BibRelEx: Exploring Bibliographic Databases by Visualization of Annotated Contents-Based Relations

Anne Brüggemann-Klein[1] Rolf Klein[2] Britta Landgraf[2]

October 31, 1997

Abstract

Traditional searching and browsing functions for bibliographic databases do no longer enable users to deal efficiently with the rapidly growing number of scientific publications. We propose to visualize contents-based relations among documents such as *cites*, *succeeds*, *improves with respect to*, and to use them for more effective exploration. In addition, we encourage users to attach annotations that may be either private or public to documents or to pairs of related documents. The aggregation of public annotations contributed by expert users represents insight into the area that exceeds the knowledge that is represented in the documents themselves. In this paper we report on the status of our project BibRelEx, and invite discussion about its future goals.

Keywords: Bibliographic database, citation, annotation, information system, information retrieval, expert knowledge, navigation, visualization.

1 Introduction

It is well known that the number of scientific publications is growing at an exponential rate. New scientific journals appear as new areas of research emerge, and numerous conferences and workshops are publishing their own proceedings. In addition, technical reports and internet publications comprise an important source of information.

How to cope with this infamous information flood has become a problem even within one's own central area of knowledge and expertise, and all the more so for non-specialists from other disciplines. As a consequence, results are being re-invented, and good solutions are not put to proper use.

In principle, bibliographic databases should be well-suited to overcome these difficulties. In practice, however, they are not. The reason for this lies mainly with the limitations of the conventionally provided access methods, namely browsing or searching.

[1] Technische Universität München, Fachbereich Informatik, Arcisstr. 21, 80290 München, Germany.

[2] FernUniversität Hagen, Praktische Informatik VI, 58084 Hagen, Germany.

Usually, the browsing function is based either on a fixed classification scheme that may be too coarse, or not concise enough, for locating a desired topic, or whose terminology is not sufficiently understood by the user. Or browsing is based on author-defined hyperlinks that are often lacking systematic structure.

The searching function, on the other hand, requires the user to be acquainted with both the query language and the terminology used. Very often, searching large databases is too inefficient, a problem well known from WWW search engines. Both browsing and searching return candidate documents that match some external criteria (for example, all documents appearing in the same subcategory of a classification scheme, or all documents containing certain terms.) Human experts, however, do not primarily think in these terms. To them, the documents' contents matter more than external criteria do. The experts know about relationships between the documents, and their view of the area is built upon them. From studying the documents' contents and their relationships, insight can be gained that may exceed the knowledge represented in the documents themselves.

Which are the essential relationships between scientific documents? The most basic one is *A cites B*. Yet, it allows some interesting questions to be answered. *Which are the survey articles in a given area?* Those who cite a big proportion of all papers in this area. *Which contributions are fundamental?* Those who are cited by many.

The answers to the previous queries could, in principle, be output in textual form, as lists. The following question demonstrates the value of a graphical visualization of relationships. *Which further work has been influenced by a given publication A?* The result set for this query could be visualized by a graph whose nodes are the papers directly or indirectly citing document A and whose edges express the *cites* relation, resulting in an at-a-glance picture of A's sphere of influence.

Whereas the *cites* relation can be extracted automatically from the documents, there are other useful relationships only an expert human reader can establish. Examples are *A improves on results from B*, *A deals with problem P*, or *A uses method M from B*. In addition, one would like to be able to attach a short summary, written by an expert, to a document, or a more detailed annotation to two documents that are related to each other.

Our project *BibRelEx* will be the first to help the user explore a bibliographic database by visualizing, and exploiting, contents-based relations among its entries. We are building BibRelEx on top of the database *geombib*, which currently contains about 8700 bibliographic references in the area of computational geometry.

Roughly, we are proceeding in three steps. First, we want to add as many interdocument relationships as possible to the database. This task involves also the design of several tools, for example for detecting multiple entries and for maintaining public and private local versions of the database under periodic updates. Next, we want to provide the users with suitable tools for visualizing, and navigating through, graphs that are defined by the above relationships. The goal is to provide a three-dimensional image of the set of all relevant documents and of the references among them. Finally, we want to enable and to encourage the users to annotate both single documents and pairs of related documents. Such annotations may either be private, and thus belong to one's personal workspace, or they may be public, representing insight shared with the scientific community.

