Joint EUROGRAPHICS - IEEE TCVG Symposium on Visualization (2004)
O. Deussen, C. Hansen, D.A. Keim, D. Saupe (Editors)

Browsing and Visualizing 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, 54286 Trier, Germany

Abstract

Access to publications is provided by conventional libraries, digital libraries operated by learned societies or commercial publishers, and a huge number of web sites maintained by the scientists themselves or their institutions. But comprehensive meta-indices in combination with a helpful graphical user interface for this increasing number of information sources are missing for most areas of science.

*Our DBLP (**D**igital **B**ibliography & **L**ibrary **P**roject) Computer Science Bibliography is a major service used by thousands of computer scientists. It provides fundamental support for scientists searching for publications or other scientists in similar communities. For better assistance we developed a new browser prototype which has a user-friendly interface and plays a central role in the search and browsing of the data. The DBL–Browser provides smart search functions and several textual and graphical visualization models. This paper gives an overview of some important research issues within the field of bibliographical information retrieval and visualization. After introducing the whole framework, the DBL–Browser itself and various visualization models are described.*

Categories and Subject Descriptors (according to ACM CCS): H.3.7 [Digital Libraries]: I.3.3 [Computer Graphics]:

1. Introduction

The rapid development of information systems throughout the last years and the advantages brought by them has led to an increasing amount of data which is available in a digital way. Even in the scientific and digital library area more and more new conferences, journals, and other publications become available on the World Wide Web. Cleverdon estimates the amount of publications of the most important scientific journals to 400,000 per year [Cle84] and INSPEC, the leading English-language bibliographic information service, is growing at the rate of 350,000 records each year [Sci03]. The enormous mass of data causes an information overload for the user. Searching for relevant and interesting publications turns out to be more like finding a needle in a haystack. It is no longer reasonable to assume that simply browsing will provide a lucky find. Especially for new users to a system or a certain domain it is difficult to map their information needs and vocabulary to the supplied vocabulary in the system [FLGD87]. Although a variety of sophisticated techniques such as thesauri and dictionaries are available to help users formulate their queries, only some of them are able to diminish the vocabulary problem [DCFQ00].

For the digital library domain we developed a combination of the query-based and browsing-based approach which in comparison with the mentioned practices is superior for solving the vocabulary problem. 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 visualized 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 domain. After a short overview of the whole framework the main ideas and features to be found in the DBL–Browser are presented. Then the individual visualizations are discussed in detail before we outline our current research issues.

2. Browsing and Visualizing Bibliographic Data

Searching for helpful information can be a difficult task when dealing with complex data. One of the parts within our research is to support users in these tasks through the use of a sophisticated user interface. Within the Digital Library domain we developed a modularized framework for querying, visualizing and browsing through bibliographic data.

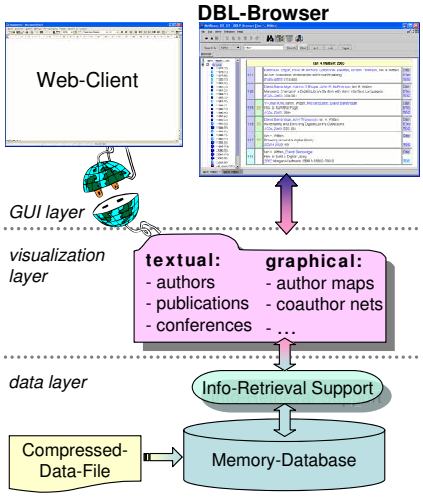

Figure 1: *Moduls of the DBL-Browser framework.*

2.1. Browser Framework

The whole framework as shown in Figure 1 is made up of three hierarchical layers for storing and retrieving, for visualizing and for browsing the data. The actual user interface, called the DBL–Browser, is at the top of the framework. The user interface and the supporting infrastructure are being developed in parallel to support the growing user and application requirements and due to its modularization additional features can easily be integrated.

