Some New Tests of Relevance Theory in Information Science

Howard D. White

whitehd@drexel.edu
College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 (USA)

Abstract
A central idea in D. Sperber & D. Wilson's relevance theory is that an individual's sense of the relevance of an input in a context varies directly with its cognitive effects and inversely with its ease of processing in that context. H. D. White has argued that this idea has an objective analogue in information science—the tf*idf (term frequency, inverse document frequency) formula used to weight indexing terms in document retrieval. Here, tf*idf is used to weight terms from five bibliometric distributions in the context of the seed terms that generated them. The distributions include the descriptors co-assigned with a descriptor, the descriptors and identifiers assigned to an author, two examples of cited authors and their co-citees, and the books and journals cited with a famous book, *The Structure of Scientific Revolutions*. In each case, the highest-ranked terms are contrasted with lowest-ranked terms. Clear qualitative differences between the sets of terms are intuitively well-explained by relevance theory.

Introduction

This paper provides new data toward assessing a basic claim from White (2007a, b)—namely that a formula used in information science for weighting search terms in relevance rankings

*Weight = term frequency * inverse document frequency*

instantiates a central idea of Sperber & Wilson's relevance theory from linguistic pragmatics

Relevance = cognitive effects / processing effort.

In other words, cognitive effects and processing effort, which S&W discuss almost exclusively as subjective experiences in individuals, have an objective analogue in the tf*idf formula at the heart of classic information retrieval. While tf*idf is far narrower in scope, it produces actual numbers that are conducive to broad rankings, and such rankings accord with S&W's claim that subjective relevance is at best ordinally scaled.

What tf*idf does is to assign higher weights to terms that are *specific* in a given context and lower weights to terms that are *vaguer* and *more general* in that context. Here, contexts are set by seed terms, such as descriptors or authors' names. These seeds are used to generate bibliometric distributions of the terms associated with them in commercial databases. Using Thomson Dialog software, one can readily obtain the counts needed to weight terms by the tf*idf formula. This permits relevance rankings of terms in the context of the seed. We know that the terms have already produced cognitive effects in that context; indexers, authors, and editors have attested as much by bringing them and the seed together in bibliographic records over time. Thus, that part of S&W's definition of relevance has been fulfilled. But what about processing effort? It is the conjecture of the present paper that, if tf*idf functions as White has argued, *terms with higher weights will be easier to relate to the seed than terms with lower weights.*

Selected terms from five bibliometric distributions have therefore been ranked by their tf*idf weights in the Results section, so that readers can intuitively test this conjecture. If they agree, it strengthens the assumption that Sperber & Wilson's relevance theory, information retrieval, and bibliometrics are not just coincidentally related.

It would of course be good to investigate the matter experimentally by running trials with

Proceedings of ISSI 2009, pages 275–285.
Edited by B. Larsen and J. Leta.

subjects and conducting formal hypothesis tests. The data do lend themselves to such word-association techniques as card-sorting (to name only that). The more modest effort on view here simply promotes the general idea of using bibliometric data in psychological research.

Background

The two long articles by White (2007a, b) were intended to show, through a complex series of examples, how Sperber & Wilson's (1995) relevance theory (RT) could be adapted from the field of linguistic pragmatics to explain and unify a wide variety of findings in information science (IS). The widely influential ideas of S&W are quite compatible with those of the cognitive wing of IS, although they emerged wholly outside it. S&W define relevance as a property of sensory inputs to cognitive contexts in individuals. Contexts in S&W's sense are sets of assumptions—people's internal representations of the world—and the inputs, needless to say, include spoken and written communications. The two determiners of relevance have most recently been set forth in Deirdre Wilson's lecture notes on RT at University College London, where she is a professor (Wilson, 2007; boldface hers):

> **Relevance to an individual**
>
> Other things being equal, the greater the cognitive effects (of an input to an individual who processes it), the greater the relevance (to that individual at that time).
>
> Other things being equal, the smaller the processing effort required to derive these effects, the greater the relevance (of the input to that individual at that time).