The rest of this paper is organized as follows. In Section 2 we discuss related work in this area. Next, in Section 3, we briefly introduce the bibliographic database *geombib*. Section 4 presents our project BibRelEx in more detail; we report on its current status and discuss its further goals.

2 Related work: citation databases

Following up the references contained in a scientific document has always been an essential part of systematic study. It is not surprising that several systems or projects exist that aim to support this activity.

Well known is the *Science Citation Index*. For each document it lists all papers that cite this document. Unfortunately, only a limited subset of journals is covered; conference proceedings or technical reports are not contained at all. A license for the Science Citation Index is quite expensive. The format of its entries is closed. It is not possible for the user to add annotations.

R. D. Cameron [Cam97] suggests building a *universal citation database* that links all publications ever written according to the citations among them. He discusses a distributed approach using different servers for different sources such as journals, conferences, etc. He proposes to form a consortium of universities, academic societies, and library associations to establish such a system.

D. M. Jones [Jon95] has originated the *Hypertext Bibliography Project*. A Web-based system employs hypertext links to establish citations, and sets up one Web page per document containing a list of links to papers that cite it. This project is already well under way. It covers a selection of major journals and conferences. Key word search and access by author name are supported. For many documents, their abstracts are available on-line.

Our project *BibRelEx* differs from these approaches in the following way. We do not limit ourselves to retrieving all documents that contain a reference to a single query document; instead, we want to display the citation graph—or suitable subgraphs—of a whole area, and we want to enable the user to navigate through this structure. In addition, the user shall be able to annotate documents and inter-document relationships, as mentioned in the introduction. On the other hand, we do not seek to cover the whole area of computer science in this project. Instead, we focus on a bibliographic database of moderate size that already exists and is well accepted and widely used. This database, *geombib*, is briefly introduced in the following section.

3 The bibliographic database geombib

The database *geombib* currently contains about 8700 bibliographic entries in the area of computational geometry. It has originated from merging two private bibliographies and is now being maintained by B. Jones of the University of Saskatchewan.

The database is updated by a community effort, in the following way. Typically, users downloads the whole database which consists of a single file. To a copy of this local version, they can now add new entries of recent publications, or correct incomplete or erroneous entries. After four months, the resulting file is compared against the original one, and the differences are submitted to the administrator. From the data submitted, a new release is compiled and distributed one month later.

This approach works surprisingly well, thanks to the effort of many researchers who do not even shrink from entering whole proceedings volumes. As a result, the database covers most of the existing publications in this field, technical reports and workshop proceedings included.

The entries are in BibTeX format, with pre-defined fields for author, title, publisher, etc. Queries may specify the contents of any subset of fields. Each database entry is assigned a unique key, derived from the authors' names, the initial characters of the words in the title, and the year. In order to cite a paper in one's own work, one simply cites its key; the bibliography is automatically generated from the list of keys by BibTeX. This useful fact, and the up-to-dateness of geombib, are among the most important reasons for its success in the computational geometry community.

4 The BibRelEx project

Once the decision for building BibRelEx on top of an existing database had been made, two consequences were immediate. First, the system has to be designed in such a way that geombib in its current form and BibRelEx can coexist; geombib as is has to remain fully operational. Second, we also have to provide a critical mass of citation information and annotations, so that geometers can use the new exploration methods to their advantage.

4.1 Getting started

In order to exploit contents-based relations among documents for navigation, the relations themselves must be present in the database. Geombib does already provide, for each database entry, optional fields named *cites, precedes, succeeds, annote*, but less than ten percent of them have so far been filled in.

Therefore, we have started our project by entering the citations contained in the contributions to the Proceedings of the Annual ACM Symposium on Computational Geometry. At present, we have covered the years '85, '90, '91, '92 and '96. While most of the papers published in the proceedings were already represented in geombib, a lot of the cited work was not; together, about 950 new entries were generated and about 2200 links filled in.

