A Directed Concept Search Environment to Visually Explore Texts Related to User-defined Concept Models

Muhammad Faisal Cheema[1], Stefan Jänicke[1], Judith Blumenstein[2] and Gerik Scheuermann[1]
[1]*Image and Signal Processing Group, Institute for Computer Science, Leipzig University, Leipzig, Germany*
[2]*Faculty of History, Arts and Oriental Studies, Leipzig University, Leipzig, Germany*

Keywords: Text Visualization, Digital Humanities, Distant Reading, Concept Search, Information Retrieval, Concept Editor, Concept Modeling.

Abstract: We introduce a concept search environment that caters for the needs of humanities scholars who want to improve the accuracy of search results when querying historical text corpora. For this purpose, we designed a so-called *Concept Editor* that allows to model historical concepts in a diagram style according to the imaginations of the humanities scholar. For the inspection of results determined in the proposed concept search, we provide a *Concept Search Results Viewer* that uses the existent layout of the underlying concept model to visualize related texts according to the relevance to the given concept. We further designed the overall system the way that the humanities scholar can iteratively refine the concept idea, which leads to a gradual improvement of search results. To illustrate the whole development pipeline, we provide a usage scenario on modeling the concept *epilepsy* with the purpose of improving the accuracy of results compared to usual applied keyword-based search methods.

1 INTRODUCTION

The retrieval of information utilizing search engines such as Google is a daily activity nowadays. When applying the usual keyword-based search, the results are often unsatisfactory as they contain too less precise hits and numerous irrelevant results. The same issue occurs for humanities scholars applying keyword searches on historical text corpora accessible via platforms like Perseus Digital Library[1] or PHI Latin Texts[2] to discover texts related to a specific topic. Then, the scholar needs to reformulate the query hoping to improve the quality and quantity of search results.

Taking the search for text passages related to the concept *epilepsy* as an example, a traditional keyword search for the corresponding Latin term *morbus comitialis* (disease of the assembly) yields incomplete results. In contrast, a truncated search for *morb* comiti** returns more but less accurate results. Also, false positives in the form of entirely unrelated text passages occur, e.g., a passage from a political text written by the Roman historian Titus Livius (*Ab urbe condita libri* 41,18,16) containing both the words *morbus* and *comitia*. Also problematic with basic keyword-based search attempts is the impossibility to receive texts paraphrasing the given concept. In the *epilepsy* scenario, related texts exist that do not use the term *morbus comitialis* explicitly. An example is the didactic poem *De rerum natura* written by the Roman poet and philosopher Lucretius, where the disease is paraphrased with occurring symptoms such as *concidere* (collapse) and *spuma* (foam). As many other terms were used to describe *epilepsy* in Latin texts, e.g., *morbus sacer* or *epilepsia*, keyword-based approaches were inadequate for the humanities scholars to discover unknown but related results.

This paper presents a system designed to improve the search capabilities on historical text corpora. Based upon a semi-automated process, we aim to extract more accurate results by extending the traditional keyword search with a so-called concept search. Exemplary, we will present all important steps from modeling the concept *epilepsy*, performing a concept search to discover texts related to *epilepsy*, analyzing the search results, and modifying the concept iteratively. In summary, the contributions of our work to the visualization and Digital Humanities communities are:

[1]http://www.perseus.tufts.edu/
[2]http://latin.packhum.org/

Cheema, M., Jänicke, S., Blumenstein, J. and Scheuermann, G.
A Directed Concept Search Environment to Visually Explore Texts Related to User-defined Concept Models.
DOI: 10.5220/0005727400720083
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 2: IVAPP*, pages 74-85
ISBN: 978-989-758-175-5

- A **Concept Editor** that borrows ideas from concept and mind mapping to allow humanities scholars to model concepts according to their imaginations.
- A **Concept Search** algorithm that retrieves all texts related to the modeled concept and clusters them in the concept model according to their relevance to the given concept.
- A **Concept Search Results Viewer** that intuitively visualizes the results of the concept search by interactively providing statistics of the occurring terms with the help of TagPies.
- A user-driven **Concept Search Environment** that gives the humanities scholar full control when creating concepts, analyzing search results, and the opportunity to gradually improve search results by iteratively modifying the underlying concept model.

Additionally, we provide insights about the collaborative, iterative development of the concept search environment, a usage scenario that illustrates the potential of the presented system, and possible future tasks.

2 RELATED WORK

The prior goal of our work is the development of a Concept Search environment that improves the humanities scholar's search capabilities on ancient text corpora while controlling the entire Concept Search process. To solve this task, we draw upon existing techniques in the fields of diagramming, natural language processing and visualization. Many diagramming tools have been developed to cater various needs, and similarly, there are several ways to find text passages related to a given topic and to visualize the search results.

2.1 Existing Concept Search Methods

Besides traditional keyword-based search methods to collect related text passages in large corpora, other techniques have evolved in past years that take advantage of the semantic and statistical information inherent in the texts with the goal to improve precision and recall of the results.

Concept search (Giunchiglia et al., 2009) is such a method that automatically retrieves data which is conceptually similar to a user-specific search query. By employing techniques like word sense disambiguation (Navigli, 2009), the result sets compared to a traditional keyword search can be improved by taking not only the spelling of a word but also its meaning into account during the search process. Similar approaches take advantage of different semantic relations within the text, e.g., synonyms and word variations (Guha et al., 2003), or search processes can be based upon an underlying ontology (Fernández et al., 2011). In contrast to considering semantic relations, topic modeling approaches apply statistical models to measure the conceptual similarity of texts (Wallach, 2006). The Latent Dirichlet allocation (LDA) is the most often used topic model (Blei et al., 2003). Within the computation process, each text is seen as a mixture of latent topics. Based on a predefined number of final topics, texts cluster iteratively according to the similarity of contained words.

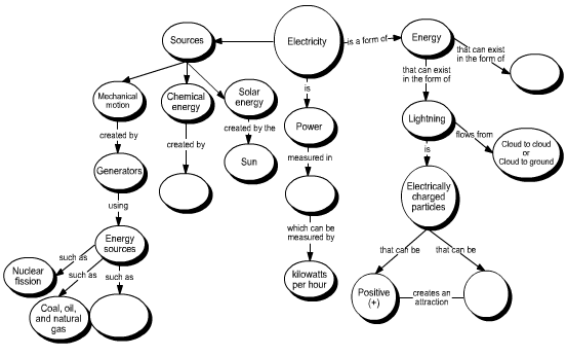


Figure 1: Example concept map.

