

Combining Bibliometrics, Information Retrieval, and Relevance Theory, Part 1: First Examples of a Synthesis

Howard D. White

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

E-mail: whitehd@drexel.edu

In Sperber and Wilson's relevance theory (RT), the ratio Cognitive Effects/Processing Effort defines the relevance of a communication. The $tf*idf$ formula from information retrieval is used to operationalize this ratio for any item co-occurring with a user-supplied seed term in bibliometric distributions. The tf weight of the item predicts its effect on the user in the context of the seed term, and its idf weight predicts the user's processing effort in relating the item to the seed term. The idf measure, also known as *statistical specificity*, is shown to have unsuspected applications in quantifying interrelated concepts such as topical and nontopical relevance, levels of user expertise, and levels of authority. A new kind of visualization, the pennant diagram, illustrates these claims. The bibliometric distributions visualized are the works cocited with a seed work (*Moby Dick*), the authors cocited with a seed author (White HD, for maximum interpretability), and the books and articles cocited with a seed article (S.A. Harter's "Psychological Relevance and Information Science," which introduced RT to information scientists in 1992). Pennant diagrams use bibliometric data and information retrieval techniques on the system side to mimic a relevance-theoretic model of cognition on the user side. Relevance theory may thus influence the design of new visual information retrieval interfaces. Generally, when information retrieval and bibliometrics are interpreted in light of RT, the implications are rich: A single sociocognitive theory may serve to integrate research on literature-based systems with research on their users, areas now largely separate.

Introduction

In this article, I integrate ideas from bibliometrics, information retrieval, and Sperber and Wilson's influential book, *Relevance: Communication and Cognition*. The synthesis is quite general, and its validity may be tested by anyone with access to standard bibliometric counts, such as those

available in many databases on Dialog. When rank-ordered, these counts form highly skewed distributions called, among other things, *empirical hyperbolic*, *core-and-scatter*, *scale-free*, *power-law*, and *reverse J*. Whatever the name, I show that when the terms in any bibliometric distribution are treated as components of the well-known $tf*idf$ formula from information retrieval, those terms are interpretable as what Sperber and Wilson (S&W) have called assumptions relevant in a context. The context is the seed term from which the bibliometric distribution was created, and the following definitions hold (Sperber & Wilson, 1986, 1995, p. 125):

"An assumption is relevant in a context to the extent that its contextual effects in this context are large."

"An assumption is relevant in a context to the extent that the effort required to process it in this context is small."

For greater clarity, S&W now use "positive cognitive effects" rather than "contextual effects" in the wording of the first assumption (Wilson & Sperber, 2002). "A positive cognitive effect," they write "is a worthwhile difference to the individual's representation of the world—a true conclusion, for example" (p. 251). They go on to say, "Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time" (p. 252). I will therefore use "cognitive effects" ("positive" being understood) in both parts of this article. "Contextual effects," the equivalent phrase, will appear when earlier writers who used it are quoted.

Sperber and Wilson write almost exclusively about how relevance is created in dialogues between persons. Information scientists focus on a different sort of dialogues—those between a person seeking information and a system designed to provide it, the system being a literature-based artifact whose human designers are absent. In the latter dialogue, both questioner and answerer are governed by views of what constitutes relevant information, but a distinction has long been made between what the nonhuman *system* deems relevant output and what the human *user* does, because they are by no means necessarily the same. Here, the "assumptions

Received October 5, 2005; revised April 22, 2006; accepted April 23, 2006

© 2007 Wiley Periodicals, Inc. • Published online 25 January 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20543

relevant in a context” are not the user’s. A measure of relevance based on term counts is a system measure, and the assumptions are the system’s, as instructed by its human designers. Suitably marshaled, however, the counts permit a responsive answer within a limited bibliographic domain—one that is qualitatively similar to what a well-informed person could supply. Thus, the metaphor of a dialogue between a human being and a system can be sustained. In the dialogue envisioned here, all the user need do is set a context with a seed term designating an interest—for example, “*Moby Dick*” (short for, “I’d like to see writings relevant to *Moby Dick* studies”). The burden of the present work is to show that the system’s response accords with our ordinary sense of the cognitive effects of a message and the effort it takes to process it.

In such matters, one cannot hide behind content-neutral formalisms, however sophisticated; one must please intuition immediately with a convincing verbal surface. That means that words must be exhibited—the right words in meaningful relations, such as a good response to the query, “*Moby Dick*.” I therefore concentrate on words, names, and phrases in a new treatment of the semantics and pragmatics of term-weighting. My broader goal is to extend the theory of responsiveness of literatures (White & McCain, 1989, 1997). Hence, I seek to link algorithmic systems research with sociocognitive studies of users (the two frameworks of information retrieval whose separation is deplored in Saracevic 1996, 1997; Ellis, 1998) and of linking both with bibliometrics. The ultimate goal is to unify literature-based information science under a dialogue metaphor—that of an answerer making relevant replies to a questioner within a context (cf. Blair, 1992, Chen & Xu, 2005; Ruthven & van Rijsbergen, 1996).

The New Synthesis

The $tf*idf$ term-weighting formula involves multiplying term frequencies (tf) by inverse document frequencies (idf). It takes several forms, but all are conventionally used to rank documents by their degrees of relevance to a query (Grossman & Frieder, 1998, pp. 13–16). Here, I use logarithmic versions of tf and idf to rank distributions of documents by their degrees of relevance to a seed term used as a query (more on this below). Seed terms need not be the keywords or descriptors on which the $tf*idf$ formula usually operates. They can themselves be—and here are—names of writings or authors’ oeuvres. In any case, ranking documents by their relevance to a seed term will produce the core-and-scatter distributions long familiar in bibliometrics. So I am indeed combining bibliometrics and retrieval, although in a novel way.

It turns out that the idea of weighting terms by their inverse document frequencies, which Sparck Jones (1972) introduced to the relevance-ranking formula as “statistical specificity,” is richer than previously realized, especially when applied to distributions of cited titles and authors from the databases of the Institute for Scientific Information (ISI,

now Thomson Scientific, Philadelphia, PA). Retrievalists typically apply idf only to subject terms, such as natural-language keywords, and seldom discuss the results in any detail. Manning and Schütze (2000, pp. 541–544) briefly explain statistical specificity as the varying “semantic focus” of terms, but qualitative examples from the literature are rare. Here, I provide new interpretations of statistical specificity and relevance in three kinds of bibliometric distributions, all taken from the records of ISI: the books and serials cocited with a book, the authors cocited with an author, and the books and articles cocited with an article. These are all densely populated distributions as a rule, and it is the business of $tf*idf$ and other procedures to break out comparatively small fractions of them as most relevant in Sperber and Wilson’s sense. I will show what $tf*idf$ weighting does with respect to the overall distribution, and readers may decide the extent to which it is beneficial. (It seems a mixed blessing to me.)

When terms in a bibliometric distribution are seen through S&W’s lens as predicting degrees of cognitive effects and degrees of processing effort, bibliometrics becomes more user-oriented, more “psychological,” and more instrumental in information retrieval. At the same time, S&W’s relevance theory (RT) gains an inexhaustible source of potentially corroborating data. (I touch on possible tests later in both Part 1 and Part 2.) Moreover, measured effects-and-effort data can be plotted and displayed by computer. Taking a cue from Bradford (1950), Saracevic (1975) called the bibliometric distributions “relevance-related.” Here I offer a new style of graphics—*pennant diagrams*—that bear out this claim. Pennant diagrams use bibliometric data and IR techniques on the system side to mimic a relevance-theoretic model of cognition on the user side. Relevance theory may thus be conducive to the design of new visual information retrieval interfaces (VIRIs).

If such a possibility leads information scientists to learn more about RT, that would not be amiss (cf. Belew, 2000, pp. 304–305; Budd, 2004, pp. 454–457; Green, 1995, p. 647). Relevance theory has had its critics in information science (Hjørland, 2000; Saracevic, 1996) and elsewhere (reviewed in Yus Ramos, 1998). However, as I suggest in Part 2 of this article (White, 2007), interpreting relevance as a compound of cognitive effects and processing effort allows various information-related behaviors to be unified in a single theory—one that yields consistent explanations of what goes on when people create or use or judge literature-based systems. In Part 2 I draw on RT to confirm several conjectures about the nature of relevance by earlier writers and contribute toward solving some longstanding puzzles of information science.

Readers may recall that Harter (1992) strove to import RT into information science as “psychological relevance.” The present work builds on Harter’s and in fact, confirms several of his insights, including his claim that “Relevance is the idea that connects [information retrieval] to bibliometrics, and understanding it in one context should aid our understanding of it in the other” (p. 613). As he foresaw, the advantage of adopting RT is that it is a psychologically plausible theory with sufficient breadth to subsume the many

accounts of “relevance,” “pertinence,” and “utility” that presently vie for attention in retrieval evaluation studies (see, e.g., Borlund, 2003; Cosijn & Ingwersen, 2000; Mizzaro, 1997; Schamber, 1994). Under cognitive effects, RT can handle not only topical relevance (as it must in indexing-based retrieval), but also other important kinds, such as evidentiary and analogical relevance (Bean & Green, 2001). Moreover, RT uniquely ties relevance to processing effort, a very desirable integration given that information scientists routinely find minimization of effort to characterize information-seekers’ behavior (Buckland & Hindle 1969; Case, 2005; Mann, 1993, pp. 91–101; Poole, 1985, pp. 86–92; White, 2001a, pp. 104–106).

In linguistic pragmatics, where Sperber and Wilson’s ideas support a large and growing literature (Yus, 2006), the two components of relevance are discussed as implicitly measurable but never, so far as I know, measured. That is because in linguistics RT is concerned not with the world of recorded information, but with brief utterances by persons (Blakemore, 1992), and different impacts are left wholly to readers’ intuitions. An example from self-communion is found in Goatley (1997, pp. 138–139, reparagraphed and slightly condensed):

You wake up thinking, (1) If it’s raining I won’t go to the lecture this morning. You look out the window and discover, (2) It’s raining. From existing assumption (1) and the new information (2) you can deduce further information (3): (3) I won’t go to the lecture this morning. (2) is relevant because, in the context of (1), it produces new information or contextually implies (3). **** You wake up thinking: (4) If it’s raining I won’t go to the lecture this morning. Then either you look out of the window and see: (5) It’s raining. *or* you look out of the window and see: (6) It’s raining and the refuse collectors are emptying the bins. In the context of (4), (5) and (6) have the same Contextual Effects. But (5) is more relevant than (6), because (6) requires more Processing Effort (Wilson & Sperber 1986, pp. 27–30). The notion of Relevance, then, which is comparative rather than absolute, can be summed up in the following formulae: (7) Other things being equal, the greater the Contextual Effects, the greater the relevance. (8) Other things being equal, the smaller the Processing Effort the greater the relevance. Or, alternatively, expressed as a fraction: (9) Relevance = Contextual Effects/Processing Effort. This equation makes it clear that if there is no Contextual Effect there will be no relevance, no matter how little the Processing Effort involved.

This capsule account emphasizes that S&W’s relevance is not a matter of *yes* or *no* but of *more* or *less*, which accords with a common conclusion in information science about the nature of relevance judgments (Greisdorf, 2000; Saracevic, 1975). It also includes an equation for computing degrees of relevance, the updated version of which is Relevance = Cognitive Effects/Processing Effort. As S&W (1996) write, “A nonarbitrary strategy available to cognitively endowed evolved organisms consists in trying to maximize the expected effect/effort ratio” (p. 532).

TABLE 1. Summary of relevance effects.

System side (measured)		Human side (not measured)
<i>tf</i>	Predicted cognitive effect	Actual cognitive effect
<i>idf</i>	Predicted ease of processing	Actual ease of processing
<i>tf*idf</i>	Predicted relevance	Actual relevance

I propose that the logged frequencies of terms co-occurring with a seed term be construed as measuring their predicted cognitive effects within the context of that seed term. The logged inverse document frequencies for the same distribution can be construed as measuring the predicted processing effort of the terms co-occurring in that context. Multiplying the term frequencies of a bibliometric distribution by their inverse document frequencies—the *tf*idf* formula seen in information retrieval textbooks—is equivalent to dividing their contextual (i.e., cognitive) effects by their processing effort, as seen in Goatley’s equation. It also resembles dividing benefits by their costs (cf. Hardy, 1982; Sperber & Wilson, 1995, pp. 123–124). The ratio is a measure of predicted relevance, as in Table 1.

Although relevance is said to vary inversely with effort in Wilson and Sperber (2002, p. 259), they disavow exact quantification. “That is,” they write (Sperber & Wilson, 1995):

Relevance is a property which need not be represented, let alone computed, in order to be achieved. When it is represented, it is represented in terms of comparative judgements and gross absolute judgements, (e.g. ‘irrelevant’, ‘weakly relevant’, ‘very relevant’), but not in terms of fine absolute judgements, i.e. quantitative ones. Since we are interested in relevance as a psychological property, we have no reason to aim for a quantitative definition of relevance. (p. 132)

Whereas the count-based procedure described here does indeed quantify their concept, that is on the system side. On the human side, where interpretation takes place, the user will surely ignore exact numbers in favor of “gross absolute judgements,” based on the orders of magnitude in logarithmic scales. It is interesting to note that this crudely ordinal perspective accords with the view of the user in information space set forth by B. C. Brookes (1980a, 1980b): Information space in this view is not linear but logarithmic.

Peter, Mary, and *Moby Dick*

To illustrate how relevance is created in conversations between human beings, S&W frequently use exchanges between an imaginary Peter and Mary, whom I shall borrow for my purposes here. Assume that Mary is an expert in Herman Melville studies, about which Peter wants to know more. Having just read *Moby Dick*, he wants to extend his knowledge of that novel. The following exchanges serve as reference points for my later discussion on how *tf* and *idf* can be used to create a “recommender system” (Furner, 2002) that simulates Mary’s expertise.

1. Peter: What should I read to follow up on *Moby Dick*?
Mary: Oh, I don't know. For criticism, maybe *Studies in Classic American Literature* by D. H. Lawrence. Or *Love and Death in the American Novel* by Leslie Fiedler. If you want to stick with Melville, try *White Jacket* or *Mardi*.

Mary infers that Peter wants something clearly related to *Moby Dick* but not too specialized. There is no doubt that her reply is relevant—that it is intended to produce informative effects within the context of Peter's question and that neither it nor her suggested readings require undue processing effort (for instance, only parts of the Lawrence and Fiedler books deal with Melville). Mary here is not like an online catalog that, given the seed term, *Moby Dick*, simply spews out direct matches, such as editions of that novel or books with *Moby Dick* in their titles; the four titles she mentions do not contain the string, "Moby Dick." Note also that she would be equally responsive if, in reply to Peter's question, she simply answered ostensibly, by holding up a copy of Lawrence or Fiedler or another Melville novel from her shelves so that their titles could be seen. That sort of ostensive communication is important to my argument, because it resembles the screen-based visual response of a computerized retrieval system in a way that a spoken reply in colloquial English does not.

Now suppose that Mary answers instead with no more than the name of a journal:

2. Peter: What should I read to follow up on *Moby Dick*?
Mary: Oh, I don't know. I have a whole run of *American Literature* over there. Why don't you go through that and look for articles?

Although Mary's response in scenario 2 is also relevant, it is much vaguer and hence less relevant than her response in scenario 1. It does not give Peter specific things to look for and suggests an open-ended browsing task through many volumes of a journal that saddles him with a huge processing effort. Even more effort would be implied if Mary answered with the extremely vague, "Oh, I don't know. How about some essays? Or some novels?"

