PERGAMON

Information Processing and Management 39 (2003) 689–706

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Real-time author co-citation mapping for online searching

Xia Lin *, Howard D. White, Jan Buzydlowski

College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA

Received 10 February 2002; accepted 29 April 2002

Abstract

Author searching is traditionally based on the matching of name strings. Special characteristics of authors as personal names and subject indicators are not considered. This makes it difficult to identify a set of related authors or to group authors by subjects in retrieval systems. In this paper, we describe the design and implementation of a prototype visualization system to enhance author searching. The system, called AuthorLink, is based on author co-citation analysis and visualization mapping algorithms such as Kohonen's feature maps and Pathfinder networks. AuthorLink produces interactive author maps in real time from a database of 1.26 million records supplied by the Institute for Scientific Information. The maps show subject groupings and more fine-grained intellectual connections among authors. Through the interactive interface the user can take advantage of such information to refine queries and retrieve documents through point-and-click manipulation of the authors' names.

Keywords: Author co-citation analysis; Information retrieval systems; Author searching; Kohonen feature maps; Pathfinder networks

1. Author searching

Whether searching in traditional libraries hundreds of years ago, or looking up a reference on the Web today, people have traced writings through the names of persons who created them. Author searching is a mode of retrieval long familiar to anyone who needs to find a work in a collection of documents or bibliographical records. It is now usually done through a computerized index that is optimized for string matching. During the lookup process, author names are treated as text strings. The computer system can quickly identify whether there is a match for a given string in the author index and deliver the records in which the match occurs.

* Corresponding author. Tel.: +1-215-895-2482; fax: +1-215-895-2494.
E-mail address: xlin@drexel.edu (X. Lin).

0306-4573/03/$ - see front matter
doi:10.1016/S0306-4573(02)00037-7

However, users of the system will need to do a lot of additional work if they want rapid identification of other authors in the same subject area as their input author. They will need to expend still more effort if they want to know the subject areas of the given author and related authors. This need not be so. With today's computational power, the computer can—and should—do much more than provide a simple string match during author searching.

Long ago, Cleveland (1976) verified that, in certain kinds of retrieval systems, author names are strong indicators of subject content in documents. Borko and Bernier (1978) observed that, if time is used as a variable, author indexes can also "show what authors are doing, whether they have changed their subject fields, and just how active they are." It is well known that certain bibliographical databases include citation counts, co-citation counts, subject areas of the authors who cited other authors, and so on. The computer should be able to exploit such facts.

We here describe a new approach that makes use of some of these facts to enhance author searching. What the user gets from our kind of retrieval is much more than records that match an author's name. Specifically, our Web-based prototype system, called AuthorLink, accepts a single author's name as input and provides as output a list of 24 authors who are most often co-cited with the input author. It then shows how these additional authors are related to each others in maps based on their co-citation counts. The maps, which are created on the fly, enhance the user's understanding of author relationships and their subject areas. Moreover, they can be used directly in retrieval.

AuthorLink produces the interactive maps in real time from a non-trivial database supplied by the Institute for Scientific Information (ISI): the Arts and Humanities Citation Index (AHCI) for 1988–1997, which has about 1.26 million records (the publisher granted this data set to the authors in ASCII text format for research purposes). The maps show subject groupings and more fine-grained intellectual connections among authors, based on the perceptions of citers in the journal literature covered by AHCI. Through the interactive interface the user can take advantage of such information to refine queries and retrieve co-citing documents through point-and-click manipulation of the authors' names.

AuthorLink is operational on a password-protected Web site run at Drexel University. It grows out the second author's research program of two decades, as stated in articles and book chapters such as White (1990a, pp. 104–105):

> Although many of its separate steps are algorithmic, co-cited author analysis is still labor intensive and time consuming. If it could be more easily done—that is, if the separate steps could be integrated as one, smooth-flowing, economical machine process—its practical applications in subject retrieval and in mapping scientific or scholarly fields would lead to wide use. The two applications meet, for example, in the preparation of reviews of literatures, a widely practiced form of scholarly writing. *** What we need is a technology that would rapidly allow reviewers anywhere to create their own maps, using known key authors as input, and to retrieve the "research front" papers that co-cite those authors. As I wrote some years ago (White, 1983, p. 312): "Co-cited author maps help verify 'natural' groupings of the literature (the clusters) upon which sections of the review can be based. Moreover, if included as graphics with published reviews, they illustrate relationships and offer discussion points that otherwise might be missed." The separate pieces of a "reviewer's technology" existed then, and they do now, but no one has yet brought them together.

AuthorLink brings them together. In the following section, we briefly review literature on author co-citation analysis (ACA), which is the foundation of our approach. We then describe our design and methodology, followed by a discussion of the prototyped system. Finally, we present some examples to illustrate how the system enhances author searching in different situations.

2. Traditional ACA

ACA was developed to study the structure of literatures (White & McCain, 1989, 1997). It has been used to explore the intellectual structure of specialties within science and scholarship for more than 20 years (White & Griffith, 1981). The basic assumption of ACA is that works by any two authors that are jointly cited (co-cited) in later works reflect some connection between the two authors. The more frequently they are co-cited, the more closely they are related. The analysis of frequency patterns of co-citations of a group of related authors will reveal their salient linkages, as well as the subject areas they represent both individually and collectively (Chen & Carr, 1999a,b; Ding, Chowdhury, & Foo, 1999).

ACA involves several procedures to collect co-citation frequencies and map those data statistically. McCain (1990) outlines typical procedures for traditional (“Drexel-style”) ACA:

- Selection of authors,
- retrieval of co-citation frequencies,
- compilation of a raw co-citation matrix,
- conversion to a correlation matrix,
- multivariate analysis of correlation matrix (using principle components analysis, cluster analysis, and multidimensional scaling),
- interpretation: authors’ names on the maps can be translated into subject terms naming the clusters. Clusters and the dimensions (axes) on which they are placed reveal the intellectual structure of a field.

The procedures involve many manual processes as well as several different computerized systems, such as DIALOG for literature searching, SPSS for statistical analysis, and DeltaGraph for visualization. Until recently, it was doubtful whether the procedures could be integrated and implemented in a real-time search environment. First of all, co-citation analyses require access to ISI citation databases and a rapid search engine. Generating co-citation frequency counts and the co-citation matrix typically involves hundreds or even thousands of ANDed-pair queries to the database. Secondly, statistical mapping procedures such as multidimensional scaling and factor analysis are computationally intensive and difficult to integrate with a search system. Thirdly, the results of statistical computation need a separate graphical system if they are to be visualized. This further complicates the building of an integrated system.

