

Term Co-occurrence Analysis as an Interface for Digital Libraries

Jan W. Buzydlowski
Drexel University
Philadelphia, PA 19104
janb@drexel.edu

Howard D. White
Drexel University
Philadelphia, PA 19104
whitehd@drexel.edu

Xia Lin
Drexel University
Philadelphia, PA 19104
linx@drexel.edu

ABSTRACT

In this paper we examine the relationship of term co-occurrence analysis and a user interface for digital libraries. We describe a current working implementation of a dynamic visual information retrieval system based on co-cited author maps that assists in browsing and retrieving records from a large-scale database, 10 years of the Arts & Humanities Citation Index, in real time.

Keywords

PFNET, SOM, Visualization, Digital Libraries, Author Co-citation Analysis.

1. INTRODUCTION

The first wave of research on digital libraries focused on making data physically available. While digital libraries promise to deliver huge amounts of content immediately, this very abundance exacerbates the problem of separating relevant from irrelevant material. The next wave of digital library research must address the ever present problem of intellectual access—that is, matching requests appropriately in information systems so that users can find what they are looking for. This often requires human intervention to represent the “raw” material (documents) through synoptic language such as descriptors, keywords, and abstracts. Given the volumes of material contained in digital libraries, such representation is a large-scale, formidable task.

There have been many efforts, described in, e.g., [1], to automate the process of briefly representing documents with suitable indexing terms or abstracts. There have also been many attempts to display those documents en masse in a format that allows users to better understand the entire collection, e.g., [2] and [3]. Much of the current work in presenting an entire collection, e.g., [4] and [5], has focused on the visualization of the materials. This is a positive trend, in that visualization helps one see patterns that cannot be readily absorbed in any other way [6].

Unfortunately, many visualizations are being built on trivial collections. Others are marred by cryptic symbolism and/or

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Joint Conference on Digital Libraries '01, June 24-28, 2001, Roanoke, VA.

Copyright 2001 ACM 1-58113-000-0/00/0000...\$5.00.

inadequate labeling in the displays. Some in the bibliometric tradition, e.g., [5], are static rather than dynamic interfaces. In this paper we present a new way of visualizing a large, “real-world” collection of bibliometric data based on dynamic queries from users. We describe an implementation of such a system and illustrate its use with some examples.

2. METHOD

In each document there are sections that contain various terms. A fruitful method for rendering a collection of documents is through analysis of their co-occurring terms—that is, the words, phrases, author names, and so on that appear together in designated spans of text in the same document. For example, a number of salient words can be extracted from the title of a work or from its abstract (eliminating stop words and perhaps using word stemming). The authors cited in a bibliography section can also be extracted, as can the journals in which those authors publish. Such co-occurrences can then be counted and statistically processed to serve as indicators of the substantive contents of the collection.

Each document must be parsed to reveal the salient terms. Fortunately, this parsing can be automated, e.g., [1]. Since the number of co-occurring terms in segments of each document can vary, they constitute a repeating field. It is easy, however, to cast this information in the form of a record containing a document number and a term for each of all of the terms in a section

The advantage of term co-occurrence analysis lies in the fact that all of the information is derived from the documents themselves and requires no human intervention. The more documents one has, the more information becomes available. This makes it ideal for digital libraries.

2.1 Term Co-occurrence Analysis

Information retrieval is often divided into two categories: searching and browsing. Searching implies that you have a good-to-perfect idea of what you want. Browsing implies that you will be able to recognize what you want when you see it.

Browsing helps a user examine a new field before doing more in-depth research. If one were starting to do research in a new specialty, it would be helpful to see what other research areas are involved and how they are related, who the specialty’s major authors are, what its major journals are, and so on. For instance, if an undergraduate wanted to explore philosophy, she would

probably have a name like Plato in mind. It would be helpful if she could use Plato's name as an input to learn the names of other philosophers, such as Aristotle, Hegel, or Kant, who are repeatedly co-mentioned with him in scholarly writings. Similarly, if a reader likes the novelist John Steinbeck, it would be helpful for him to know the other novelists, critics, and scholars who are most often mentioned in Steinbeck studies.