Here is one final example. Suppose Mary suggests readings that are as specific as those in scenario 1 and that require about the same amount of reading time, but whose connection with *Moby Dick* is much less direct:

3. Peter: What should I read to follow up on *Moby Dick*?
Mary: Oh, I don't know. How about *Hamlet*? Or *The Odyssey*?

Mary's reply in the third scenario would be appropriate if she were suggesting readings in a course on Great Books for Peter, but we are assuming that Peter, having just read *Moby Dick*, wants to continue reading items obviously related to it. Now it is true that literary people are good at extracting meanings from comparisons of very different works, and it is also true that *Moby Dick* has been studied in light of many other classics of world literature. But even a professor of literature would find it easier to relate *Moby Dick* to critical

works like *Call Me Ishmael* or Melville's *Quarrel with God* than to *Hamlet* or *The Odyssey*, and Peter is not a professor (S&W identify him as a surgeon). So, again, because of processing effort, Mary's reply in scenario 3 would be less relevant than her reply in the first scenario.

Mary's replies above could be reduced to the titles she recommends. That is inescapable, because what this Mary portends is, of course, a bibliometrically-based VIRI, a pseudo-Mary. The plight of both bibliometrics and information retrieval is that, at present, they are limited to dealing with noun phrases, of which titles are one example, rather than with language in full (cf. White, 2002). Only noun phrases occur to retrieval system designers as indexing terms, and only noun phrases occur to users as search terms. Only noun phrases pile up as countable tokens across texts (identical sentences occur rarely across different texts and do not pile up in the same way). Without noun-phrase token-counts, there are no term-weights and no core-and-scatter distributions, no relevance-ranked retrievals and no bibliometrics. Until someone devises an artificial intelligence that can go beyond using more or less sophisticated forms of term-matching to answer questions, information science will be a science of noun phrases.

Background and Methods

The exchanges between Peter and Mary find parallels in the demonstrations below. The first reveals tf and idf effects in the bibliometric distribution of works cocited with *Moby Dick*. The seed term, *Moby Dick* creates the context in which, to use S&W's language, the other cited works are "assumptions"—that is, items the system assumes to be relevant in varying degrees. *Moby Dick* was chosen because it is a cultural icon known to persons from many backgrounds. Nevertheless, *mutatis mutandis*, the effects shown here are present across all the relevance-related bibliometric distributions, including those for scientific and technical articles. Subsequent demonstrations in this article will use a cited author and a cited article as seed terms. Other kinds of seed terms appear in Part 2.

The data used in Part 1 involve cited-reference (CR) strings from ISI. For articles, these consist of author, year of publication, volume, initial page, and serial title. An article by Nina Baym in *Publications of the Modern Language Association* is coded as BAYM N, 1979, V94, P909, PMLA.

Baym's title is "Melville's Quarrel with Fiction," but ISI minimizes data entry by never using titles of cited articles as an identifying feature. For books, the CR strings consist of author, publication year, and title, e.g., DOUGLAS A, 1977, FEMINIZATION AM CULT, which designates *The Feminization of American Culture* by Ann Douglas.

Data from parts of the CR strings can be retrieved. Searches on cited authors (CA) return only authors' surnames and initials. (Only the sole or first cited authors of works are returned from ISI data on Dialog.) Searches on cited works (CW) return only serial titles or book titles (often abbreviated, as in the two examples above).

TABLE 2. Results of selecting *Moby Dick* as a cited work (CW) and ranking the first nine works cocited with it.

? Select CW=Moby Dick				
	S1			708 CW=MOBY DICK
? Rank CW Cont Detail				
DIALOG RANK Results (Detailed Display)				

RANK: S1/1-708 Field: CW= Files: 439				
Rank fields found in 708 records—10585 unique terms)				
RANK No.	Items in File	Items Ranked	%Items Ranked	Term
1	708	706	99.7%	MOBY DICK
2	2013	120	16.9%	AM LIT
3	1375	87	12.3%	PIERRE
4	8247	83	11.7%	PMLA
5	314	65	09.2%	AM RENAISSANCE ART E
6	188	65	09.2%	MARDI
7	130	60	08.5%	MELVILLE LOG DOCUMEN
8	148	59	08.3%	CONFIDENCE MAN
9	9073	55	07.8%	LETTERS
10	85	54	07.6%	WHITE JACKET

Particulars of my June 1998 search are given in Table 2. The data are from ISI's Arts & Humanities Search (AHS, file 439 in Dialog). Searching on *Moby Dick* as a cited work (CW) retrieves 708 articles that cited it during 1980–1998 in humanities journals covered by AHS. Dialog's Rank command is invoked to rank the cited works (CW) in the 708 articles by their frequency of cocitation with Melville's novel. The listing is requested to be in continuous (Cont) descending order with full details (Detail). Dialog returns 10,585 cocited items; the first 10 are shown. The "RANK No." at left corresponds to the descending counts in the "Items Ranked" column. *Moby Dick* appears as the top item, as the seed term in a CW listing always will (it was cited twice as *Moby-Dick* and 706 times without the hyphen). The "details" beyond what would appear by default are the counts labeled "Items in File" (used here) and the percentages labeled "% Items Ranked" (not used here).

Because in this case the terms themselves name works (as opposed to, say, authors or subjects), I have actually performed a huge retrieval of documents (which, below, are interchangeably called "works," "items," "documents," and "titles"). As seen in Table 2, the top 10 (some abbreviated ISI-style) are *Moby Dick* itself, four other Melville novels (*Pierre*, *Mardi*, *The Confidence Man*, and *White Jacket*), two journals (*American Literature* and *PMLA*), a work of criticism (F. O. Matthiessen's *American Renaissance: Art and Expression in the Age of Emerson and Whitman*), a biography (Jay Leyda's *The Melville Log, A Documentary Life of Herman Melville, 1819–1891*), and "Letters"—a generic term that presumably refers most often to Melville's published letters, but that could refer to any other author's letters as well. "Letters" exemplifies what Manning and Schütze (2000) call a "semantically unfocussed term"; others include "Works," "Memoirs," "Essays," and "Poems."

In the present analysis, the "Items Ranked" become the term frequencies (tf). The "Items in File" are converted to the inverse document frequencies (idf). Dialog programmers must have realized the potential importance of the "Items in File" counts to make them routinely available for further analysis, but they apparently do nothing further with them in their own system.

Conventionally, term frequencies are counts of the number of times a query term appears in a retrievable document, and document frequencies are the number of documents in a collection that contain that given term. My approach uses another version of term frequencies, based on Dialog's capability for generating bibliometric distributions from a tagged field within retrieved records. The set of records retrieved by a seed term may be regarded as a single large-scale document, for which the Rank command can count the frequencies (i.e., tokens) of all unique terms (i.e., types) in a field of the records, as in Table 2. In effect, each of these terms is ANDed with the seed term, and the frequencies—the tfs—are simply intersection counts. The corresponding document frequencies—the dfs—are the counts of these terms in the collection whether they are intersected with the seed term or not.

The $tf \cdot idf$ weighting formula used here for the i th term in document j is Manning and Schütze's (2000, p. 543):

$$\text{weight}(i,j) = (1 + \log(tf_{i,j}))\log(N/df_i)$$

where $tf_{i,j} \geq 1$. If $tf_{i,j} = 0$, the weight is 0. Bibliometric distributions are noted for having long tails of terms that co-occur with the seed term only once, and because $\log(1) = 0$, adding 1 to $\log(tf)$ restores them. The tfs are converted to logarithms to dampen (or "squash") them "because more occurrences of a word indicate higher importance, but not as much relative importance as the undampened count would suggest" (Manning & Schütze, 2000, p. 542). Base 10 logarithms are used, and N is the total number of documents in the collection. N may be obtained for each ISI database by expanding its Subfile (SF) field in Dialog. AHCI had roughly 2.24 million records as of mid-2005; for 1998, I used the estimate 2 million.

Regarding the inverse document frequency measure N/df , Jurafsky and Martin (2000, p. 653) write, "Due to the large number of documents in many collections, this measure is usually squashed with a log function..." The logic of idf is that the more frequently a term appears across a collection of documents, the less "semantic focus" it has and the less good it is at differentiating them (Robertson, 2004; Sparck Jones, 2004). The measure gives greatest weight, or $\log N$, to a term that occurs only once and a weight of zero to a term that occurs in all documents in the collection.

The titles, *American Literature*, *Pierre*, *PMLA*, and especially "Letters" from Table 2 are examples of terms that move sharply downward when weighted by their idfs. For example, the weight for the journal *American Literature* is $(1 + \log(120)) * \log(2000000/120) = 9.23$, which moves it from second in the Dialog ranking to ninety-third in the $tf \cdot idf$ ranking. In other words, as terms become increasingly common across documents, idf penalizes them for being less

“statistically specific.” I show below how, as terms become less statistically specific, they also require more effort psychologically to process. However, because idf is an inverse measure, it requires terms that cost increasingly *greater* effort to be weighted increasingly *less*. To avoid the cognitive flip-flops this causes, I will reverse the processing effort scale without otherwise changing it by calling it *ease* of processing, so that high values on it mean relatively easy processing and low values, relatively hard processing. Then it will parallel the cognitive effects scale, which is straightforwardly positive, with high values corresponding to high cognitive effects.

Ease of processing has to do with how easy it is to see a connection between a given term and the seed term. For example, it is presumably easy to relate the words “Melville” or “Whale” or “Sea” to *Moby Dick*, but harder to relate the words “Letters” or “Thought” or “Africa” to it. If explicit literary works are being considered, it is easier to relate *Two Years Before the Mast* to *Moby Dick* than to relate *Hamlet* to it.

The connections being processed are obviously superficial. The ease of processing scale does not measure how hard a work is to read, or how hard it would be to incorporate in a new contribution to Melville studies.

High or low values on the cognitive effects scale are determined by the judgments of citers. Compared to relevance judgments in typical retrieval evaluation trials, which tend to be elicited from students or hirelings not actually requiring the retrievals, citations in the ISI databases reveal what real researchers with real projects found worth mentioning to document their claims. As Harter writes (1992, pp. 612–613), “An author who includes particular citations in his list of references is announcing to readers the historical relevance of these citations to the research; at some point in the research or writing process the author found each reference relevant.” Of even greater significance is the piling up of references by multiple authors in the same context—for example, repeated references to Melville’s novella, *Billy Budd*, in the context of references to *Moby Dick*. Repeated references by professionally engaged scholars or scientists are much stronger evidence of relevance than judgments elicited in laboratory settings, and repeated references are what determine values on the cognitive effects scale here. (To some extent, they put the “socio” in “sociocognitive.”)

There is a parallel between references to works in the context of a title-phrase like *Moby Dick* and references to journals in the context of a subject-phrase like *Lubrication*. As is well known, S. C. Bradford (1950) studied the distribution of journals ranked on a logarithmic scale by the counts of the articles they had published in two subjects, lubrication and applied geophysics. This can now be interpreted as a tf scale predicting cognitive effects. For example, a mark of what came to be called a Bradford distribution is that relatively few top-ranked journals yield a disproportionately large set of articles on a subject. In S&W’s language, the journals at the high end of the scale—the “core” journals—produce their greatest effects in the context of a subject term and hence are most relevant to it. More precisely, they are

the system’s assumptions as to what is most relevant (see the extended discussion in Part 2).

Pennant Diagrams

The meanings of tf and idf rankings in a bibliometric environment are clarified by a kind of graphic I have not seen before, although its mechanics will be familiar enough. It is simply a scatterplot of the logged data in the *Moby Dick* retrieval, the first 10 cases of which appeared in Table 2. Here, tf values of the Items Ranked are placed on the *x* axis as cognitive effects, and idf values of the Items in File are placed on the *y* axis as ease of processing. Figure 1 (made with DeltaGraph, Red Rock Software, 2005) shows the resulting shape. It will be obvious why I call it a pennant diagram.

The $tf \times idf$ formula calls for a multiplication of two sets of values and a single product for each item. In contrast, pennant diagrams plot each item on separate coordinates, leaving tf and idf unmultiplied. As a result, the effects of the weights are visible before they enter the formula, which can be informative. However, pennants can also show the effect of multiplying tf by idf, as in Table 4 and Figure 3 below.

Each point in Figure 1 represents an item from the *Moby Dick* literature or from other literatures that include references to *Moby Dick*. To lay bare the shape of the distribution, the points are as yet unlabeled. The rightmost point is *Moby Dick* itself because (mixing metaphors) the seed term will always be at the tip of the pennant. Whatever the seed term is—a book, an article, an author, a journal, a descriptor—it will always have the greatest effect in its own pennant, where it is itself the object of study. (In *Moby Dick* studies, the item most relevant to read is *Moby Dick*, as many students have discovered the day of the test.)

A word more on the nature of the seed term: In recent years, I have worked on literature visualization systems that

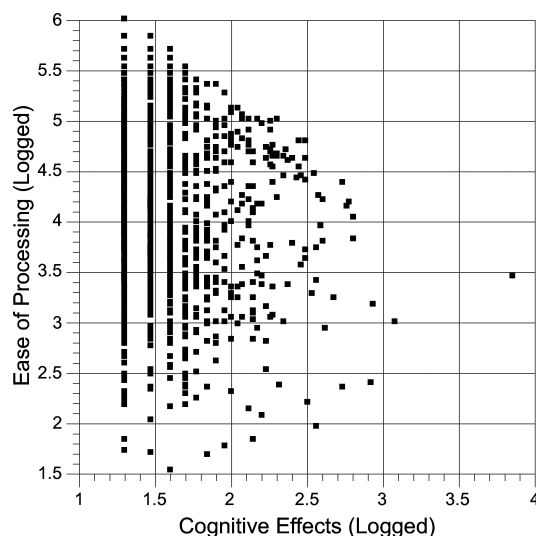


FIG. 1. Pennant diagram for *Moby Dick* studies.

create displays on the basis of a single name or descriptor, the intent being to minimize the cognitive load on users (see, e.g., White, Lin, & Buzydlowski, 2004). Given the well-documented difficulties many people have with online searching, including Web searching, the minimization of cognitive effort should be a primary goal of design. An attractive feature of pennant diagrams is that they can be generated from single terms that many users should be able to supply, such as the name of a work (e.g., *Moby Dick*) or the name of an author (e.g., Herman Melville). They can also be generated from a single descriptor or keyword, in which case the rest of the pennant might consist of the other descriptors that co-occur with the seed term—descriptors that can be *recognized* for inclusion in a search strategy rather than requiring lookup in a thesaurus. Furthermore, a pennant diagram can be generated from any combination of terms (e.g., a statement with ANDs, ORs, and NOTs) that yields enough documents to make plotting the set worthwhile. Therefore, pennants are a flexible kind of graphic.

Broadly speaking, the documents represented by the points in Figure 1 become increasingly relevant to *Moby Dick* as the pennant narrows rightward. The leftmost columns are least relevant; the points nearest the seed term are most relevant. In like fashion, the points along the top of the pennant are relatively easy to process, in the sense of discerning their relevance to *Moby Dick*; the ones along the bottom are relatively difficult.

For simplicity of presentation, the leftmost column in Figure 1 comprises items cocited twice with *Moby Dick*. Almost 8900 items cocited with it only once have been omitted, which cuts the number of items to 1711. The march of the columns rightward across the figure represents the items that are cocited with Melville's novel three times, four times, and so on. The columns are formed by large numbers of items tied in rank on cognitive effects, and they diminish in length because the documents cocited more and more frequently with *Moby Dick* become fewer and fewer. As one moves rightward on the cognitive effects scale, the numbers of ties progressively thin out, until finally tied values are replaced by unique values. (Such a shift is a well-known feature of bibliometric distributions; see, e.g., Nelson & Tague, 1985.)

