Visualising and Mining Digital Bibliographic Data

Stefan Klink, Michael Ley, Emma Rabbidge,
Patrick Reuther, Bernd Walter and Alexander Weber

Department of Database and Information Systems (DBIS),
University of Trier, D–54286 Trier, Germany

Abstract: Finding related publications and their correct bibliographical data is getting harder and harder due to the mass of unrequested information and the decreasing precision of many information providers. The DBLP Computer Science Bibliography is a service used by thousands of computer scientists which provides fundamental support for scientists searching for publications or other scientists in similar communities.

This paper describes a user–friendly interface which plays the central role in searching, browsing, and mining of bibliographical data. After introducing the concept of multi–layered browsing and an integrated smart component for entering new data, *DBL–Browser* itself and various visualisations are described.

1 Introduction

The modern information society faces a severe dilemma. More information than ever is available, but accessing relevant information is still a very challenging task [Kl04]. Cleverdon estimates the amount of publications of the most important scientific journals to 400,000 per year [Cl84] and INSPEC, the leading English–language bibliographic information service, is growing at the rate of 350,000 records each year [Sc03].With the *DBL–Browser* we have attempted to work against this problem by providing an efficient way for browsing a given data base. By combining both textual and visual browsing functionality we established a browsing–based retrieval and visualisation system which enables users to better understand their search domain and consequently offers the opportunity to expand their original query. As already indicated by [DCFQ00] users like both the graphical nature of information organisation and multi–level browsing systems. Both features are central parts of the browser. Additionally, the browser uses a combined query–based and browsing–based approach for the exploration and navigation of the digital library domain. Starting from an unspecific query or using hyperlinks to browse from the 'homepage' users can browse through the bibliographical data. During the browsing process all data is visualised by appropriate graphical techniques which help users to understand their search domain, helps them find relevant authors or publications and above all provides information about further researchers or important conferences or journals. The following paragraphs introduce the *DBL–Browser* which supports a searching and browsing–based approach within the Digital Library.

2 Browsing Digital Bibliographic Data

There are several things, that make the *DBL–Browser* an easy–to–use everyday application: One of the main aspects is it’s straightforward user interface. Anybody using a common web–browser is able to use the *DBL–Browser*. All essential features are at hand – like searching and filtering the data. The search system has all typical functions, with additional features such as combined searches or vague searches, based on the Levenshtein distance [Le65]. In addition to the classic navigation, the user interface offers the concept of *Tabs*, thus the user is able to put different stages of a search session into different tabs. These tabs can be visually aligned to show more than one part of information at one time or for giving different visualisations of the same information – so you can have a histogram view of an author as well as a view showing authors that are related to the current author. We call this *multi–layered* browsing, because the user is always able to get different views (layers) to the same data. For example, they can switch between a chronological text representation of all publications from a given author to a graphical histogram, showing different aspects of the same data in a different representation. The other main feature of the browser is the additional navigation provided by the *everything–is–clickable* concept. Every piece of information shown by the browser provides a link to additional associated information. I. e., a search or a Table of Contents (TOC) page.

An additional component, the *DBL–Editor*, can be plugged-in to the *DBL–Browser* to allow new data to be entered with fast access to the existing database. This brings the task of entering data to a new level of quality, because the data maintainer can easily cross–check the new publications against the old publications. This is especially important for keeping names consistent, or to see if two people with the same name are different people. The data is entered XML with on–the–fly syntax highlighting and validation against the given DTD on saving.

3 Visualisations

The *DBL–Browser* was originally developed to browse the textual visualisations [AFG+03] of the DBLP [Le97][Le02]. As the browser has evolved, so have the textual visualisations. The browser now includes, in addition to the author pages, textual visualisations of search results, BibTeX pages and TOC pages [KLR+04]. Both the BibTeX and TOC pages allow the user to continue to explore the knowledge domain to find other ’similar’ documents or related authors. The author pages themselves have also evolved, allowing for a clear, consistent layout of information across all textual visualisation.

The filter [KLR+04] has evolved to include the use of icons on leaf nodes. These icons help to expose features about the next level of data, namely how much data is contained. The same filter style is used with textual visualisations of search results. Structuring the data in this way ensures that the user is not ’overloaded’ with data that must otherwise be mentally filtered. BibTeX and TOC pages make use of the filter to navigate within the pages, providing links to session headings, abstracts, citations and references.