How, then, does one use term co-occurrence analysis to help in searching or browsing? Our technique makes use of *term seeding*. The general principle is that given a single word or phrase as a "seed" or starting point, we can retrieve the other terms that most frequently occur with it in designated fields across a large collection. Documents that contain the seed term are systematically examined to return the related terms, rank ordered by frequency of occurrence. For instance, we can retrieve the other authors that co-occur most frequently with the seed author of Plato in the references of journal articles.

It is useful to have the list of top-ranked related terms, given a single input term. However, once a list is retrieved, it is also of interest to know how each item in the list is related to every other. All of the elements of the list are related, but terms within the list have sub-relations as well when they are systematically paired. For instance, although the four philosophers mentioned above are studied in various combinations, most educated persons would expect Plato and Aristotle to be conjoined more often than, say, Aristotle and Hegel; conversely, Hegel and Kant would be conjoined more often than Plato and Kant. Our goal is to make explicit all such sub-relations, which means extending a ranked list of term-counts into a square matrix of term-counts for every pair in a set of terms

To determine the various relationships between all of the terms returned from the term seed, the analysis of co-occurrences will give a metric to determine the strength of an association. The strength of the co-occurring terms comes from the number of times two terms occur together within the collection. In our example, Plato and Aristotle will presumably co-occur a large number of times and so will Hegel and Kant. Plato and Cher will occur less frequently (if at all—hopefully not). The more times one author co-occurs with another author, the stronger the association those two authors have; the less frequent, the less strong. In this example, we are working within the framework of Author Co-citation Analysis (ACA), developed over a 20-year period in, e.g., [7], [8], and [9]. The same methodology and arguments for ACA's validity can be applied to appropriate repeated terms to form a more general term co-occurrence analysis.

2.2 Display Techniques

The data structure that represents the co-occurrence of terms is called a co-occurrence matrix. The row and column headings represent the terms of interest, and the intersections of the rows and columns hold counts of the number of times the pair of terms co-occur. Since the order of the terms is not significant in determining the frequency of co-occurrence, the matrix can be represented by either its upper or lower half.

Once the pairwise co-occurrences have been derived, it is possible to examine the raw co-occurrence frequencies directly. However, due to the large number of them (e.g., 25 terms produce $25(24)/2 = 300$ co-occurrences), it is difficult to grasp what all of the numbers jointly mean. Fortunately, there are several visualization techniques that allow one to present all of the co-occurrences simultaneously, tapping into the human ability to process pictures better than numeric matrices. Three visualization, or mapping, techniques that have been used to render multiple co-occurrences are multidimensional scaling (MDS), Kohonen Self-Organizing Maps, and Pathfinder Networks.

MDS is a mathematical technique that reduces the high dimensionality of the co-occurrence matrix to a more visually representable two or three dimensions. It is a well-established methodology that produces maps in which authors are positioned as points on a page. The basic metaphor in such mapping is that greater similarity of authors, as measured by their co-occurrence counts, is rendered spatially as greater proximity of their points. MDS is often used in conjunction with clustering algorithms that permit cluster boundaries to be drawn around groups of related points.

A second mapping algorithm used is that of self-organizing maps (SOMs), also called Kohonen Maps after their inventor. This display technique uses self-training neural networks to determine the placement of terms in two or three dimensions, e.g., [10]. A SOM is similar to a MDS map in that authors are represented as points on a page, but it also automatically groups similar authors into explicit word or concept areas. Again, closeness of points or point areas implies relatively strong relationships in the raw data.