A similar effect is visible in the columns of pennants. Many of the points down the columns are in fact, thick pile-ups of points representing items tied in value on the ease of processing scale. However, as one moves lower on that scale, the ties again thin out. While the points along the upper edge of the pennant are tightly bunched, those along the bottom break into a loose fringe. The latter represent items with unique rather than tied values on the processing scale (just like the ones at the right of the cognitive effects scale). These items occur most frequently in the entire collection. Tightly serried columns loosening rightward and downward into unique values are typical of pennant diagrams made from bibliometric distributions.

Pennant shapes occur for all the bibliometric distributions I have plotted to date: works cocited with a work, authors cocited with an author, references cocited with a reference,

journal titles co-occurring with a subject heading (Bradford's distribution), and descriptors co-occurring with a descriptor. Allowing for differences in the units of analysis, the pennants are all interpretable in the same way. The structures explained below seem very durable, and they flow directly from Dialog data with simple algorithmic processing.

In the case of bibliographic distributions of cited documents, such as we have in Figure 1, the pennant diagram can be interpreted as the record of individual differences in citers' perceptions in various domains over time. It aggregates what hundreds of citers saw as relevant connections for the seed work, here *Moby Dick*, over a multiyear period. It could also aggregate the perceptions of journal editors choosing authors and works to publish, or of indexers choosing authors and works to bring under subject headings. All pennant diagrams have this psychological side, as will be discussed later. If pennant diagrams render literatures better than displays such as the bibliograph (Brookes 1973), it is because they convey more information and are intended from the outset to be intelligible to nonmathematicians. Within the limits of present interface technology, points in the pennant can be labeled for easy recognition—an enormous advantage if bibliometric data are to be widely useful.

Even so, it should be remembered that the items plotted in Figure 1 are merely unedited verbal strings. The computer cannot tell that two nonidentical strings in ISI abbreviation form, such as *Anatomy Criticism* and *Anatomy Criticism 4*, are really references to the same work (Northrop Frye's *Anatomy of Criticism: Four Essays*). They appear as two separate points in the pennant diagram, based on their own separate counts; ideally, they should be combined. White (2001a) introduced the term "allonyms" for such strings, which complicate almost any study that relies on data from ISI. Allonyms are caused by inadvertencies like misspellings, omissions of publication year, inconsistent forms of abbreviation or pagination, slight differences in punctuation, and so on. Were such allonyms merged and their counts combined (by me or, better, by ISI), it would considerably cut down on the 10,585 items to be plotted in the retrieval in Figure 1. References to different editions or printings of the same work (which differ only because of publication year) are also candidates for having their counts combined. I have combined only a very few counts here.

Sectors of the Pennant

Figure 2 introduces the structure of meaning in pennant diagrams. It shows Figure 1 overlaid with hand-drawn lines creating three sectors, A, B, and C, enclosing, respectively, *subordinate*, *coordinate*, and *superordinate* items relevant to *Moby Dick*. Although the sectors are not precisely measured and not every item in them clearly fits my labels, the sectors suggest broad, qualitative gradations within the diagram. (If the qualitative differences of the sectors can be preserved, it seems worth trying to produce the sector boundaries algorithmically based on different tf and idf ratios.)

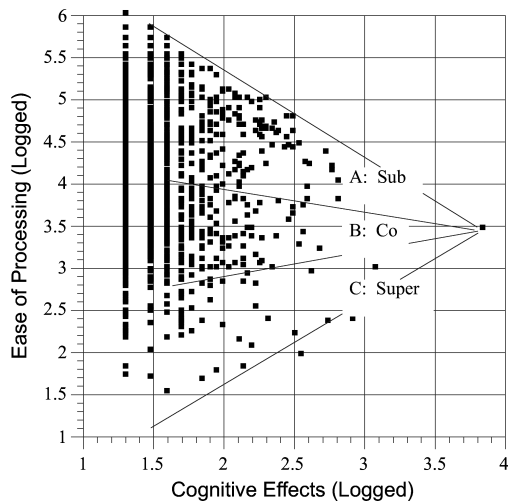


FIG. 2. Semantic sectors for *Moby Dick* studies.

It has long been known that citations, in this case references to books and serials, can be understood as a form of subject indexing (Garfield, 1979, pp. 2–10). Not surprisingly, they here exhibit a hierarchical structure similar to what one finds with descriptors in thesauri. However, we are not dealing here with conventional semantic hierarchies of the sort seen in thesauri or WordNet. If we think of *Moby Dick* as having many different cognitive implications for thousands of citers, it is more accurate to say that the subordinate documents in sector A narrow its cognitive implications, whereas the coordinate and superordinate documents in sectors B and C increasingly widen them. With respect to ease of processing, it is easiest to see the relevance to *Moby Dick* of documents in sector A, and increasingly difficult to see the relevance of documents as we move through sectors B and C. The latter documents are by no means irrelevant to *Moby Dick*; they simply require more expertise, imagination, or effort to connect to it.

The documents in sector A are the ones most statistically specific to the seed term, the ones most cohesive and coherent with *Moby Dick* studies proper. “Cohesive” and “coherent” are used here as they are in discourse analysis. “Cohesive” means that the phrase “*Moby Dick*” and, by implication, its subtitle “*The Whale*” and author “Herman Melville” are explicitly repeated across documents (which is the “lexical cohesion” of Halliday & Hasan, 1976). “Coherent” means that there is an implicit connectivity of sense or logic across documents that we can readily understand from what we know of the world, even if we know little about *Moby Dick* and Melville studies per se (cf. Beaugrande & Dressler 1981; Brown & Yule, 1983). One terminological change: in discourse analysis, “cohesion” and “coherence” refer to text-forming relations within texts, whereas I will here be referring to relations across texts. To maintain this distinction, I will use “intercohesion” and “intercoherence” for cross-textual relations (White, 2002).

In the context of Melville studies, the most relevant items are those that can be related to *Moby Dick* with little effort or expertise; their titles or subtitles alone are informative enough

to make the determination. These appear in sector A, the sector of topical relevance in which persons without special subject competence can operate. A book called *The Trying Out of Moby Dick* is about *Moby Dick*, as your cat can see. A slightly less revealing title, *Call Me Ishmael*, duplicates the famous first sentence of the novel. Books with poetic titles like *The Wake of the Gods* are brought into the fold through their subtitles (*Melville’s Mythology*). Also appearing here are other writings by Melville, some prominently involving ships, sailors, and voyages. Other items in A are about whales as animals and whaling as an industry. Such writings preserve the implications of *Moby Dick* within relatively narrow confines, and in that sense they are “subordinate” to it. To recognize them, one need not have any special claim about what *Moby Dick* means as a work of art. One need not even have read it (or seen a movie version); given the titles, a knowledge of English and a bit of cultural literacy will suffice.

Sector B begins where people in general can no longer easily see relations of intercohesion and intercoherence among items. It contains artistic or critical works that require insight or special knowledge to relate to the seed term; connections are not obvious from their titles or authors. Many sector B documents exhibit relevance relations that are probative or evidentiary rather than topical—that is, they count for or against a claim (cf. P. Wilson, 1968, pp. 43–44; Walton, 1989, pp. 78–79), including claims of analogy. (In practice, this means that items in B tend to be linked to the seed by citation-bearing sentences rather than by subject headings—a difference with big implications for retrieval.) Only by claiming that something is the case can one relate *Moby Dick* to “coordinate” writings in sector B such as *The Scarlet Letter* or *Walden* or *Huckleberry Finn*, which are clearly very different from it in topic. If it is discussed with them, then the grounds must be at some deeper level than superficial aboutness.

Scholars and scientists (and some students) exist to make claims at these deeper levels. Minimally, one needs to know such things as that critical works in sector B like F. O. Matthiessen’s *American Renaissance* or Leslie Fiedler’s *Love and Death in the American Novel* contain discussions of *Moby Dick* even though their titles (and their library subject headings) do not say so. To go further, a certain creative leap is needed, of the sort that literary scholars and critics routinely make. For instance, if one believes with Fiedler that, in classic American literature, sexual love between men and women is largely replaced by thinly veiled homoeroticism between light-skinned and dark-skinned men, then *The Last of the Mohicans*, *The Narrative of A. Gordon Pym*, *Moby Dick*, and *Huckleberry Finn* may all count toward that claim, and *The Scarlet Letter* may count against it (or will need explaining). In a work substantiating such a claim, they would all be cocited (as they are in Fiedler’s book).

In the humanities, acres of exposition are devoted to establishing relevance relations of this deep or nonobvious sort among imaginative works. The necessary insights involve what Koestler (1964) called “bisociation”—the fusion of concepts from hitherto separate matrices of thought that marks creativity in all the arts and sciences. This is just

another way of saying that relevance of the deeper, nontopical sort is created, not simply gleaned by knowing English and living in the world. Don R. Swanson, for one, has made exactly this point (Swanson, 1977), and his own bisociations of hitherto unconnected writings (most famously, writings on the effects of fish oil and writings on Raynaud's syndrome) have created new relevance relations in medicine just as Leslie Fiedler's did in literary studies. Swanson the scientist explores hidden causal relations, whereas Fiedler the humanist explores hidden analogical relations, but both exhibit considerable originality of mind. Bisociations like theirs across texts are not something one can routinely expect from students—or from indexers (see Part 2).

Quantitatively, items appear in sector B rather than A because the inverse of their higher counts under Items in File gives them lower scores on the ease of processing scale. What this means qualitatively is that they admit of more various interpretations than items in A. For example, works like *The Trying Out of Moby Dick* and *Call Me Ishmael* are pretty much confined to *Moby Dick* studies, but works like *The Scarlet Letter* or *Huckleberry Finn* have obvious uses in contexts well beyond *Moby Dick* and are cited accordingly. So are critical surveys like the Matthiessen and Fiedler books. Being cited in these additional contexts drives up their counts in the Items in File column. Thus, although the documents in sector A have identical or similar counts under Items in File and Items Ranked, the documents in sector B have counts under Items in File that are high in relation to their counts under Items Ranked.

This simple disparity can produce interesting intelligence. If a work appears in many contexts beyond that of the seed term, its connection to the seed term will be relatively hard to see, hard to process. If, nevertheless, multiple citers have in fact seen it, as witnessed by the work's relatively high value on the cognitive effects scale, then the connection is probably worth seeing—more worth seeing, literary types would argue, than the obvious connections in sector A. In other words, it takes commonsense to relate *Moby Dick* to *Call Me Ishmael*; it takes deeper insight to relate *Moby Dick* to *Huckleberry Finn*. "Relate" here means something like "Explain the connection, preferably thematically, while standing on one foot."

Sector C begins where the terms become still less specific, shading over into titles of journals, names of genres (and other broad-gauge nouns) and world classics. Here is where one finds the aforementioned *American Literature* and *PMLA* and terms like "Letters," "Essays," "Collected Works," "Art," and "Nature." Here also are *Job*, the *Iliad* and the *Odyssey*, and *Paradise Lost*. Documents in sector C have even higher counts under Items in File (perhaps an order of magnitude higher) than documents in sector B. Such items are superordinate to *Moby Dick* in the sense that they represent its linkages to generic literary vocabulary and to highly cited titles in the world canon. Because of their broad scope, the relevance of sector C items to *Moby Dick* is not at all apparent, and uncovering their actual relationships to Melville's novel takes considerable effort.

It could be, of course, that *Moby Dick* is being cited with these latter items in a disjoint, uninteresting way. If not, it would seem to require a Mortimer Adler or a Harold Bloom, not any ordinary intellect, to relate *Moby Dick* to *Hamlet* or Joyce's *Ulysses*, both of which appear in sector C. (One tie, says Williams, 1963, is the use of interior monologue, "The flow of meditation in *Moby-Dick* points back to *Hamlet* but also forward to James Joyce," p. 255.) In general, the more an item is used in contexts other than that of the seed term, the more difficult it will be to relate to the seed term. Moreover, items in C imply physical as well as intellectual effort. For instance, as noted above, *Moby Dick* is cocited with 120 items in the journal, *American Literature*. Reasons why can be guessed at, but actually tracking down those 120 items to learn what they are and how they relate to *Moby Dick* would demand considerable literature searching, library-going, and downloading from the Web. So would tracking down and evaluating the generically titled works (e.g., "Letters," "Narrative," and "Essays") with which *Moby Dick* is cocited. It follows that we can demote such hard-to-process terms as less relevant to *Moby Dick* than those in the first two sectors, while still admitting that they are relevant to a degree—or so the system predicts.

Deirdre Wilson, one of the originators of RT, may seem to contradict this argument when she says that statistically frequent words like "brothers and sisters" are easier to process than statistically rare words like "siblings" (Wilson, 1994, p. 46). However, she is using word counts as an indicator of familiarity of vocabulary and implying that, in ordinary conversations, familiar words are more understandable than jargon. I, on the other hand, am using word counts as an indicator of specificity in judging the relevance of titles to a seed term. In this more abstruse area, familiar words are harder to process because their relation to the seed term is vaguer and less obvious.

Some of the difficulties in processing items in sector C are simply artifacts of analyzing cited works (CW) rather than cited references (CR). Serial titles, for example, are relatively opaque at best, and serial titles by themselves crop up only in CW analyses. The same is true of names of genres, like "Letters," "Poems," or "Memoirs." In a CR analysis, these would be identifiable as, e.g., Herman Melville's *Letters* or Emily Dickinson's *Complete Poems* or Tennessee Williams's *Memoirs*, but in a CW analysis, such disambiguating information is lost. In like fashion, a CW analysis cannot distinguish homonymic names of works (e.g., Emerson's essay, "Nature" is conflated with the word "Nature" appearing in other titles).

The opposite problem of allonymic fragmentation also occurs. For example, the journal *American Literature* is sometimes abbreviated *Am Lit* and sometimes *AL*; these abbreviations have separate counts and thus separate points on the pennant. However, a CW analysis works well enough for present purposes, because it shows the range of effects when tf and idf counts are used with unedited terms from ISI databases. Editing the terms and counts would change the placement of some points, but not the overall meaning of the sectors.

All of what has been said in this section should be easy enough to link to the earlier dialogue between Peter and Mary, to which we will return below.

Content Analysis of the Pennant

A content analysis of the pennant diagram reveals more concretely, what is going on in its different parts. While this analysis is devoted to *Moby Dick*, a corresponding one can be performed for any relevance-related bibliometric distribution, and comparable effects should be visible.

Table 3 sets up the discussion. It uses integer values from the axes of Figures 1 and 2 to define cells, and shows the number of items within them. The integer values serve to identify sets of titles in the cells (row number is given first, column second). The single boldfaced item at 3–4, for example, is *Moby Dick* itself. The counts in (approximate) sector A are italicized; those in B are unmarked; those in C are underlined.


The verbal data are displayed in Table 4. It is structured like Table 3, except that fewer than one tenth of the documents in Table 3 are listed, and cell 3–4 for *Moby Dick* has been replaced with a whale doodle so that listings in the other cells will fit into a single column. Titles are abbreviated just as they came from ISI (allonyms for several works are visible). If a cell has no more than 12 documents, all are given; otherwise only the top 12 appear, ranked by their tf*idf scores. Although this greatly understates the content of cells with hundreds of documents, it suggests the semantics of the pennant reasonably well.

The tf and idf rankings powerfully elevate Melville studies to the upper part of Table 4. Items that can be readily linked to *Moby Dick* or its author on the basis of titles alone have been italicized. (Many ISI abbreviations need expansion—e.g., *Pacifism Rebellion W* is John Bernstein's *Pacifism and Rebellion in the Writings of Herman Melville*. The perhaps unfamiliar *John Marr*, *After [the] Pleasure Party*, *Clarel*, and *Battle Pieces* are titles from Melville's poetry.) Cell 4–3 at mid-right contains the 11 items highest on both the cognitive effects and ease of processing scales and hence most relevant to *Moby Dick*. Up and leftward, the cells 5–2, 6–2, and 6–1 are strongly identifiable with Melville studies (including hundreds of the items not shown). Cell 6–2 has only 10 documents, which may be checked in their entirety. Among the 168 items in 5–2, the six most common words are (spelling out abbreviations) Melville, 30 occurrences; Melville's, 24; American, 21; Moby, 7; Dick, 7; and Literature, 6. Among the 232 items in 6–1, the six most common words are Melville, 22;

TABLE 3. Counts in cells of the *Moby Dick* pennant.