Because of the difficulties of processing and mapping author co-citation data in real time, ACA maps have not been used in past retrieval systems. Nevertheless, it has long been clear that ACA maps could serve users who want to retrieve works citing the interconnected authors. Properly implemented, they could be an effective aid to subject retrieval, usually leading to documents that

conventional subject searches would not find. The challenge is how to utilize ACA in a practical retrieval system and generate ACA maps in real time.

To the best of our knowledge, no practical retrieval systems except AuthorLink do this (cf. the account in Boyack, Wylie, and Davidson (2001)). Chen (1999) applied ACA to several collections to visualize the "semantic space" of the collections, including both authors and subjects. While the mapping results are useful, he used different systems for indexing and for mapping, not aiming to link ACA maps to a real-time retrieval system. The BIRS interface of Ding, Chowdhury, Foo, and Qian (2000) was designed to work with practical search engines. However, from the description in their paper, it is not clear whether their mapping is done in real time.

More distantly related is the Butterfly system of Mackinlay, Rao, and Card (1995). This was a live interface to real-time retrievals based on ISI citation data, but it used individual papers as its unit of presentation, showing for any paper the references it cited (one "butterfly wing") and the later-published items citing it (the other "wing"). This system seems to have remained at the experimental stage.

3. Design principles and features

Our demonstration project applies ACA and information visualization (IV) to a practical retrieval system. Some compromises and choices had to be made during the design stage. To guide our design, we imposed these principles:

- Minimal cognitive load on the user, who should be able to create maps by inputting a single author's name—not a long list of authors.
- Exploration of co-cited authors based on recognition and learning instead of foreknowledge.
- On-the-fly generation of an author co-citation matrix.
- Completion of mapping procedures within seconds.
- Abbreviated representations of an author's two-dimensional space (only his or her top 24 co-citees are shown).
- Practical visualization interface for content exploration and discovery.
- Tight integration of mapping procedures and search engine.
- Interactivity: the user can manipulate screen objects to achieve goals.
- Exploitation of a rich, real-world database that covers substantial literatures over a long period of time.
- Record retrieval from the full database.
- Applicability of interface to databases of any size, with little correlation between processing time and size of databases; good chance of scaling up to other non-toy applications.
- Immediate understandability of information on screen; no extensive explanations, codes, or legends.
- Optionally displayable counts of the documents co-citing pairs of authors in the map.

Many decisions of the system design are based on these principles. For example, we limited the number of names on the screen to 25 after considering processing time and clarity of display. We

applied mapping procedures to search results, rather than the whole database, to ensure the upward scalability. For the sake of immediate understandability, we used unadorned authors' names, simultaneously displayed, as icons on the graphical display, as opposed to, e.g., geometric objects with different shapes or colors that have pop-up labeling. This last decision pays homage to design criteria stated by Tufte (1983); cf. White and McCain (1989).

Many interactive functions are implemented on the interface. For example, the user can drag and adjust author labels that overlap or that are partially out-of-bounds (both may happen when labels on a map faithfully represent underlying data structure). The user can create a second map based on *two* focal authors rather than one by clicking on one of the co-citees in the first map. The user can also move author labels to a search box (by clicking or drag-and-drop) to construct a query automatically. This simple and practical interaction allows the user to take advantage of visual displays to retrieve documents.

An important design feature in such retrievals is that the focal author whose name was used to create the map is automatically placed in the "Main Author" box at top right of the interface and will always be ANDed with any other names placed in the "Additional Authors" box just below it when a search is carried out (White, Buzydlowski, & Lin, 2000). The effect is that any search on any name in the map retrieves articles that also cite the focal author. This is particularly useful in maps of major creative artists. For example, as described in White, Lin, and Buzydlowski (2001), any authors chosen as a basis for retrieval in Oscar Wilde's map will be co-cited in the context of Wilde studies (or in contexts that at least mention his work).

4. A sample search

AuthorLink uncovers the complex associative relationships of author co-citation, simplifies them, and makes them understandable. An example will illustrate how such associative relationships help the user.

A college student wants to learn more about Nobel Laureate Hebert A. Simon. Through a traditional string-match search in an ISI database, she would get either a list of documents authored by Simon, or the papers citing Simon's papers. She would need to read a lot of those papers to understand that Simon made significant contributions in many areas, including computer science, psychology, economics, and linguistics.

If she used AuthorLink, she would start the same search by entering Simon-HA (in fact, she could enter either "Hebert A. Simon" or "Simon, Herbert A."; the AuthorLink interface will automatically convert them to Simon-HA, which is ISI's author format; we use this format for retrieval and display). What she would get is a map like Fig. 1. This map shows Simon as seen through the eyes of many researchers who cite his work over the 10-year period covered in our database. It identifies those authors co-cited most often with him, and groups them by their specialties (although the specialties are not automatically labeled). It thus invites the searcher to explore and interact with the specialty literatures that draw on Simon's writings in different subject areas. For example, the searcher could single-click on a name, say "Newell-A," to add it to the query box. On hitting the Submit button, she would get a new map as in Fig. 2(a). It sets Simon in the context of authors who clearly represent artificial intelligence and cognitive science, because that is what the Newell–Simon connection connotes.

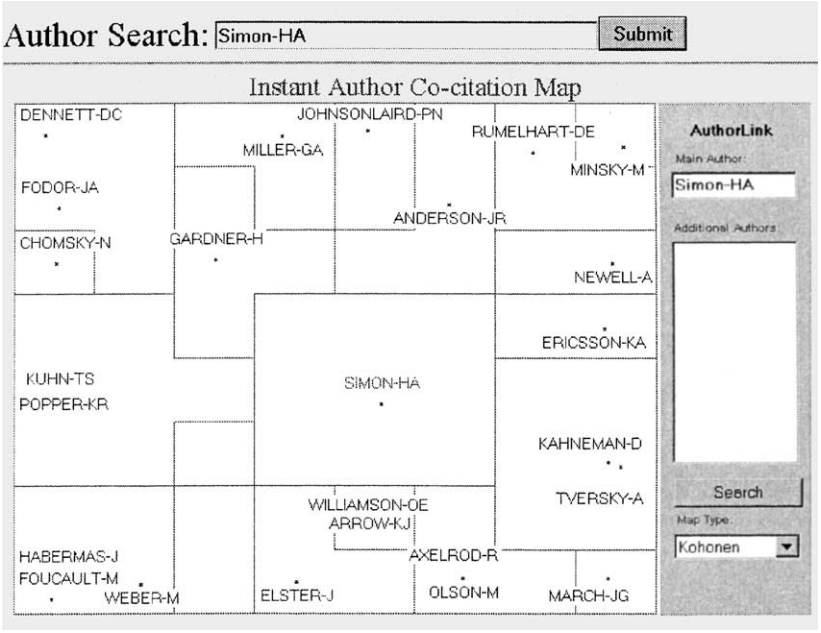


Fig. 1. An author map for Nobel Laureate Herbert Simon, displayed in Kohonen format.