A third mapping algorithm is Pathfinder Networks (PFNETs). The algorithm was developed by Schvaneveldt and other cognitive scientists [11] to eliminate less salient links and retain the more salient links in verbal association networks. The latter were often generated by having human subjects make similarity choices between paired stimuli and then simplifying the pairwise data as a PFNET. Since any co-occurrence matrix can be represented as a network, all of the PFNET techniques also apply to term co-occurrence analysis and have been used in our project to generate maps. Additional algorithms, e.g., [12], [13], are needed to render or embed the PFNET on a screen or page, and the resulting visualization positions terms as points on a page with the additional feature of explicitly connecting with lines the terms that are most directly associated with each other (and not connecting terms that are linked via intermediate terms).

The format of SOMs and the presentation algorithm of PFNETs arguably convey at least as much information as the MDS algorithm [14], and they are highly tractable for real-time implementation. Thus, our current research in visualization of co-occurrence terms favors SOMs and PFNETs.

3.0 IMPLEMENTATION

We have created a system based on the methodology described above. It is a generalized Web-based interface that can work with any term-based collection. The data are contained in a specialized database, the visualization algorithms are coded in C for speed, and the interface is coded in Java for portability and Web-browser execution.

There are currently two data sources on which the system implemented. One, which we presently call ConceptLink, maps co-occurring descriptors from the Medline database of the National Library of Medicine. Another, presently called AuthorLink, is based on a citation index from which we can extract co-cited authors. This paper will describe the latter in detail.

Our system has three tiers. The front tier is a Web interface in Java and HTML; the middle tier is an application server including various implementations of data mapping procedures (in Java Servlets, C and CGI). Due to the modularity of the architecture, the back-end tier can be any database or search engine. The initial implementation used the BRS Search Engine. Currently, the back end for AuthorLink is a specialized co-occurrence database system called Noah. This database was developed especially for the storage and quick retrieval of co-occurrence data. The back end for ConceptLink is Oracle8i, which stores and processes a large co-occurrence database and the PUBMED search engine.

The data source for AuthorLink is the Arts & Humanities Citation Index (AHCI) for the decade 1988 to 1997. The data were given our college by the Institute for Scientific Information, based near Drexel University in Philadelphia, as a research grant in 1998. There are approximately 1.26 million bibliographic records. The resultant database consists of approximately 7 million terms. Although the size of the database is nontrivial, the results below are achieved in real time.

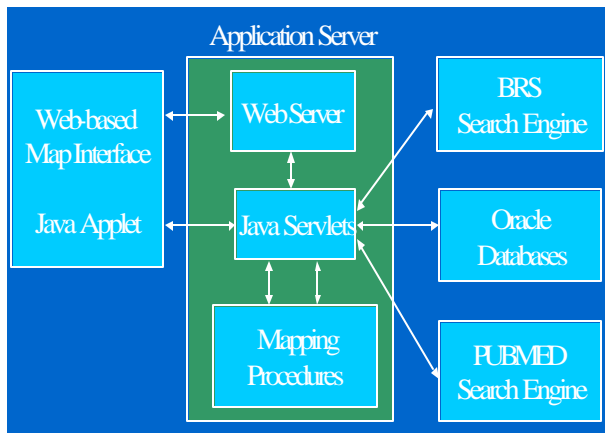


Figure 1. System Architecture

3.1 The Interface

We have simplified the input to ACA from what it has been in the past. Rather than requiring a whole list of authors as input, we require only a single name, such as Plato or Steinbeck, to generate the analysis. This greatly lightens the cognitive load on the user, who gets considerable information back for minimal input. We can map not only world giants such as Plato and Shakespeare, who are known to most educated people, but literally thousands of other people in the humanities as well, as long as they are cited in the journals covered by AHCI.

Initially the user is presented with a screen explaining the system and with a blank text box in which to enter an author name, the

named seed, to retrieve a list of most highly associated authors, as in Figure 2.

