BIBLIOMETRIC FACTOR MAPS FOR KNOWLEDGE DISCOVERY IN DIGITAL LIBRARIES

Andreas Strotmann[1]*; Dangzhi Zhao*[2].

[1]School of Public Health, University of Alberta
Edmonton, AB, T6G 2G3, Canada
e-mail: andreas.strotmann@ualberta.ca
[2]School of Library and Information Studies, University of Alberta
Edmonton, AB, T6G 2J4, Alberta, Canada
e-mail: dzhao@ualberta.ca

Abstract

In this paper we describe the architecture of a visual bibliometric browsing plug-in for the growing number of digital libraries that provide cited references in their document meta-data, using a simple but effective visualization method for citation network analyses we recently introduced. Citation-based network analysis methods such as co-citation analysis have long been recognized as effective tools for gaining insight into the intellectual structure of a field through its literature. Visualizations of these networks can help the user get an intuitive aggregated overview of the field and the interrelationships between documents or authors, which in turn can aid query expansion, search refinement, and exploratory browsing. Our design calls for a visualization of the results of a multivariate factor analysis of a bibliometric similarity matrix calculated from a user's search results and/or from documents that are closely related to them. This provides the user a digital library with an interactive map of the literature that the user is interested in, where each visual element aggregates different aspects of the search result (authors and/or subfields). By helping the user see the forest for the trees (i.e., a structured visual landscape of the intellectual domain covered by the user's search and its bibliometric vicinity rather than a long list of search results), these maps and the relevant links they contain promise to provide a valuable aggregated browsing tool for digital libraries.

Keywords: bibliometric information retrieval, aggregated search, citation indexing, citation analysis, factor analysis.

1. Introduction

As full-text documents that include reference information, along with autonomous citation indexing tools, are becoming available, digital libraries and repositories are starting to provide cited references as part of their metadata, and the time has come to study ways to enhance users' experiences in open digital libraries and repositories via bibliometric analyses. Citation analysis results can not only help understand scholarly communication structures but also add significant value to information retrieval (IR) in digital libraries. For example, evaluative citation analysis results can help retrieve high quality documents and publications by core players (authors, institutes, countries, etc.), and relational citation analysis results can help expand queries to resulting clusters of documents, authors and subareas.

A growing number of literature retrieval systems, e.g., those of the Institute for Scientific Information (ISI), Scopus, Google Scholar, and CiteSeer, have demonstrated the value of incorporating citation analysis results into IR systems by providing the number of citations each document receives, various indicators of journal quality, and links to document sets and/or authors that are related to the current search through cocitation or cited-by links. However, they do so in the form of lists rather than visual aggregates.

In this paper, we propose to add automatic *visualizations* of bibliometric co-citation and bibliographic coupling analysis results as a compact and informative visual enhancement to digital library systems such as the Eprint archive system. Using the user's current search result – a set of documents or authors – as a basis for bibliometric analysis, we propose to provide the user with visual maps of the search result and its bibliometric vicinity in the digital library (Figure 1). Specifically, following a technique introduced in [1,2], we propose to visualize the results of a multi-variate factor analysis of core authors whose work is provided by a digital library, based on co-citation or on bibliographic coupling similarity matrices. Author nodes in these visualizations link to information that a user likely was not aware of when formulating the original query, aiding knowledge discovery. It is possible to use these visualizations as a complete browsing tool for a digital library by visualizing analysis results for the full database, and making it available on the digital library's home page.

Our approach is unique among bibliometric IR systems in providing *aggregated* visual browsing, where search results are visualized at two levels of aggregation simultaneously, namely on the oeuvre level, which aggregates documents by their authors, and on the intellectual subfield level, which aggregates authors by the degree to which they are cited together in the literature, for example. This is made possible by performing a factor analysis and visualizing factor nodes along with author or document nodes.