She could explore Simon's economic side, or learn about it for the first time, by adding Kenneth Arrow to the search box (like Simon, Arrow is a Nobel prizewinner in economics). The map in Fig. 2(b) now shows an almost completely different set of authors in which economists predominate; psychologists and AI people have almost completely disappeared.

In both of these cases, Newell's and Arrow's names are the only ones that need to be chosen for a search; Simon's is already ANDed in. The reason for this is a design principle mentioned above: we do not want authors to be combined independently of the focal author. To illustrate, the names of two figures who are themselves very highly co-cited, Thomas Kuhn and Karl Popper, appear in Fig. 1. If a user put them in the "Additional Authors" box for searching and our system entered them without Simon, hundreds of documents unrelated to Simon would be retrieved. If the search is from *Simon's* map, we require that any search also involve him; this greatly reduces the large set of documents co-citing the Kuhn–Popper pair.

Note that the maps synthesize information that is not otherwise readily available in the database. In fact, as described in the next section, the system needs to run hundreds of queries against the database and put the results through several analytical procedures to map the major associative relationships. To acquire similar understanding of those relationships without the maps, a user would need to search, read, and interpret hundreds of documents—a burdensome effort.

Finally, she can simply click on the "Search" button to retrieve a list of all the documents citing Simon and Arrow (Fig. 2(c)). If, after reviewing the map, she finds another author interesting to her, she can add the author's name to the search box, and receive a new list of documents citing all three authors.

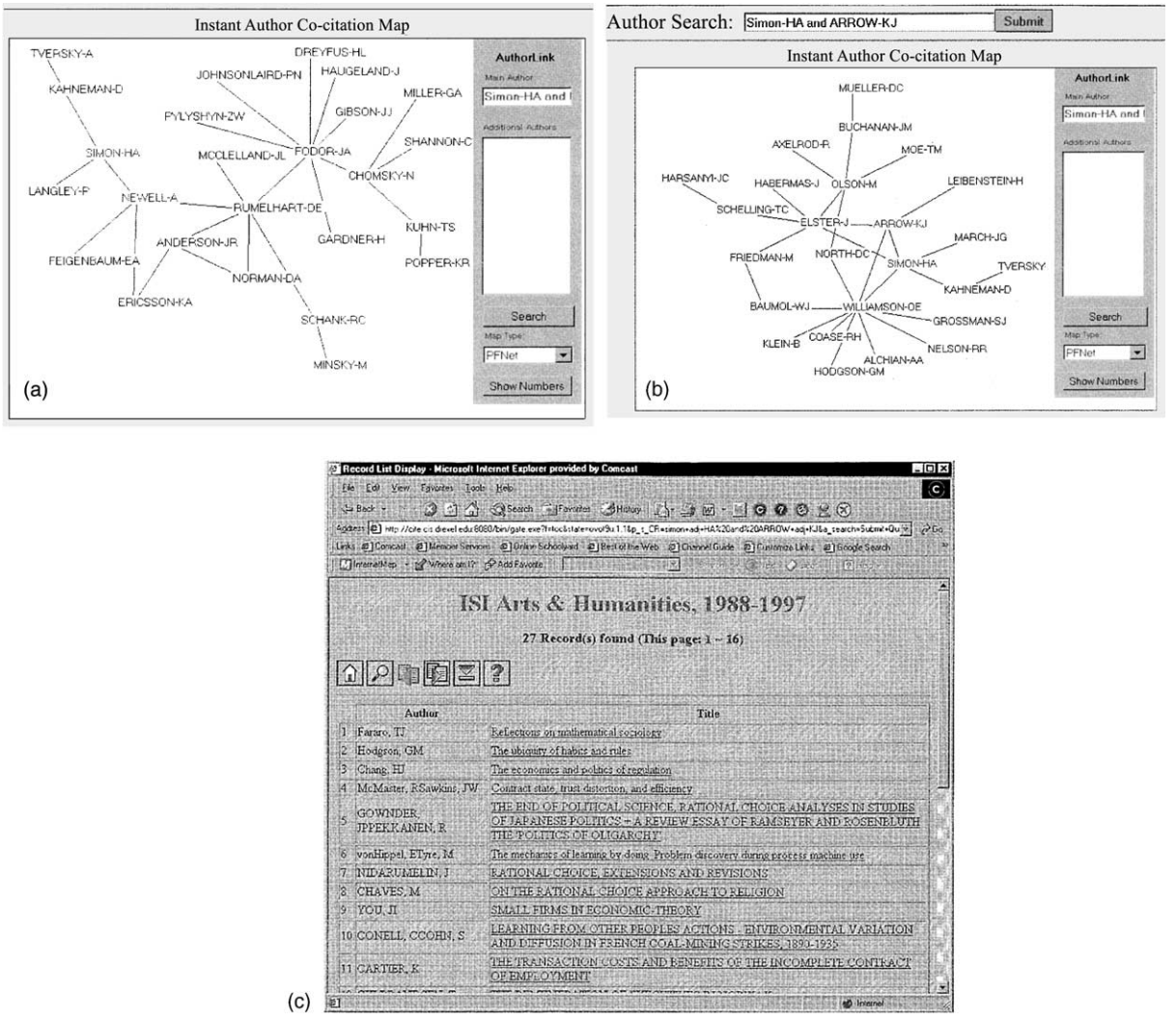


Fig. 2. (a) PFNET for authors co-cited with Simon and Newell. Many of these authors write on AI and cognitive science—one area of Simon's contributions. (b) A PFNET for authors co-cited with Simon and Arrow. These authors are mostly economic theorists—another area to which Simon contributed. (c) A list of documents citing Simon and Arrow.

