VisNavi: Citation Context Visualization and Navigation

Farag Saad and Brigitte Mathiak

GESIS - Leibniz Institute for the Social Sciences,
Unter Sachsenhausen 6-8, 50667 Cologne, Germany
{farag.saad,brigitte.mathiak}@gesis.org

Abstract. The process of retrieving information for literature review purposes differs from traditional web information retrieval. Literature reviews differentiate between the weightiness of the retrieved data segments. For example, citations and their accompanying information, such as cited author, citation context etc., are a very important consideration when searching for relevant information in literature. However, this information is integrated into a scientific paper, in rich interrelationships, making it very complicated for standard search systems to present and track them efficiently. In this paper, we demonstrate a system, *VisNavi*, in the form of a visualized star-centered approach that introduces the rich citation interrelationships to the searchers in an effective and navigational appearance.

Keywords: digital libraries, citation context, visualization, navigation, information retrieval, text extraction.

1 Introduction

The dramatic increase of literature, on a daily basis, in all science fields creates the possibility that traditional literature search systems might be inefficient in supporting a clearer view of citation interrelationships, such as, citation networks, citations context etc. For a deeper understanding of a cited paper, the extraction of the context, in which it has been cited, is an essential step in performance (the utilization of citation contexts). Citation context refers to the textual information that surrounds the citation key, which is used to refer to the cited paper, inside the citing paper body. In the recent past, some work has made good progress to tackling this problem, such as, the prototype Action Science Explorer [1] and CircleView [2]. However, these prototypes either paid little attention in supporting searchers in interpreting the intensive information flow or were deficient in providing and presenting citation information to the searchers. This demonstration tries to overcome or alleviate this deficiency by combining interactive visualization and at the same time abstract textual information, so that the searchers can explore and navigate between related papers efficiently.

Paper Collection: Our paper collection came from the DGS (German Society for the Social Sciences[1]) corpus. It represents a full set of digitized proceedings

[1] http://www.soziologie.de/

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 849–852, 2013.

of the German Society for the Social Sciences (DGS), spanning 100 years. It consists currently, of 7,000 social science papers presented as PDF documents with their corresponding Metadata and this corpus is continually growing.

2 Functionality Overview

The Prototype VisNavi consists of two main components that are integrated to perform the search task. The *off-line* component that involves PDF text and citation information extraction and the *on-line* component that involves the searching, visualizing and navigating of scientific articles. The off-line component is responsible for extracting and cleaning up text from the PDF files e.g., correcting any misspelled OCR (Object Character Recognition) words. Furthermore, it includes a citation context extraction which is used to extract all citations along with their context. The on-line component is used to present a paper of interest in a visualized manner to the searcher. For smoothly integrating an interactive visualization, we made the use of the open source visualization library infoVis[2].

2.1 Off-Line: PDF Text and Citation Context Extraction

In order to extract the text from the OCRed papers, we used one of the PDF extraction open source tools[3]. However, PDF extraction tools extract the text from the PDF files, as it is, regardless of any noise in the text i.e., a considerable number of original documents in the DGS corpus are deteriorated and can't be clearly OCRed. To tackle this issue, approaches have been implemented, for example, we corrected mistakenly extracted word/words by using our previously developed n-gram approach [3]. In order to extract the citation context, we make the use of the open source tool ParsCit [4] to automatically extract the reference list and its corresponding context in a given paper. ParsCit is used to allocate reference strings inside the text of the paper, parsing them and extracting their citation contexts. In order to achieve high citation extraction accuracy, ParsCit employs state-of-the art machine learning models, in order to obtain information from the reference string e.g., authors names, paper title, conference, etc. Furthermore, it applies heuristic rules to find and bind reference keys in the paper body text and its citation context. Since the DGS corpus spans over 100 years, signifies that a substantial number of original documents have deteriorated. Furthermore, historical documents sometimes contain unrecognized fonts, fragmented letters, shaded backgrounds, unrecognized line breaks, overlapping letters or skewed text, etc. Therefore, a clean-up process has been necessary.

2.2 On-Line: Searching, Visualization and Navigation

The search process starts by submitting a query e.g., (Kapitalismus Widersprüche Ökonomisierung "capitalism contradictions economization") through the full-text