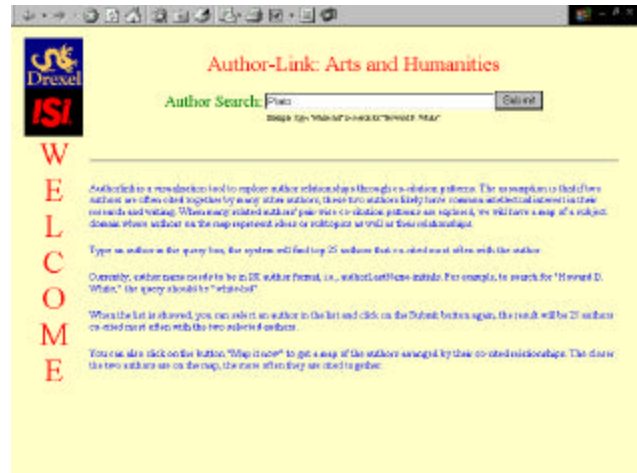


Figure 2. Initial Screen

After the user enters the named seed, the database is queried to determine the authors who are most often co-cited with the named seed author. The list of author names is returned, in order of frequency, along with the number of times each author co-occurred. (The default return is the first 25 authors, the seed plus 24 others, but this can be easily adjusted.) For instance, if the user selected “Plato” as the named seed, then the resulting co-related authors would be returned as shown in the left box in Figure 3.



Figure 3. Associated Authors with Plato

To determine the strength of associations for the 300 co-occurrences of the top 25 authors, the user selects the “Map It Now” button and the database is again queried. The resulting co-occurrence matrix is passed to the mapping algorithm.

The initial map displayed is that of a SOM. An example of what the map looks like is shown in Figure 4.

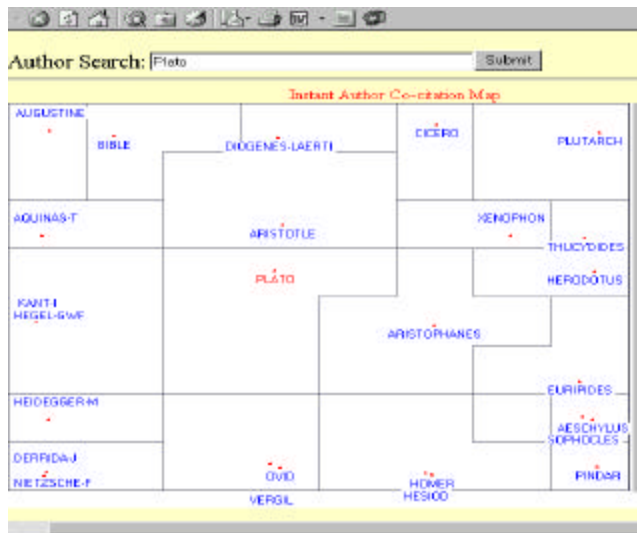


Figure 4. A SOM based on Plato

Plato appears in red near the center of the map. (Seed authors are not always central.) Near him are closely connected other authors, such as Aristotle, his only rival in Greek philosophy. Plato's eminence in the Western intellectual tradition is such that he is implicated in the 1988-97 AHCI not only with medieval and modern philosophers (Augustine, Aquinas, Hegel, Kant, Nietzsche, Heidegger, Derrida) but with Greek dramatists (Aristophanes, Aeschylus, Sophocles, Euripides), Greek and Roman poets (Homer, Hesiod, Pindar, Ovid, Vergil), Greek and Roman historians (Herodotus, Thucydides, Xenophon, Plutarch, Cicero), and the Bible. Plato's biographer, Diogenes Laertius, is also present. We can show that all of these names are automatically placed in appropriate groupings by the Kohonen algorithm.

The user can choose from a list box to display the same co-occurrence matrix as a PFNET. The PFNET for Plato is shown in Figure 5.

