10] and the author information is embedded based on node-embedding of the co-author network using Node2Vec and LINE [GL16, TQW*15, GF18, CWPZ19]. For each article we calculate 5 different embeddings of the abstract text and 2 different embeddings of the co-author data. The embeddings and the corresponding pairwise similarity scores are asynchronously calculated using a Python/Tensorflow [ABC*16] backend which produces precalculated score files, one for each embedding type. The frontend visualization is implemented in HTML/D3 [D311] and loads the score files upon start which allows for high responsiveness since only minor calculations are performed within the browser.

3. Visualization Approach

The design seeks to reuse already well-proven visual metaphors (such as circles for clusters, heatmaps for value display and word-highlighting for text similarity) and it also provides a custom design for the target-to-all comparisons (see Figure 1 [C]). When the visualization is loaded the articles are represented as unclustered article icons in the *Clustering View*. The user may specify the desired criteria to use, and may also specify which of the 5 different text embeddings and which of the 2 different author embeddings to use. Each change of the settings results in an animated sequence where each article is clustered together with the article that it is most similar to (if any). If the similarity is symmetrical (i.e., article X is most similar to article Y, and Y is most similar to X) this will result in a cluster with only 2 items, and if the similarity is asymmetric (i.e, X is most similar to Y, but Y is most similar to Z) the cluster can contain any number of items. Clusters are represented as circles where size encodes the number of articles in each cluster, and spatial position and color encode both the average pairwise similarity score within the cluster, as shown in Figure 1 [A].

Clicking a cluster displays the *Intra Cluster View* where the pairwise similarity scores are displayed in a heatmap matrix (see Figure 1 [B]). This view allows for analysis of the overall pairwise similarity to determine the level of homogeneity in the cluster. High homogeneity would suggest that all articles are treating a similar topic, while low homogeneity would suggest that there are several different topics (or angles of related topics) that have been clustered together due to asymmetric similarity.

Clicking an article icon in the *Intra Cluster View* displays the *Target-to-all View* where an overview of the similarity details of the Top 10 best matches (for the selected article) are displayed as a radial bar chart diagram annotated with a word cloud of matching words (see Figure 1 [C]). The layout aims to provide an efficient at-a-glance overview and the best match is displayed at the section starting at "12 o'clock", and then the scores fall off in clockwise order. Furthermore at the bottom, the text pairs are displayed for close reading in a view that uses yellow color to highlight co-occurrences of words to facilitate assessment of the similarity (see Figure 1 [D]). This view supports the understanding of how similar the selected target article is to any of the other articles in the data set.

To support advanced search, the user may track articles on publication author and/or keyword in order to see if the selected articles will cluster together. Tracked articles will be highlighted in green-colored frames throughout the views. In the *Clustering View*, green color indicates the fraction of tracked articles within the cluster (the darker the green the higher the fraction). This feature is helpful for answering questions such as (1) if a specific author tend to write articles on similar subjects, or (2) if articles that mention certain keywords also show high overall similarity.

4. Conclusion and Future Work

In this poster abstract, we have briefly introduced the ongoing work on SimBaTex, a visual analytics application which enables the user to identify items of interest in a large set of text documents. Our intention is to extend our work to the domain of multivariate networks and provide the possibility for a wider range of features to be included in the similarity specification [WJMK21]. We also intend to do a formal evaluation with a user study. Another possible path for future work is to explore the possibilities to use similarity-based clustering as a base for topic extraction.

References

[ABC*16] ABADI M., BARHAM P., CHEN J., CHEN Z., DAVIS A., DEAN J., DEVIN M., GHEMAWAT S., IRVING G., ISARD M., KUDLUR M., LEVENBERG J., MONGA R., MOORE S., MURRAY D. G., STEINER B., TUCKER P., VASUDEVAN V., WARDEN P., WICKE M., YU Y., ZHENG X.: Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (2016), pp. 265–283. URL: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf. 2

[BG19] BELINKOV Y., GLASS J.: Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics 7* (Mar. 2019), 49–72. URL: https://www.aclweb.org/anthology/Q19-1004, doi:10.1162/tacl_a_00254. 2

[CWPZ19] CUI P., WANG X., PEI J., ZHU W.: A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering 31*, 5 (May 2019), 833–852. doi:10.1109/TKDE.2018.2849727. 2