Cognitive effects are changes in the assumptions a person holds at a given time. Inputs cause changes by (1) strengthening an existing assumption, (2) weakening or overturning it, or (3) combining with it to produce a new assumption through inference. (Inferential effects in conversations are the main subject of S&W's book.) *Processing effort* depends on variables such as how recently an item of communication has been used, how frequently it has been used, its linguistic complexity, and its logical complexity (see Wilson's lecture notes for examples). We use processing effort to *stop* choosing among multiple possible interpretations of an input; as soon as one possible interpretation attains a satisfactory degree of relevance without undue effort, we accept it and move on.

Goatly (1997: 139) notes that the key RT formulation can be expressed as a ratio in which the factors operate simultaneously:

$$\text{Relevance} = \text{Cognitive Effects} / \text{Processing Effort}$$

We realize, of course, that this ratio is not represented by hard numbers in our heads; both S&W and many studies in IS agree that the relevance of inputs can be rank-ordered but not measured more precisely. Even so, the ratio makes intuitive sense. If something we hear or read has no cognitive effects, it is not relevant to us. Nor will it be relevant, whatever its potential effects, if the effort of processing it is too great. On the other hand, an input that has major effects and is readily processed will be experienced as highly relevant. According to S&W, a propensity to screen inputs by testing their degree of relevance is universal in the human species. What varies are the contexts in which the inputs occur. So the ratio above should always be seen as operating in a context.

White claimed that the ratio above can be operationalized for analysis by being applied to the standard bibliometric distributions—for example, to lists of terms produced by the Rank command in bibliographic databases on Dialog. He argued that, given those lists, the well-known term-weighting formula from information retrieval, *term frequency* times *inverse document frequency* (tf*idf), could be computed for each term and interpreted as its RT ratio, giving bibliometrics a novel psychological tinge. He saw the results as providing linguistic evidence that RT and information retrieval (as represented in IS by researchers such as Gerard

Salton and Karen Sparck Jones) are mutually reinforcing. So are RT and bibliometrics, whose term distributions Saracevic (1975) called "relevance-related"—an insight later expanded in Harter (1992), the article that first brought RT to the attention of information science. In both RT and IS, however, such ideas are unconventional and require explanation.

The tf measure corresponds to a term's predicted cognitive effects in the context of the seed; the idf measure, to its predicted processing effort. Since high idf values somewhat confusingly correspond to low processing effort, White (2007a, b) relabeled the latter variable "ease of processing"; that is, high=easy and low=hard. Multiplying tf by the inverse factor idf is analogous to dividing values for cognitive effects by values for processing effort. One can then examine terms ranked by their tf*idf scores and judge how well the relevance-theoretic predictions are borne out.

As a way of demonstrating these points, White (2007a, b) offered a new kind of graphic, pennant diagrams. These show that relevance-ranked terms of all sorts can be meaningfully displayed on axes that predict their cognitive effects and ease of processing in the context of a seed term. (For a brief, lucid introduction to pennants, see Schneider, Larsen & Ingwersen, 2007.) Toward the same end, the present paper contrasts terms from bibliometric distributions, but only in simple tables rather than pennants.

Bibliometric distributions are generated by listing all the terms co-occurring with a seed term and ranking them by their frequency of co-occurrence. For example, if the seed term is a subject heading, one can list all the journal names co-occurring with it when it has been used to index at least one of their articles. The journals can then be ranked high to low by the number of such articles they contain (which creates a Bradford distribution). Similarly, if the seed term is a cited author's name, one can list all the authors co-cited with that author and rank them by frequency of co-citation. It follows that the co-occurring terms are in varying degrees relevant to the seed, as evidenced by their term frequencies (tf), and that the terms that co-occur most frequently with the seed are most relevant to it. A person entering a seed term into a bibliographic system sets a context, so to speak, in which the relevance of terms can be judged. The system responds with a list of terms ranked by predicted relevance to the seed, as if offering its own "assumptions" in that context, S&W-style.