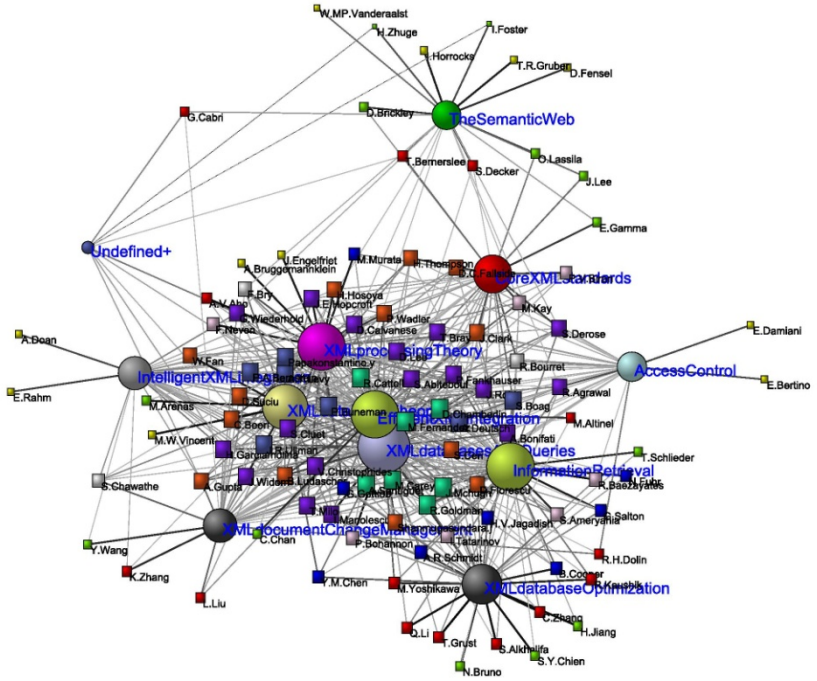

Figure 1: A bibliometric factor map of the XML field 2000-2005(cited authors): visualization of an author co-citation analysis (structure matrix).

2. Review of bibliometric information retrieval systems

Citation indexes such as SCI, Scopus, Google Scholar, and CiteSeer play two important roles: they are both a unique information retrieval system and a data source for citation study of science and technology. Results from citation analysis studies cannot only help understand scholarly communication structures and processes but can also aid information retrieval. For example, evaluative citation analysis results can help retrieve high quality documents and publications by core players (authors, institutes, countries, etc.), and relational citation analysis results can help expand queries through resulting clusters of documents, authors and subareas.

A growing number of literature retrieval systems, e.g., those of the Institute for Scientific Information (now Thomson), Scopus (by Elsevier), Google Scholar, and CiteSeer, demonstrate the value of incorporating citation analysis results into information retrieval systems by providing such information as the number of citations each document receives, various indicators of journal quality, and links to document sets and/or authors that are related to the current search through strong

cocitation or cited-by links. However, they do so largely in the form of long lists rather than visual maps.

Bibliometric information retrieval systems, by contrast, seek to make full use of bibliometric techniques and results to help solve problems currently facing IR systems [3]. Most frequently, these systems use visualizations to dynamically present concept networks produced through word analysis, or to show document or author networks [3,4,5,6,7]. These networks can help the user get an overview of a field or the interrelationships between concepts, documents or authors, which helps query expansion and search refinement. However, a fully "bibliometrics aware" IR system that combines evaluative with relational bibliometric analysis to aid searching is still not available.

Although it is not difficult to understand the benefit of bibliometric analysis (especially citation analysis) to IR systems, research on bibliometric IR systems has so far been focused on word analysis and concept networks. The application of the best-known bibliometric technique – citation analysis – and the oldest one – bibliometric coupling – has yet to be explored, partly because most IR systems do not include references, and partly because almost all databases that do include references are proprietary.

One notable exception to this rule is *AuthorLink* [7,8,9], a system designed for visual browsing of authors linked via co-citations to their works in a subset of the Humanities database by the Institute for Scientific Information (ISI). This system allows the user to search the database by an author's name, identifies in the entire database those authors who are highly related (i.e., highly co-cited) with this author, and displays a visual map of these authors based on a co-citation graph, using Pathfinder network pruning or clustering via Kohonen self-organizing maps. Although this system shows promising results in enhancing users' IR experience via a visual interface to an IR system based on co-citation network analysis, its adoption of a particular commercial database (i.e., ISI databases) has limited its impact.

