Knowing References

Bernard Bayer, The Ohio State University, USA

Abstract: By looking at how scientific papers are related to each other, you can theoretically reconstruct how knowledge, information and data were gained, developed, and gathered. The relationships are outlined through cited references, which furnish the raw materials for reconstructing the research. In theory, researchers pay their intellectual debts through the references; in practice, not all bibliographic linkages are created equal. Still, knowing references is important in determining how a scientist, especially one in the hard sciences, develops information that may, in some cases, become knowledge.

1. Introduction

Some view the relationships between data, information and knowledge as a hierarchy. Data are organized into information, which, with other input, can enter the mind and become knowledge. The process does not have to stop there. Insight from environs unknown can enter the picture and knowledge becomes wisdom. Finally, there is enlightenment. Somewhere between data and enlightenment are the papers at this conference.

A Librarian of Congress made this distinction between information and knowledge: "...knowledge is orderly & cumulative, information is random & miscellaneous" (Boorstin, 1979). The brain seems to handle information and knowledge differently, with parallel systems for item memory and for category knowledge (Knowlton and Squire, 1993).

A formalized setting for presenting information in the hard sciences is the journal paper. Each paper includes references to "help the author to describe...(his or her) contribution by comparing and contrasting it with previously reported findings...(and) permit the author to build on information that colleagues have demonstrated to be correct elsewhere in the formal literature" (Carter, 1987, pp. 20-21).

The importance of references is recognized by the editors of professional journals. In advice to authors, the staff of one such journal recommends that authors, before submitting a paper, ask themselves whether all sources cited are "genuinely relevant and necessary" (Staff, 1994, p. 23).

After publication, papers and their cited references are analyzed by information scientists and others outside the hard sciences to look for sociological and psychological meanings and bibliometric truths behind the patterns of usage. For example, some researchers have looked within the text of individual articles to
determine whether degrees of relatedness between cited and citing works can be predicted by analyzing where in the paper each citation is referenced and how many times the same citation is used (Bonzi, 1982).

Others have tried to find patterns of relatedness between two papers where the connection is not a direct citation. For example, source documents can be linked through bibliographic coupling, which occurs when two papers have one or more references in common (Kessler, 1963). In addition, cited documents can be linked by co-citation, which occurs when two documents are jointly cited by a third. The linkage becomes stronger as more documents cite each of the original two papers (Small, 1973).

Clusters and maps of specific fields of knowledge can be generated on the basis of the strengths of co-citation linkages. The results are series of knowledge domains that can be generated automatically with citation data (Garfield, 1983, pp. 98-134).

In this paper, the focus narrows back to the individual paper, its references, and the information contained in each. Here, we are back from the macro-world of domains and into the micro world of the individual paper in the field of physics. We want to show that cited references are much more than reading lists for more information.

2. Question and Methodology

The question to be explored is: can you determine what exactly, within each cited work, was used by the author of the citing article? In the exploration, the complete text of cited works will be examined to identify what specific pieces of referenced information were used to produce the new knowledge discussed in the citing paper.

Two studies have shown that such an approach is possible with book materials. The first showed that two-thirds of the citations in scholarly books are to information on a single page in the referenced books (Kilgour, 1991, p. 96). The second showed that scholars use direct quotations in sufficient quantities to warrant more study to try to determine patterns in that usage (Kilgour and Feder, 1992, p. 269). Another study had extended similar research to journals as well as books, with mixed results (Haas, 1993).

The focus here is on articles published in physics journals. In hard sciences, such as physics, the journal paper is the major knowledge representation. The basic methodology used in this investigation was adapted from the first two studies of book materials, as follows:
(1) five journal titles in the field of physics were selected from the journals with the highest impact factor, as determined from the listings in the ISI Journal Citation Reports (Institute for Scientific Information, 1992, pp. 107-108). (Letters and review journals were excluded. The five selected were: Journal of Chemical Physics, Nuclear Physics, Nuclear Physics B, Physica C, and Physical Review B.)

(2) an article from a recent issue of each of the journals was selected randomly, and photocopied.

(3) each cited paper in a journal article was located and photocopied.

(4) text preceding the reference in the source paper was matched manually against all the text of the referenced papers to determine how much of the referenced information was extracted in the source paper.

For example, for one paper from a January 1993 issue of Physical Review B, thirty-four additional papers (the cited works) had to be retrieved and examined word by word, and compared to the text of the citing paper.