Ease	Effects				Total
	1	2	3	4	
6	<i>132</i>	<i>10</i>			142
5	397	168			565
4	519	182	<i>11</i>		712
3	119	127	6	1	253
2	<u>17</u>	<u>18</u>	<u>4</u>		39
Total	1184	505	21	1	1711

TABLE 4. Selected items in cells of the *Moby Dick* pennant.

6-1: 232 items <i>Moby Dick Doubloon E</i> <i>Benito Cereno Hdb</i> <i>S Bartleby Scrivener</i> Crisis Life Writings <i>John Marr</i> <i>Extracts Occasional</i> <i>Melville Moby Dick O</i> <i>Melville par Lui Mem</i> <i>H Melville Moby Dick</i> Modern Anthology <i>Checklist Editions M</i> <i>Melville Piazza Tale</i>	6-2: 10 items <i>Moby Dick Doubloon</i> <i>Natural Hist Sperm W</i> <i>H Melville Annotated</i> <i>Wake Gods Melville M</i> <i>After Pleasure Party</i> <i>Pursuing Melville</i> <i>Clarel Poem And Pilg</i> <i>Pacifism Rebellion W</i> <i>White Jacket Or Wort</i> <i>Melvilles Marginalia</i>	
5-1: 397 items <i>Moby Dick Whale</i> There Was Child Went Indian Antiquities <i>Redburn His First Vo</i> <i>Melville Collection</i> <i>H Melville Represent</i> Themes Directions Am <i>Monsieur Melville</i> <i>Am Whaleman Study Li</i> <i>Melvilles Reading Re</i> <i>Battle Pieces</i> <i>Account Arctic Regio</i>	5-2: 168 items <i>Trying Out Moby Dick</i> <i>Ishmaels White World</i> <i>Melville Moby Dick</i> <i>Salt Sea Mastodon Re</i> <i>Melvilles Reading Ch</i> <i>Moby Dick Centennial</i> <i>Melvilles Reading</i> <i>New Perspectives Mel</i> <i>Melvilles Thematics</i> <i>H Melville Tragedy M</i> <i>Omoo</i> <i>Melvilles Orienda</i>	
4-1: 519 items Devil Deep Blue Sea Disparition Essays Lectures Life RW Emerson Return Vanishing Am Gnostic Relig Castle of Otranto New England Lit Cult American Notebooks Middle Passage In American Grain Lord of Flies	4-2: 182 items <i>Piazza Tales</i> <i>H Melville</i> <i>Redburn</i> <i>H Melville Biography</i> <i>Benito Cereno</i> <i>Bartleby Scrivener</i> Symbolism Am Lit Am Hieroglyphics Love Death Am Novel Am Transcendental Q Am Adam Great Gatsby	4-3: 11 items <i>White Jacket</i> <i>Melville Log Document</i> <i>Confidence Man</i> <i>Mardi</i> <i>Lett H Melville</i> <i>Typee</i> <i>Subversive Genealogy</i> Am Renaissance Art E <i>Billy Budd</i> Studies Classic Am L Scarlet Letter
3-1: 119 items Consequences Pragmat Ecclesiastes Romeo and Juliet Georgia Rev Ency Philos Plague U Toronto Q New York Rev Books Language Counter Mem Glyph Awakening European Lit Latin M Anxiety Influence	3-2: 127 items Complete Poems Am Scholar Texas Studies Lit La Modern Language Q <i>Melville</i> Collected Poems Anatomy Criticism Sewanee Rev Complete Works Truth Method Modern Language Note Poetical Works Rhetoric Fiction	3-3: 6 items <i>Pierre</i> Am Lit  Walden New England Q 19th Century Fiction Am Q
2-1: 17 items Poetics Crisis Africa J Philos Interpretation Discourse Poetry Novel Ethics Opera Natural Thought	2-2: 18 items Poet Writings Poems Communications AL Experience Narrative Republic Criticism Nation Works Style	2-3: 4 items PMLA Letters Cited Indirectly Nature

American, 6; Life, 5; Moby, 5; Dick, 5; and Whaling, 5. Some counts would be even higher if the full titles of works were not reduced to ISI's abbreviations.

These items exemplify sector A, whose lower border runs roughly from *Billy Budd* in 4–3 through *Bartleby [the] Scrivener* in 4–2 to *Account Arctic Regio* in 5–1 (the latter is William Scoresby's *An Account of the Arctic Regions with a History and Description of the Northern Whale-Fishery*, first published in 1820, 31 years before *Moby Dick*).

Idf weights for the original Items in File counts put *Billy Budd* down into sector B and *Pierre* down into sector C, near *Am Lit*. These are errors; ISI conflates Melville's *Billy Budd* with the opera Benjamin Britten based on it, and Melville's *Pierre* with many other works having "Pierre" in their titles. I have corrected the weights here and in subsequent figures.

Cells 4–2 and 5–1 are populous border cells in which the top-ranked items include some Melville studies but which quickly become heterogeneous in character as they shade into sector B. The six commonest words in 4–2 show that the strong Melville identification has disappeared: they are American, 28; Stud[-y or -ies], 11; Review, 8; Fiction, 5; Novel, 5; and Literature, 4. In 5–1, they are American, 45; Literature, 17; History, 13; Melville, 9; Poe, 9; and Review, 9. Because 5–1 is higher on the ease of processing scale than 4–2, it should be easier to find obviously "Melvillean" items in it, and the nine occurrences of Melville's name bear this out.

Sector B items, left in roman, are those in which the link to Melville is no longer plain to the uninitiated. Along the upper border of B might be put three items from 4–3: *Am Renaissance Art E* (Matthiessen's book, noted above), *Studies Classic Am L* (D. H. Lawrence's *Studies in Classic American Literature*) and *[The] Scarlet Letter*. Continuing this trend are several works in 4–2, from *Symbolism Am Lit* (Charles Feidelson's *Symbolism and American Literature*) to *[The] Great Gatsby*. These "coordinate" works project cognitive implications of *Moby Dick* quite different from those of the relatively briny sector A, with its items like *Nat Hist Sperm W[hale]* and *Salt Sea Mastodon*. Sector B contains works of criticism and history that relate *Moby Dick* to other American or foreign classics. It also contains the classics themselves, e.g., *The Scarlet Letter* and *The Great Gatsby*. In cell 4–1, which holds a 519-item miscellany, one notes that *Moby Dick* has been cocited with works as diverse as Horace Walpole's *The Castle of Otranto*, William Carlos Williams's *In the American Grain*, and William Golding's *Lord of the Flies*.

A key point about sectors B and C is that they greatly extend the notion of relevance. Table 5 attempts to drive this home by presenting from the invisible parts of cells 4–2 and 3–2 some 98 classics of various kinds that have been cocited multiple times with *Moby Dick*. They were arbitrarily selected from hundreds more in the table. Despite ISI's title abbreviations, most will be recognizable to readers literate in the humanities. Citers in humanities journals appear to be an intellectually promiscuous lot; obviously *Moby Dick* is not a unitary concept symbol like the papers Small (1978) uncovered in chemistry. But even in the sciences, cocitation creates relevance relations that extend far beyond identity or even similarity of topic (cf. Part 2).

TABLE 5. Selected literary works cocited with *Moby Dick*.

<i>Cell 4-2</i>	<i>continued</i>	<i>continued</i>
Red Badge of Courage	Crying of Lot 49	Prairie
Leaves of Grass	Rime of Ancient Mari	Interpretation Dream
House of Seven Gable	Great Chain Being	Grammatology
Blithedale Romance	Doctor Faustus	Waste Land
Wise Blood	Four Quartets	King Lear
Mosses from Old Mans	To Lighthouse	Golden Bough
Invisible Man	Hero with 1000 Faces	Portrait of Artist a
Huckleberry Finn	Shadow and Act	Is There Text This C
Palm at End of Mind	Souls Black Folk	Faust
Week Concord Merrima	Anatomy Melancholy	Paradiso
Self Reliance	Against Interpretation	Iliad
Look Homeward Angel	Liberal Tradition Am	Cantos
Common Sense	Writing Degree Zero	A la Recherche du Te
Uncle Toms Cabin	Bear	Faerie Queene
Democracy in America	Portrait of Lady	Pensees
Religio Medici	Will to Power	Odyssey
Unbearable Lightness	Merchant of Venice	Archaeology Knowledge
Slaughterhouse Five	Lolita	Speech Acts
Heart of Darkness	Virgin Land	Being and Time
Sartor Resartus	Man His Symbols	Illuminations
Sound and Fury	Brothers Karamazov	Order Things Archaeo
Go Down Moses	Great Expectations	Macbeth
Education H Adams	Richard III	Philos Investigation
Armies of Night	Mirror Lamp Romanti	Essay Human Understa
Mr Sammlers Planet	Divine Comedy	Orientalism
Gravitys Rainbow	Protestant Ethic	Hist Sexuality
Light in August	Civilization Its Dis	Isaiah
Catch 22	Principles Psychol	Confessions
Absalom Absalom	<i>Cell 3-2</i>	Job
Raven	Paradise Lost	Hamlet
Marble Faun	Finnegans Wake	Metamorphoses
Ambassadors	Ulysses	Laws
Moll Flanders	Biographia Literaria	Genesis

Table 5 was made simply by listing famous books (minus serials and books with generic titles like *Autobiography*). One effect of the idf values on statistical specificity was therefore not apparent until after the table was made. *Moby Dick* is a classic of *American* literature, and most of the items listed from cell 4–2, which is the higher of the two on the ease of processing scale, are American in origin, whereas most of the items from cell 3–2 are foreign. The implication is that, in the context of *Moby Dick* studies, American works are more relevant than foreign works. This is filtering of a degree of subtlety probably unforeseen by proponents of tf*idf weighting, and it depends solely on the interplay of their counts under Items Listed and Items in File.

The border between coordinate works in sector B and superordinate works in sector C appears where serials and generic works begin to be plentiful. Examples may be found in the bottom rows of Table 3 (many other serials in rows 3 and 2 are not shown). The thing to notice about serial titles in this context is their opacity. Titles like *American Scholar*, *Texas Studies in Literature and Language*, *Modern Language Quarterly*, and *Sewanee Review* from cell 3–2 are indeed relevant to *Moby Dick*, but in ways that, without further lookups, are not guessable. The relevance of titles like cell 3–3's *American Literature*, the *New England Quarterly*, and *19th Century Fiction* is plainer (*Moby Dick* is a 19th-century American fiction about

a New England whaling ship) but still only very generic. In this, the latter titles resemble the single terms that characterize row 2 at the bottom of the table. These single terms are actually references to works—for instance, “Nature” often refers to Emerson’s essay by that name; “Republic” often refers to Plato’s *Republic*—but their connection to *Moby Dick* is not easily inferred, as I have noted, and many people, on seeing them, would not even know that they are titles.

What go to the bottom in $tf*idf$ filtering are items that, although they are titles, seem like ordinary words—that is, vocabulary that everyone shares. Thus, nonexperts who knew nothing of the specific works in the top rows of Table 3 (that is, works like *Pacifism and Rebellion in the Writings of Herman Melville* or *The Wake of the Gods: Melville’s Mythology*) would know, in a sense, the titles at the bottom simply by knowing English. The point bears on something to be discussed again in Part 2—the way people who lack the names of specific writings search the Web or ask for guidance at a reference desk. Many use words that are vague and generic, which is what makes their searches difficult. That is exactly what the relatively low values of such words on the ease of processing scale would predict. For instance, superordinate words from titles like *Criticism* and *Novel* are not good search terms when coordinate *Symbolism in American Literature* or subordinate *Melville’s Reading* would capture the real interest. Pennant data may thus prove useful in modeling search vocabularies, from the precise terms used by the most sophisticated searchers to the vague terms used by the least.

Tf*idf and Retrieval

What does $tf*idf$ weighting do with respect to overall retrieval? Sparck Jones and Willett (1997, p. 307) justify it as a corrective to user behavior:

The basis for IDF weighting is the observation that people tend to express their information needs using rather broadly defined, frequently occurring terms, whereas it is the more specific, i.e., low-frequency terms that are likely to be of particular importance in identifying relevant material. This is because the number of documents relevant to a query is generally small, and thus any frequently occurring terms must necessarily occur in many irrelevant documents; infrequently occurring terms have a greater probability of occurring in relevant documents—and should thus be considered as being of greater potential when searching a database.

This nicely states the reasoning behind the idf factor, and it is correct as far as it goes. However, we will need to qualify it after looking closer at *relevance*, a notoriously tricky notion.

When $tf*idf$ weighting is used to rank the 1711 items of the *Moby Dick* distribution by relevance, the results are much like what we have already seen. Those highest in statistical specificity go to the top of the list, with *Moby Dick* at the head, and those lower in specificity go progressively to the bottom. The idf factor suppresses items that are less obviously related to *Moby Dick* and elevates items that are more obviously related to it. If the list is plotted in stages as a pennant diagram—say, 100 items at a time—items high on

the ease of processing scale are plotted first as we proceed down the ranking. For example, if the top 100 items in the $tf*idf$ ranking are plotted, the items are overwhelmingly high on ease of processing, as Figure 3 shows.

$tf*idf$ weighting does permit intermingling of items from different sectors, but items lower on the ease of processing scale must be very high on the cognitive effects scale to make the top 100. Here, only eight items qualify—seven in sector B (from *Am Hieroglyphics* at upper left to *Am Renaissance Art E* at lower right) and one in sector C (the journal *Am Lit*, bottom right). Because of the idf effect, sector A will be filled much faster than the other two in progressive plotting. For example, if one includes the top-ranked 300 items in the pennant, one finds items still being disproportionately added to sector A.

A few adjustments in Figure 3 need explaining. The problem of overlapping points and labels worsens when pennant diagrams are reduced for journal publication. If true scale values are preserved, the labels must be in extremely small type, and, leftward, grow unpresentably dense. Here, by shortening the axes at both ends, moving *Moby Dick* from 3.85 to 3.30 on the horizontal axis, and disentangling label pile-ups, I can display the items with the greatest cognitive effects in the three sectors.

It is worth poring over the packed upper labels to appreciate just how successful the idf effect is. Several works with *Moby Dick* specifically in their titles are highest in sector A. Below them are many works with Melville in their titles (more would be added if titles were given in full). Along the bottom of the sector are major works by Melville, from *Pierre* leftward to *Clarel*.

However, Figure 3 shows that $tf*idf$ weighting markedly privileges topical relevance over any other kind, and this kind of relevance is not the whole story. Were the *Moby Dick* distribution presented simply as a list rather than as a pennant diagram, most users would scan the list from the top down. If we imagine, generously, that they are willing to scan 100 titles, they would still encounter mostly works from Melville studies and nothing like the full variety of items relevant to *Moby Dick*. In that sense the $tf*idf$ result is deceptive, because most people break off scanning well before 100 items (perhaps when they reach a “problematical” item like *The Scarlet Letter*), as studies of responses to lengthy Web retrievals reveal (see Yang’s 2005 review).