In [3], the authors describe *BIRS* (for bibliometric information retrieval system). Like AuthorLink, BIRS is designed as a stand-alone IR system, accessed via the web. In addition to word net and document citation network visualizations made available through their system, BIRS implements a classic author co-citation analysis visualization in the form of a Multi-Dimensional Scaling map, the kind of map traditionally used in ACA, e.g. in [10]. The bibliometric visualizations offered by BIRS as an aid for researching the literature available on the web are quite impressive. However, its very power makes it hard to imagine its use as a search plugin for an e-print server, although it might work as a separate search interface for one. As [4] points out, however, the MDS scaling maps provided by BIRS as a visualization of an author co-citation analysis impose unnecessary limitations.

In [4], the author describes a system for visualizing rather large-scale author co-citation analysis results. This system's visualization is based on a multivariate factor analysis of an author co-citation count matrix, to which it applies Pathfinder network pruning. This approach allows the system to visualize an impressive number of authors in an intellectual map which successfully highlights important "bridging" authors in a field.

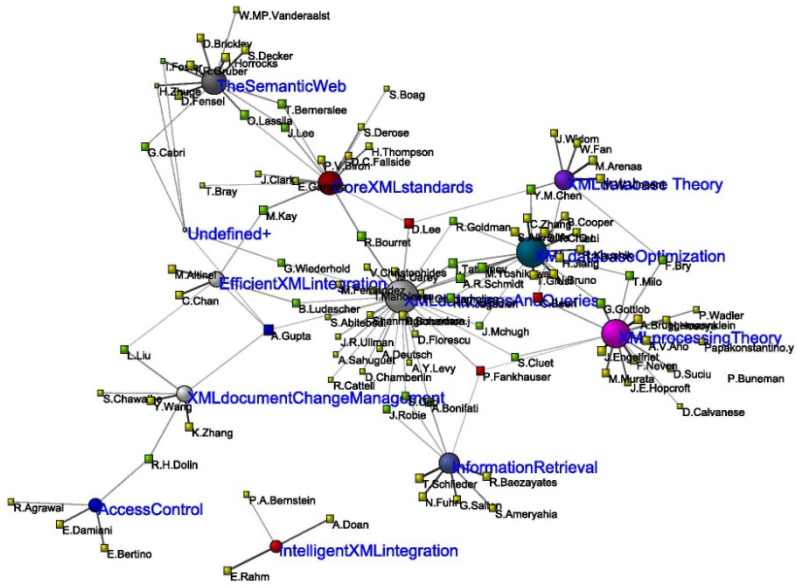

Figure 2: A bibliometric factor map of the XML field 2000-2005(cited authors): visualization of an author co-citation analysis (pattern matrix).

However, that map, even though laid out in two dimensions, reduces the intellectual structure to an almost one-dimensional tree view, which removes all the intricate interrelations between authors and factors that characterized the original factor analysis results. Pathfinder pruning may trim the links in the visualization to *the* most important ones, but it does so, at least in those presented in [4], at the expense of removing *all* the secondary links, including those that provide potentially interesting bridges between the different areas.

All three systems, and admittedly the sample visualizations we present in this paper, are based on data extracted from commercial databases. However, as full-text documents that include references, along with autonomous citation indexing tools, are becoming available, digital libraries and repositories are beginning to provide cited references as metadata, and the obstacle imposed to the development of bibliometric IR systems by data sources is fading away. The time has come to study

ways to enhance users' IR experience in open digital libraries and repositories through integrating bibliometric analysis results in their web-based user interfaces.

3. Bibliometric Factor Map Examples

In this paper, we propose one such approach which adds bibliographic coupling and co-citation analysis as a compact and informative visual tool for browsing digital libraries. This approach is unique among the bibliometric IR systems reviewed here in that it provides aggregated visual browsing, where search results are visualized at two interrelated levels of aggregation simultaneously, namely at the oeuvre level, which aggregates documents by their authors, and at the intellectual subfield level, which aggregates authors by the degree to which they are cited together in the literature. To illustrate the idea, we describe an example of such an IR enhancement from the user's perspective. We will then discuss its architecture and design further below.

Figures 1-4 show visualizations of search results in ISI's Web of Science for the topic "XML" during 1996-2005. Of these, figures 1 & 2 visualize results of an author co-citation analysis, and figures 3 & 4 those of an author bibliographic coupling analysis. Authors on the maps are "core authors" in the search results set – the top 120 authors ranked by the number of times they each have been cited (figures 1-2) or the top 120 authors ranked by number of publications (figures 3-4).