[CYK*18] CER D., YANG Y., KONG S., HUA N., LIMTIACO N., JOHN R. S., CONSTANT N., GUAJARDO-CESPEDES M., YUAN S., TAR C., SUNG Y., STROPE B., KURZWEIL R.: Universal Sentence Encoder. *CoRR abs/1803.11175* (2018). arXiv:1803.11175. 2

[D311] D3 — Data-driven documents. https://d3js.org/, 2011. Accessed March 29, 2021. 2

[FHKM17] FEDERICO P., HEIMERL F., KOCH S., MIKSCH S.: A survey on visual approaches for analyzing scientific literature and patents. *IEEE Transactions on Visualization and Computer Graphics 23*, 9 (Sept. 2017), 2179–2198. doi:10.1109/TVCG.2016.2610422. 2

[GF13] GOMAA W., FAHMY A.: A survey of text similarity approaches. *international journal of Computer Applications 68* (04 2013). doi:10.5120/11638-7118. 2

[GF18] GOYAL P., FERRARA E.: Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems 151* (July 2018), 78–94. doi:10.1016/j.knosys.2018.03.022. 2

[GL16] GROVER A., LESKOVEC J.: Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, Association for Computing Machinery, p. 855–864. URL: https://doi.org/10.1145/2939672.2939754, doi:10.1145/2939672.2939754. 2

[IHK*17] ISENBERG P., HEIMERL F., KOCH S., ISENBERG T., XU P., STOLPER C. D., SEDLMAIR M., CHEN J., MÖLLER T., STASKO J.: Vispubdata.org: A metadata collection about IEEE Visualization (VIS) publications. *IEEE Transactions on Visualization and Computer Graphics 23*, 9 (Sept. 2017), 2199–2206. doi:10.1109/TVCG.2016.2615308. 2

[JFCS15] JÄNICKE S., FRANZINI G., CHEEMA M. F., SCHEUERMANN G.: On close and distant reading in digital humanities: A survey and future challenges. In *Proceedings of the EG/VGTC Conference on Visualization — STARs* (2015), EuroVis '15, The Eurographics Association. doi:10.2312/eurovisstar.20151113. 2

[JKLZ14] JUSUFI I., KERREN A., LIU J., ZIMMER B.: Visual exploration of relationships between document clusters. In *2014 International Conference on Information Visualization Theory and Applications (IVAPP)* (2014), pp. 195–203. 2

[KK15] KUCHER K., KERREN A.: Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proceedings of the IEEE Pacific Visualization Symposium* (2015), PacificVis '15, IEEE, pp. 117–121. doi:10.1109/PACIFICVIS.2015.7156366. 2

[KKLS17] KERREN A., KUCHER K., LI Y.-F., SCHREIBER F.: Biovis explorer: A visual guide for biological data visualization techniques. *PLOS ONE 12*, 11 (11 2017), 1–14. URL: https://doi.org/10.1371/journal.pone.0187341, doi:10.1371/journal.pone.0187341. 2

[LTW*18] LIU J., TANG T., WANG W., XU B., KONG X., XIA F.: A survey of scholarly data visualization. *IEEE Access 6* (Mar. 2018), 19205–19221. doi:10.1109/ACCESS.2018.2815030. 2

[ML10] MITCHELL J., LAPATA M.: Composition in distributional models of semantics. *Cognitive Science 34*, 8 (Nov. 2010), 1388–1429. doi:https://doi.org/10.1111/j.1551-6709.2010.01106.x. 2

[TQW*15] TANG J., QU M., WANG M., ZHANG M., YAN J., MEI Q.: Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web* (Republic and Canton of Geneva, CHE, 2015), WWW '15, International World Wide Web Conferences Steering Committee, p. 1067–1077. URL: https://doi.org/10.1145/2736277.2741093, doi:10.1145/2736277.2741093. 2

[WJMK21] WITSCHARD D., JUSUFI I., MARTINS R. M., KERREN A.: A statement report on the use of multiple embeddings for visual analytics of multivariate networks. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP '21) — Volume 3: IVAPP* (2021), IVAPP '21, INSTICC, SciTePress, pp. 219–223. doi:10.5220/0010314602190223. 2