2.1.1. Data Layer

The fundamental part of the whole framework, as shown in Figure 1, is the data layer as the basis. All requests coming from the user interface are answered by the Information Retrieval Support accessing the Main Memory Database and it in turn supplies all the data used for visualization. Rapid development and high performance are important issues.

The underlying data is first represented as plain XML files of a medium size, almost 200MB. As of now (February 2004), the bibliography lists nearly 480,000 publications and for more than 175,000 of them it provides direct links to the abstracts and/or full texts stored on the publishers' Web sites. For storing and copyright reasons some abstracts but no full text data of the publications are stored in the database. This 'raw' DBLP XML data is used as a test set by several researchers, for example [LS00, PG02].

For efficiency and performance reasons these files are converted into a compressed format which is then read into the main memory. From the main memory the compressed data is directly used within specially developed data structures. An adapted version of Huffman coding [WMB99] and other complex methods are used for the compression of text parts or numbers, respectively [Ley02].

With these techniques it is possible to load the complete bibliographical data of the DBLP database into the main memory of a computer and it requires just 65MB without any losses and without changing the structure of the data.

2.1.2. Visualization Layer

Even in the Main Memory Database the data is stored in a compressed form. Anytime data is requested it must be processed by the Visualization Layer to generate a human interpretable visualization. In this way the Visualization Layer serves as the interface between the user and the data and acts as a higher level entry point for the use of the DBLP data.

The Visualization layer allows the user interface to supply a servlet like query and provides a Visualization Result object for the user interface to display. Thus way the visualizations and user interface can be easily added to or exchanged.

2.1.3. Graphical User Interface Layer

The DBL–Browser is just one example of an interface for viewing and browsing the complete DBLP data. The browser offers some additional features which can not be realized in the current web page and is designed for use on a stand-alone computer (or laptop) without the necessity of a network.

Another interface example is that of a data acquisition module, which not only visualizes the existing data but gives also the opportunity to change and extend the data base by authorized users. Finally, a user interface could also be implemented using a client/server architecture where a browser is on the client-side and the complete retrieval and visualization package with the underlaying data is on the server-side.

2.2. DBL-Browser

Our first application to give comfortable, easy-to-use access to the complete DBLP data is the DBL–Browser. The DBL–Browser is implemented as a NetBeans module and is run on the NetBeans Platform. Through the use of modules, additional functionality can easily be added or removed from the interface. Additional functionality could be written by anyone wishing to use the DBLP data. Figure 2 shows the interface and an example visualization. The different aspects of the DBL–Browser will now be discussed, with all labels referring to the figure.

2.2.1. Viewing Data

The DBLP data and query results are displayed and browsed in a *browser container*, which is divided into two parts – the *filter* **(F)** and the *view window* **(V)**. When a query is performed the resulting data is visualized in the view window. The filter allows the data to be structured by different criteria, which is then visualized in the view window. In effect the Filter acts as a query refinement tool. Example criteria used by the filter are publication years **(1)**, coauthors **(2)** or publications in certain journals **(3)** or conferences **(4)**.

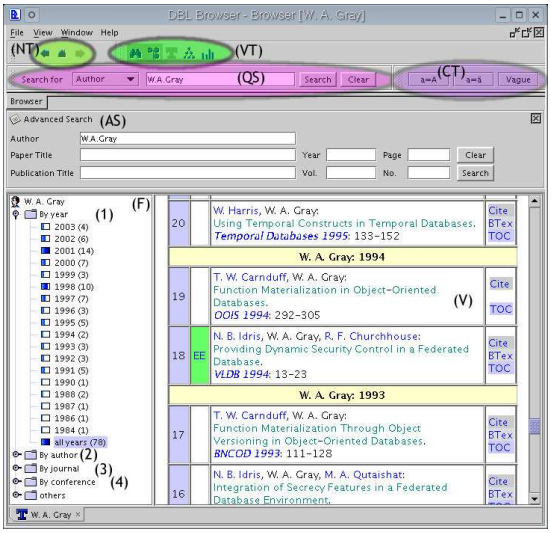

Figure 2: *Example view of the browser application*

Tabs are used to present multiple browser containers, allowing the user to independently browse different areas of the data at once, or the same data in different views. Browser containers can also be moved alongside each other so that a maximum of 5 containers are visible at the same time.