This means that figures 1 & 2 include authors who have been crucial in laying the theoretical or practical foundations of this field without necessarily ever having published in it, thus providing opportunities for the user to focus on high-impact authors, or to widen the search to related research fields (e.g., from a search for XML to the field of IR in semi-structured databases). In the author bibliographic coupling maps (figures 3 & 4), by contrast, it is the citing authors of the topic area that are mapped, so that these maps provide tools for focusing on the research areas that many currently active authors share (e.g., the Semantic Web in this example).

Figures 2 and 4 visualize the pattern matrix of a factor analysis with oblique rotation, which shows the active research areas in the XML research field and the core authors in each area, and figures 1 and 3 the structure matrix, which show the overall structure of the XML research field and the interrelationships between the active research areas. The factor nodes (i.e., research areas) in figures 1 and 2 have been hand-labeled; in figures 3 and 4, we show that even without such labeling, the visual structure of the field is available to the user, including the factor nodes that, in this case, simply denote the field that is made up of the authors they connect to.

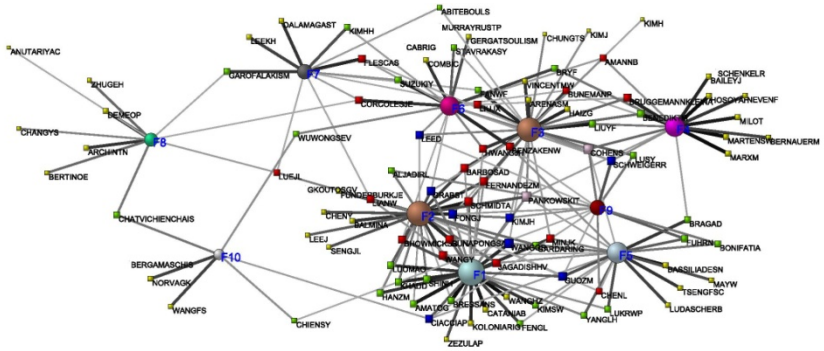

Figure 3: Author bibliometric coupling analysis visualization of an SCI subject search for *XML* in 2000-2005 (citing authors) - structure matrix.

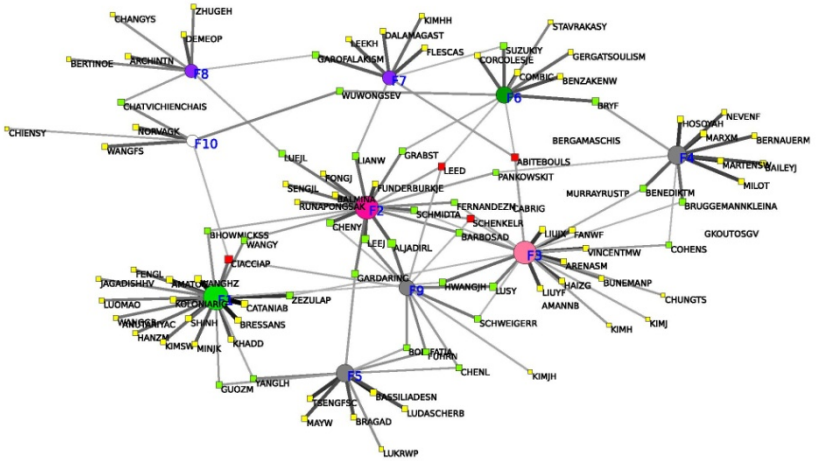

Figure 4: Author bibliometric coupling analysis visualization of the same search (citing authors) - pattern matrix

When the user performs the search *XML or eXtensible Markup Language* in a digital library (in the sample visualizations, in SCI), the user will get to a "topic page" that displays, in addition to the standard list of document hits as in most current digital libraries, one of the four maps shown in Figures 1-4, along with a concise description of the meaning of the map, and with a clear indication of the option to show any of the other three maps.

When an author node is clicked, the user will be taken to an "author page," the structure of which is similar to a "topic page" but lists all articles written by this author, along with lists of authors who are highly related to this author by co-citation or bibliographic coupling. The maps on the author page visualize the interrelationships between these related authors.