The real significance of $tf*idf$ weighting thus lies in how it directs attention. It is designed to put the items whose relevance is easiest to see where people are most likely to see them and to put titles whose relevance is harder to see down where people are less likely to look. It fills up the plausible browsing space with items that anyone can match on noun phrases—items that are topically relevant—and not with items that are relevant only to a special claim. It thus seems fair to say that lists ranked by $tf*idf$ weighting are designed to appeal to people without special claims, people who can make only the easier relevance judgments—students, librarians, readers unfamiliar with a literature, hired judges in information retrieval experiments. Presumably, the designers of document retrieval systems want this outcome because it makes their systems look good to anyone, expert or not. In this, architects of $tf*idf$

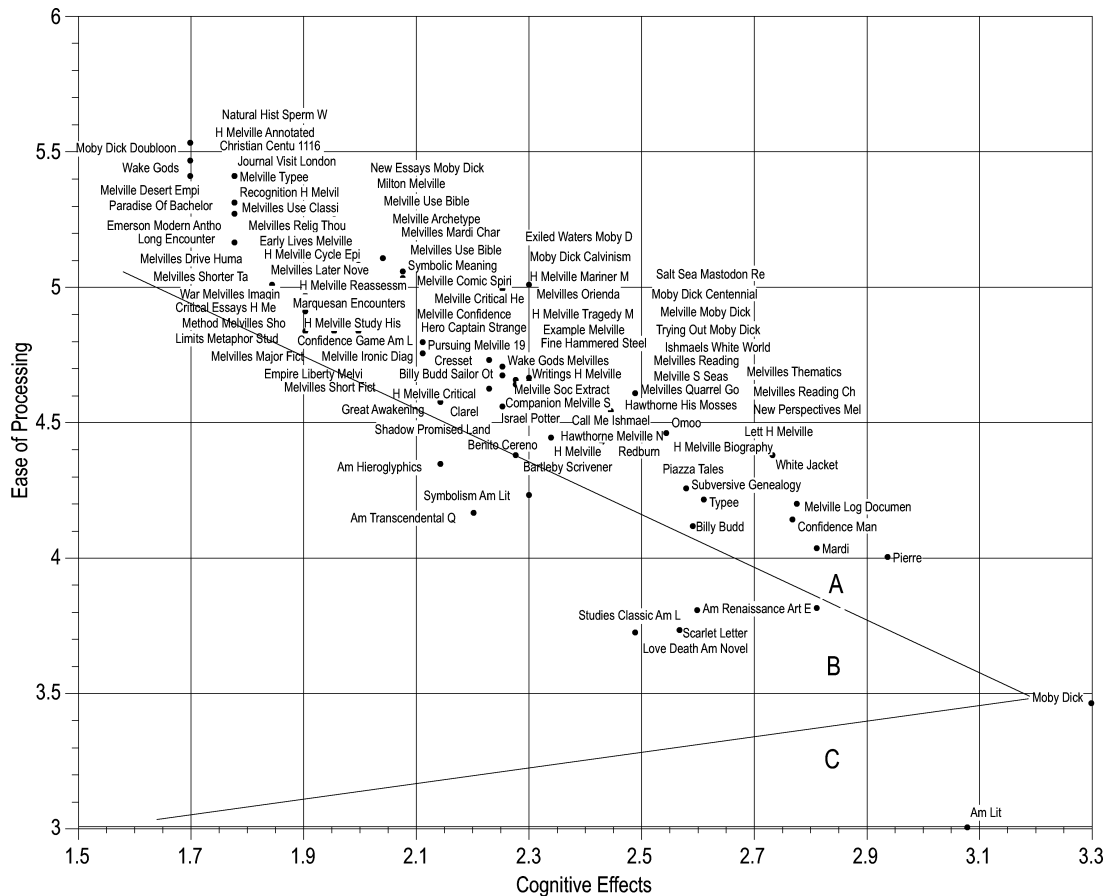


FIG. 3. Pennant diagram of top 100 items in *Moby Dick* distribution after $tf*idf$ weighting.

schemes are no different from librarians, and in fact $tf*idf$ weighting is obviously related to library classification: see if Figure 3's automatically generated upper sector does not resemble the Melville stacks of a university library.

The ISI citation record reveals, however, that scholars and scientists repeatedly make relevance judgments of less obvious kinds. Melville scholars cite items from other literatures, and specialists in other literatures cite items from Melville studies. The results are found in sectors B and C. The pennant diagram can be refashioned to direct attention to these harder-to-judge coordinate and superordinate items as well as the easy-to-judge subordinate items of sector A.

Figure 4 is a sketch of a version that might appear as a VIRI on a computer screen. It balances the top 40 items in sector A with roughly equivalent numbers of items in sectors B and C. (The B and C items are not bunched immediately below the top 100; they are scattered and take us far down the list of 1711 items.) Again, the log scales have been truncated and *Moby Dick* repositioned. Figure 4 shows the rightward part of the pennant where cognitive effects are greatest. Some leftward points have been left unlabeled to suggest the many hundreds of items that are excluded from the figure. When the same literary work seems to be in different positions in Figures 3 and 4, it is not because of any movement of the underlying point but because I tweaked its label to correct overlaps.

Figure 4 compactly illustrates much of what has already been said about pennant diagrams. The descending orders of

specificity, all algorithmically produced, are quite plain: broadly, sector A is Melville studies, sector B is American studies, and sector C is world literature, serials, and generically titled works. The two-dimensional layout of the pennant makes the increasingly salient associations in all sectors simultaneously visible, which they would not be in a one-dimensional list. To say more about how pennant diagrams organize their constituent items and to give that discussion a psychological cast, I will return to Peter, Mary, and relevance theory.

Mary and Pseudo-Mary

The dialogue between Peter and Mary can be reconceptualized in terms of Figure 4. Recall that Mary is here cast as an expert in Melville studies whom Peter asks, "What should I read to follow up on *Moby Dick*?" As a literary person, Mary knows many things that qualify her to answer Peter's question. The pennant diagram models part of this knowledge. Peter's question is represented by the point for *Moby Dick* at extreme right. Everything else in the diagram simulates Mary's response, based on her knowledge of titles relevant to Melville studies. Because this Mary looks suspiciously like a recommender system rather than a woman, let us call her pseudo-Mary. Her mind consists of bibliographic data (here, the ISI citation record) and some computer programs. (She is obviously simpler than Maria, the robot in the 1927 film, *Metropolis*, or Helen, the artificial intelligence in Richard

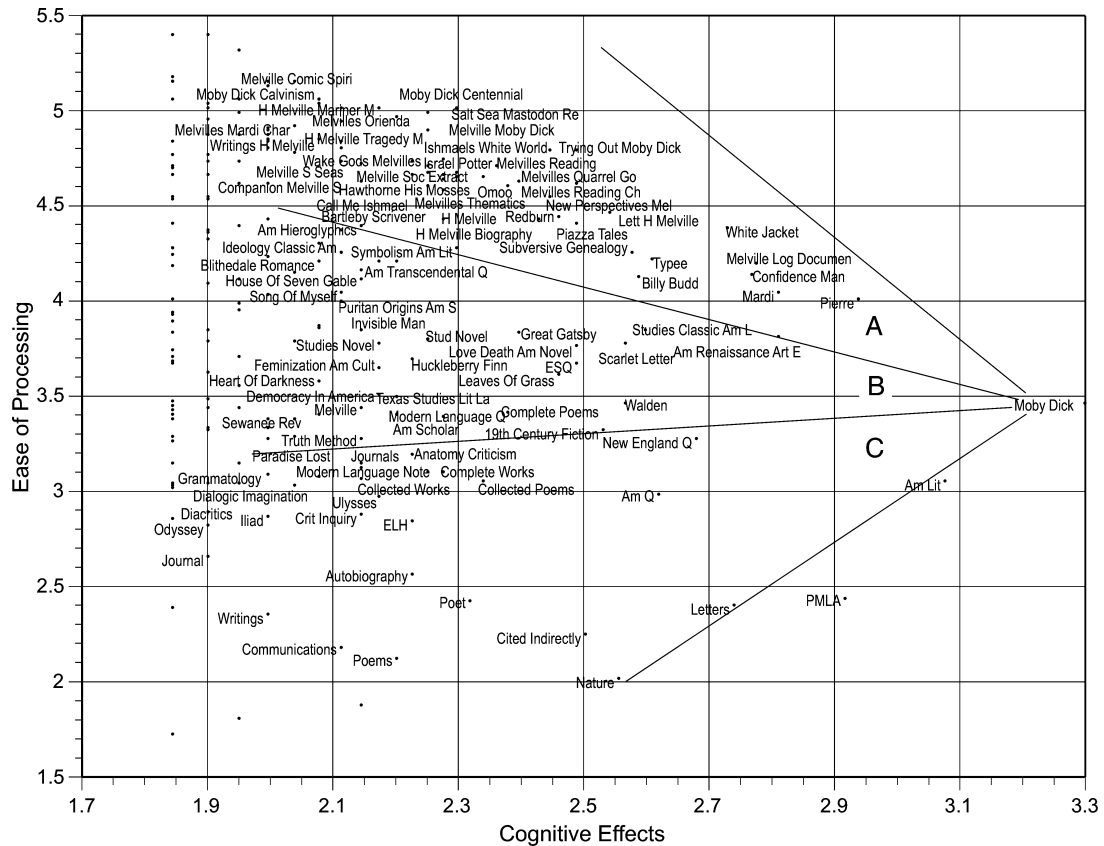


FIG. 4. "Balanced" pennant diagram of selected titles from *Moby Dick* distribution in three sectors.

Powers's 1995 novel, *Galatea 2.2*, or the simulated movie star who gives her name to the 2002 film, *SImpOne*). Prompted by a term that sets a context, pseudo-Mary displays bibliographic items relevant to that context on Sperber and Wilson's two dimensions. That is, she answers by arranging items on the effort scale to indicate how easy it is to perceive their relevance and by moving some items closer to the seed term to predict their greater cognitive effects. Peter simply has to know how to interpret the display. Table 6 compares items in Mary's mind with their counterparts in pseudo-Mary's.

As a brief alteration of perspective, imagine Peter walking from the point for *Moby Dick* in Figure 4 toward the vast library of documents fanned out before him. He moves in "Brookesean" information space, gathering or ignoring items as he goes. Because the display is ordered, he needs only a few guidelines to answer his own question: (a) Nearest items are most relevant. (b) Obvious choices are at right in sector A. (c) Interesting nonobvious choices are straight ahead in sector B. (d) Don't wander off into sector C. Following these guidelines, he would find pseudo-Mary recommending the same things (along with others) that the real Mary did earlier.

Note that he can proceed by what Sperber and Wilson called "gross absolute judgments"; that is, he need not concern himself with the actual logarithmic measures by which items are placed. Merely by looking, he can see that items up ahead are increasingly dense, like library stacks, which signals both diminishing relevance and ever-greater demands on his time. Leftward, in sector C, are those forbidding

serials with their vague titles. If he stays in sector A or B and picks items relatively close to *Moby Dick*, which are two orders of magnitude higher on the cognitive effects scale than the items farthest from him, he will quickly have more than enough to read and can quit with most of the items in the pennant unvisited. That, in fact, seems a likely outcome for all but the ambitious scholar.

Is this merely to reinvent library classification, as hinted above? It is true that library classificationists such as Dewey and Cutter wanted to create the pseudo-Marys of their day; the groupings produced under their schemes communicate like her by means of ordered ostension. However, neither of their schemes makes use of citation-based (i.e., use-historical) measures to *recommend* some items over others. Rather, classifiers simply group items by intuited similarities. Their intuitions do lead to groupings like those seen in sector A; for example, they put different works of fiction and poetry by Melville in the same place, and they put works of criticism and commentary on Melville close by. Nevertheless, they never produce the groupings seen in sectors B or C as part of the Melville stacks.

Even in sector A, groupings based on intuition would not imply, as the pennant diagram does, that one document is more relevant than another (for example, that *Pierre* is more relevant to *Moby Dick* in contemporary scholarship than is, say, *Typee* or *Billy Budd*). Nor would traditional library stacks single out critical studies such as Matthiessen's or Lawrence's or Fiedler's in sector B as being particularly relevant to Melville's novel (or deny that status to other possible

TABLE 6. Comparison of human expertise and some artificial equivalents in a pennant.

What Mary knows	Pseudo-Mary's equivalent
<i>Moby Dick</i> is a gigantic novel by world standards, and it sets a context in which a vast number of earlier and later works are cocited.	<i>The size of the entire pennant diagram, the recognizability of many of the works in it, the wide span of their publication dates</i>
<i>Moby Dick</i> has been mentioned with thousands of other imaginative works and intensively studied with a relatively small number.	<i>Overall shape of pennant diagram; rightward marshaling of the "small number"</i>
Highly important in the context of <i>Moby Dick</i> are Melville's other novels, shorter fiction, and poems. Also important are his personal writings, such as his letters.	<i>Sector A</i>
<i>Moby Dick</i> has spawned a whole library of criticism and commentary. There are also many studies of Melville's whole <i>oeuvre</i> and several biographies of him.	<i>Sector A</i>
In writing <i>Moby Dick</i> , Melville drew on writings about whaling.	<i>Sector A</i>
<i>Moby Dick</i> is a sea novel, like Melville's other novels <i>Typee</i> , <i>Omoo</i> , <i>Mardi</i> , <i>Redburn</i> , and <i>White Jacket</i> and his novellas <i>Benito Cereno</i> and <i>Billy Budd</i> .	<i>Rightmost part of Sector A</i>
Melville wrote <i>Mardi</i> (1849), <i>Moby Dick</i> (1851), and <i>Pierre</i> (1852) as a trilogy. In recent years <i>Mardi</i> and <i>Pierre</i> , both difficult novels in their ways, have been frequently studied in relation to <i>Moby Dick</i> .	<i>Rightmost part of Sector A</i>
The most important other writer in Melville's life is his friend Nathaniel Hawthorne, to whom <i>Moby Dick</i> is dedicated. Melville and Hawthorne are conjoined in many studies of American literature.	<i>Sector A, Sector B</i>
<i>Moby Dick</i> is often discussed in books about American literature, particularly if they focus on classics of the 19th century. These books have global titles and devote only sections to Melville and <i>Moby Dick</i> .	<i>Sector B</i>
Many journals have published scholarly articles that cite <i>Moby Dick</i> .	<i>Sector B, Sector C</i>
In writing <i>Moby Dick</i> , Melville drew on literary classics such as Shakespeare, Milton, and the Bible.	<i>Sector C</i>
The letters, journals, and poems of other writers, such as Emerson, Dickinson and Whitman, figure in Melville studies.	<i>Sector C</i>

contenders, such as Camille Paglia's *Sexual Personae*). They would not show the particular relevance of certain American novels, such as *The Scarlet Letter* and *The Great Gatsby*, in sector B, or certain serials, such as *American Literature*, in sector C. More broadly, they would not break out the mass of documents seen in the pennant diagram from the incomparably larger mass of documents in a big research library or a union catalog. The full retrieval from which the pennant is generated is actually quite small compared to the latter undifferentiated mass, and it is structured so that the items beyond any desired limit can be ignored, just as library stacks can be exited or a conversation broken off.

Returning to psychological metaphors, one may interpret the pennant diagram as simulating what comes to Mary's mind when she hears Peter's question. The simulation through pseudo-Mary differs from library stack arrangements not only in being more complex and nuanced, but also in being instantly able to be rearranged as different questions are asked. That ability, though in a highly circumscribed domain, is what makes pseudo-Mary resemble a human adviser. Unlike the stacks, she can reconsider the relevance of titles to any new seed term. Different pennant diagrams depict the different associations she makes.

Her associations, moreover, are open for inspection. In human dialogues, not everything that occurs to the speakers gets said; much remains veiled. Remarkably, in pennant

diagrams we see not only pseudo-Mary's most salient associations but also the ordered masses of phrases from which they emerged—her different levels of consciousness, as it were. It is not wholly farfetched to liken pennant diagrams to cross-sections of a mind as it creates relevant responses to different verbal stimuli. The variable *tf* and *idf* weights have their analog in the different strengths of the mind's neuronal connections.