5. Technical details of AuthorLink

Web-based AuthorLink has four major components: (1) an html-based front end for searching, (2) ACA procedures, (3) mapping procedures, and (4) an interactive graphical interface (Java applet). All four components are controlled through a Java application server.

The html-based search front end is generated by Java servlets. It accepts the user's input for an author search, converts the input to a query in the ISI-data required format, and sends the query

to BRS/Search, a commercial search engine that we purchased for the AuthorLink project. ISI data include only last names and initials for authors. As noted, the front end accepts queries with full names, like "Herbert A. Simon," "Simon, Herbert A.," and "Simon, HA." They are all automatically converted into "SIMON-HA" which is the query format for BRS/Search.

The ACA procedures are a set of C programs that interact with BRS/Search. When a focal author is entered, the set of 24 other authors who are co-cited most often with the focal author will be generated, using the BRS command TALLY for rank ordering. When the user chooses to map the 25 authors, another $25(24)/2 = 300$ BRS queries will be automatically sent to BRS to obtain author co-citation counts for every non-duplicated pair of authors in the list. The result is a symmetric author co-citation matrix that can be used for mapping.

Mapping procedures take the co-citation matrix as input and generate maps as output. A Java applet interface visually represents the mapping results. Currently, we have implemented two mapping procedures—a self-organizing map (SOM) and a Pathfinder network (PFNET). Both algorithms reduce the complexity of input data while retaining informative relationships.

In PFNET, explicit links are drawn to show co-citation relationships (Appendix A). The PFNET algorithm creates a fully connected graph in which only the "least-cost" paths between authors are drawn. The algorithm regards authors as nodes and assumes a graph in which all nodes are completely connected by weighted paths. The weights in this case are the co-citation counts for each pair of authors. After summing the weights, PFNET eliminates all paths except those with the lowest weights. The latter turn out to be the highest (or tied-highest) *single* co-citation counts for any pair, because combinations of counts as path weights sum to more. These least-weight or least-cost paths are drawn as links, throwing into relief citers' perceptions of what authors go together most strongly. This usefully simplifies the network.

The SOM (Kohonen, 1997) is a type of neural network in which similar data are gradually moved toward each other to form "semantic neighborhoods" (Appendix B). On a SOM display, each author is represented by a dot and a text label. Author clusters and neighborhoods are represented by geographical regions and their proximities. Authors in the same regions are co-cited most often. Authors in regions next to each other are co-cited more frequently than those authors in regions further apart. The size of the regions indicates relative citation frequencies among authors on the map. The larger the regions, the more frequently the authors inside the regions are co-cited with other authors on the map. The regions, sizes, and locations are all automatically determined by the algorithm based on input co-citation patterns.

6. Further search examples

In this section, we demonstrate in several ways how AuthorLink might be used to help searchers. Advantages include:

- AuthorLink puts a selected author in his or her most relevant neighborhood of authors.
- AuthorLink presents an overview of a field or a subject area.
- AuthorLink lets the user observe intellectual connections of authors and discover unsuspected linkages.

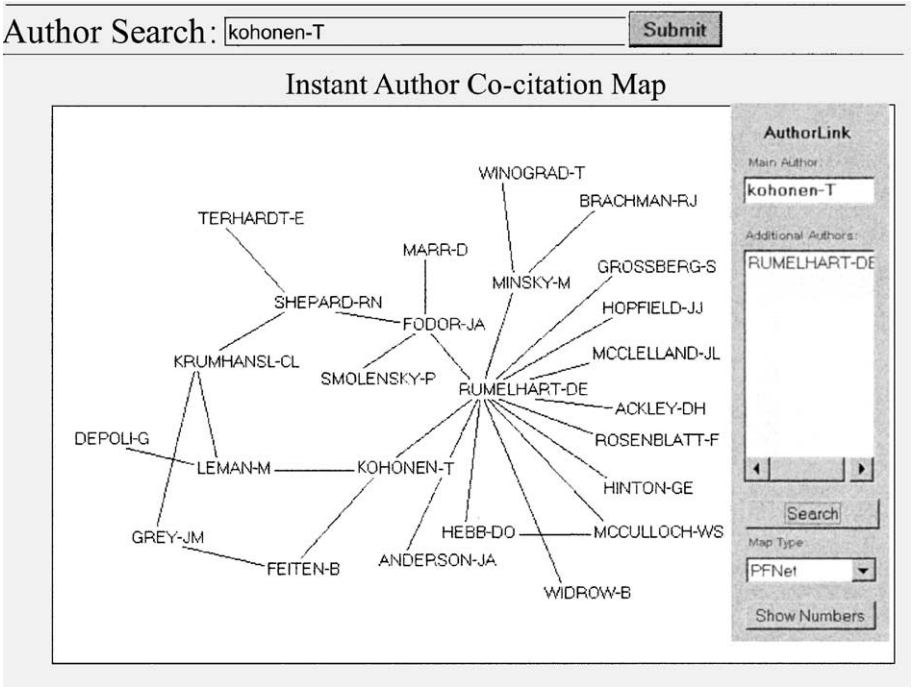


Fig. 3. The PFNET map for T. Kohonen.

- AuthorLink can distinguish similar author names that are otherwise conflated in ISI data.
- AuthorLink makes it easy for the user to explore unfamiliar territories.

AuthorLink puts an author in most relevant neighborhood of related authors: A user unfamiliar with a particular author can map him or her and try to recognize some of the authors nearby. Thus, a map can identify someone's intellectual home area by implying subject-matters related to the selected author's writings. For example, a user unfamiliar with Teuvo Kohonen's work requested an author map for "Kohonen-T" through AuthorLink (Fig. 3). Recognizing some of the other authors, he remarked, "Oh, there are many connectionists here." He quickly concluded that Kohonen's work was related to connectionism and neural networks. As one instance, this author map reflects the fact that *Parallel Distributed Processing*, the two-volume classic by David E. Rumelhart and James L. McClelland, was often cited in the context of references from arts and humanities articles to neural networks and connectionist research.

