Introducing a User Interface with an Entity-Strategy-based Approach for Exploring Document Collections

Daniel Hienert[1] and Wilko van Hoek[1]

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
{daniel.hienert,wilko.vanhoek}@gesis.org

Abstract. In this paper we present a first sketch of an alternative approach for searching and exploring document collections. The traditional approach applied in Digital Libraries and Web Search Engines is based on search forms and result lists. The user enters a keyword and is presented with a list of document metadata with authors, titles and descriptions. We propose an alternative approach that is based on entities in a document collection like authors, documents and topics. The user can search for these entities and can then choose from a set of highly abstracted search strategies, e.g. to get highly cited papers from an author. The approach is applied in a zoomable and infinite user interface that enables the user to explore freely and where the search history is always present.

Keywords: Visual Interface, Exploratory Search, Visual Exploration, Search Strategies.

1   Introduction

Today's Digital Libraries (DLs) still make use of the standard paradigm of query-response. Users can enter a query and are presented with a list of relevant documents which they have to inspect and filter according to their information need. Already Bates presented a list of alternative search strategies from the real-world such as 'Citation Searching' or 'Journal Run' [1] which partly have been adopted in modern scholarly database systems such as Scopus or Web of Science. However, this is far from being usual practice in DLs, where full-text indexing of documents opens new possibilities. Exploratory Search [5] proposes a model beyond query-response, with a focus on the learn and investigation step in the search process. Highly-interactive search systems can support these steps whereby interactive visual search systems are a part of it. Therefore already a number of visual search tools have been proposed for the exploration of DL content. Early attempts experimented with different visual metaphors or tried to gain insight with the visualisation of the distribution of information facets. More recent tools are for example the INVISQUE system [4] which supports the search and manipulation of results on an infinite panel or PivotPaths [2] which show relations between concepts, resources and people. Another important aspect for learning and investigation while searching is the visualisation of the search process

itself. Scientists spend much time with literature search; their search history can enlarge quickly over months to even years. Today's DLs only support to save search results or documents, other artefacts such as document inspection are lost. Research showed that search histories support revisitation [6] in web search and support the user's orientation within a search session [3]. In the following we want to present a User Interface (UI) concept which combines visual information search with different search strategies and a visible search history.

2   Concept Overview

The concept consists of four core ideas:
1. The UI is an infinite panel with zoom & pan functionality. The user can start a search session from every point on the surface. Starting point is a simple search form, where one can search for artefacts of a document collection such as topics, persons, documents, journals similar to a search in a standard DL.
2. The UI does not only return result lists with document metadata, but small interactive elements which represent entities and artefacts like persons, documents, topics and results of applied search strategies like a list of highly-cited papers.
3. For any of these entities, the user can choose a search strategy or functionality from a select box to initiate the next search step. A search strategy for the entity person can be e.g. 'Highly-cited papers', for the entity document e.g. 'Similar Topics'. Compare Table 1 for more search strategies based on the entity type.
4. Over time a search graph is shown on the surface that represents the whole user's search history. Users can select certain regions or search paths to (a) minimise/expand them, (b) to label, categorise or annotate them or (c) to set up an alert via email for new results, (d) to save/export/share them or for any other action that may be applied to a search path.

Table 1. Examples for search strategies based on the entity type. In brackets the appropriate search strategy of Bates [1] are shown.

| Based on person	Based on document
• Highly-cited/highly-referenced papers • Main/important co-authors • Main/sub research topics	• Cited by (Citation Searching) • Referenced by • Similar Topics (Subject search) • Main/co-authors (Author searching) • Documents from footnotes (Footnote chasing)
Based on topic	• Same Journal (Journal run) • Same category/classification (Area scanning) • Related/Similar papers (e.g. by content, topics, references, subject etc.)
• By relevance (tf-idf, BM25) • Highly-cited papers • Important authors/journals • Journal/author productivity • Author centrality	
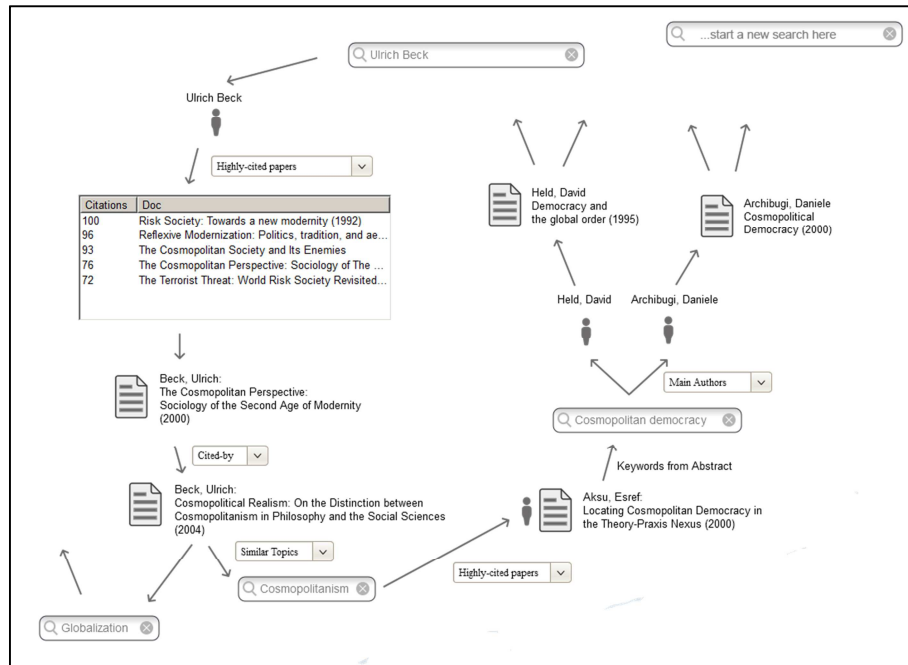
Fig. 1. Exploring social science literature starting from the author 'Ulrich Beck'