All the above mentioned techniques suffer from one major disadvantage: the user has a very limited control over the detection of conceptual similarities, which are based on a prior semantic or statistical model. Only scholars having advanced knowledge of information retrieval can tweak the parameters to obtain better results. In contrast to these automated processes that use existing semantic knowledge from thesauri, topic maps or semantic networks, our concept modeling approach allows the humanities scholar to express her understanding of a concept without the need of any existing semantic or statistical knowledge. Teevan (Teevan et al., 2004) argues that even with a perfect search engine, a poorly constructed search question may not lead to the right answer. So, the user needs to be provided with a directed search system. As semantic and statistical models would take the control away from the humanities scholar, we designed a directed search environment. This capability of retaining the control over the modeling and search process was an important requirement in our project as it supports the iterative refinement of models to steadily retrieve better results.

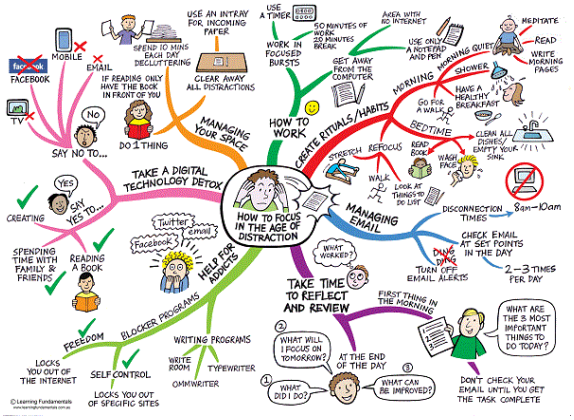


Figure 2: Example mind map (Kanter, 2015) (Figure under CC BY 2.0 license, see https://creativecommons.org/licenses/by/2.0/ for details).

2.2 Diagram Types and Diagramming Tools

According to Ware, half of the brain is trained in interpreting graphical patterns (Ware, 2010). Thus, visual thinking tools such as diagrams utilize this capability in helping humans to comprehend the visualized facts faster.

There are several ways to transform an idea or a concept into a diagram style. One such method is concept maps (Hager et al., 1997) that illustrate the relationships between multiple concepts. Shapes are used to model ideas, images, or words as concepts, and arrows among these shapes communicate relations; an example is given in Figure 1. In software engineering, the idea of concept mapping is used to visualize the design of a system in form of a diagram with the Unified Modeling Language (UML) (Booch et al., 2005).

A similar method to visually organize information are so called mind maps (Budd, 2004). In contrast to concept maps, a mind map represents thoughts and ideas about a central concept (see Figure 2). The concepts generated by the collaborating humanities scholars of our projects borrow from both diagram types, as they focus a central concept that is sometimes connected to correlated sub-concepts.

Several freely or proprietary diagramming tools such as Inkscape (Bah, 2009), MSVisio[3] or yEd[4] enable scholars to create diagrams. Most of these tools are used to only represent and organize information, which is insufficient for our needs. The Concept Editor proposed in this paper enables the scholar not only to model particular concepts, but also to manually attach semantic information for further processing.

2.3 Concept Search Visualizations in the Digital Humanities

Many works with a Digital Humanities motivation attend to the matter of visualizing the results of Concept Search processes or comparable methods.

The VarifocalReader, which facilitates the work with individual, potentially large historical texts, applies a topic segmentation to cluster text parts according to the user's requirements, e.g., to extract and highlight sections dealing with certain topics (Koch et al., 2014).

Topic modeling is an often used basis for the analysis of topic changes within news corpora (Cui et al., 2011; Dou et al., 2012). After automatic topic extraction, the temporal evolution of the found topics can be observed in stream graphs. To minimize the computational effort, the user can be integrated into this process by picking topics of interest (Cui et al., 2014). A rather abstract method to visualize topic changes in news corpora is the dust-and-magnet visualization (Yi et al., 2005). Applied by Eisenstein, topics exert magnetic forces on various historical newspapers that generate temporal trajectories in the form of "dust trails" to illustrate topic change (Eisenstein et al., 2014).

Other approaches apply automated topic modeling techniques to cluster texts in large text collections. The tool Serendip uses probabilistic topic models to support the semi-automated exploration of such corpora (Alexander et al., 2014). The words belonging to specific topics can be analyzed both in close and distant reading views. That juxtaposed tag clouds are beneficial to explore automatically extracted topics from various text corpora has been shown at the recent Digital Humanities conference (Jähnichen et al., 2015; Montague et al., 2015). Also, the utilization of graphs to communicate the relationships among discovered topics has been proven useful (Kaufman, 2015).

3 DIGITAL HUMANITIES BACKGROUND

As a consequence of many digitization projects, humanities scholars nowadays have access to large historical text corpora. Within the digital humanities project *eXChange*[5] computer scientists and

[3] http://www.visiotoolbox.com/

[4] http://www.yworks.com/en/products/yfiles/yed/

[5] http://www.exchange-projekt.de/

humanities scholars collaboratively develop strategies to query such text corpora. The *eXChange* database contains numerous ancient Greek and Latin texts from various sources (e.g., Perseus Digital Library (Per, 2015) and Bibliotheca Teubneriana Latina (Btl, 2015)).

The usual approach of humanities scholars querying a digital corpus is a keyword based search. Quite often, they get lots of results, which are hard to revise, so that the generation of valuable hypotheses is nearly impossible. Moreover, a simple keyword search causes incomplete result sets with partially inaccurate search results as it does not include word forms or synonyms. Especially when observing ancient terms that were used in different domains – politics and medicine – the humanities scholars require a mechanism that filters for results relevant for the research question at hand. Therefore, the prior goal of the project was to develop a system capable of extracting more accurate results by extending the traditional keyword search with a so called concept search.