In general, when a factor node on the map is clicked, the user will reach a ``topic page'' that displays the list and maps of a *subset* of the original search results. Selecting a factor node thus allows the user to zoom in to a particular research area. Alternatively, in a "zoom-out" mode, clicking the factor node could present the user with an author-search page that lists all authors with significant loadings on that factor.

Such a visual IR interface adds to the current ones that always provide search results as ranked lists in a number of ways – first, it is visual and thus more compact and, for some people, more intuitive than a pure list view; second, it lays out the members of a list in a map that informs the user about how these people are related to each other with respect to the intellectual structure of the topic area that the user queried; and third, it adds, via factor nodes, the ability for the user to focus on a subfield of the topic area. These aggregates are mapped out intuitively, based on significant interrelationships between them, and both levels of aggregation are displayed simultaneously.

For a visual browsing plugin to a digital library search engine, screen space is a serious issue, of course, so that a plug-in of this kind will lay out a smaller number of top-ranked authors than are available in the list view, but it can link to a larger and more detailed visualization and can display detailed figures as shown here with about 100 author and a dozen factor nodes each.

4. Architecture of a Bibliometric Visual User Interface

Regardless of the specific analysis provided, the core flow of information in a bibliometric visual user interface for a digital library along the lines that we outlined above will be largely the same:

- the current view of the user is determined, along with a set of documents that correspond to that view,
- a citation graph is constructed from this document set,
- the nodes in the graph are ranked using bibliometric measures,
- a number of top-ranked nodes are chosen (the number depends on the size of the visualization on the screen),
- a matrix of bibliometric similarity measures between nodes in this graph is calculated,
- a multivariate factor analysis with oblique rotation is performed on the matrix
- the resulting factorization is visualized as a graph of richly interlinked nodes. Author or document nodes are of a size determined by their bibliometric ranks, and links between these nodes and factor nodes of a thickness and/or grayscale

value proportional to the loading of an author or document to a factor.
- the resulting 2-dimensional map is rendered and annotated with hyperlinks that correspond to appropriate digital library search queries for each node in the graph.

Different bibliometric relationships correspond to different kinds of maps of intellectual field structures. A map created from an author co-citation analysis based on the current search result set as citing papers, for example, maps authors that most heavily influenced the research covered in the current result set. With a document bibliographic coupling analysis, on the other hand, we map the interrelatedness structure between the documents in the current result set; in this case, the factor nodes of the graph allow the user to focus her search on a subset of the current search result.

There are a number of different maps one can produce in this general manner, by varying the bibliometric relationship measured (co-citation, co-phrase, bibliometric coupling), the unit of analysis (document, author, or others), or the set of documents analysed (current search result as citing papers, or papers referencing those in the current search result).

We are currently undertaking a series of investigations regarding the relative merits of the different choices available to the implementer of a bibliometric mapping extension to a digital library system, and it is not a priori clear which choice will work best for the user of such a service.

Both document and author level analyses have their merits – document-based analyses providing a more detailed map than author-based ones, while author-based analysis may make it easier to see the forest for the trees. Any one of the different bibliometric relationships is of value, as each answers a different question: co-citation measures who or which paper has been influential in this area; bibliometric coupling, who has similar research interests or which papers are on a similar topic; and co authorship, which research groups/invisible colleges can be identified.

5. System Design

In general, digital library systems such as e-print servers or institutional repositories serve as IR and archiving systems for scientific literature, often offering extensive meta data on the contents of the archive or repository. Recently, interest has been surging in providing metadata of a particularly interesting kind – cited references from the literature the digital library offers.

It is therefore now possible for an independent researcher to create a real bibli-

ometric IR system, and to deploy it in an environment that makes it possible to evaluate its usability and usefulness in a realistic setting. A proof-of-concept implementation can be realized using off-the-shelf open software systems – our first target digital library is E-LIS, the Library and Information Science e-Print Archive, which is realized using the open EPrints server.

Our initial target implementation will therefore be in the form of an EPrints server plugin, which enriches the search and browse interfaces of that system with the visual browsing tools described above.