A Cocited Author Pennant

Pennant diagrams for cocited authors resemble those for cocited works. Their cognitive effects and their ease of processing again determine the relevance of authors to the seed author. Cognitive effects in this case are predicted by authors' cocitation counts with the seed author (*tf* values), whereas ease of processing depends on authors' overall citation counts in the file (*idf* values): the lower their counts, the narrower their implications for the seed author and the less effort in processing them. However, whereas anyone can read different levels of ease into the titles of the *Moby Dick* example, names of authors are inherently harder to interpret, and their implications are apparent only to a domain insider. Moreover, authors' names can designate both *oeuvres* and persons, which complicates the analysis. For those reasons, I used my own name as the seed in the diagram in Figure 5 so as to draw on as much domain knowledge as possible.

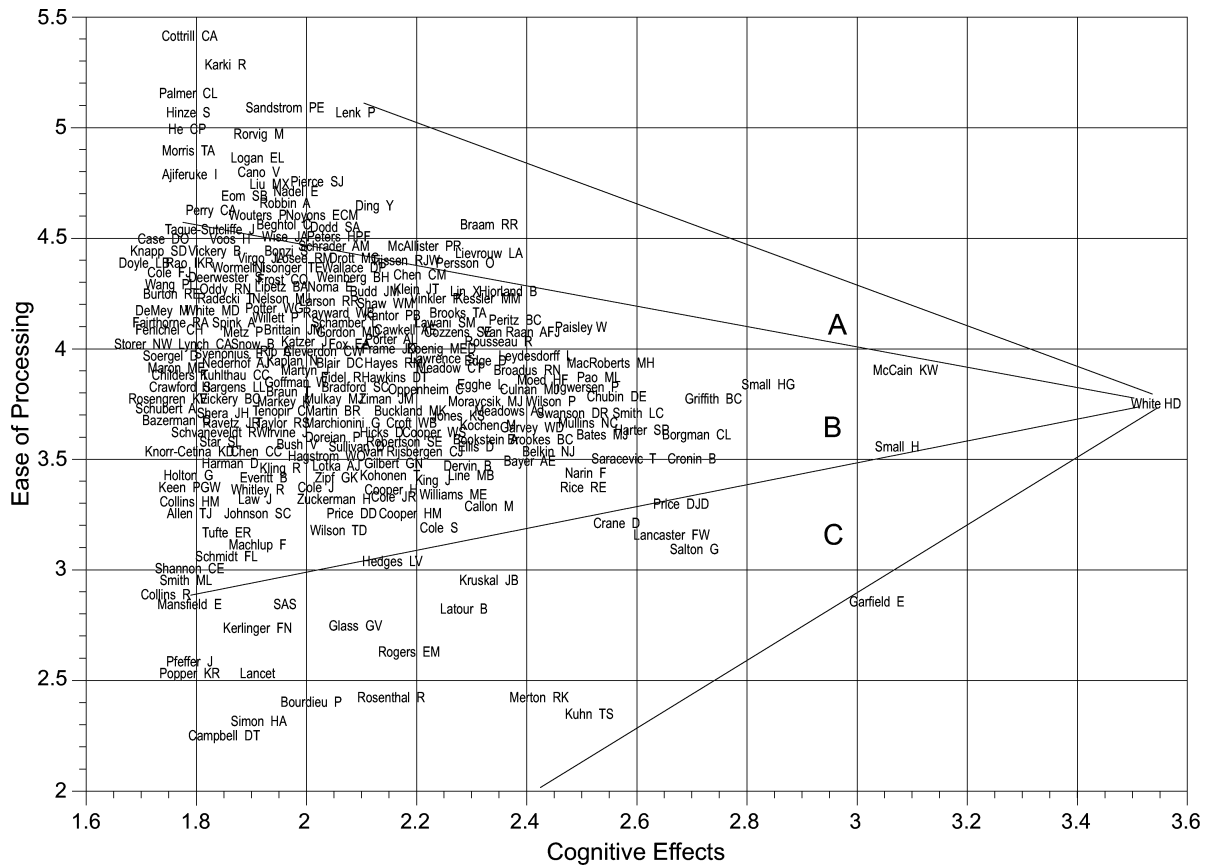


FIG. 5. Pennant diagram of authors cocited with White HD.

On examining the results, I found I could divide my cocitees into interpretable sectors. Anyone with appropriate expertise can check whether my perceptions hold in pennant diagrams for other cocited authors. The evidence presented is still quite sketchy, and readers will need to augment my remarks with domain knowledge of their own. Where the requisite data are lacking, however, my account suggests specific hypotheses about what the sectors contain and possible methods for testing these hypotheses.

In Figure 5, sector boundaries are again placed by hand. They mark off writings whose differences can be demonstrated through content analysis of titles (or fuller texts), as in the *Moby Dick* example above. The sectors in fact differentiate writings in several ways simultaneously, all of which a human adviser might take into account when discussing the authors of a field. Accordingly, this section presents a variety of qualitative meanings that can be extracted from cocited author pennants, especially from idf weightings. Not all of them have counterparts in pennants for cocited works, such as *Moby Dick*. The most interesting is

the emergence of a broad *authority* effect that coexists with statistical specificity and ease of processing. The significance of the authority effect is that it enables an artifact like pseudo-Mary to answer a greater variety of questions, which is useful if we are to simulate even a small part of a human adviser's knowledge.

Table 7 is a bridging summary. It implies that, as seeds, a cocited work and a cocited author will produce pennant diagrams that are informative in roughly the same way. In both cases, the items are progressively less easy to interpret in the context of the seed name as one moves down the sectors. Just as it is easiest to relate cocited works to *Moby Dick* when they are by Melville or from Melville studies, less easy when they are other American classics of fiction, nonfiction, and criticism, and least easy when they are world classics, serials, or generic titles, so, in the cocited author parallel, it is easiest to guess how I am connected to authors who contribute to my own scholarly sub-specialties, less easy to guess my connections with names from the entire discipline of information science, and least easy to guess how my name fits with those of authors from other

TABLE 7. Statistical specificity and ease of processing of terms associated with a cocited work and a cocited author.

Sector	Specificity	Ease of processing	Seed: <i>Moby Dick</i>	Seed: White HD
A	High	High	Melville's oeuvre, Melville studies	"White HD-related" subspecialties
B	Medium	Medium	American literature, American studies	Information science
C	Low	Low	World classics, generic or serial titles	Science studies, research methods

disciplines. I would expect these relations to hold, *mutatis mutandis*, in the pennant diagram for any cocited author.

The authors cocited with a seed author make up his or her “citation image” (White 2001a, 2001b). Most images have far too many names to present in an article such as this, and Figure 5 is no exception. Made with DeltaGraph (Red Rock Software, 2005) from data gathered in Social Scisearch in 2002, it contains only the authors cocited with me at least six times, and even these have been pruned by about a third to reduce the problem of overlapping labels. For example, I cut some names that turned up elsewhere in the diagram in a second form (a few dual allonyms still appear—e.g., Small H, Small HG; Price DD, Price DJD); even so, the tiny typeface is necessary. Labels are centered on their points where possible. Many names have been teased from pileups and so are slightly displaced from their true positions. The Items in File counts for Wilson P and Wilson TD were inflated by homonymic authors outside of information science, and I substituted more plausible counts for them (like my earlier corrections for *Pierre* and *Billy Budd*). There are also other authors named White HD, and I made sure that only cocitations with the right White HD were tallied.

Cognitive Effects

How good is Figure 5 as an algorithmic response to the query on my name? Actually, quite good. The great majority of names in Figure 5 are information scientists, as is appropriate. If Mary happened to know a lot about information science and Peter for some reason asked her to identify Howard D. White, she might give particulars like, “He’s part of that Philadelphia crowd that does citation analysis—Henry Small, Eugene Garfield, the ISI people. I think he and Kate McCain are at Drexel, which is right near ISI. Belver Griffith was at Drexel; he was part of that, too.” That is what pseudo-Mary also is doing, laconically, by moving certain names rightward toward the seed name in Figure 5.

More precisely, pseudo-Mary is saying that the oeuvres of McCain, Small, Garfield, and Griffith are the ones most relevant to mine. As often happens, these frequently cocited authors have also worked in the same places, and that physical proximity is what the human Mary brings out. (Were computerized author-affiliation data available, pseudo-Mary might bring it out as well.) Nevertheless, whatever the accompanying account, maps of my own citation image always put me close to McCain, Small, and Garfield (cf. White 2003). As an expert witness, I would agree that selected works of theirs will indeed have cognitive effects if read with mine, in the sense that readers will see that we are dealing with strongly reinforcing topics. Moreover, read jointly, our works may imply more than they do singly, which is how RT defines relevance. The same is true of other names drawn rightward in Figure 5, such as Griffith BC, Borgman CL, and Cronin B. (Bibliographic and anecdotal evidence could readily be given.)

Someone steeped in bibliometrics might object that my connections with McCain, Small, Garfield, and Griffith are already well known—that pseudo-Mary is saying nothing new. However, because she must perform based on the citation record, she cannot be more creative than her input allows. (Striking new bisociations must be left to users.) And of course the perception of novelty is not constant across persons; what is old hat for one is fresh news for another.

It is, to repeat, the *conjunction* of the seed author and the cocited authors in the pennant that is important. As put by Wilson and Sperber (2002), “The most important type of cognitive effect achieved by processing an input in a context is a CONTEXTUAL IMPLICATION, a conclusion deducible from the input and the context together, but from neither input nor context alone” (p. 251). Thus, considered separately, any author in Figure 5 has a large and heterogeneous set of implications. However, if we take White HD as context and, say, Borgman CL as input, I would quickly conclude that reference is being made in the underlying data to *Scholarly Communication and Bibliometrics*, a 1990 book that she edited and I contributed to, or perhaps to more recent work we have both done in topical areas implied by that title. The ease of all inferences is of course seed-specific. Persons not in my position could check my inferences—or anyone else’s—in the bibliographic records underlying the pennant.

Figure 5 illustrates how separate tf and idf positioning can inform. For example, given White HD as seed, McCain KW and Garfield E have roughly similar cognitive effects (McCain has the second highest and Garfield the fourth highest tf count with me). However, McCain’s works are displayed as being more relevant to “White HD studies” because it is harder to know how Garfield’s much larger oeuvre, with its lower statistical specificity, relates to mine. Garfield published hundreds of weekly columns and many other works as president of ISI, and these jointly have been cited several thousand times; hence his low idf.

Figure 5 also illustrates how pennants complement lists. If the authors in Figure 5 are ranked one-dimensionally by their tf*idf products and the top 100 are plotted in a pennant, only the names in sector A and the upper and rightmost parts of sector B appear. The same thing happened in Figure 3 with *Moby Dick* studies. As noted, tf*idf promotes *easy-to-see* relevance: many of the top 100 share my research specialties. So does Garfield, but the idf filter removes him from the top 100, along with giants from outside information science like Kuhn and Merton. Pennants thus add value to tf*idf by showing what the filter might otherwise hide from view. The latter, relatively hidden connections could be the very ones that most interest domain experts.

Statistical Specificity

The statistical specificity of authors’ names, as opposed to topical phrases, needs clarification. Recall that specificity varies with the idf ranking, which here is based on authors’ overall citation counts. However, the counts obviously do not make authors’ names more or less specific when they

designate *persons*. In Figure 5, for example, the woman, Charlotte A. Cottrill (Cottrill CA) at top left is not more specific than the man Donald T. Campbell (Campbell DT) at bottom left. The specificity dimension applies, rather, when such names stand as proxies for writings in oeuvres. In the latter case, what varies in specificity are other bibliographic details of works in these oeuvres, which the idf ranking draws into a newly meaningful order. Here, the details implied by authors' names are titles of works, and these hidden titles must be uncovered if the pennant is to make full sense.

To complicate matters, the titles are not independently interpretable but must always be compared to titles in the seed author's writings if specificity is to be judged. As in the content analysis carried out earlier on titles cocited with *Moby Dick*, this tests the intercohesion and intercoherence of words and phrases associated with individual oeuvres. Absent such comparisons, one cannot see how authors' names vary in ease of processing because there is no prima facie reason why, say, Cottrill CA should be easier to process than Campbell DT. As a general principle, an oeuvre high on the idf scale should contain titles that share features with titles in the seed author's oeuvre. That makes the relation between them easier to process than one in which the titles have little or nothing in common. Moreover, for any seed author, content analyses of oeuvres should reveal a transition down the sectors, from works that are intercohesive with works by the seed to works that are not.

In Figure 5, titles implied by the seed name do indeed vary in their closeness to titles implied by the cocitees. Authors at top left and I are close at the level of specific works. For example, several articles of mine have titles that contain some version of "author cocitation analysis" or "cocited author mapping." A content analysis would find echoing phrases in titles implied by the names Karki R ("Searching for Bridges between Disciplines: An Author Co-Citation Analysis on the Research into Scholarly Communication"), Cottrill CA ("Co-Citation Analysis of the Scientific Literature of Innovation Research Traditions"), and Sandstrom PE ("Information Foraging among Anthropologists in the Invisible College of Human Behavioral Ecology: An Author Co-Citation Analysis"). The effect extends to subject matter not apparent in titles, such as the dissertation by Perry CA, "Scholarly Communication in Developmental Dyslexia: Network Influences on Change in a Hybrid Problem Area," which contains cocitation maps of authors, or the article by Lenk P, "Mapping of Fields Based on Nominations," which compares cocitation of authors with nomination of authors by experts as bases for maps.

In contrast, the name Campbell DT at lower left stands for the oeuvre of the much-cited methodologist, with whom my connection is distant at best. A content analysis of titles of his well-known works—e.g., "Experimental and Quasi-Experimental Designs for Research" (with J. C. Stanley), "Reforms as Experiments," and "Ethnocentrism of Disciplines and the Fish-Scale Model of Omniscience"—suggests nothing intercohesive or obviously intercoherent with any-

thing of mine, and it would be hard to link the two of us plausibly with subject indexing (unless it were something quite global, like "Interdisciplinarity"; we are linked through his "fish-scale model" article).

Ease of Processing

If specificity is a topical dimension measured through content analyses of works, ease of processing is a psychological variable that is measurable through trials with people—more precisely, domain experts, including seed authors themselves. The experimental stimulus would be cocited author pairs—that is, the seed author systematically paired with other authors drawn from the pennant, such as White HD–McCain KW or White HD–Campbell DT. The desired associational response could be a subject phrase that describes the pair, a third author likely to be cocited with them, or some other statement as to why they are related. Responses could be scored for the *speed* with which they are made, for *accuracy* when checked against bibliographic or content-analytic data, and for *agreement* across multiple judges. The hypothesis would be, of course, that responses will differ significantly by sector or region of the pennant diagram, with judges finding it easiest to process authors who appear in sector A or near the seed author. Such trials can be imagined even if they are not carried out, because they resemble countless word-association studies of the past and lend themselves to classic analysis-of-variance techniques. In other words, relevance theory as adapted here can be tested with techniques familiar to experimental psychologists, thus aligning possible future studies with a particular research tradition.

For the present, I will illustrate ease of processing simply by considering a few sample authors from the two extreme sectors of Figure 5. When I am cocited with names in sector A, the reason will tend to be that their work is being related to mine, and the context my work sets will be relatively focused and specific. In that sense, the connection will be easy for me (or another insider) to explain: Specificity drives ease. I know without lookups, for example, that the authors at upper left, such as Cottrill CA, Karki R, Sandstrom PE, Lenk P, or Ding Y, are cocited with me because they are contributing in some way to author cocitation analysis, the line of research I started in White and Griffith (1981) and pursued in many later articles. Similarly, I know that Robbin A and Dodd SA are cocited with me because the three of us wrote about social science data archives in the 1970s. Rorvig ME and Hinze S share my interest in mapping citation data; Palmer C, my interest in communication across disciplines. Thus, "Author cocitation analysis," "Data archives," "Literature mapping," and "Interdisciplinarity" are topical phrases by which my name and theirs could be jointly indexed. Furthermore, I could sometimes correctly predict the exact works of ours that are being cocited (cf. White, 2002).