Following the suggestions from other visualization researchers who worked together with humanities scholars (Jänicke et al., 2015b), we closely collaborated within the *eXChange* project when designing the concept search system. Initially, we discussed the needs of humanities scholars of our project, their workflows, research questions and challenges in several meetings. As a *concept* is a broad term used with different meanings in a variety of fields (Margolis and Laurence, 2014), one of the first steps was to define the *concept* term with respect to our context. The concepts generated by the humanities scholars in our project borrow ideas from mind maps as well as concept maps, as they focus on a central idea that is sometimes related to other (sub-)concepts. Overall, a concept is defined through a bunch of terms and their word forms, which distribute over (sub-)concepts. A prior concern of the scholars was to have a fully controllable search process. To meet this need, we designed the visualizations in a way that each data transformation step is comprehensible. A Concept Editor was proposed, where the scholars can model their ideas of concepts. Based on these user-defined concept models, we developed a concept search and designed a Concept Search Results Viewer. The development of the whole concept search system is explained in the following section.

4 CONCEPT SEARCH PIPELINE

For the sake of simplicity and better understanding, a concept is modeled in form of a tree. Related sub-

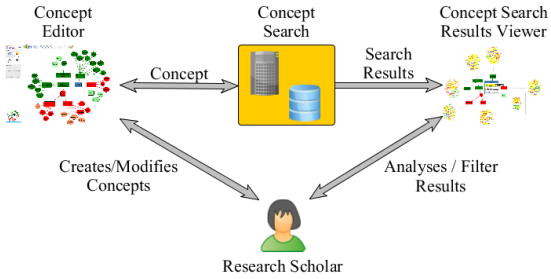


Figure 3: Concept Search Pipeline.

concepts and specific terms and word forms are combined in a hierarchical fashion to the central concept, which is represented with the root node of the tree. Internal nodes represent sub-concepts, and leaf nodes are words defining the associated (sub-)concepts. An example model for the concept *epilepsy* is given in Figure 4.

As illustrated in Figure 3, the approach primarily consists of three steps. At first, the scholar generates a concept model using Concept Editor. The resultant concept model is then passed on to the proposed Concept Search that ranks the search results, which are retrieved related texts, and places them into the concept model. Finally, the search results are arranged in the Concept Search Results Viewer, which aggregates search results and serves details on demand by interactively visualizing statistics about occurring terms related to the underlying concept model in tag clouds.

4.1 Concept Editor

The Concept Editor is an interface for humanities scholars to encode their ideas of concepts on the screen. Figure 4 shows a screenshot of the Concept Editor containing the concept *epilepsy*. Rectangular nodes represent (sub-)concepts while oval nodes denote concrete terms. Nodes can be drawn via drag and drop and the scholar can structure the concept according to her imagination by connecting related nodes. The node colors indicate whether a sub-concept and its associated terms are supportive (in green) or contradictory (in red) to the central concept.

In the example concept model for *epilepsy*, the root node *epilepsy* is the central concept in blue color. The green colored shapes indicate supportive sub-concepts and terms (e.g., *labels for epilepsy*), whereas red shapes indicate contradictions (e.g., *political terms* unrelated to epilepsy). Darker shades (e.g., *morbus*, *febrire* etc.) represent definite knowledge of the scholars, whereas light shades (e.g., *vinium*, *cotidiana* etc.) represent assumptions. According to the model, texts containing the terms

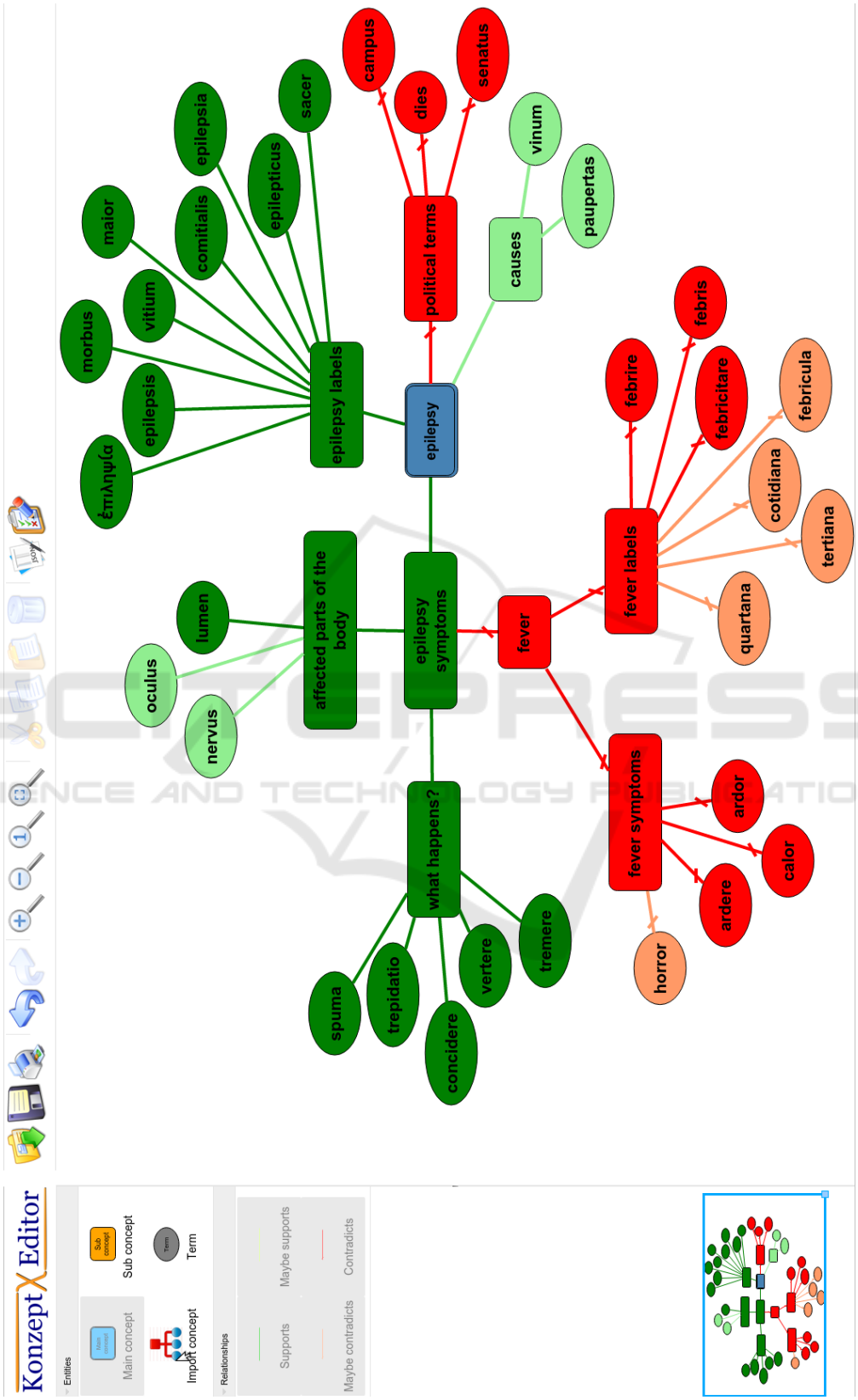


Figure 4: Concept Editor interface showing the concept *epilepsy*.