2.2.2. Features

The most important features of the DBL–Browser are the search fields. The browser provides search access in two ways – a *Quick Search* **(QS)** and an *Advanced Search* **(AS)**. The Quick Search is located on the toolbar to provide permanent access to search capabilities. It allows users to enter a query on a single data type, ie: author, title, journal or conference. The Advanced Search allows the user to enter more sophisticated queries, querying a combination of data types, as well as supplementary information such as the year or page number. Searches can be configured by the *configuration* toolbar **(CT)**. The user can specify whether the search should be case (in)sensitive and/or umlaut sensitive (a=ä) and/or exactly match the entered query. The configuration applies to the Quick Search and the Advanced Search.

The user can interact with the browser in two main ways. Firstly, a *navigation* toolbar **(NT)** consisting of the standard 'Back', 'Forward' and 'Home' buttons allows the user to navigate through their query history. In this way the browser supports and promotes the 'query process' rather than facilitating individual queries. The query process, search–browse–navigate, enables the user to fully explore a knowledge domain. The second interaction area is the *view* toolbar **(VT)**. From here users can control whether the Advance Search is visible (or not), whether the filter is visible or which visualization of the data they wish to see.

3. Visualizations

The DBL–Browser does not only use textual visualization to communicate data to users. The browser also includes a range of graphical visualizations. Some basic graphical visualizations show the textual data in a graphical form. More advanced graphical visualizations aim to expose *underlying relationships* in the data. At any stage the user can switch backwards and forwards between the different visualizations.

3.1. Textual Visualizations

The textual visualization presents data as HTML and consists of two main types – BibTeX/Table of Contents (TOC) pages and Author Pages. The data displayed by the text visualization corresponds in a large extent to the web pages of the DBLP (see also http://dblp.uni-trier.de).

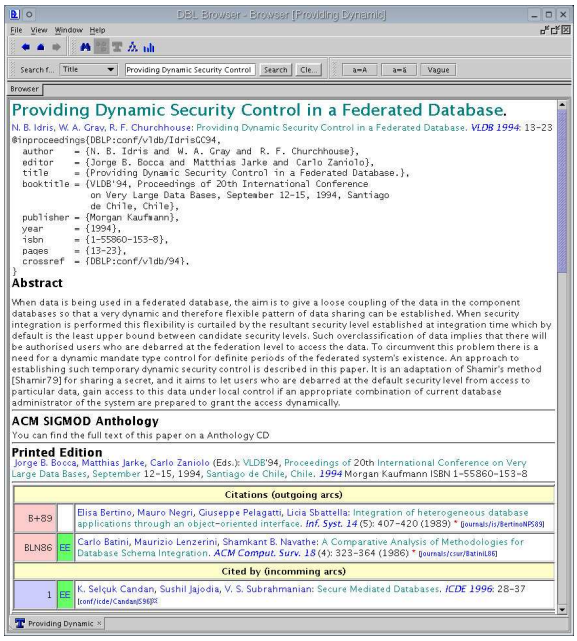

Figure 3: *BibTeX entry with abstract and citations*

BibTeX pages show all the available information for a publication (see figure 3). This includes the properly formatted BibTeX record, the abstract and citations/references. The TOC page visualization shows the formatted BibTeX record as well as session headings and the publication details of the accepted papers. Both the BibTeX and TOC pages are accessible from any visualization of publication details.

The Author Page, as in figure 2, is the most important visualization. It presents a number of publications in a HTML table of four columns. The first column shows the number of the publication for that author; the second indicates the presence of a link to an electronic edition; the third column gives the publication details; and finally, the fourth column

provides links to the BibTeX page and TOC page. Both the BibTeX and TOC pages allow the user to continue to explore the knowledge domain and find other 'similar' documents and related authors.

