Integrating Interactive Visualizations in the Search Process of Digital Libraries and IR Systems

Daniel Hienert, Frank Sawitzki, Philipp Schaer, Philipp Mayr

GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Köln, Germany
{daniel.hienert, frank.sawitzki, philipp.schaer,
philipp.mayr}@gesis.org

Abstract. Interactive visualizations for exploring and retrieval have not yet become an integral part of digital libraries and information retrieval systems. We have integrated a set of interactive graphics in a real world social science digital library. These visualizations support the exploration of search queries, results and authors, can filter search results, show trends in the database and can support the creation of new search queries. The use of weighted brushing supports the identification of related metadata for search facets. In a user study we verify that users can gain insights from statistical graphics intuitively and can adopt interaction techniques.

Keywords: Visualization, IR System, Digital Library, Interactivity.

1 Introduction

Visualizing information in digital libraries is an ongoing research field in the information retrieval community. But until now visualization has not become an integral part of real world digital libraries (DL) and information retrieval (IR) systems. This may be due to complexity, abstractness and missing interactivity of graphics. Also a lot of systems are stand-alone where the visualization takes most of the space and is not integrated in the standard *search field - result list* paradigm. In our approach we have integrated interactive visualizations directly in the search process. This allows the combination of advanced queries, IR features, different information types and interactive graphics.

2 Related Work

Graphics have been used for different aspects of DL like visualization of metadata, queries, results, documents, co-authors and citations. Most recent tools are BiblioViz [7], PaperLens [6], PaperCube [2] and INVISQUE [8]. Most of these tools have in common that they provide an alternative view and access to a document corpora, but are not integrated in the standard *search field - result list* paradigm of Digital Libraries. Gopubmed [1] is a semantic-based search engine for medical papers. Analog to our solution it offers statistical visualizations in the search results for the actual query.

3 Embedding Visualizations in a Real World DL

We chose the social science information portal Sowiport[1] as a real-world environment for embedding interactive visualizations. Sowiport integrates literature references, persons, institutions, research projects, events and studies from all areas of the social sciences. It currently contains about 7 million information items from 18 databases. The technical basis for the search functionality is a SOLR system, an open-source search server based on the Lucene library. Technical basis for visualizations is a web-based visualization toolkit with the support of various interaction techniques [5].

The temporal and spatial distribution of search results is an interesting aspect to discover i.e. where and when to a specific topic has been published or what databases cover which countries. A line/bar chart for temporal distribution and a map for spatial distribution are visualizations to grasp this information quickly. The temporal chart shows the distribution of results over the last 50 years. The Google map takes location information from metadata fields like *publication place* for documents or *place of the institute* for research projects.

For every possible facet of a search result we can offer visualizations that show facets frequencies in a horizontal bar chart. For example, the bar chart for the facet *subject* shows up to 50 thesauri terms sorted by frequency. The chart allows the comparison of frequency and importance of certain terms in an easy way. Two more visualizations base on an analysis of co-authorships for top frequent authors and on several controlled vocabularies including a recommender service to find related search terms. A co-author network for the top 50 authors of a search query is presented in a network graph. The resulting network is presented as a graph with a node for each author and the edge labeled with the count of co-authorships. Related search terms are presented in a network graph to provide search term recommendations [4]. Data basis can be different vocabulary services the user can choose from.

Any visual item in the graphics can be equipped with a search icon that can initiate a new search or filter the actual results. For example, facets in the top lists like persons or keywords can filter results down to a specific topic or author or clicking the icon in the co-author graph performs a new search on an author. In the related keywords graph clicking the icon leads to new queries for related or suggested keywords.

We also use the approach of *weighted brushing* as introduced by VisGets [3]. Visual elements are highlighted according to the count of shared associations. If a visual element A is marked in a graphic, information items are identified that contain that attribute. If another visual element B shares three information items with A, it is highlighted with a higher opacity than element C which shares only one information item with A.

4 User Study

We have conducted a user study to verify the following hypotheses: (1) Can users gain insights from statistical graphics that could not be retrieved from other resources

[1] www.gesis.org/sowiport

within search and can users utilize these graphics intuitively? (2) Can new interaction techniques be adapted and understood without any instructions to filter results and find related metadata? We asked 19 students (22 to 30 years; 6 male/13 female; average rating of 2.47 (from 1=very good to 6=very bad) in experience with scientific information portals) from a university course of information science to carry out certain tasks in our system and after each task to fill out a questionnaire.