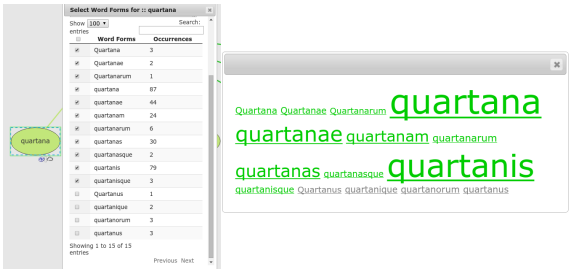


Figure 5: Word form selection interface showing the term *quartana*.

morbus or *comitialis* were most likely used to address the concept epilepsy. Terms like *horror* (dread) and *quartana* (fever) show contradiction.

Each term is connected to all possible spellings and word-forms contained in the database. A popup on demand provides a list and a tag cloud that allow the scholar to observe and select word forms inherent in the corpus, which are potentially relevant for the corresponding concept. Figure 5 shows the selected word forms for the term *quartana*. As fever is a female term in Latin, the male form of the adjective *quartanus* is excluded in that scenario. After the concept is built, it is stored for persistence and forwarded to a Concept Search module.

4.2 Concept Search

The Concept Search method is a two-stage process. First, a hierarchical keyword search is performed on the basis of the observed concept model. Second, each search result is assigned to the topically best fitting node in the concept tree.

4.2.1 Hierarchical Keyword Search

The selected word-forms of all terms are the basis for the search. The procedure of the keyword hierarchy search was intentionally kept as simple as possible to minimize pre-assumptions on how a fitting concept search would function.

Beginning with a list of all words that have been added to the concept, each entry is searched for with a classical string lookup on a normalized version of the texts. Per word, all occurrences are located, grouped with respect to the containing text and then counted within the groups. Those individually found texts are collected across all words. This forms a non-negative (and usually quite sparse) integer matrix where every row-sum and every column-sum has a value of at least 1. This representation can then be used to rank the results according to the concept hierarchy. Each occurring “positive term” (green and light green) increases the rank of a text, whereas each occurring “negative term” (red and light red) decreases it.

4.2.2 Assigning the Results to Concept-Nodes

According to their relationship to the underlying concept, the results – texts within the database – are positioned in the concept hierarchy. Instead of showing only a list of search results, we put all results in the concept hierarchy the scholar is familiar with. This helps to get a notion of a text’s relevance to the concept.

The assignment of a search result to a concept node is performed as follows. Initially, the search result is assigned to each positive leaf concept node (green colored) for which the corresponding text contains one of the concept’s “positive terms.” We do not assign a search result to negative leaf concepts (red colored) as they are unimportant matches for the concept. Then, the concept tree is traversed bottom-up in depth first search fashion. A search result is moved up the tree and placed in the parent concept node if it meets two criteria: (1) all positive child concept nodes contain the search result, and (2) the ratio of positive to negative words is greater than a configurable threshold. A result fulfilling both criteria in all traversal steps will be moved to the root concept node. If this is not the case, a search result may be placed in multiple concept nodes. We remove duplicate entries and keep the best fitting position for a search result, which is the concept node with the closest distance to the root node. If there are multiple candidates with the same distance, we keep the position where the search result obtains the highest ratio of positive to negative terms. So, each search result is assigned to its best fitting concept node. The higher a search result is placed in the hierarchy, the closer it is considered to the concept.

4.3 Concept Search Results Viewer

The Concept Search Results Viewer supports the analysis of the concept-related retrieved texts in two views. The first view shows the aggregated results in the concept hierarchy. In the second view, the relationships between individual texts and word forms can be interactively explored. To preserve the mental map, both views are derived from the user-defined concept layout and need to be recalculated every time the user changes the underlying concept.

4.3.1 Layout Calculation

After the search results are assigned to the concept nodes (see Subsection 4.2.2), the node sizes increase