AuthorLink presents an overview of a field or a subject area: This is the focus of most past ACA research. In AuthorLink, there are only 10 years of data from AHCI, and so the disciplinary scope is relatively limited. Nevertheless, the maps provide instant, content-rich overviews. For example, a user interested in citation indexing studies started by entering "Eugene Garfield," the creator of citation indexing. After looking at a list of the 24 authors co-cited most often with Garfield, she decided to add "Small-H" to the query, since Henry Small is known for his contributions to citation analysis. Then she mapped the result. The authors co-cited most often with Garfield and Small in AHCI surround them in Fig. 4 (Small is cited as both "H" and "HG"). Several clusters

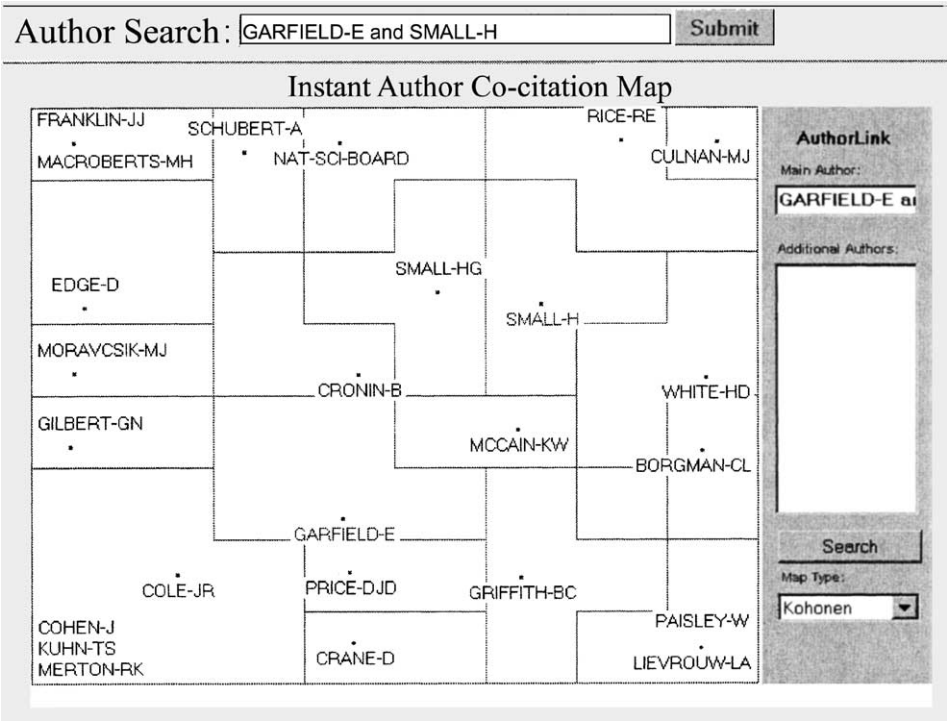


Fig. 4. Overview of authors co-cited most often with E. Garfield and H. Small.

of authors are seen from the viewpoint of the humanities. By selecting different authors to add to the search box (the large blank panel at right), the searcher could explore several research areas. For example, at lower left are sociologists and historians of science influential in citation studies (Moravcsik, Gilbert, Cohen, Kuhn, Merton, the Coles, Crane, and Price). At upper left are names associated with the use of citations as science indicators (Schubert and the National Science Board) and also some authors who have debated the merits of citation analysis (Franklin, pro; MacRoberts and Edge, con).

The interests of the citationists in the right half of the map are implied by the fact that they all contributed to the book *Scholarly Communication and Bibliometrics* by Borgman (1990) or the special ACA issue of the *Journal of the American Society for Information Science* (White, 1990b).

AuthorLink lets the user observe intellectual connections of authors and discover unsuspected linkages: We have shown authors their own maps and asked for their reactions. They typically say things like "This is very interesting; I knew many of the authors near me except so-and-so," or "This is what I expect, except that I was surprised that so-and-so was not here." As a result, they usually dragged the unfamiliar names on the maps to the search box and retrieved the articles. This let them discover new connections and citations they were not aware of before. For example, the second author of this article, who appears at right in Fig. 4, knew all of the names in the Garfield-Small map and can make reasonable conjectures as to why they are appear as they do.

However, he had to do a search to identify one of the co-citees in his own map, the anthropologist Nicholas G. Blurton-Jones, linked to him by Pamela Sandstrom in her studies of information foraging. Needless to say, this broadened his horizons.

AuthorLink can distinguish similar author names that are otherwise conflated in ISI data: ISI renders authors solely as surnames and initials. This makes it impossible to distinguish homographic authors through their names alone. For example, Albert Einstein the physicist and Alfred Einstein the music historian would be searched by same query, "Einstein-A," and the result always mixes their co-citees. One user used the query "Einstein-A" to search for Albert Einstein. He saw that Mozart was one of the names most often co-cited with Einstein. Out of curiosity, he added Mozart to the query. He observed that the map changed significantly—most of the names were now related to music rather than to physics (Fig. 5(a)). Through the search engine, he learned that one "A. Einstein" is a musicologist and Mozart scholar (Einstein, 1962). Going back to search on "Einstein-A," he added another physicist, Niels Bohr, to the query. The resulting map was what he was looking for—the intellectual world of the other "A. Einstein," the famous, bushy-haired father of relativity theory (Fig. 5b).

AuthorLink makes it easy for the user to explore unfamiliar territories: A graduate student needs to write a paper on Lao Tzu, the ancient Chinese philosopher and author of the *Tao te ching.* Coming fresh to the matter, he uses AuthorLink to explore the scope of readings on this topic. In his author map (Fig. 6), he sees three groups. One centers on Lao Tzu, including Chuang-Tzu and Huai-Nan-Tzu. The second centers on Confucius, with Mencius, Han-Fei-Tzu, Arthur Waley (a famous translator of Chinese classics), and so on. The third includes some western scholars like A.C. Graham and Joseph Needham. Through interacting with the search engine and retrieving some of the journal articles citing these scholars, the student quickly learns that they translated or commented on major works of early Chinese philosophers. Their views have had a major impact on how Chinese thought is viewed in the Western world.