[2] `http://philogb.github.com/jit/docs.html`
[3] `http://pdfbox.apache.org/`

search engine integrated in VisNavi. Next, a set of relevant papers with their Metadata are displayed in the search engine interface. If the searcher is interested in visualizing one of the retrieved papers, the searcher must indicate this by clicking on the visualization icon displayed along with each search result. For searchers who have no experience using the system, assisted information is displayed by placing the mouse cursor over the visualization icon, to provide the user with information describing the next stage of interaction i.e., which next action will be provided by the system. Once the user decides to visualize a paper, this paper is visualized and its author is placed in the center and its citations are displayed around it (See Figure 1). Furthermore, a citation context for each cited paper is displayed on the right side of the system's interface (See Figure 1 (A)). Using Parcit, the citation contexts are extracted based on the window size of the surrounding words around a citation key, which may lead to having an incomplete citation context. Therefore, we provided the user with a feature to explore wider context by placing the mouse cursor on "view the entire citation" link. Thereafter, the system reacts and extends the already displayed context to the user e.g., by extending the citation context window size (See Figure 1 (B)). In order to give the user confidence in a citation context provided by the system, he/she can see the highlighted citation key, color-coded in green, inside the displayed context (See Figure 1 (A)). If the user would like to shift the focus to a new paper, he/she needs to click on the paper of interest, which then will be shifted to the center and its citations will automatically be displayed around it. If the user is solely interested in viewing a unique citation, he/she can click on the desired cited paper (cited author node) and the system will respond by displaying only the selected citation. The cited author node and the selected citation node are color-coded white (on a black background) and the rest of the citation nodes are color-coded grey.

2.3 Evaluation: A Pilot User Study

We undertook a user study (with 10 participants), considering four main points of interest, such as ease of use, visualization and navigation efficiency, assisted information usefulness and the appearance of the tool. Users were requested to use the tool by submitting an information need and visualizing a paper of interest. Thereafter, they were asked to start navigating between related papers. After using the tool for a few search tasks, the users were requested to rate each functionality of the tool by giving a specific score between 0 (low) and 5 (high). Overall evaluation average of the proposed tool was 3.95/5 (79%). General comments have been reported by few users. For example, the citing author (paper of current focus) should be represented by a clearer, bigger node with different colors to the nodes representing the cited authors (cited papers). Furthermore, users might be interested in having the possibility of viewing the full text (PDF file) for each citation, at any level of interaction. In addition, many requests to integrate a smooth zooming feature (users would have some control in the appearance of the tool) were emphasized. These comments are achievable and planned to be integrated in the improved version of the tool.

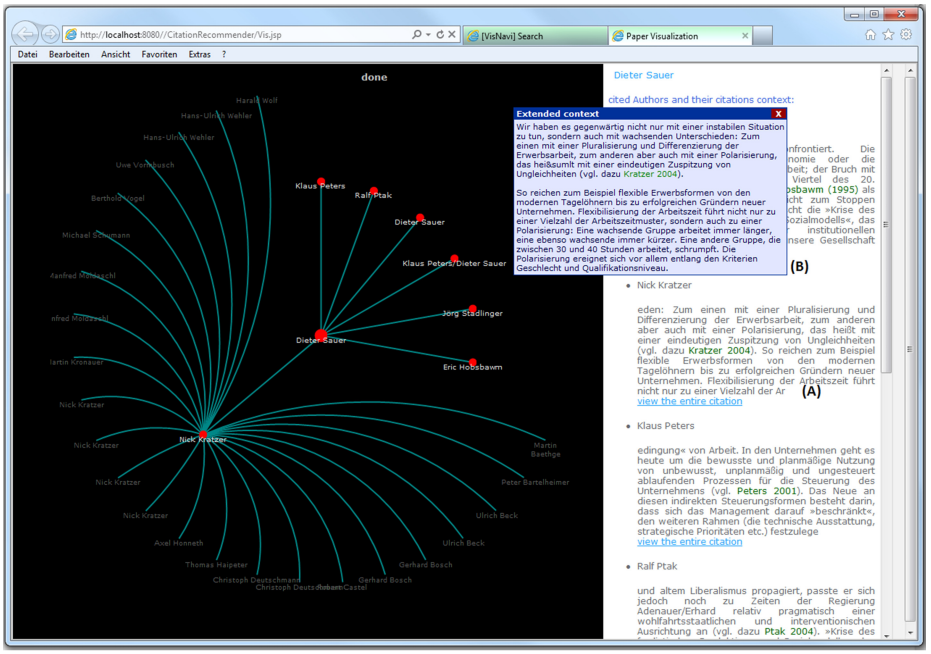

Fig. 1. Paper visualization

3 Conclusion

Advances in visualization and string fragment allocations made it possible to combine both, to create a useful system that integrates interactive visualization and abstract textual information that together achieve remarkable support for researchers in their literature review. We consider the proposed system an opportunity for searchers to smoothly obtain literature information in a new way that may be difficult to achieve employing standard search systems.

References

1. Gove, R., Dunne, C., Shneiderman, B., Klavans, J., Dorr, B.J.: Evaluating visual and statistical exploration of scientific literature networks. In: Proceedings of the Visual Languages and Human-Centric Computing (VL/HCC), pp. 217–224 (2011)
2. Bergström, P., James Whitehead Jr., E.: Circleview: Scalable visualization and navigation of citation networks. In: Proceedings of the 2006 Symposium on Interactive Visual Information Collections and Activity (IVICA) (2006)
3. Ahmed, F., Nürnberger, A.: Evaluation of n-gram conflation approaches for arabic text retrieval. JASIST: American Society for Information Science 60(7), 1448–1465 (2009)
4. Councill, I., Giles, C.L., Kan, M.Y.: Parscit: an open-source crf reference string parsing package. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008) (2008)