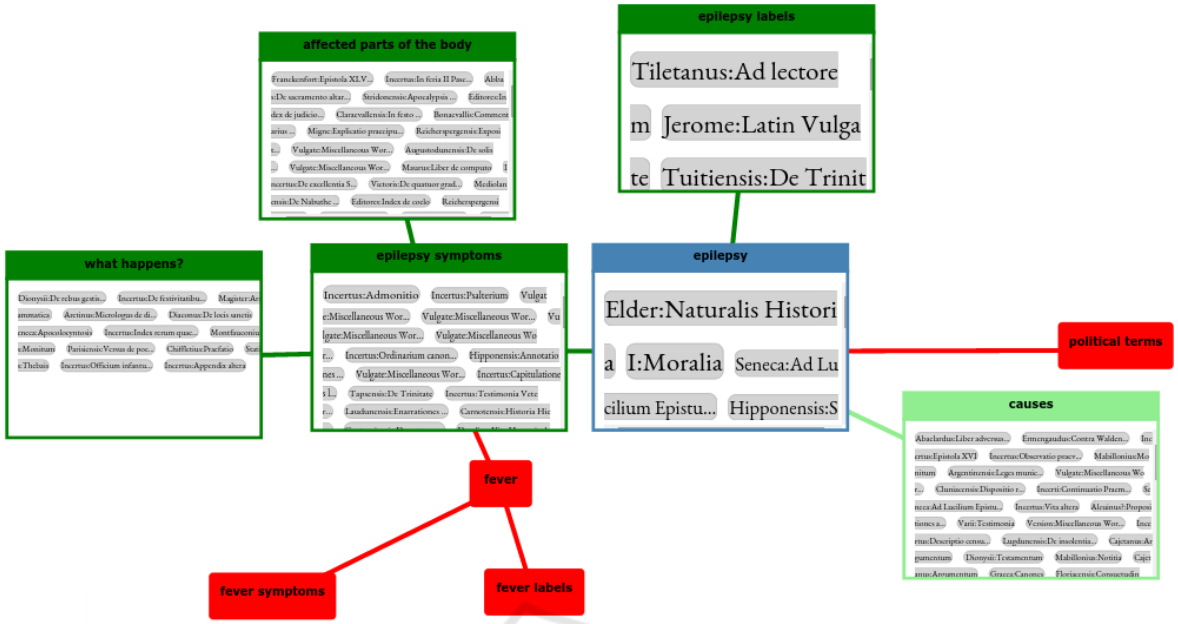


Figure 6: Concept Search Results Viewer showing the aggregate results view of the concept *epilepsy*.

with respect to the number of labels that need to be shown. To keep the results layout consistent to the user-defined concept layout, all node sizes are equal and remain fixed (large result sets can be inspected via scrolling). The increase of node sizes creates overlaps between nodes. To remove these overlaps the layout needs to be refined while preserving the original topology of the concept tree.

The overlap problem is solved heuristically. Initially, the original coordinate positions are assigned according to the concept tree. Then, the concept tree is traversed bottom up. For each node, we check if an overlap exists with its parent node. If an overlap exists, the length of the edge between parent and child node is increased by moving the child node till the overlap resolves. This step is first done for all leaf concept nodes to remove overlaps with the corresponding parent nodes. When applying this heuristic to internal nodes, we take the bounding box of the whole sub-tree into account. This ensures that no overlap exists between a node's sibling and it's children. After all the nodes are traversed, we obtain a layout that preserves the topology of user-defined concept tree layout.

4.3.2 Visualizing Aggregate Result Summaries

In this view, all retrieved texts are shown in their corresponding node containers with a label – author and title of the text – in a font size reflecting the determined rank during the concept search. The results assigned to a node are ordered by decreasing rank, so that the most relevant text labels are shown at the top of lists. The aggregate result summaries view for the *epilepsy* scenario is shown in Figure 6.

As this is the first view on the results of the concept search, the humanities scholar can get an idea how the texts are placed in the concept hierarchy and which texts seem to be most relevant to the concept.

4.3.3 Visualizing Word Summaries of Individual Texts

To support the exploration process of the concept search results, we provide a detailed view showing the relationships between a partial result set (texts of an individual node) and the occurring word forms in these texts. To save screen space, we collapse all other nodes except the selected one. This helps users to concentrate on the task at hand.

The word forms occurring in the associated texts are shown as TagPies (Jänicke et al., 2015a), which provide an intuitive, comparative view of the term distribution of a concept node. We selected TagPies as the humanities scholars are used to work with this visualization within the *eXChange* project, and it turned out to be valuable for their workflows (Jänicke et al., 2015a). Figure 8 shows a TagPie of word forms of the terms *spuma, trepidatio, concidere, vertere* and *tremere*, which are assigned to the *what happens?* sub-concept in *epilepsy symptoms*.