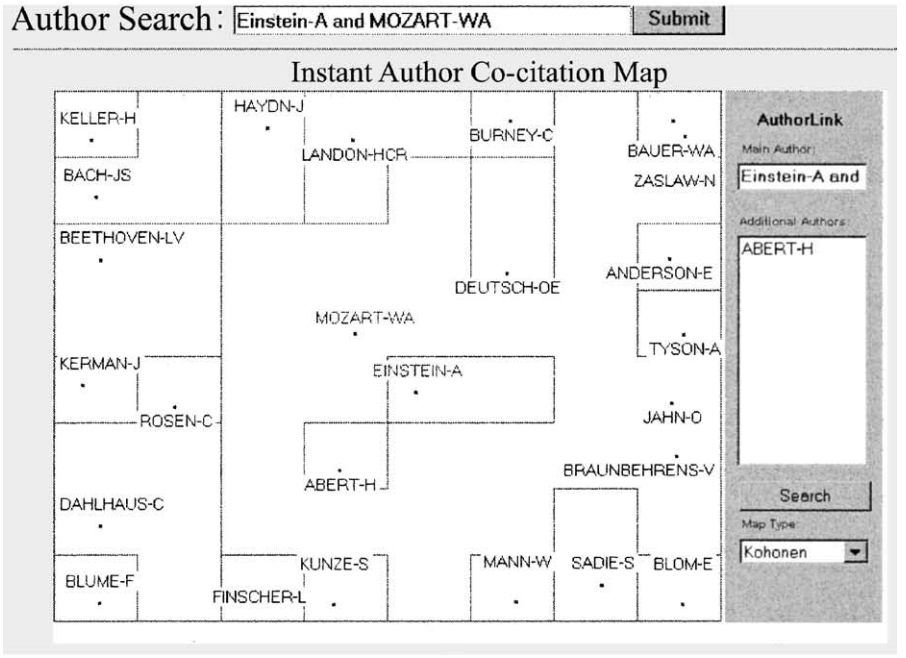
The student continues to interact with the search engine to collect readings for each of the three clusters. In the end, he is able to develop a good reading list on an unfamiliar field, as if guided by experts.

7. Conclusion and future work

Author searching was important in the past for libraries and bibliographical systems. It will remain important in future document retrieval systems and the Web. To better utilize the computational power, however, we need to make author searching go much deeper and provide more information. Our study suggests that author searching should no longer be limited to obtaining works *by* an author. Much more information can be provided to the user, particularly when author co-citation information is available. The AuthorLink system described here opens the possibility of using ACA in a practical search system. With AuthorLink, co-citation ties among authors could become one of the best guides for finding and understanding intellectually related works.

AuthorLink still needs to be tested by users. In Buzydlowski's dissertation (in preparation), he is comparing how subjects interact with the Kohonen map and PFNET map. He is also collecting

(a)



(b)

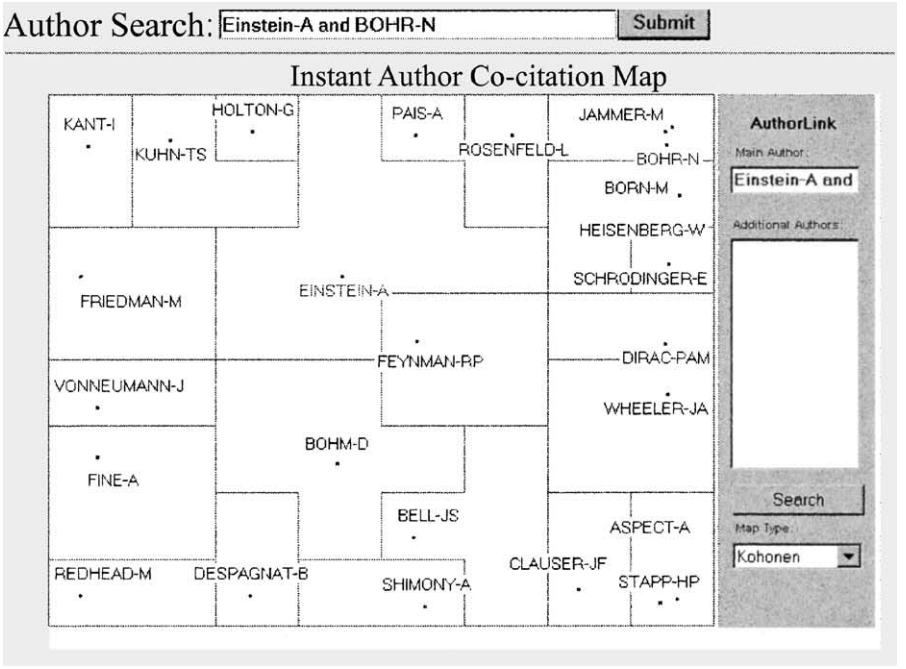


Fig. 5. Author maps for music historian Einstein-A and for physicist Einstein-A.

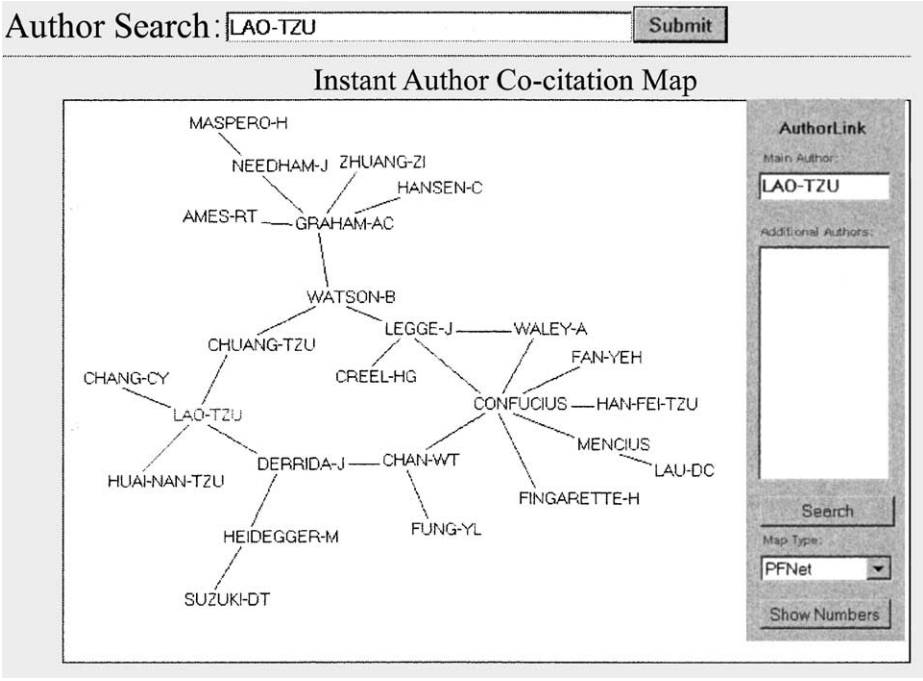


Fig. 6. PFNET of the Chinese Philosopher Lao-Tzu.

quantitative data to test the overall goodness of fit of each visual map to the user's cognitive map. Results of his research will help to improve our system.