### 3. Results

The results from the analyses did not yield the anticipated results. It was not possible to determine the exact referenced information used by the physicists in doing their research. No easily identifiable direct quotes or specific data could be matched from the cited reference to the citing paper. Each cited document presented another set of arbitrary decisions about what might have been used to build a knowledge structure in the mind of the author.

Specifically, the problems included:

(1) the relationship between reference information and the source paper was not explicitly stated by the authors in terms of direct quotes
(2) cue words were not used in the text of the articles to let the reader determine what referenced information was used in the research (cue words include: "according to...", "...showed that...")
(3) terms in the citations did not always match terms in the citing paper. For example, a cited reference might use "gas-gas," while the citing paper would use "binary."
(4) page numbers did not help to narrow the location of referenced information because the number of the first page of each cited article is always referenced, no matter on which page the referenced information is located.
(5) references to review articles do not indicate which, if any, of the reviewed papers were used.

The results were not encouraging, and therefore this line of investigation was dropped.
4. Discussion and Conclusions

In summary, when an author cites a particular work, it is pertinent in some, or many, ways to the research. However, it is not possible to determine which subset of the text within the cited works furnished necessary information for the research. Did the author cite the entire paper? a section of it? a paragraph? a specific datum? The text of the citing paper does not supply enough clues to reconstruct the original knowledge structure.

In addition, some authors and editors are not careful enough when checking citations for accuracy. Anywhere from 14% to 50% of the cited references may be in error, at least in medical journals (Goodrich and Roland, 1977, p. 18). Not all the errors are serious ones. In our exploratory study of physics journals, the error rate was closer to 10%. For example, in the paper from Physical Review B mentioned earlier, four of thirty-four citations were incorrect: two had wrong pages, one had a wrong author, and one was a completely wrong citation.

Further, some references are included for financial, rhetorical, and political reasons. (There are other debts that need to be paid to keep research ongoing.) On the other hand, there are citations that are omitted through carelessness, poor searching of the literature, or on purpose for reasons less than honorable. In addition, some cited references are review articles and, as such, are used as a shotgun approach to paying intellectual debts.

However, despite all the acknowledged problems, the cited references are relevant. The citing process is a very efficient document retrieval system. If precision and recall were used to measure the system's retrieval effectiveness, both would be close to 90%, even with a 10% error rate. All the references are in some way relevant, otherwise they would not have been cited. It is clear from the study that the cited references are much more than a supplemental readings list.

Reading a paper and following up on its references is a means of reconstructing the structure of knowledge. The reader's structure may not be the author's structure, but it does not matter. "It is an axiom of modern learning theory that effective complex learning involves the learner's active reconstruction of knowledge, rather than passive receipt of the information components of understanding" (Ford and Ford, 1993, p. 579). In teaching physics, the aim is "to get the student actively engaged in working with the subject matter, rather than passively hearing about it" (Hestenes, D. 1979, p. 242).

The citation process is still not complete. An author pays only one of his or her intellectual debts with references. It is the major one, but there are still others to be acknowledged.
Acknowledgements

Some of the research reported herein was performed by Professor Frederick G. Kilgour and the author, under a grant from OCLC, Inc., Dublin, Ohio, USA. (Acknowledgements such as these pay more of an author's intellectual debts, and give you more information for your knowledge structure. You find out who else did the work and who paid the bills.)

Authors also should always ask themselves: "Are all acknowledgments of indebtedness made? Has permission to acknowledge personal assistance been obtained?" (Staff, 1994, p.23)

Proper citing of references and proper acknowledgements are how authors pay their full intellectual debts to others.

Notes

1. In another study, the references to a work were analyzed to determine if an examination of references isolated from their citing document could reveal anything about that document. Here are some of the data for one work:
   (1) number of references: 57
   (2) most heavily cited work: Euclid's Elements (3 times)
   (3) most heavily cited journal: Nature
   (4) number of physicists, mathematicians and engineers cited: 7
   (5) number of humanist, including biblical authors, and sociologists cited: 8.
   What can you tell about the citing work? I'll give some additional information at my presentation and disclose the name of the work in question. Hint: the author is Danish, in honor of the location of this conference.

2. Finally, citations are very practical. A newspaper columnist in Washington D.C., USA, suggested that the best places to catch big fish are those where the number of fish citations issued by the Virginia Department of Game and Inland Fisheries is highest (Mueller, 1989, p. 2). Here citation means "court summons", but the principle is the same: the higher the citation count, the better the area. That is also the way ScienceWatch picks the superhot fields of science (Taubes, 1994, p.1), the hottest scientist of 1993 (Editor, 1993, p.1), and the hottest papers in physics (Mitton, 1993, p. 6).

References