In contrast, when I am cocited with names at the left of sector C, the reason will be that my work is being mentioned with theirs in the same article but rarely linked closely to

their ideas, and the context thus set will be relatively broad and vague. In that sense the nature of the connection will be hard to process: Even as an insider, I will find it hard to subsume under a topical phrase. I frankly do not know why I am being cocited with prolific polymaths like Simon HA, Popper K, and Bourdieu P at lower left (or with corporate authors like SAS and *Lancet*). I can draw inferences from what these names connote, but that is all, and the connections are likely to be indirect.

My connection with sector C names such as Cooper HM, Hedges LV, Glass GV, and Rosenthal R is a bit easier to infer (and a content analysis of titles could probably detect it). It comes about because I contributed a chapter on literature retrieval to a handbook on meta-analysis (White, 1994), and the methods and statistical techniques of meta-analysis are their specialties. But I am contributing to their field, not they to mine, and I could not confidently predict what work of theirs is being cited with mine as I sometimes can in the case of authors in sector A.

To avoid possible confusion, let me repeat that names low on the ease of processing scale are not necessarily hard to process by themselves; quite the opposite. It is relating them to the seed term that is difficult. Thus, if Mary is well read not in my field but in Melville and the humanities, as we earlier assumed, she would probably know names like Popper, Kuhn, Bourdieu, and Latour in sector C as part of her general cultural literacy. But she would not know how to relate them to the context set by an obscure White HD in information science when I myself do not know how without lookups and further reading. Conversely, the names in sector A would be unknown to Mary and most other people, but I and a few of my readers not only know them, but can immediately say why those names and mine appear together. The same is true of names drawn rightward on the cognitive effects scale, such as the abovementioned Borgman CL. Thus, bibliometrics and information retrieval can be linked to a specific cognitive variable, domain literacy or expertise, which has definite implications for literature searching (White 2002; Wildemuth, 2004).

Age and Authority

Table 8 predicts that, in any cocited author's pennant, the idf ranking will produce meaningful stratifications other than relative ease of processing. All are predicated on some principled way of demarcating the sectors, whether tf and idf counts or verbal phrases are used as raw data.

One kind of stratification is by age of works in the cocited oeuvres. It takes years for most works to become

highly cited. Thus, works with low citation counts will tend to be newer than works with high counts. In Figure 5, idf ranking assures that the oeuvres in sector A have lower citation counts than mine, and the contrasting oeuvres in sector C will have higher counts. Therefore, works in sector A will presumably have more recent publication dates than works of mine, and works in Sector C less recent dates. As one example, two of the Sector A articles mentioned above—Karki (1996) and Cottrill, Rogers, and Mills (1989)—are newer than White and Griffith (1981), the article with which they are cocited. In contrast, the latter is newer than two sector C works with which it is cocited—Garfield (1979) and Crane (1972). Overall, the average age of authors' oeuvres is predicted to be sector A < sector B < sector C.

There is also a sociological dimension. Taken as persons rather than as oeuvres, the authors in sector A tend to be younger than I am, both by year of birth and in years since the doctorate or first publication. Like Katherine McCain, who was my doctoral student in the 1980s, several wrote their dissertations or first articles using citation analysis (some learned it from McCain and me directly), and this is still work for which they are being cited. In my eyes, sector A is thus primarily the sector of "the newer people." For several authors there, I have been asked to write letters toward promotion or tenure or grants; for others, I have been sent articles to referee: sociologically speaking, they could be called my juniors. Following this logic, sector B contains peers in varying degrees, and sector C contains my seniors. The latter (some of whom are dead) tend to be both older and much better known than I am; most were well established when I was a doctoral student. I might be asked to contribute to a festschrift for one of them (e.g., White, 2000), but never to referee their work, much less to write one a letter of recommendation.

Formally, the idf ordering predicts that the average age of the authors across the sectors will follow the pattern Juniors < Peers < Seniors. It further predicts that their intellectual seniority—their fame or reputation—will follow the same pattern.

The seniority rankings coincide with subject-matter partitions already encountered. Juniors in sector A will tend to be from the seed's own specialty (or specialties). Peers in sector B will tend to be from the seed's discipline but quite possibly from other specialties. Seniors in sector C—particularly those far from the seed on the cognitive effects scale—will tend to be from other disciplines altogether. Three names from Figure 5's leftmost column—Morris TA, Childers T,

TABLE 8. Predictions about sectors when seed is a cocited author.

Sector	Cocitee is seed's	Cocitee's generation is	Works cocited are	Cocitee's oeuvre is	Cocitee's fame is	Cocitee is identified with	Main flows of intercitation
A	Junior	Younger	Newer	Smaller	Less	Seed's subspecialties	Juniors → seed
B	Peer	Roughly same	Mixed ages	Roughly same	Equal	Seed's discipline	Seed ↔ peers
C	Senior	Older	Older	Bigger	Greater	Other disciplines	Seed → seniors

and Collins R—illustrate this interpretation of the sectors. Theodore Morris in sector A took his doctorate at my college a few years ago with a citation-analytic dissertation (I was on his committee). Thomas Childers in sector B is a long-time Drexel colleague whose research interests, such as reference librarianship, partially overlap mine. Randall Collins in sector C is the consummate synthesizer whom I have only briefly met but regard as one of the greatest living sociologists.

Thus interpreted, the seniority dimension provides another reason why oeuvres by juniors are easier to relate to the seed's than oeuvres by peers or, especially, seniors. It is not merely that individual works in them are easier to match on specific topics. Sector A oeuvres also tend to be smaller; their authors have not been writing as long, and there is less to read. For example, Ted Morris's work consists of a dissertation and a few articles; Tom Childers's, of many articles and several short books; Randall Collins's, of many articles and many large books, including, recently, *The Sociology of Philosophies*, a magisterial work that runs well over 1000 pages.

Elsewhere, I have discussed citees as juniors, peers, and seniors but suggested no measure of intellectual seniority other than citers' opinions: "One may hesitate to define cohorts precisely in terms of birthdates, but most scholars and scientists have a keen sense of the authors in their fields who arrived, in the reputational sense, some years before they themselves did—the seniors. Scholars and scientists also know persons in their own cohorts whose reputations were made concurrently with their own—their peers. The final grouping—their juniors—consists of persons who were coming up when their own reputations were already made" (White, 2001b, p. 625). It would appear that the idf scale and the sectors now allow us to measure intellectual seniority objectively. The idf scale is really just a version of authors' citation counts, a time-honored indicator of reputation or prestige (White, 2004, has many examples). Among the other measures of fame that might be tapped as convergent validity checks are how often authors are correctly identified on name-recognition tests, the space devoted to them in popular media, and their visibility on the Web (cf. Posner, 2001).

Last, as shown in Table 8, the idf ranking may predict the directions in which, for a given seed, authority runs. The pennant is based on cocitation, but authority can be measured by *intercitation*—that is, by who cites whom among names in the diagram. The general prediction is that people will tend to cite *across* or *up*, not *down*, in the idf stratification system. That is, the seed will tend to cite peers and seniors. Although peers may reciprocate—the "double-arrow" or mutual relation in Table 8—seniors usually will not (or will less frequently than they are cited). Correspondingly, juniors will tend to cite the seed more than the seed cites them (cf. White, 2000; 2001b). The flow of authority is thus asymmetric—from seniors to the seed, and from the seed to juniors. Predicted intercitation flow in Table 8 is just the reverse of the flow of authority.

Pseudo-Mary's Answers

We saw in Table 6 how pseudo-Mary could simulate, through ordered ostension, some of Mary's knowledge about Melville and *Moby Dick*. It should now be apparent that pseudo-Mary can handle cocited authors as easily as she did Melville's cocited work. In the pennant diagram for a cocited author, some names will be higher than others on the tf or cognitive effects scale, and the idf scale will also stratify them by ease of processing, seniority, and reputation. All of these kinds of knowledge can be used in a system that simulates human question-answering abilities in a limited domain. For example, authors' reputations are one of the things that learned people know and one of the things that lets them speak relevantly in a scholarly or scientific context.

Tables 7 and 8, though not set up like Table 6, suggest the kinds of answers that pseudo-Mary could give if properly programmed. Presumably, such programming could make use of intercitation data as well as the cocitation data that is the sole basis of Figure 5. Continuing with White HD as the specimen name (many thousands of others could be substituted), we could ask and expect reasonable answers for such questions as:

- What authors are most relevant to White HD?
- Who are his approximate peers in information science?
- Who are the junior authors he has influenced?
- Who are the senior authors who have influenced him?
- Who are the authors whose connections to him are less obvious—whose work could be connected to his only by creative reading?

Moreover, because these authors are oeuvres as well as persons, the proper programming would let them be disaggregated into works, whose relationships and time-sequencing could be further examined (cf. the systems described in Garfield, Pudovkin, & Istomin, 2003, and Morris, Yen, Wu, & Asnake, 2003). Through content analysis of terms from titles and abstracts, it might also be possible to indicate the specialties and disciplines to which a seed author has been linked. Such facts can now be laboriously gleaned from Dialog retrievals (if one knows what to look for and how to process it), but one can imagine a system that would present them much more directly, in response to conversational input.

A Pennant for an Article

Passing to my concluding example, Table 9 is another comparison of the content of sectors A, B, and C. It exhibits the different yet parallel types of entities that are found in the sectors of pennants for cocited works (i.e., books and serials), authors, and, now, articles. In the column headings are reminders of the ISI-Dialog commands for obtaining cocitation data for the three kinds of seeds. The seed article analyzed here is Harter (1992), published in this journal as "Psychological Relevance and Information Science."

TABLE 9. Items in sectors of three kinds of pennants made with cocitation data.

Sector	Works (Rank CW)	Authors/Oeuvres (Rank CA)	Articles (Rank CR)
A	Subject-specialized books	Subject-specialized juniors	Subject-specialized articles, theses
B	Coordinate books	Peers	Coordinate articles
C	Serials, generic titles, world classics	Seniors, culture heroes	Books, classic articles

Harter's pennant appears as Figure 6. Like an unkempt bulletin board, it is cluttered with strings of text. Downloaded in ISI's cited-reference (CR) format from Social Scisearch in 2002, these are items cocited with Harter's article in at least two later works. I give only a judgment sample of them, clipped to their first 20 characters; others remain numeric IDs. Nevertheless, small type is again necessary, and the pennant is hard to read (a recurring problem in visualizations of bibliometric data). It nevertheless corroborates several points made earlier. All reinforce the idea of an artificially intelligent system like pseudo-Mary.

Because ISI's cited reference strings combine both cited works and cited authors, many of the effects seen thus far are present in Figure 6. As a preliminary to the discussion, recall that ISI's template for a cited journal article (such as BAYM N, 1979, V94, P909, PMLA) differs from its template for a cited book (such as DOUGLAS A, 1977, FEMINIZATION AM CULT). Thus, genre can be detected in a cited reference;

moreover, ISI specifically uses "thesis" to designate doctoral dissertations.

Author Effects: Intercoherence

Harter's article is a contribution to cognitive information science, and the articles that pseudo-Mary predicts as having the greatest cognitive effects are all from the discipline's cognitive wing. Their authors include Carol Barry, Nicholas Belkin, William S. Cooper, Brenda Derwin, Peter Ingwersen, Carol Kuhlthau, Tefko Saracevic, Linda Schamber, Don R. Swanson, Robert S. Taylor, and Patrick Wilson—a highly intelligible result to domain experts. One can usually infer how works of theirs (and others) in sectors A and B can be connected with the Harter article once the bibliographic details are known. The retrieval is thus quite coherent. (Scientific articles in general should produce more obviously coherent retrievals in sector B than novels like *Moby Dick*.)

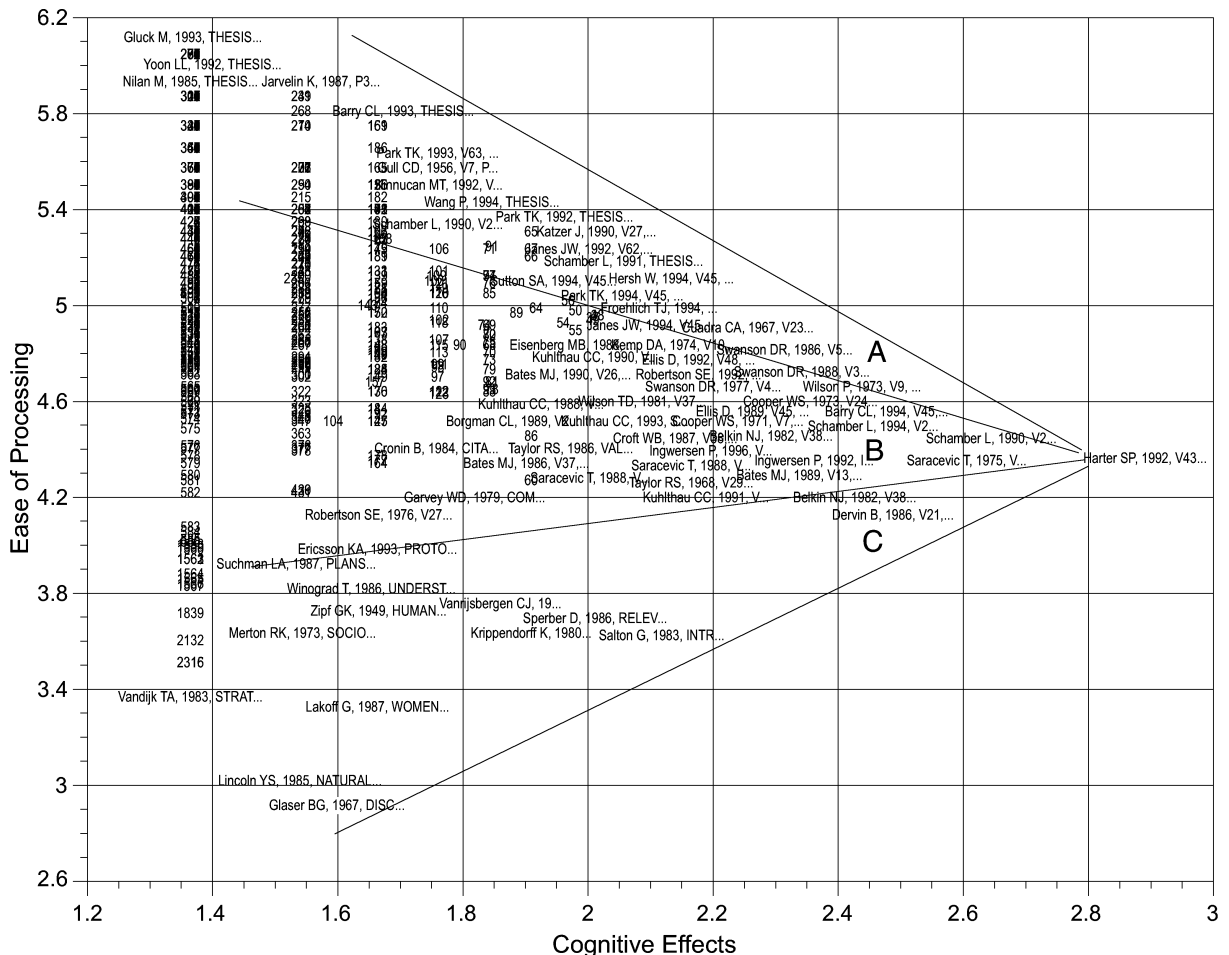


FIG. 6. Pennant diagram of items cocited with Harter's (1992) "Psychological Relevance and Information Science."