Anther critical test for our system is whether it will scale upward to even larger data sets, especially in an interactive environment, when ACA and IV are applied. In early trials, we have successfully mapped data in real time from the full-scale ISI databases on DIALOG, although the maps are not yet live interfaces for point-and-click retrieval.

Finally, a lot more can be done to enhance the interactive design of the author co-citation maps. For example, users should be able to observe how the author clusters are formed gradually, should be able to follow links to see how two authors are related, and should have tools to label the maps with their own understanding of how authors are grouped. These are some examples of the features we plan to add to the next version of the AuthorLink system.

Appendix A. Pathfinder networks

A PFNET is a computational procedure to devise a simplified network model for a set of interrelated data. Similar to multidimensional scaling techniques, the procedure examines all the data relationships and creates *paths* for only the most efficient connections among the data. As a result, the data and data relationships can be represented as a graph of nodes and edges. There are many properties and features of PFNETs; readers are referred to the book edited by Schvaneveldt (1990) for further reading on PFNETs. The following is a brief description of the procedure using our data as an example.

Table 1
Sample co-citation data for three authors

	Smith	Brown	Jones
Smith	10	3	2
Brown	3	23	5
Jones	2	5	9

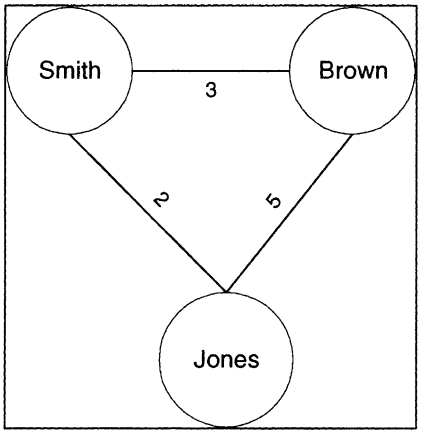


Fig. 7. A network representation of author co-citation data in Table 1.

In an author co-occurrence matrix, the names themselves represent the nodes and the co-occurrence values represent the edge weights. For instance, given a simple 3×3 matrix in Table 1, a corresponding network representation is shown in Fig. 7.

In order to keep this example simple, only three nodes are used permitting a maximum of three edges (or links). However, in larger networks the visual complexity overwhelms the viewer and presents no advantage over the simple raw co-occurrence matrix. For example, a network of 25 authors (nodes) has 300 edges! (While not every author of the 25 is co-cited with every other author, a co-citation count of 0 is also considered an edge by the program). A PFNET is used to remove some of the redundant or less-salient links to show a more understandable network.

A PFNET is created by examining each link between each node pair. Alternate paths around the two nodes are examined to see if there is a shorter path. If there is, then the link between the two nodes is eliminated. The walk length, q, of the alternate path (i.e., how many links to examine) and the metric used to measure the distance of the alternate path, r, (e.g., city-block distance, Euclidean distance, or maximum link length) are specified by the user. For our research, we use $q = n - 1$ and $r =$ maximum link length as this reveals a network with the fewest links.

A PFNET based on Fig. 7 is shown in Fig. 8. In this example, the link between Smith and Jones (weight of 2) is not removed, as the maximum weight of the alternate path, Smith–Brown–Jones, is 5, which is greater than 2. However, the link between Jones and Brown (weight of 5) is removed, as the maximum weight of the alternate path, Brown–Smith–Jones, is 3, which is less than 5.

The reasoning for the removal of the link is that Jones is better related to Brown through the association of Smith rather than directly. Thus in our example the network with three links is reduced to two. In large networks, this reduction results in readily interpretable networks.

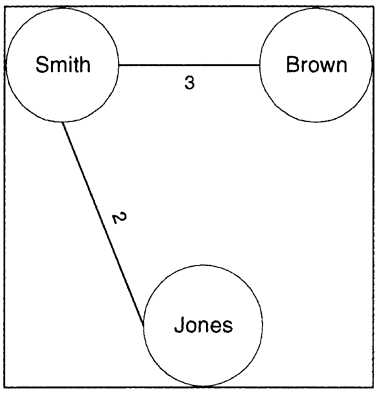


Fig. 8. A PFNET representation for data in Table 1.

Finally, an additional step is needed to display, or embed, the resulting network. We chose the embedding algorithm by Kamada and Kawai (1989) for AuthorLink. This algorithm was chosen for its stability, aesthetics, the ability to display links of proportional lengths, and its close correspondence to multidimensional scaling.

Appendix B. Self-organizing feature maps

Artificial neural networks algorithms are divided into two major types: supervised learning and unsupervised learning. Kohonen's self-organizing feature map (SOM) is a major unsupervised learning algorithm. The algorithm takes input data from a high-dimensional space and maps them into a lower dimensional space while maintaining proximally the same data relationships in the two spaces. The way this is achieved is through a self-organizing process. Initially, data are mapped randomly; learning takes place when all the data are presented to the network through many iterative cycles. Each time, if two input data elements are measured having a strong relationship in the high-dimensional space, their mapping images on the lower dimensional space will be moved closer. This process continues until certain stability is established.

The algorithm has been used widely as a visualization technique for high-dimensional data. Readers are referred to Kohonen's book, *SOMs*, (Kohonen, 1997) for details of the algorithm, its mathematical background, and its applications. Here again we provide a simple description of the algorithm, using author co-citation data as an example.

The SOM algorithm implemented in AuthorLink is a two-dimensional grid of 8×8 evenly distributed nodes. The number of input nodes is 25 (Fig. 9). Every output node is connected to every input node (The figure does not show all the links). Thus, each output node corresponds to a vector of values (called weights). Initially, all the weights are assigned a small random number.