However, the list of terms has not yet been adjusted for processing effort. Information scientists have long known that terms that occur relatively rarely in a collection of documents tend to yield more discriminating retrievals than terms that occur more widely. This concept was introduced in Sparck Jones (1972) as "statistical specificity," because the terms that frequency counts identify as relatively rare also tend to be more specific than commoner terms. When presenting documents for judgment, system designers want the documents retrieved through relatively specific terms to be ranked on top, where their relevance is easier to see. In other words, they require less effort to process—less mulling over—and thus are easier to accept. The inverse document frequency measure in tf*idf is designed to push such documents up and their opposites down. It can be understood as a counterpart to processing effort in the RT ratio.

Method

As noted above, versions of tf*idf are used in information retrieval to weight a user's search terms for use in the relevance ranking of documents. The tf function begins as a count of how frequently a search term appears in a document (the more times, the higher that term's weight). The df function begins as a count of the frequency of a term in the overall collection of documents. When made inverse by being divided into the number of documents in the collection, it gives high weights to relatively rare terms, and low weights to relatively common terms. When ranked by tf*idf, documents with many rare terms go up, while documents with a few widely occurring terms go down.

Both tf and idf functions may be damped by being converted to logs (Jurafsky & Martin, 2000). The tf*idf weighting formula used in White (2007a, b) is from Manning & Schütze (1999). For the *i*th term in document *j*:

$$\text{weight}(_{i,j}) = (1 + \log(\text{tf}_{i,j}))*\log(N/\text{df}_i)$$

where all term counts ≥ 1, logarithms are base 10, and *N* is the total number of documents in the collection. The same formula is used here.

Table 1 outcomes will be discussed substantively in Results, but serve now to introduce the ways in which all the data of this study were obtained and processed. The table contains raw data and derived weights (rounded) from an online search done in Social Scisearch on Dialog in early 2008. The seed in this case was the information scientist Blaise Cronin as a cited author (CA).

Table 1. Sample bibliometric data and computations for weighting them

Name	Count with seed	Count overall	Log tf	Log idf	Weight
Cronin B (seed)	1085	1085	4.04	3.44	13.89
Björneborn L	49	78	2.69	4.59	12.33
Almind TC	51	98	2.71	4.49	12.15
Thelwall M	79	206	2.9	4.16	12.06
Bar-Ilan J	67	174	2.83	4.24	11.97
Davenport E	72	215	2.86	4.14	11.84
Rousseau R	110	465	3.04	3.81	11.59
MacRoberts MH	89	356	2.95	3.93	11.58
Brooks TA	59	204	2.77	4.17	11.55
Cano V	26	60	2.41	4.7	11.35
Baird LM	24	52	2.38	4.76	11.33
McCain KW	80	376	2.9	3.9	11.33
Vinkler P	51	199	2.71	4.18	11.31
Goodrum AA	23	50	2.36	4.78	11.28
Peritz BC	49	208	2.69	4.16	11.19
White HD	114	665	3.06	3.65	11.17
	Raw tf	Raw df	1 +Log(tf)	Log(3 mil/df)	tf*idf

When entered, "CA=Cronin B" forms the set of articles (including his own) that cite him, which then numbered 1,085. Operating on this set, the command "Rank CA cont" produces a list of the authors co-cited with him, ranked in *cont*inuous descending order of frequency. The co-citation counts, now out of order, are given in the second column, labeled "Count with seed" at top and "Raw tf" at bottom. The entire retrieved set of bibliographic records is treated as a single document in which co-occurrences of author names are counted as raw term frequencies. (Standard information retrieval would obtain tf values for multiple documents separately.)

When the command is entered as "Rank CA cont detail," each author's total citation count in the database is added to the listing. Those counts are in the third column, labeled "Count overall" and "Raw df." (Raw df are *occurrence* counts, and raw tf are *co-occurrence* counts.) The fourth and fifth columns show the raw tf and df values converted to logged weights, according to the Manning & Schütze formulas. (For computing idf, the number of bibliographic records in the Social Scisearch collection was arbitrarily set at three million.)

The sixth column gives, in descending rank order, the tf*idf weights as computed for Cronin and his co-citees.

It is instructive to compare the individual log tf and log idf scores of Laura M. Baird and Katherine W. McCain, who both have the same tf*idf weight (11.33) in Table 1. McCain's work when read in conjunction with Cronin's has a greater predicted cognitive impact than Baird's (2.9 > 2.38), but Baird's is easier to process (4.76 > 3.9). That is because "Baird" here stands for one paper, "Do citations matter?" (Baird & Oppenheim 1993), whereas "McCain" stands for multiple papers that would require greater effort to read and assess. The tf*idf formula rewards specificity of implication over less specific breadth.