We have already developed software, written as Python scripts, that constructs author citation graphs from document sets and produces an author co-citation or author bibliometric coupling matrix. So far, we have been using SPSS for the subsequent factor analysis, but open-source statistical analysis software like GNU R can be adapted to this purpose. For the automatic labeling of factor nodes in a first prototype, we plan to use a simple *tf*idf* significance weight to rank phrases in a factor's most significant document set. More sophisticated automatic labeling techniques may be added later. For the visualization, we have so far been using the Pajek implementation of the Kamada Kawai layout algorithm. For server-side realization, general-purpose graph handling libraries are available from a number of resources.

In order to reduce the server-side load, and in order to increase the response time on the user's side, it is possible to off-load both the graph visualization and the factor analysis to the user's browser, although in the latter case automatic labeling would be difficult to do.

Our initial target implementation is somewhat complicated by the fact that reference linking and citation indexing are not yet fully standardardized ingredients of the EPrints software system and of the OAI protocol that it implements. Nevertheless, the E-LIS archive, an EPrints server, does index the citations of documents in its archive; it just does so very simply, namely by extracting the cited reference strings from papers, and storing them as metadata with each document's database entry. We aim for an add-on to E-LIS which adds E-LIS-internal visual browsing, and expect, in case of a successful implementation, that this will provide some impetus towards more thorough E-LIS internal citation indexing, which in turn will allow us to provide more complete visual bibliometric browsing to its users.

6. Conclusions

As digital libraries or institutional repositories for the scientific literature begin offering cited reference data from their holdings, it becomes possible to add visual

browsing based on well-established bibliometric analysis methods. By visualizing the intellectual structure of a field in the aggregate, factor analysis based author citation network analysis visualizations offer an intuitive overview, and by including factor nodes in the visualization, they offer a unique knowledge discovery tool for users of a digital library.

A proof-of-concept implementation can be realized using off-the-shelf open software systems – our first target digital library is E-LIS, the Library and Information Science e-Print Archive, which is realized using the open EPrints server, to which we propose to add a factor map generation plug-in.

Notes and References

[1] ZHAO, D. and STROTMANN, A. Information Science during the first decade of the Web: an enriched author co-citation analysis. Journal of the American Society for Information Science and Technology, 59 (6), 2008, p. 916-937.

[2] ZHAO, D. and STROTMANN, A. All-author vs. first-author co-citation analysis of the Information Science field using Scopus. In *Joining Research and Practice: Social Computing and Information Science;* Proceedings of the American Society for Information Science and Technology 2007 Annual Meeting, October 19 - 24, 2007, Milwaukee, Wisconsin, USA.

[3] DING, Y., CHOWDHURY, G.G., FOO, S. and QIAN, W. Bibliometric information retrieval system (BIRS) : a web search interface utilizing bibliometric research results. Journal of the American Society for Information Science and Technology, 51, 2000, p. 1190-1204.

[4] CHEN, C.M. Visualizing semantic spaces and author cocitation networks in digital libraries. Information Processing & Management, 35, 1999, p. 401-420.

[5] CHEN, C.M. et al. Alleviating search uncertainty through concept associations: automatic indexing, co-occurrence analysis, and parallel computing. Journal of the American Society for Information Science, 49, 1998, p. 206-216.

[6] CHEN, C.M. et al. Internet browsing and searching : user evaluations of category map and concept space techniques. Journal of the American Society for Information Science, 49, 2007, p. 582-603.

[7] LIN, X., WHITE, H. and BUZYDLOWSKI, J. Real-time author co-citation mapping for online searching. Information Processing and Management, 39, 2003, p. 689-706.

[8] WHITE, H.D., BUZYDLOWSKI, J. and LIN, X. Co-cited author maps as interfaces to digital libraries : designing pathfinder networks in the humanities. In *IEEE International Conference on Information Visualization*, London, July 18-22, 2000, p. 25-30.

[9] WHITE, H.D. User-controlled mapping of significant literatures. Proceedings

of the National Academy of Sciences, 101, 2004, p. 5297-5302.

[10] WHITE, H.D. and MCCAIN, K. Visualizing a discipline: an author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science, 49, 2007, p. 327-355.

June 2009
Printed on demand
by *"Nuova Cultura"*
www.nuovacultura.it

Book orders: ordini@nuovacultura.it