Given an author and the author's 24 most frequently co-cited authors, a 25×25 co-citation matrix can be obtained from the co-citation frequencies of every pair of the authors. This co-citation matrix is used to train the SOM in our system. Repeated, a row from the co-citation matrix is randomly selected and compared to every output node to determine a "winner." Weights of the winning output nodes then are updated so that the next time this input node is presented,

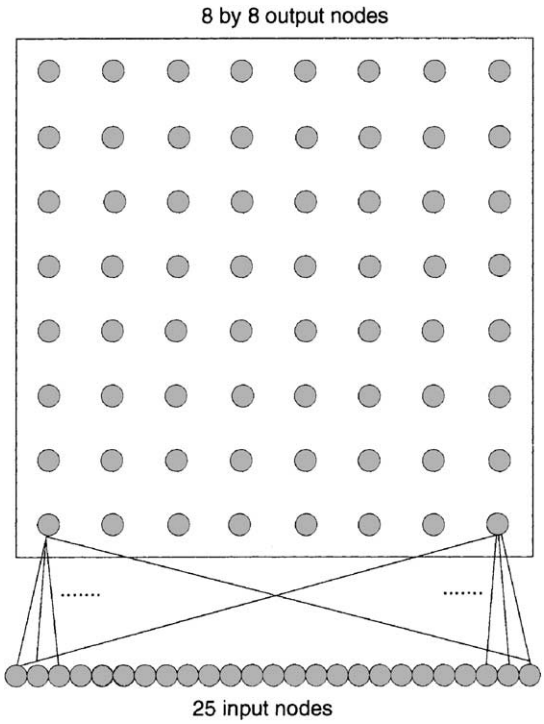


Fig. 9. The input nodes and output nodes of Kohonen's feature map.

this output node will likely be selected again as the "winner." In the meantime, nodes surrounding the winning node are similarly adjusted. This learning process generally needs to go slowly with many numbers of iterations. The number of iterations needed to train a SOM is often determined empirically (in our case, we optimize the number of training cycles to 2500).

After the training, input vectors closest in the input space will map to the same regions in the output map. The regions are delineated by areas of nodes in which the elements with the highest value on the vectors are the same. For example, if four output nodes in the upper-right hand corner of a network all have the highest values when author Brown's co-citation pattern is presented, the four nodes would be considered as a region for author Brown (Fig. 10).

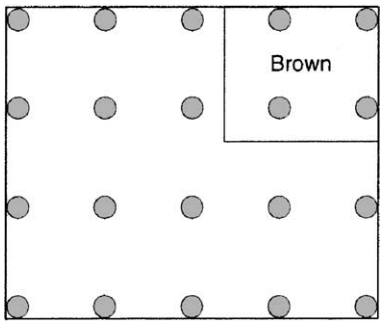


Fig. 10. The region for author Brown is formed.

The finished networks visualize many relationships of the input data. Data mapped to the same region are considered closely related. Regions next to each other are considered having stronger relationships than regions apart. Data corresponding to large regions are considered having stronger influences than those data corresponding to small regions. Such kinds of relationships are much easier to see in the SOM than in the original input matrix.

References

Borgman, C. L. (Ed.). (1990). *Scholarly communication and bibliometrics*. Newbury Park, CA: Sage Publications.

Borko, H., & Bernier, C. L. (1978). *Indexing concepts and methods*. New York: Academic Press.

Boyack, K. W., Wylie, B. N., & Davidson, S. D. (2001). A call to researchers: Digital libraries need collaboration across disciplines. *D-Lib Magazine, 7*(10), http://www.dlib.org/dlib/october01/10contents.html.

Buzydlowski, J. A comparison of self-organizing maps to pathfinder networks for the mapping of author co-citation analysis. Ph.D. dissertation, Drexel University, in preparation.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management, 35*(3), 401–420.

Chen, C., & Carr, L. (1999a). Trailblazing the literature of hypertext: An author co-citation analysis 1989–1998. In *Hypertext '99, Proceedings of the 10th ACM Conference on Hypertext,* Darmstadt, Germany (pp. 51–60).

Chen, C., & Carr, L. (1999b). Visualizing the evolution of a subject domain: A case study. In *Proceedings of IEEE Visualization 99,* San Francisco, California, USA (pp. 499–502).

Cleveland, D. B. (1976). An *n*-dimensional retrieval method. *Journal of the American Society for Information Science, 27*(5), 342–347.

Ding, Y., Chowdhury, G. G., & Foo, G. (1999). Mapping the intellectual structure of information retrieval studies: An author cocitation analysis, 1987–1997. *Journal of Information Science, 25*(1), 67–78.

Ding, Y., Chowdhury, G. G., Foo, G., & Qian, W. (2000). 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.

Einstein, A. (1962). *Mozart, his character, his work*. New York: Oxford University Press.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31*(1), 7–15.

Kohonen, T. (1997). *Self-organizing maps* (2nd ed.). New York: Springer.

Mackinlay, J. D., Rao, R., & Card, S. K. (1995). Organic user interface for searching citation links. In *Mosaic of Creativity: CHI '95: Proceedings of the Association for Computing Machinery Special Interest Group on Human-Computer Interaction ACM/SIGCHI Conference on Human Factors in Computing Systems, Denver, CO* (pp. 67–73). New York: ACM.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41*(6), 433–443.

Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder associative networks: studies in knowledge organization*. Norwood, NJ: Ablex.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

White, H. D. (1983). A cocitation map of the social indicators movement. *Journal of the American Society for Information Science, 34*(5), 307–312.

White, H. D. (1990a). Author cocitation analysis: overview and defense. In C. L. Borgman (Ed.), *Bibliometrics and scholarly communication* (pp. 84–106). Newbury Park, CA: Sage.

White, H. D. (Ed.). (1990b). *Perspectives on . . . author cocitation analysis*. *Journal of the American Society for Information Science, 41*(6), 429–468 [Special section].

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163–171.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. In M. E. Williams (Ed.), *Annual review of information science and technology* (vol. 32, pp. 99–168). Medford, NJ: Information Today.

White, H. D., & McCain, K. W. (1989). Bibliometrics. In M. E. Williams (Ed.), *Annual review of information science and technology* (vol. 24, pp. 119–186). Amsterdam: Elsevier.

White, H. D., Buzydlowski, J., & Lin, X. (2000). Co-cited author maps as interfaces to digital libraries: designing pathfinder networks in the humanities. In *Proceedings, IEEE International Conference on Information Visualization, London, England* (pp. 25–30). Los Alamitos, CA: IEEE Computer Society.

White, H. D., Lin, X., & Buzydlowski, J. (2001). The endless gallery: Visualizing authors' citation images in the humanities. In *Proceedings of the American Society for Information Science and Technology* (pp. 182–189). Washington, DC.