When the user hovers a text label in the node con-

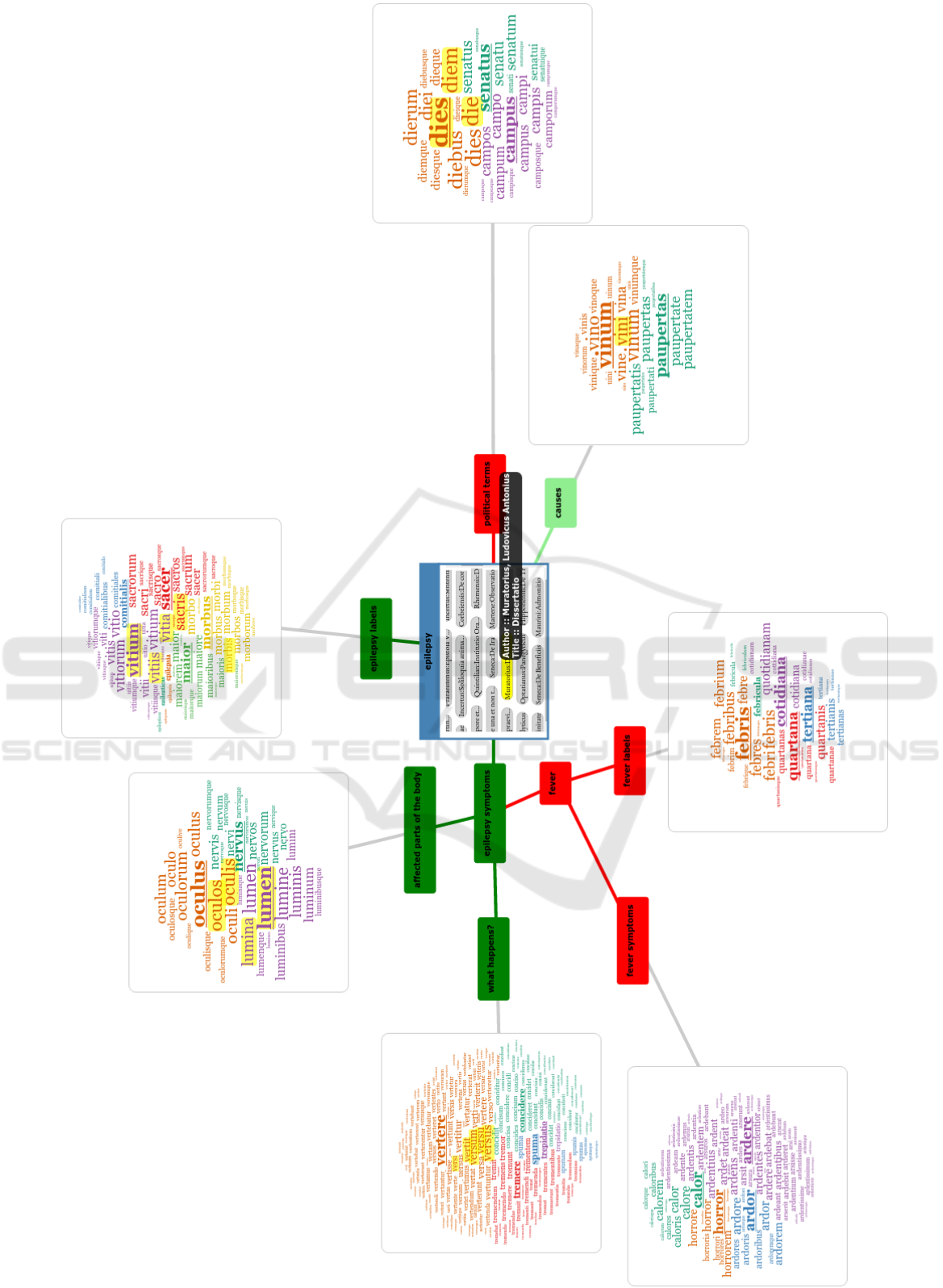


Figure 7: Concept Search Results Viewer showing the word forms occurring in texts belonging to the selected node. The terms of a selected text are highlighted.

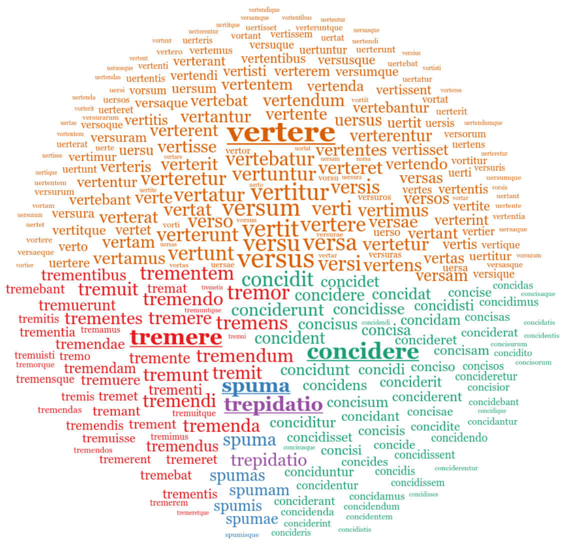


Figure 8: TagPies belonging to the subconcept *what happens?* that is related to *epilepsy symptoms*.

tainer, all associated word forms are highlighted to give the user an idea of the topical context of the selected text in accordance to the given concept (as shown in Figure 7).

5 DISCUSSION

When developing the concept search environment, we closely collaborated with the humanities scholars – historians and classical philologists – involved in the digital humanities project *eXChange*. This section provides a typical usage scenario and takes a look at the iterative development and evaluation of the system.

5.1 Usage Scenario

The leading example of this paper – working with the ancient concept *epilepsy* – was an object of interest for one of the collaborating classical philologists. She modeled the concept using the provided editor as shown in Figure 4. Looking at the results of the concept search (see Figure 6), she observed the discovered texts assigned to the various nodes. Clicking the main concept node for *epilepsy* reveals the related terms by displaying the found word forms in the corresponding TagPies. In that scenario, she was scrolling through the results and used the highlighting functionality to get a quick overview of the terms contained in various historical texts (see Figure 7). Thereby, she discovered that the frequent occurrence of the term *vitium* (defect) is related to a large num-

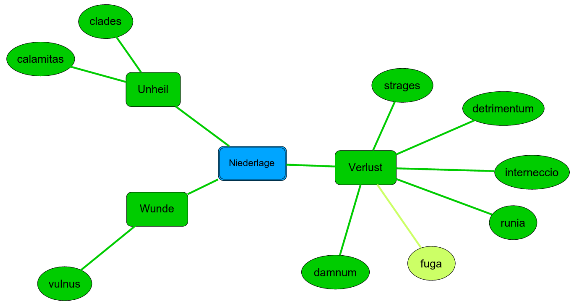
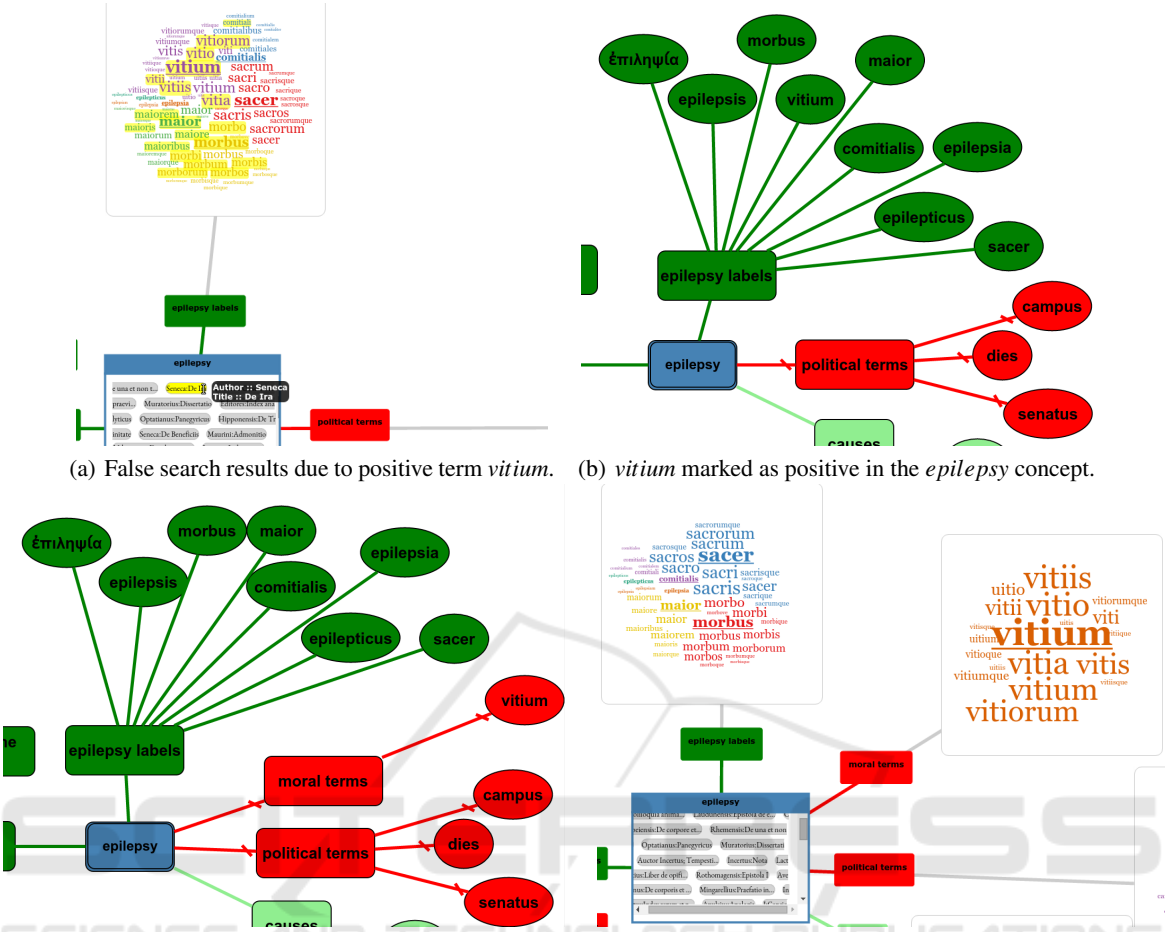