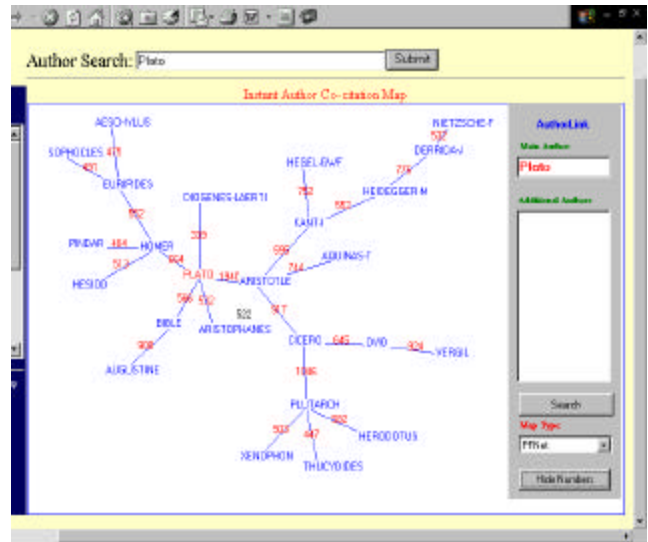


Figure 5. A PFNET based on Plato

The user can toggle on or off the number of co-citations for each linked pair. If the cursor is placed on an author's name, that author's co-citation count with Plato pops up.

Again, the names directly linked by the PFNET form highly cogent groupings. For example, the relationships between Plato and Aristotle and between Hegel and Kant that were presumed above are borne out by the actual links that are formed from the 1988-97 AHCI. More generally, the PFNET suggests a virtual "Great Books Course" built around Plato. We take this as evidence that AuthorLink has considerable promise as an aid to browsing in a digital library space. Experts on a given author will be able to see that author's intellectual affiliates, some of which may be new to them, as derived from contemporary citation patterns. This is indexing by use, similar to that of "recommender systems." Novices on an author may not be able to interpret all the connections (which are based on the perceptions of citing scholars), but they at least have rich leads as to authors that are worth reading together.

But what of retrieval? The AuthorLink interface is connected to the full 1.26 million bibliographic records in 1988-97 AHCI. If the user wants to pursue Plato's connections with any other author(s) in the map, the literature making those connections can be retrieved. In the Additional Authors list box to the right of the map (as shown in Figure 5), Plato's name is automatically entered as the seed author. The user can select one or more additional names, say Ovid and Vergil, to search with Plato in a Boolean AND relationship. The user drags-and-drops the names into the Additional Authors list box, clicks on Search, and the articles in AHCI that contain those names are returned, as in Figure 6, which displays a BRS/Search Web interface.

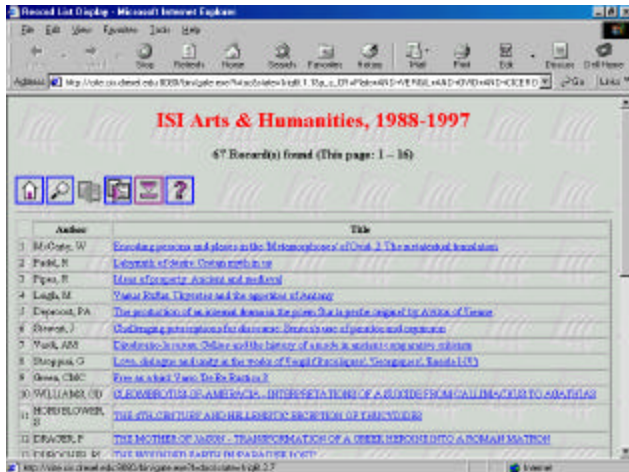


Figure 6. Retrieved Articles citing Plato, Ovid, Vergil

Clicking on the hyperlinked title of any article will bring up its full bibliographic record plus a listing of the works it cites. In the latter, the cited works by the ANDed authors (e.g., Plato, Ovid, and Vergil jointly) are highlighted. Eventually it should be possible to bring up full-text copies of many of the cited articles rather than their bibliographic records. This will come about as the Institute for Scientific Information and its many academic library customers move from print to fully digitized and hyperlinked resources.