The publication details lists firstly all authors in the sequence given in the original paper. With a single click on a coauthor a new query is generated and all publications are searched and visualized analogously. With this behavior the user can easily overview coauthor networks. The next line shows the complete title. By clicking on a word, a search is initiated which returns all publications where this word occurs in the title. After the title, the name of the journal/conference and the page numbers are listed. By clicking on the journal/conference name it's TOC page is shown.

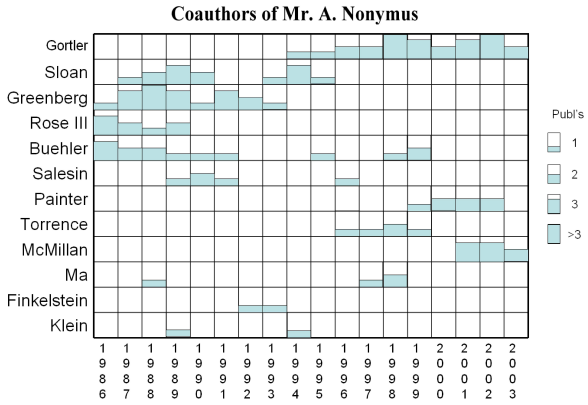

Figure 4: *Time Line of coauthors*

3.2. Filter

As introduced in section 2.2.1, the filter groups data by different criteria, the main criteria being publication years, coauthors, journals and conferences. Structuring the data in this way ensures that the user is not 'overloaded' with data that must otherwise be mentally filtered.

The filter uses different item (a leaf node) icons, where the icon used depends on the number of publications in that item. The filter in figure 2 shows the use of icons. The icons help to expose features about the next level of data, namely *how much* data is contained. In support of the icons, the actual number of publications in the item is included in brackets after the item name, exposing more specific details about the underlying data. Both features of the filter provide users with useful insights into the data. Firstly, the ability to view the range of the group. For example, to see whether the bulk of the publications were published recently (the researcher/research area is active) or a number of years ago (the researcher is not active). Secondly, the top item of the group can be identified. This highlights other interesting authors or important conferences in that research area and can lead the user to find more 'similar' publications.

Filter grouping of this type is used with the Author Pages and also when presenting search results containing publications. The filter is also used to provide links within the TOC and BibTeX pages to items such as session headings or abstracts, citations and references. This filter is not used to expose features of the data, but rather to aid in the internal navigation of a page.

3.3. Graphically Enhanced Visualization

3.3.1. Time Lines

Time line diagrams give an overview of progression over time and can be used for several topics. One example is a coauthor time line. Figure 4 illustrates the coauthor time line for the last 3 decades of Mr. A. Nonymus, with his 12 most common coauthors. The order of the coauthors can be sorted alphabetically or by the number of common publications.

Another example is a time line diagram of journal/conference streams within a specific subject where all related journals/conferences are visualized within one diagram. Again, the items can be sorted alphabetically, by the number of publications or by similarity to the subject of the group. A similar time line diagram is available for each author where all publications through time are displayed.

3.3.2. Histograms

One of the oldest graphical visualizations of the DBL–Browser is the histogram view. Originally the histograms were used to show a break down of an authors publications over time and to show the growth of the DBLP data [Ley02]. Histograms are also used to show all publications over time or publications over time for a specific publication type (conference, journals, books series etc.)

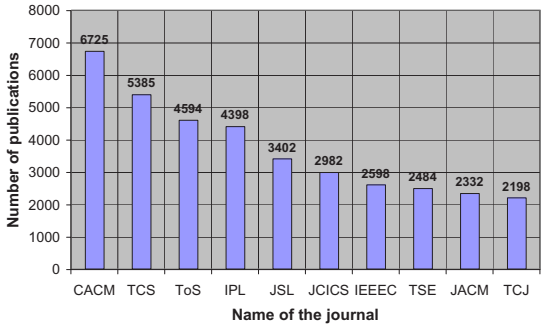

Figure 5: *Publications per journal.*

Another variation is to show the total number of publications in each journal or conference across the range of all

journals or conferences (figure 5). A histogram of this type allows a user to identify the top themes of the data set.