Figure 9: Concept *Niederlage* (defeat) modeled in the Concept Editor.

ber of texts written by Seneca the Younger (see Figure 10(a)), a philosopher of the first century AD who wrote about ethics. Close reading underpins the expectation that Seneca the Younger used the term *vitium* in a moral sense. When modeling the concept, she put the term to the subconcept *epilepsy labels* (see Figure 10(b)) as the archaic author Plautus used it in a physical sense with the intention to discover similar texts. But the frequent usage in moral contexts distorts the intentional meaning of the concept *epilepsy*. The iterative design of the concept search environment allows her to modify the *epilepsy* concept by removing the term *vitium* from the positive subconcept *epilepsy labels* to a new, negative subconcept *moral terms* (see Figure 10(c)). After re-running the concept search based on the improved concept (see Figure 10(d)), the false positives do not appear in the main concept node anymore and she can focus on potentially more related texts.

5.2 Evaluation and Feedback

We integrated the humanities scholars of our project at each development stage, on the one hand, to communicate the functionality of the system, on the other hand, to ensure designing a usable and comprehensible tool. Seven humanities scholars (6 female, 1 male) working in the project participated in five workshops that we organized to iteratively evaluate system design as well the current prototypes. The first two workshops were focused on the design of the concept modeling and concept search. In the later workshops the scholars were able to play with the prototypes, pose questions when issues occurred, and suggest usability as well as conceptual improvements of the system design. Also, we collected general impressions of the system in interviews at the end of the workshops. Besides the workshops, there were several meetings held to evaluate and discuss the results of the concept search system. Another example thematizing the ancient concept *Niederlage* (defeat) modeled during



(a) False search results due to positive term *vitium*.

(b) *vitium* marked as positive in the *epilepsy* concept.

(c) *vitium* moved to negative sub-concept *moralterms* in the updated *epilepsy* concept.

(d) Missing false positive search results due to relocation of *vitium*.

Figure 10: Use case showing how iterative modeling and searching improve the concept model and search results.

these sessions is shown in Figure 9, and the results are presented in Figure 11.

Overall, the humanities scholars liked that the visualization system stepwise turned away from being "just a working black box." The major reason for this statement was that the user retains full control of the concept creation and search process. Especially, the iterative modification of concepts, which supports the user to "finally build a good concept," was seen very powerful. Within the above outlined usage scenario, the humanities scholar stated that the "iterative process shows, how the imagination of the *epilepsy* concept becomes more and more precise to the humanist." Often mentioned was the capability of the system to support the humanities scholar in detecting conceptual errors when modeling concepts. The results shown in the Concept Search Results Viewer lead to "rethinking the concept model and to apply some changes." The combinatorial display of text titles, authors and corresponding terms was seen important when analyzing the results. Especially, the integration of TagPies helps to receive an immediate picture of the distribution of the terms and word forms the underlying concept is built of.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented a novel concept search environment to be used by humanities scholars to discover historical texts containing concepts modeled according to their ideas. The system design allows for an iterative modification of the humanities scholar's concept, so that a gradual improvement of results is possible. Although the system is particularly designed for dealing with ancient texts, we designed it the way that it is applicable to other domains using different text genres.