4.1 Tasks and Questions

Participants had to submit a query for the keyword *information society* and then had to answer the following questions only with the help of the visualizations:

a. When the most documents for the topic *information society* had been published and how long the upward trend before lasts based on the number of publications?
b. How many documents were published in Spain?
c. What are the three most frequent keywords and persons for this search?
d. Who has published together with Ullrich Bauer and how many papers were published together?
e. What is the broader term of the search term *information society*?
f. What keywords are associated with the person Rudi Schmiede?
g. Which persons use the keyword *internet*?
h. Use interactive elements in the visualizations to filter search results to the person Rainer Kuhlen.

Tasks (a)-(e) could be answered with the help of a single visualization, for the tasks (f) and (g) the interaction techniques weighted brushing had to be used and for task (h) the users had to click the arrow icon. Instructions about how to use the new interaction techniques comes only from the legend. For each question users should write down the answer in a free text box, answer how long the task took, rate how difficult they perceive the task with a five-point scale (2=very easy, 1=easy, 0=normal, -1=difficult, -2=very difficult) and could give comments or suggestions. Table 1 shows the summarized results.

Table 1. Summarized results of the tasks.

Question	Correct answers (in %)	Average time needed (in seconds)	Difficulty level (2=very easy, 1=easy, 0=normal, -1=difficult, -2=very difficult)
a.	68	43	1
b.	89	21	1.33
c.	100	15	1.53
d.	100	16	1.17
e.	100	23	1.05
f.	89	53	0.41
g.	89	19	0.84
h.	89	54	0.50

After conducting the tasks, users could assess their overall impression of statistical visualizations supporting the search process: 18 from 19 users found visualizations useful to support the search process and also 18 from 19 users found that visualizations could support them in their personal research. Users rated the interaction technique to filter search results or start a new search with the green arrow icon with an average rating of helpful (1.05) and the interaction technique to highlight metadata from common documents with weighted brushing also with helpful (1.00).

5 Conclusion

Most users could use the graphics intuitively without any instructions to answer questions for maxima, trends, spatial distribution, top keywords and persons and co-authorships for the actual search query. Even novel interaction techniques for digital libraries like weighted brushing and filtering from graphics could be used and understand only from the legend's explanation. Perceived difficulty level was relatively stable for tasks that could be answered with the help of one individual visualization (>=1) and on a lower level for interaction techniques (>= 0.4 and <=0.84) triggered by that they were introduced from the scratch. Users found visualizations useful for search in general and for their personal search and rated both interaction techniques as helpful.

References

1. Alexopoulou, D. et al.: A2-D7 GoPubMed: ontology-based literature search. (2007).
2. Bergstrom, P., Atkinson, D.C.: Augmenting the exploration of digital libraries with web-based visualizations. 2009 Fourth International Conference on Digital Information Management. 1-7 (2009).
3. Dork, M. et al.: VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. IEEE Trans. Visual. Comput. Graphics. 14, 6, 1205-1212 (2008).
4. Hienert, D. et al.: A Novel Combined Term Suggestion Service for Domain-Specific Digital Libraries. In: Gradmann, S. et al. (eds.) Research and Advanced Technology for Digital Libraries. pp. 192-203 Springer, Berlin, Heidelberg (2011).
5. Hienert, D. et al.: VIZGR: Combining Data on a Visual Level. Proceedings of the 7th International Conference on Web Information Systems and Technologies (WEBIST). (2011).
6. Lee, B.: Understanding research trends in conferences using PaperLens. Extended Abstracts of the ACM Conference on Human Factors in Computing Systems 2005. pp. 1969–1972 ACM Press (2005).
7. Shen, Z. et al.: BiblioViz: A System for Visualizing Bibliography Information. In Proceedings of Asia-Pacific Symposium on Information Visualization. pp. 93–102 (2006).
8. Wong, W. et al.: INVISQUE: intuitive information exploration through interactive visualization. Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems. pp. 311–316 ACM, New York, NY, USA (2011).