Additionally, histograms can be used to visualize search results. This allows the user to identify, among other observations, which results contain the most number of publications and thus provide a greater possibility of fulfilling their requirements. Just like the text visualizations, histograms can be browsed. For example, clicking on a journal (bar) in figure 5 links to a histogram showing the publications of that journal over time, and so on.

3.3.3. Graphs

Our work with graph visualizations has currently focused on two types – the Coauthor Relationship graph and the Conference/Journal Relationship graph.

Coauthor Relationship Graph

Work is currently continuing on visualizing the coauthor relationship. Because of the mass of authors listed in DBLP, it is more bewildering than useful to display the complete coauthor graph. In our approach we focus on local extracts of the graph. Using only a small part of the coauthor graph may be helpful to comprehend the surroundings of an author. The center of the built graph is a node representing the current author a_0 (see figure 6).

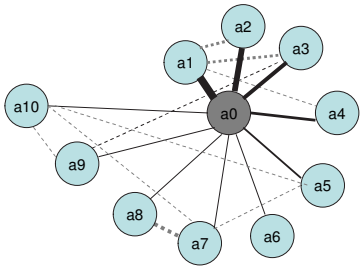

Figure 6: *Relations between the author and the coauthors.*

This node is connected to the most important coauthors $a_1 \ldots a_n$ by edges, where the thickness of an edge represents the similarity. In addition to displaying the similarity with the thickness of the edges, the similarity is also visualized by the layout of the surrounding coauthors. The nearer a coauthor is to the actual author, the more similar they are.

In accordance with the idea of a browsable user interface the constructed graph enables a user to navigate through the coauthor graph. By clicking on a coauthor, focus switches and the most important coauthors connected to the newly focused author are displayed.

Conference/Journal Relationship Graph

The Conference/Journal Relationship graph, figure 7, shows the conference or journal of interest (the principle) in the middle of the visualization. Located to the right are the conferences or journals that the principle is related to. On the left are the conferences or journals that are related to the principle. Section 4 further describes relationships with in the DBLP data. The numbers associated with each arrow reflects the ranking of the conference/journal at the arrow head when the conference/journal at the arrow tail is principle. The higher the ranking the closer the relationship, thus a ranking of 0 reflects that the conference/journals has the closest relationship to the principle. The user may switch to the graph view to find related conferences or journals. Clicking on the graphical object representing a conference/journal displays the corresponding Conference/Journal Relationship graph. The user may then switch back to the textual TOC page and look for publications or authors of interest.

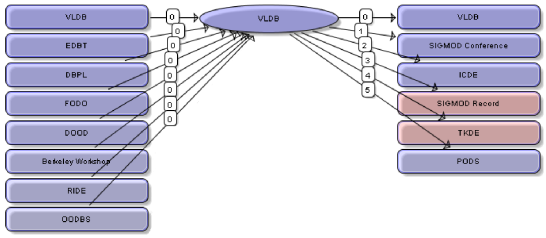

Figure 7: *Conference/Journal relationships*

4. Research on Journal and Conference Similarity

While the main focus of our research is visualization techniques, the data layer of the DBL–Browser framework is also subject to research.

A database of the size of DBLP contains an enormous amount of data, which can be subject of a thorough analysis. Probably there are many not yet discovered or obvious relationships which remain hidden in the database. Current research at our DBIS department tries to extract new information and relationships from the database.

Based on the *Vector Space Model*, the research tries to find a similarity measurement for journals or conferences which itself can be used to improve the users needs to find relevant documents [SM83]. Because of the rich bibliometric information supplied in DBLP there are many different possibilities to create a similarity measure.