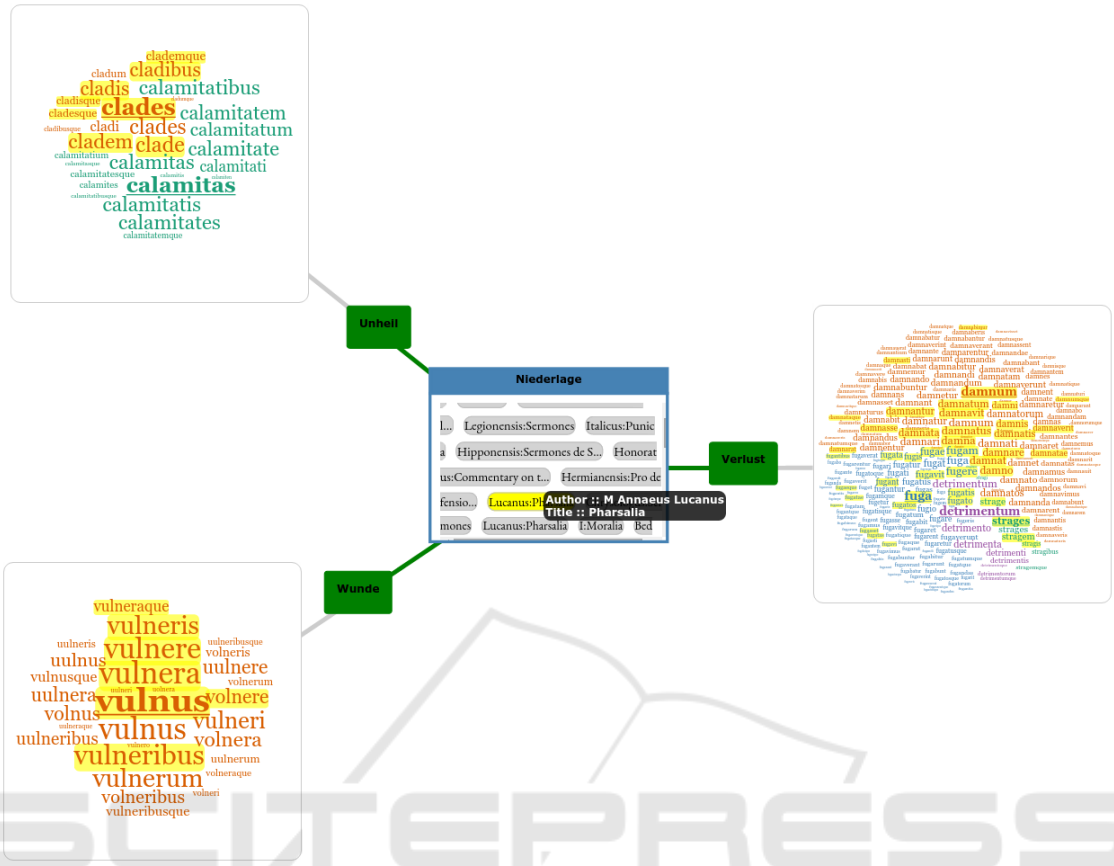


Figure 11: Search results for concept *Niederlage* shown in the Concept Search Results Viewer.

For the collaborating humanities scholars, the concept search system is valuable at current stage, but several tasks to improve the search capabilities exist. While quite fit for the visualization purpose, a simple counting of word occurrences does of course not yet constitute a reasonable ranking for result texts in the sense of Information Retrieval. Such a method should eventually take some form of statistical significance of word occurrences into account and perhaps use globally-normalized measures like TF/IDF (Baeza-Yates et al., 1999) or related ranking methods such as Okapi BM25 (Robertson et al., 1995). Even more important, it should somehow find a way to regard "negated" sub-concepts (and their - possibly also negated - (sub-)concepts) as intended by the Concept Editor. This means that special additions have to be made to tailor the ranking functions to boost the desired combinations of occurrences according to the concept – maybe even in an interactive way with feedback from the visualization.

Interestingly, the usage scenario outlined in Section 5.1 triggered a new research question for the humanities scholar: "Was the term *vitium* originally used in a physical or in a moral sense?" Although the current system helps to generate hypotheses, the change of concepts and their meanings over time in dependency of authors using specific terminologies cannot be tracked, but is very interesting for the humanities scholars. The planned extension of the concept search environment with a faceted browsing mechanism could support answering these types of research questions.

ACKNOWLEDGEMENTS

The authors thank Thomas Efer for maintaining the project back-end and implementing the concept search algorithm. This research was funded by the German Federal Ministry of Education and Research.

REFERENCES

(2015). Bibliotheca Teubneriana Latina. Walter de Gruyter. http://www.degruyter.com/db/btl (accessed March 19, 2015).

(2015). Perseus Digital Library. Ed. Gregory R. Crane. Tufts University. http://www.perseus.tufts.edu (accessed March 19, 2015).

Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., and Gleicher, M. (2014). Serendip: Topic Model-Driven Visual Exploration of Text Corpora. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 173–182.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern Information Retrieval*, volume 463. ACM press New York.

Bah, T. (2009). *Inkscape: Guide to a Vector Drawing Program (Digital Short Cut)*. Pearson Education.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *the Journal of machine Learning research*, 3:993–1022.

Booch, G., Rumbaugh, J., and Jacobson, I. (2005). *Unified Modeling Language User Guide, The (2Nd Edition) (Addison-Wesley Object Technology Series)*. Addison-Wesley Professional.

Budd, J. W. (2004). Mind Maps as Classroom Exercises. *The Journal of Economic Education*, 35(1):35–46.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., and Tong, X. (2011). TextFlow: Towards Better Understanding of Evolving Topics in Text. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2412–2421.

Cui, W., Liu, S., Wu, Z., and Wei, H. (2014). How Hierarchical Topics Evolve in Large Text Corpora. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2281–2290.

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

Eisenstein, J., Sun, I., and Klein, L. F. (2014). Exploratory Thematic Analysis for Historical Newspaper Archives. In *Proceedings of the Digital Humanities 2014*.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. (2011). Semantically enhanced information retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):434–452.

Giunchiglia, F., Kharkevich, U., and Zaihrayeu, I. (2009). Concept search. In *The Semantic Web: Research and Applications*, pages 429–444. Springer.

Guha, R., McCool, R., and Miller, E. (2003). Semantic search. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 700–709, New York, NY, USA. ACM.

Hager, P. J., Scheiber, H. J., and Corbin, N. C. (1997). *Designing & Delivering: Scientific, Technical, and Managerial Presentations*. John Wiley & Sons.

Jähnichen, P., Oesterling, P., Liebmann, T., Heyer, G., Kuras, C., and Scheuermann, G. (2015). Exploratory Search Through Interactive Visualization of Topic Models. In *Proceedings of the Digital Humanities 2015*.

Jänicke, S., Blumenstein, J., Rücker, M., Zeckzer, D., and Scheuermann, G. (2015a). Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies. *Digital Humanities Quarterly*. To appear.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015b). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In Borgo, R., Ganovelli, F., and Viola, I., editors, *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association.

Kanter, B. (2015). Cambodia4kids.org. https://www.flickr.com/photos/cambodia4kidsorg/6195211411 (Retrieved 2015-11-25).

Kaufman, M. (2015). 'Everything on Paper Will Be Used Against Me': Quantifying Kissinger. In *Proceedings of the Digital Humanities 2015*.

Koch, S., John, M., Worner, M., Muller, A., and Ertl, T. (2014). VarifocalReader – In-Depth Visual Analysis of Large Text Documents. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1723–1732.

Margolis, E. and Laurence, S. (2014). Concepts. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition.

Montague, J., Simpson, J., Rockwell, G., Ruecker, S., and Brown, S. (2015). Exploring Large Datasets with Topic Model Visualizations. In *Proceedings of the Digital Humanities 2015*.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). Okapi at TREC-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109.

Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 415–422, New York, NY, USA. ACM.

Wallach, H. M. (2006). Topic Modeling: Beyond Bag-of-Words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM.

Ware, C. (2010). *Visual Thinking for Design*. Morgan Kaufmann.

Yi, J. S., Melton, R., Stasko, J., and Jacko, J. A. (2005). Dust & Magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256.