Figure 1 shows the core idea of our approach applied to an example from the field of the social sciences. The mockup shows one possible exploration path based on a real-world document collection from the social science portal Sowiport[1].

First, the user searches for the author 'Ulrich Beck' and can then choose 'Highly-cited papers' from the strategies menu to get an overview of his most influencing work. As a result the most-cited papers are presented in a small list. After inspecting the abstracts in the document view, the user classifies the third paper as interesting and chooses 'Cited by' from the methods menu to show the latest papers which influenced it. Based on this paper the user is interested in the topic 'Cosmopolitanism' and wants to see highly-cited papers for this topic. She/he arrives at the author 'Esref Aksu' and grabs the keywords 'Cosmopolitan Democracy' from the abstract and initiates a new search. Choosing 'Main Authors' from the strategies menu shows two authors and their papers with which the search process can continue.

Because the search history is always visible on the user interface, the user can return to a previous search step such as a search, a person or a document and can continue the search there. In the above stated example, the user returned to the topic 'Globalization' from Beck's 'Cosmopolitical Realism' and initiated a new search.

---

3   Discussion & Future Work

The proposed approach has several benefits over the standard search form/result list paradigm. Based on the four core ideas of our approach these are:

1. *UI as infinite Panel*: The use of an UI panel with infinite space is a prerequisite that has been used in other visual exploration tools as well [4]. In our context it removes the limitation of showing only one search step, but can represent whole search sessions and multiple sessions.

2. *Entities*: In a standard DL, search results are limited to a list of documents ordered by relevance. The use of entities such as persons, documents or topics allows the intuitive application of search strategies.

3. *Search Strategy*: Complex search strategies are encapsulated in one-click UI elements and can be applied easily. This allows alternative exploration and views on the document collection. New search strategies can be implemented easily.

4. *Search Graph*: Every step in the search process is visible on the UI and forms a search graph over time. That allows an overview of the whole search session, but also over a set of search sessions. Therefore a prior search path can be continued, but also a search session can be shared with another person, e.g. among colleagues in a research group.

However, most search strategies require complex computation and a rich data set. For example, "Highly-cited papers" for an author needs a separate citation index, which may not always be present in nowadays DLs or the real-time computation of metrics like author centrality can be a challenge. In a next step we want to implement a system prototype that can be used for exploring different document collections such as the arXiv[2] corpus for the natural sciences or Sowiport for the social sciences which contain this rich information. Based on that, we will perform various user tests to verify the basic plausibility of our approach.

References

1. Bates, M.J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. Online Rev. 13, 5, 407–424 (1989).
2. Dörk, M. et al.: PivotPaths: Strolling through Faceted Information Spaces. IEEE Trans Vis Comput Graph. 18, 12, 2709–2718 (2012).
3. Imko, J. et al.: Semantic History Map: Graphs Aiding Web Revisitation Support. Presented at the August (2010).
4. Kodagoda, N. et al.: Using Interactive Visual Reasoning to Support Sense-Making: Implications for Design. IEEE Trans. Vis. Comput. Graph. 19, 12, 2217–2226 (2013).
5. Marchionini, G.: Exploratory search: from finding to understanding. Commun. ACM. 49, 4, 41–46 (2006).
6. Mayer, M.: Web History Tools and Revisitation Support: A Survey of Existing Approaches and Directions. Found. Trends® Hum.-Comput. Interact. 2, 3, 173–278 (2007).

---

[2] http://arxiv.org