A first approach for similarity-measurement of journals is to compare the terms appearing in the titles of the journals. This however is not very promising: Doing so, the journals "Digital Libraries" and "D–Lib", for example, would not be identified as similar, although they have the focus on the same research topic. More promising is a more complex procedure in which a journal is represented by the terms in the titles of the articles published within. First research shows reasonable results.

Besides the representation of a journal or conference by the articles, a representation based on the authors of the publication could be considered. By doing so, conferences or journals are supposed to be similar, if there are many publications from the same authors in the different journals.

Research has shown, instead of using the "author-based" and "term-based" approaches separately, a combination of both methods leads to better results. Finding a good weight for each possibility will be a challenging task. If one has a certain issue of a journal in mind, finding similar journals is probably more important for a researcher than finding out that the journal *A* is generally very similar to journal *B*. Therefore the above mentioned approaches should be applied to different levels.

Starting at a basic level, one could compute the similarity for an issue of a journal by using the "term-based" and "author-based" similarity measurements on the issue basis instead of the whole journal in general. Furthermore a hierarchical similarity measurement calculating the similarity of journals by first calculating the similarity of its issues could be taken into consideration.

The DBL–Browser currently uses a similarity measure between conferences and journals which is "author-based". A screenshot of such a relation can be seen in figure 7.

5. Summary

The modern information society faces a severe dilemma. More information than ever is available, but accessing relevant information is still a very challenging task. One reason for this problem is that users, especially novice users, are not able to formulate their query such that the system will interpret it the way the users want it to. With the DBL–Browser we have attempted to work against this problem.

By combining both textual and visual browsing functionality we established a browsing-based retrieval and visualization 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 organization and multilevel browsing systems. Both mentioned features are – as shown in the previous sections – central parts in our browser.

Encouragement

In a recently approved project we cooperate with the producer of CompuScience (FIZ Karlsruhe), the German Computer Society (GI, http://www.gi-ev.de/) and others. This research is supported by Federal Ministry of Education and Research (bmb+f) in the SemIPort project [AFGO*03]. The objective is to provide an open portal with improved coverage and additional services.

Our intention is to provide the DBL–Browser as a framework for experiments. Due to its modularization, it is an easy challenge for anyone interested to integrate his or her visualization 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 (http://dbis.uni-trier.de/DBL-Browser/). We encourage all of you to use and/or improve it. Feedback and further ideas are also welcome.

References

[AFGO*03] AGARWAL S., FANKHAUSER P., GONZALEZ-OLLALA 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., WEBER A.: Semantic Methods and Tools for Information Portals. In *GI Jahrestagung (1)* (2003), pp. 116–131. 6

[Cle84] CLEVERDON C. W.: Optimizing convenient online access to bibliographic databases. *Information Services and Use 4* (1984), 37–47. 1

[DCFQ00] DING Y., CHOWDHURY G. G., FOO S., 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 (2000), 1190–1204. 1, 6

[FLGD87] FURNAS G. W., LANDAUER T. K., GOMEZ L. M., DUMAIS S. T.: The Vocabulary Problem in Human-System Communication. *CACM 30*, 11 (1987), 964–971. 1

[Ley02] LEY M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *SPIRE* (2002), pp. 1–10. 2, 4

[LS00] LIEFKE H., SUCIU D.: XMILL: An Efficient Compressor for XML Data. In *SIGMOD Conference* (2000), pp. 153–164. 2

[PG02] POLYZOTIS N., GAROFALAKIS M. N.: Statistical synopses for graph-structured XML databases. In *SIGMOD Conference* (2002), pp. 358–369. 2

[Sci03] SCIENCEDIRECT: About the Abstract Databases–INSPEC. http://help.sciencedirect.com/robo/projects/sdhelp/about/inspec.htm, '03. 1

[SM83] SALTON G., MCGILL M. J.: *Introduction to Modern Info. Retrieval*. McGraw-Hill, 1983. 5

[WMB99] WITTEN I. H., MOFFAT A., BELL T. C.: *Managing Gigabytes: Compressing and Indexing Document and Images*, second ed. Morgan Kaufmann, San Fransisco, California, 1999. 2