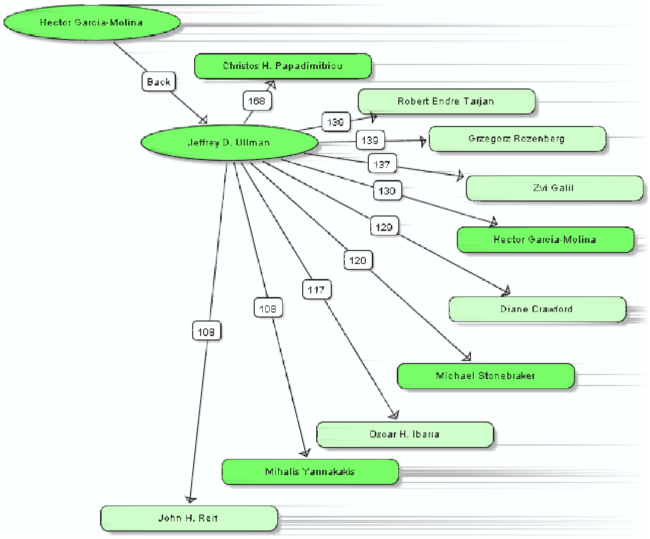

Figure 1: Author Relationship Graph

Graphs are used to visualise relationships between the data, relationships which are not apparent when searching and browsing the data. Graphs are currently provided to show relationships between conferences and journals or between authors.

The Conference/Journal Relationship Graph [KLR$^+$04] uses an author based similarity measure. That is, a stream[1] is related to another stream if there are many authors who have published in both streams. The relationships used to construct the graph are built on start up. The Conference/Journal Relationship Graph shows incoming and outgoing relationships, which the user can browse to find other streams that may contain more publications of interest. By switch back to the textual view specific information on publications can be obtained.

Relationships used to create the Author Relationship Graph are based on a stream similarity measurement. One author is related to another author if they have published in the same streams. Research continues to investigate the best way to calculate and weight the relationship. Unlike the Conference/Journal Relationship Graph, the relationships required for the Author Relationship Graph are calculated upon request.

The Author Relationship Graph, figure 1, shows the ten most related authors for a current author (Jeffrey D. Ullman). The related authors are positioned such that an author with a stronger relationship to the current author lies closer, while weaker related authors lie further away. Related authors may, or may not be co–authors of the current author, but when they are, they are represented by a darker node colour. Users can browse the graph by clicking on a related author, which moves that author to be the current author and subsequently shows the new set of related authors. As a navigational aid, the previous

[1] A stream is either a conference or a journal.

current author (Hector Garcia–Molina) is shown as an incoming relationship. This allows the user to easily browse back to the previous graph.

Each textual visualisation has a corresponding histogram visualisation. In the same way that the textual visualisations can be browsed, so too can the histogram visualisations. Histograms allow a user to identify which member of a group contains the most publications. For example, which conference or journal has the most publications for the data set; or which author in a set of search results has the most publications and could there for be interesting to look at.

4 Further Research

Scientific publications can be classified in one of three major categories, based on the amount of information available and the text length. The first category is a full text publication, which has the most available information. With a full text publication a researcher interested in finding information on a special topic can easily tell whether the considered paper is relevant for his purposes or not. Consequently, the average information content of a full text publication is rather high. Less information content can be extracted if full text is unavailable, but only an abstract or an automatically generated summary of the full text is stored. This is the second category. The average information content of an abstract is in general lower than that of a full text publication. The final category consists of publications were full text or abstract information is not available. In this case researchers can only draw conclusions for the usefulness of a text by means of bibliographic information such as title of the publication, authors of the paper, place of publication or year of publishing. Due to the normally very short titles (in DBLP about 8 words per title) and, consequently, the low average information content it is difficult to draw conclusions as to the usefulness of a paper dealing with a specific topic. Analogous to the decrease of average information content the performance of a search engine measured through recall/precision decreases if less information is available. This decrease of search performance shows the need to develop strategies in order to increase search performance when less content is available.