Author Effects: Seniority and Reputation

Although individual documents rather than oeuvres are mapped in Figure 6, the genres of the documents are consistent with my earlier account of authors stratified by seniority. The word “thesis”—the archetypal junior product—appears several times in sector A, among articles whose citation counts are relatively low. Sector B contains journal articles with counts in the midrange, and sector C contains heavily cited articles, research reviews, and books. Effects noted previously are thus recapitulated. “Seed-junior” relations, for example, are illustrated by two of the theses in sector A. They are by Taemin Kim Park and Lanju Lee Yoon, who were Harter’s students at Indiana University. Sector A also has theses by, e.g., Carol Barry, Michael Eisenberg, Linda Schamber, and Peiling Wang, which were developed into articles published in information science journals, and these articles turn up as “peers” of the seed document in sector B. Note also that these authors have contributed to a topical specialty of Harter’s in information science. Contrasting “seed-senior” relations are illustrated by works in sector C, which include books by well-known figures in disciplines (and topical areas) outside information science, such as Barney Glaser, George Lakoff, Robert K. Merton, and Teun van Dijk.

Title Effects: Intercohesion

Harter’s pennant and all others are created solely from information implicit in tf and idf counts—that is, use-historical numeric data—and not from computerized matching of terms in titles. The latter is not even possible; as noted above, ISI omits titles altogether from its cited-reference strings. (The seed of the pennant in Figure 6, for example, is HARTER SP, 1992, V43, P602, J AM SOC INFORM SC). Nevertheless, the titles in sector A and the rightward portions of sector B tend to be intercohesive with Harter’s title (“Psychological Relevance and Information Science”) once they and it are known. Consider the readily assimilable titles of several of the doctoral theses in sector A: Barry’s “The Identification of User Criteria of Relevance and Document Characteristics,” Eisenberg’s “Magnitude Estimation and the Nature of Relevance,” and Park’s “The Nature of Relevance in Information Retrieval.” Other theses in sector A are also intercohesive with Harter if one augments their titles with abstracts; for example, Schamber writes that, in her thesis, user criteria for multimedia selection “were considered to be dimensions of relevance, the central concept in information science.”

Similarly, the rightmost articles in sector B—the ones with the greatest effects in the context of Harter’s article—all have titles that repeat or suggest its key terms. The 1990 Schamber, coauthored with Eisenberg and Nilan, is “A Re-examination of Relevance: Toward a Dynamic, Situational Definition,” whereas her 1994 piece is “Relevance and Information Behavior.” The well-known Saracevic is “Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science.” Barry is “User-Defined Relevance Criteria: An Exploratory Study.” Wilson P is “Situational Relevance.”

Title Effects: Diminishing Ease of Processing

In contrast, two sector-C titles high on predicted cognitive effects but lower on predicted ease of processing are Belkin’s “ASK for Information Retrieval” (with Oddy and Brooks) and Dervin’s “Information Needs and Uses” (with Nilan). Both seem clearly more difficult to assimilate to Harter than the titles mentioned above. Although both bear on his topic once they are read, neither would be called up in a typical title or subject search that simply used “relevance” as the input term.

Still harder to process (and with lesser cognitive effects) are the four books farther leftward in sector C. These require more effort than articles to read simply because they are longer. However, they are also more difficult to relate to Harter’s piece because of their more varied subject matter, much of it bearing only indirectly on relevance in Harter’s sense. The four include Krippendorff’s introductory textbook, *Content Analysis*, along with two classic textbooks of experimental information science—van Rijsbergen’s *Information Retrieval* and Salton’s *Introduction to Modern Information Retrieval* (with McGill).

The remaining book is Sperber and Wilson’s *Relevance* itself. Although Harter aimed at transferring relevance theory from their pages to a new discipline and repeated their main word in his title, the label for their book is quite far from his in the pennant. The reason is that their RT differs greatly in both style and substance from the work on relevance in Harter’s world. His synthesis actually took considerable creativity, and to this day information scientists have not done much with the connections he made.

The point here is that *that is what pseudo-Mary is saying*, in her fashion, as she manifests the patterns hidden in cocitation counts. She is saying that information scientists—the principal citers in this case—indeed see S&W’s book as relevant to Harter’s article, but that they have been more comfortable linking the latter to standard pieces from their own shop, like the articles by Schamber or Saracevic. A canny interpreter of the pennant might find significance in its very lack of other writings from S&W’s tradition. Linguistic pragmatics and information science have barely touched.

As a last example, pseudo-Mary is saying that other well-known books appear in sector C, but that they are even harder to relate to the seed article than S&W’s book (though still relevant to it on some grounds). Their titles include Glaser and Strauss’s *The Discovery of Grounded Theory*, Merton’s *The Sociology of Science*, van Dijk and Kintsch’s *Strategies of Discourse Comprehension*, and Lakoff’s *Women, Fire, and Dangerous Things*. In suggesting such facts, pseudo-Mary is being informative in ways that exceed the capacity of lesser recommender systems. Her two-dimensional displays can reveal what one-dimensional tf*idf rankings, as typically looked at, do not.

Concluding Note

The plenitude of pennant diagrams is impressive in a way, and they manage to suggest major variables of information science with remarkable compactness, as Part 2 of this article

will show. However, even on short acquaintance, their weaknesses are apparent. For purposes of dialogue, they are simply *too much*—that is, too far from the conversational ideal of a concise, informative reply to a question. They exhibit echelons of answers that *could* be given rather than the single best answer that most questioners want. Put another way, most persons would want pseudo-Mary to be more like Mary—someone who responds appropriately to specific questions, but does not exhibit everything she knows. That, of course, is a matter of design; pseudo-Mary could be re-fashioned so that what she knows is implicit unless asked for, and her answers are as brief as possible, not giant textual flags.

Nevertheless, pennant diagrams are wieldy enough for my present purposes, and I shall carry them over into Part 2. I have already likened them to “cross-sections of a mind”—an artificial intelligence in this case—and that metaphor will remain useful as I draw out some of their implications for information science. The result evokes a forecast I made in White and McCain (1989). After noting that bibliometrics and information retrieval, long tenuously connected, seemed to be drawing closer, I predicted another convergence:

Bibliometrics models literatures, yes; but its distinctive displays can also be thought of as modeling the structure of human interests. When viewed in this psychological light, its implications go beyond information retrieval to bear on learning, knowing, and creating. We may yet see it as part of a cognitive science that is only beginning to emerge (p. 164).

Some notes on that emergent science follow in Part 2.

References

- Bean, C.A., & Green, R. (2001). Relevance relationships. In C.A. Bean & R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 115–132). Dordrecht: Kluwer.
- Beaugrande, R. de, & Dressler, W. (1981). *Introduction to text linguistics*. London: Longman.
- Belew, R.K. (2000). *Finding out about: A cognitive perspective on search engine technology and the WWW*. Cambridge, UK: Cambridge University Press.
- Blair, D.C. (1992). Information retrieval and the philosophy of language. *Computer Journal*, 35, 200–207.
- Blakemore, D. (1992). *Understanding utterances: An introduction to pragmatics*. Oxford, UK: Blackwell.
- Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54, 913–925.
- Bradford, S.C. (1950). *Documentation*. Washington, DC: Public Affairs Press.
- Brookes, B.C. (1973). Numerical methods of bibliographic analysis. *Library Trends*, 22, 18–43.
- Brookes, B.C. (1980a). Measurement in information science: Objective and subjective metrical space. *Journal of the American Society for Information Science*, 31, 248–255.
- Brookes, B.C. (1980b). Information space. *Canadian Journal of Information Science*, 5, 199–211.
- Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge, UK: Cambridge University Press.
- Buckland, M.K., & Hindle, A. (1969). Library Zipf. *Journal of Documentation*, 25, 52–57.
- Budd, J.M. (2004). Relevance: Language, semantics, philosophy. *Library Trends*, 52, 447–462.
- Case, D.O. (2005). Principle of least effort. In K.E. Fisher, S. Erdelez, & L.E.F. McKechnie (Eds.), *Theories of information behavior* (pp. 289–292). Medford, NJ: Information Today.
- Chen, Z., & Xu, Y. (2005). User-oriented relevance judgment: A conceptual model. *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS'05)*. IEEE Computer Society. Retrieved April 7, 2006, from <http://csdl2.computer.org/comp/proceedings/hicss/2005/2268/04/22680101b.pdf>
- Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, 31, 191–213.
- Cottrill, C., Rogers, E.M., & Mills, T. (1989). Co-citation analysis of the scientific literature of innovation research traditions. *Knowledge: Creation, Diffusion, Utilization*, 11, 181–208.
- Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.
- Ellis, D. (1998). Paradigms and research traditions in information retrieval research. *Information Services & Use*, 18, 225–241.
- Furner, J. (2002). On recommending. *Journal of the American Society for Information Science and Technology*, 53, 747–763.
- Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. New York: Wiley.
- Garfield, E., Pudovkin, A.I., & Istomin, V.S. (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology*, 54, 400–412.
- Goatley, A. (1997). *The language of metaphors*. London: Routledge.
- Green, R. (1995). Topical relevance relationships. I. Why topic matching fails. *Journal of the American Society for Information Science*, 46, 646–653.
- Greisdorf, H. (2000). Relevance: An interdisciplinary and information science perspective. *Informing Science*, 3(2), 67–71. Retrieved April 9, 2006, from <http://inform.nu/Articles/Vol3/indexv3n2.htm>
- Grossman, D.A., & Frieder, O. (1998). *Information retrieval: Algorithms and heuristics*. Boston, MA: Kluwer.
- Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Hardy, A.P. (1982). The selection of channels when seeking information: Cost/benefit vs. least effort. *Information Processing & Management*, 18, 289–293.
- Harter, S.A. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43, 602–615.
- Hjørland, B. (2000). Relevance research: The missing perspective(s): “Non-relevance” and “epistemological relevance.” *Journal of the American Society for Information Science*, 51, 209–211.
- Jurafsky, D., & Martin, J.H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.
- Karki, R. (1996). Searching for bridges between disciplines: An author co-citation analysis on the research into scholarly communication. *Journal of Information Science*, 22, 323–334.
- Koestler, A. (1964). *The act of creation: A study of the conscious and unconscious in science and art*. New York: Macmillan.
- Mann, T. (1993). *Library research models: A guide to classification, cataloging, and computers*. New York: Oxford University Press.
- Manning, C.D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48, 810–832.
- Morris, S.A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, 54, 413–422.
- Nelson, M.J., & Tague, J.M. (1985). Split size-rank models for the distribution of index terms. *Journal of the American Society for Information Science*, 36, 283–296.
- Poole, H.L. (1985). *Theories of the middle range*. Norwood, NJ: Ablex.
- Posner, R.A. (2001). *Public intellectuals: A study in decline*. Cambridge, MA: Harvard University Press.

- Red Rock Software. (2005). Deltagraph (Version 5.6) [Computer software]. Salt Lake City, UT.
- Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60, 503–520.
- Ruthven, I., & van Rijsbergen, C.J. (1996). Context generation in information retrieval. In J.H. Stewman (Ed.), *Proceedings of the Ninth Florida Artificial Intelligence Research Symposium (FLAIRS)* (pp. 380–384). Key West, FL: Florida AI Research Society.
- Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26, 321–343.
- Saracevic, T. (1996). Relevance reconsidered. In P. Ingwersen & N.O. Pors (Eds.), *Integration in perspective: Proceedings of the Second International Conference on Conceptions in Library and Information Science (CoLIS 2)* (pp. 201–218). Copenhagen, Denmark: Royal School of Librarianship.
- Saracevic, T. (1997). Users lost: Reflections on the past, future, and limits of information science. *SIGIR Forum*, 31(2), 16–27.
- Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 3–48.
- Small, H.G. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8, 327–340.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, 28, 11–21.
- Sparck Jones, K. (2004). IDF term weighting and IR research lessons. *Journal of Documentation*, 60, 521–523.
- Sparck Jones, K., & Willett, P. (1997). Techniques. In K. Sparck Jones & P. Willett (Eds.), *Readings in information retrieval* (pp. 305–312). San Francisco, CA: Morgan Kaufmann.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.) Oxford: Blackwell.
- Sperber, D., & Wilson, D. (1996). Fodor's frame problem and relevance theory (reply to Chiappe & Kukla). *Behavioral and Brain Sciences*, 19, 530–532.
- Swanson, D.R. (1977). Information retrieval as a trial-and-error process. *Library Quarterly*, 47, 128–148.
- Walton, D.N. (1989). *Informal logic: A handbook for critical argumentation*. Cambridge, UK: Cambridge University Press.
- White, H.D. (1994). Scientific communication and literature retrieval. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 41–56). New York: Russell Sage.
- White, H.D. (2000). Toward ego-centered citation analysis. In B. Cronin & H.B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 475–496). Medford, NJ: Information Today.
- White, H.D. (2001a). Authors as citers over time. *Journal of the American Society for Information Science and Technology*, 52, 87–108.
- White, H.D. (2001b). Author-centered bibliometrics through CAMEOs: Characterizations automatically made and edited online. *Scientometrics*, 51, 607–637.
- White, H.D. (2002). Cross-textual cohesion and coherence. The CHI 2002 Discourse Architectures Workshop. Retrieved April 9, 2006, from http://pliant.org/personal/Tom_Erickson/DA_White.pdf
- White, H.D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54, 423–434.
- White, H.D. (2004). Reward, persuasion, and the Sokal hoax: A study in citation identities. *Scientometrics*, 60, 93–120.
- White, H.D. (2007). Combining bibliometrics, information retrieval, and relevance theory, Part 2: Some implications for information science. *Journal of the American Society for Information Science and Technology*, 58(4), 583–605.
- White, H.D., & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163–172.
- White, H.D., Lin, X., Buzydlowski, J., & Chen, C. (2004). User-controlled mapping of significant literatures. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1), 5297–5302.
- White, H.D., & McCain, K.W. (1989). Bibliometrics. *Annual Review of Information Science and Technology*, 24, 119–186.
- White, H.D., & McCain, K.W. (1997). Visualization of literatures. *Annual Review of Information Science and Technology*, 32, 99–168.
- Wildemuth, B.M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science & Technology*, 55, 246–258.
- Williams, S.T. (1963). Melville. In F. Stovall (Ed.), *Eight American authors: A review of research and criticism* (pp. 207–270). New York: Norton.
- Wilson, D. (1994). Relevance and understanding. In G. Brown, K. Malmkjaer, A. Pollitt, & J. Williams (Eds.), *Language and understanding* (pp. 35–58). Oxford: Oxford University Press.
- Wilson, D., & Sperber, D. (1986). An outline of relevance theory. In H.O. Alves (Ed.), *Encontro de linguistas actas* (pp. 19–42). Minho, Portugal: University of Minho.
- Wilson, D., & Sperber, D. (2002). Relevance theory. *UCL Working Papers in Linguistics*, 14, 249–290. Retrieved April 9, 2006, from http://www.phon.ucl.ac.uk/publications/WPL/02papers/wilson_sperber.pdf
- Wilson, P. (1968). *Two kinds of power: An essay on bibliographical control*. Berkeley: University of California Press.
- Yang, K. (2005). Information retrieval on the Web. *Annual Review of Information Science and Technology*, 39, 33–80. Retrieved April 8, 2006, from http://elvis.slis.indiana.edu/kiyang/pubs/webir_arist.pdf
- Yus, F. (2006). Relevance theory online bibliographic service. Retrieved April 9, 2006, from <http://www.ua.es/personal/francisco.yus/rt.html>
- Yus Ramos, F. (1998). A decade of relevance theory. *Journal of Pragmatics*, 30, 305–345.