It must again be emphasized that our maps are produced from the data alone without any human intervention. The data source is substantive and important. The queries are dynamic, based on any user-supplied seed, and the results (list of names and maps) are returned in seconds. This allows for the real-time interaction with the nameseeds, associated author lists, and maps to support online browsing and searching.

4.0 CONCLUSION

White and McCain [3] suggest the following list (abridged) for evaluating a visual information retrieval interface:

1. Is the display an improvement over a simple list?
2. Does it provide new capabilities?
3. Is it rapidly intelligible?
4. Is it helpful in real time (or with an acceptable wait)?
5. Is it tied to an important collection?
6. Is it scalable upward to collections greater in size?

We believe the system we have built allows us to answer these questions in the affirmative.

For Item 1, a map does provide more information than a simple list. The information provided in Figure 4 or Figure 5 is significantly greater than the information provided in Figure 3.

For Items 2 and 4, our system does provide new capabilities by being dynamic and working in real time so that a user can perform term analysis and retrieval iteratively. Because of its

responsiveness, it is quite fun to use for exploring the ties of authors in whom one is interested.

For Item 3, the authors of this paper believe this to be true. However, research is currently being done to examine the usability, usefulness, and the degree of intelligibility of the map types, and this question will hopefully be answered soon.

For Item 6, the current implementation is based on a data source of significant size, e.g., 7 million terms. We are seeking data sources of even large size, e.g., the entire AHCI collection, to test the bounds of the system.

Finally, for Item 5, the AHCI database for 1988-97 is already of significant size and importance. When it is combined with AuthorLink, one can readily explore the bibliographic relationships of many thousands of writers, of all magnitudes of eminence, in the humanities. This is a potentially valuable resource for students and researchers everywhere, whether they want to browse or retrieve documents or both. Moreover, the technology AuthorLink represents is transferable to other significant digital library collections.

5.0 REFERENCES

[1] Salton, Gerard. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley.

[2] Chen, Chaomei. 1999. Visualizing Semantic Spaces and Author Co-Citation Networks in Digital Libraries. *Information Processing & Management* 35: 401-420.

[3] White, Howard D., McCain, Katherine W. 1997. Visualization of Literatures. *Annual Review of Information Science and Technology* 32: 99 – 168.

[4] Chen, Chaomei. 1999. Information Visualization and Virtual Environments. Springer.

[5] Ding, Ying, et al. 2000. Bibliometric Information Retrieval System (BIRS): A Web Search Interface Utilizing Bibliometric Research Results. *Journal of the American Society for Information Science* 51: 1190-1204.

[6] Cleveland, William S. 1993. *Visualizing Data*. Hobart Press.

[7] White, Howard D. 1990. Author Co-Citation Analysis: Overview and Defense. In *Scholarly Communication and Bibliometrics*, Christine L. Borgman, ed. Sage Publications. 84-106.

[8] McCain, Katherine W. 1990. Mapping Authors in Intellectual Space: A Technical Overview. *Journal of the American Society for Information Science* 41: 433-443.

[9] White, Howard D., and McCain, Katherine W. 1998. Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995. *Journal of the American Society for Information Science* 49: 327-355.

- [10] Lin, Xia. 1997. Map Displays for Information Retrieval. *Journal of the American Society for Information Science* 48: 40-54.
- [11] Schvaneveldt, Roger W., ed. 1990. *Pathfinder Associative Networks*. Ablex.
- [12] Kamada, Tomihisa and Kawai, Satoru. 1989. An Algorithm for Drawing General Undirected Graphs, *Information Processing Letters*, 31, 7 – 15.
- [13] Fruchterman, Thomas and Reingold, Edward. 1991. Graph Drawing by Force-Directed Placement, *Software—Practice and Experience*, 21, 11, 1129 – 1164.
- [14] White, Howard D., Buzydlowski, Jan, Lin, Xia. 2000. Co-Cited Author Maps as Interfaces to Digital Libraries: Designing Pathfinder Networks in the Humanities. *Proceedings, IEEE International Conference on Information Visualization*. 25 – 30.