While entering new entries into the database it turned out that one needs a tool for discovering multiple entries that refer to the same document. During a geombib update, duplicate keys are detected automatically. However, it is not uncommon that data-entry errors result in duplicates that give rise to different keys. This happens, for example, if the authors' names have been entered in different order, so that the keys generated by geombib differ. Also, misspelling the paper's title can result in different keys.

In order to detect such duplicates, we have developed the tool *BibConsist* [Lan97]. It checks if the corresponding fields of two entries have similar contents. Here, D. Knuth's soundex code [Knu73] is used for measuring the phonetic similarity of words. Using BibConsist we were able to detect about 70 duplicates in geombib. The real value of BibConsist, however, is in preventing the user from newly entering faulty records into the database.

In the four-months period between two geombib releases the user typically maintains a local version of the database that contains all updates suggested by the user; compare Section 3. In addition, many users are maintaining their own bibliographic databases containing the data of such publications that are interesting to them, and likely to be cited in their own work, but do not necessarily belong to the area of computational geometry (for example, a textbook on topology from which but one theorem is quoted).

In order to keep these different databases consistent, we are currently developing the tool *BibManage*. Each time a new release of geombib appears BibManage checks if it contains all updates suggested by the user, and if some documents contained in the personal bibliography do now appear in geombib, due to some other user's request. Between updates, BibManage guarantees transparency, in that all local databases are treated like a single one. In addition, a user interface for entering new records into either database is provided [Nau96].

4.2 Further work

With respect to the technical aspects of this project, the next step is to design the system's architecture. We will have to decide on a suitable visualization tool that can be installed locally and enables the user to visualize, and navigate through graph structures. To this end, we would like to make use of standard WWW browsers and their graphical capabilities.

In order to guarantee sufficient speed it will be necessary to store the geombib contents locally in a suitable database system. We will have to provide functionality for updating this database with new geombib releases and for exporting changes on database entries as input for a geombib update.

With respect to the contents, we will start entering annotations to papers, and to pairs of related papers, in the areas of Voronoi diagrams and on-line algorithms. In particular, we are adding as much as possible of the information contained in the survey by Aurenhammer and Klein [AK97].

Once a critical mass of information is available, we intend to advertise our system in the computational geometry community, and invite researchers to enter annotations of their own. Given the interest the community has so far been taking in geombib, and the reactions to our first presentation of the BibRelEx project at this year's Dagstuhl seminar on computational geometry, we are confident that the response will be positive.

Another interesting question is how to maintain the database contents over time. We feel that this task should no longer be left to volunteers. Rather, we suggest that the authors themselves submit, along with their papers, a proposal for an annotated geombib entry to a conference or to a journal. The referees could, without additional effort, check whether the proposed entry is correct and complete, and forward it to the database manager. This approach, which has also been suggested by Cameron [Cam97], is also in the authors' interest.

References

[AK97] F. Aurenhammer and R. Klein. Voronoi diagrams. To appear in J.-R. Sack and J. Urrutia (ed), Handbook on Computational Geometry, Elsevier Science Publishers B.V., 1997.

[Cam97] R.D. Cameron. A universal citation database as a catalyst for reform in scholarly communication. *First Monday*, 2(4), 1997. URL http://www.firstmonday.dk/issues/issue2_4/cameron/index.html.

[Jon95] D.M. Jones. The hypertext bibliography project. URL http://theory.lcs.mit.edu/~dmjones/hbp/info.html, 1995.

[Knu73] D.E. Knuth. *The Art of Computer Programming*, volume 3: Sorting and Searching. Addison-Wesley, Reading, Massachussetts, 1973.

[Lan97] B. Landgraf. BibConsist: A program to check BibTeX files for inconsistencies. URL http://wwwpi6.fernuni-hagen.de/wwwpi6/Forschung/BibRelEx/BibConsist.html.en #BIBCONS, 1997.

[Nau96] G. Naumann. Ein System zur Verwaltung von benutzereigenen Datenbeständen als Ergänzung zu einer öffentlichen Literaturdatenbank unter periodischen Updates. Masters thesis, Fernuniversität Hagen, Fachbereich Informatik, 1996.