A promising approach to improve retrieval performance is the theory of social networks, i. e. co–author networks. The co–author relationship can be displayed through a co–author graph, in which the nodes represent authors and the edges represent co–authorship between the authors. This environment can be used for a hierarchically search or result set expansion. Starting with a normal term–based search on the titles, the results to this query can be used to initiate a query in the co–author network. By following this procedure relevant documents which do not explicitly contain the keywords of the former search could be identified as relevant because of the similarity of authors. In this context, establishing a combined similarity measurement between the original term search similarity and the co–author relationship similarity is a challenging task. Besides the above strategy a number of other strategies could be considered to make use of social networks [Mu01]. These strategies include central author search, author network browsing, query expansion using central authors or a document ranking based on author centrality.

Our intention is to provide the *DBL–Browser* as a framework for experiments. Due to its modularisation, it is an easy challenge for anyone interested to integrate his or her visualisation ideas and algorithms. The XML and compressed version of the DBLP data and the source code of the browser are available on our web server. We encourage all of you to use and/or improve it. Feedback and further ideas are also welcome [2]. See the project homepage: **`http://dbis.uni-trier.de/DBL-Browser/`**.

References

[AFG+03] Agarwal, S., Fankhauser, P., GonzalezOllala, J., Hartmann, J., Hollfelder, S., Jameson, A., Klink, S., Lehti, P., Ley, M., Rabbidge, E., Schwarzkopf, E., Shrestha, N., Stojanovic, N., Studer, R., Stumme, G., Walter, B., und Weber, A.: Semantic Methods and Tools for Information Portals. In: Dittrich, K., König, W., Oberweis, A., Rannenberg, K., und Wahlster, W. (Hrsg.), *Proceedings of INFORMATIK 2003 - Innovative Informatikanwendungen*. volume 34 of *Lecture Notes in Informatics*. S. 116–131. Frankfurt, Germany. October 2003. Gesellschaft für Informatik e.V. (GI).

[Cl84] Cleverdon, C. W.: Optimizing convenient online access to bibliographic databases. *Information Services and Use*. 4:37–47. 1984.

[DCFQ00] Ding, Y., Chowdhury, G. G., Foo, S., und Qian, W.: Bibliometric information retrieval system (BIRS): A web search interface utilizing bibliometric research results. *Journal of the American Society for Information Science*. 51(13):1190–1204. 2000.

[Kl04] Klink, S.: Improving Document Transformation Techniques with Collaborative Learned Term-based Concepts. In: Dengel, A., Junker, M., und Weisbecker, A. (Hrsg.), *Reading and Learning: Adaptive Content Recognition*. volume 2956 of *Lecture Notes in Computer Science*. S. 281–305. Berlin, Heidelberg, New York. April 2004. Springer-Verlag.

[KLR+04] Klink, S., Ley, M., Rabbidge, E., Reuther, P., Walter, B., und Weber, A.: Browsing and visualizing digital bibliographic data. In: Deussen, O., Hansen, C., Keim, D. A., und Saupe, D. (Hrsg.), *Symposium on Visualization, Joint EUROGRAPHICS – IEEE TCVG*. Konstanz, Germany. May 19-21 2004.

[Le65] Levenshtein, V. I.: Binary codes capable of correcting spurious insertions and deletions of ones (original in Russian). *Russian Problemy Peredachi Informatsii*. 1:12–25. 1965.

[Le97] Ley, M.: Die Trierer Informatik–Bibliographie DBLP. In: *GI Jahrestagung*. S. 257–266. 1997.

[Le02] Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In: *Proceedings of the 9th International Symposium on String Processing and Information Retrieval (SPIRE)*. volume 2476 of *Lecture Notes in Computer Science*. S. 1–10. Lisbon, Portugal. September 11-13 2002. Springer.

[Mu01] Mutschke, P.: Enhancing information retrieval in federated bibliographic data sources using author network based stratagems. In: *Proceedings of the 5th European Conference of Digital Libraries (ECDL)*. volume 2163 of *Lecture Notes in Computer Science*. S. 287–299. Darmstadt, Germany. September 4-9 2001. Springer.

[Sc03] ScienceDirect. About the Abstract Databases–INSPEC. http://help.sciencedirect.com/robo/projects/sdhelp/about/inspec.htm. 20003.

[2] This work is supported by the German Ministry for Education and Research (bmb+f) in the SemIPort project.