Dialog and its Rank command can be used to gather bibliometric data such as these for any type of seed: descriptors co-occurring with a descriptor, descriptors or identifiers co-occurring with (i.e., assigned to) an author, journal names co-occurring with a descriptor, works co-cited with a work, and so on. Examples will be seen below. The searches underlying the tables were done at various times since 2000; because the tables are meant to make psychological points that are not time-bound, high currency does not matter.

Results

Sperber & Wilson's RT would predict that, in judging the relevance of communications, relative ease of processing will affect what is judged most relevant in a given context. In IS, the same prediction is implicit in the tf*idf formula in the context of a given seed term. Table 2 uses the ERIC descriptor "Information Needs" as a seed and displays 30 terms (out of many hundreds co-assigned with it), ranked by their tf*idf scores.

Table 2. ERIC descriptors co-occurring with the descriptor Information Needs

tf*idf	Top 15 terms	tf*idf	Bottom 15 terms
8.05	User Needs (Information)	1.28	Attitudes
7.71	Information Management	1.27	Employment
7.59	Information Seeking	1.27	Age
7.35	Access To Information	1.22	Administration
7.26	Information Transfer	1.17	Needs
7.23	Users (Information)	1.17	Design
7.22	Relevance (Information Retrieval)	1.13	Teachers
7.19	Information Literacy	1.06	Data
7.1	Information Utilization	1.04	Role
7.06	Information Scientists	1.04	Behavior
6.88	Community Information Services	0.98	Groups
6.85	User Satisfaction (Information)	0.96	Research
6.66	Management Information Systems	0.92	Community
6.63	Information Policy	0.76	Relationship
6.63	Online Searching	0.74	Schools

Note that, while gradations in tf*idf scores *within* the columns have relatively little qualitative import, the qualitative differences *between* the terms in the two columns are very pronounced. This is the result of selecting terms from the opposite ends of the tf*idf distribution for contrast, but it is also the aim of algorithmic retrievalists. If one imagines a card-sorting task in which subjects were asked to put in two piles (without further ranking) these 30 descriptors according to their ease of association with the concept of "Information Needs," it appears quite probable that the consensus would resemble the outcome in Table 2. The left column echoes "information" in the phrasing of 14 out of 15 descriptors. It contains several close

synonyms of the seed term, and the rest all seem more plausible choices than the terms in the right column. This is not to say that the terms at right cannot be combined with the seed to suggest possible research scenarios—e.g., "Information Needs of Teachers in Schools"—but none of them cohere with it conceptually like the terms at left. The terms at right are not irrelevant, merely less relevant.

It remains to be stressed that the terms at left are *not* necessarily the ones most closely associated with "Information Needs" when Dialog simply ranks terms by their co-occurrence frequencies (the tf listing). Then, the top 10 include "Higher Education," "Foreign Countries," "Elementary Secondary Education," and "Library Services." But these are demoted to much lower ranks by the idf part of the formula, which blindly seeks associations that are easier to process. The way it operationalizes "easier"—a kind of limiting and focusing of sense with respect to the seed—may be seen in Table 2. Lower processing effort means higher relevance.

Table 3. INSPEC descriptors and identifiers co-occurring with Katy Börner as an author or co-author

tf*idf	Top 15 terms	tf*idf	Bottom 15 terms
6.77	Data Visualisation	3.11	Internet
6.68	Audio User Interfaces	3.09	Feature Extraction
6.13	Citation Analysis	3.07	Query Processing
6.05	Digital Libraries	3.06	Art
6.01	Realistic Images	3.02	Image Coding
5.89	Program Visualisation	2.8	Distributed Processing
5.77	Augmented Reality	2.79	Inference Mechanisms
5.7	Computer Animation	2.73	Knowledge Based Systems
5.6	Online Front-Ends	2.65	Data Compression
5.47	Haptic Interfaces	2.65	Object-Oriented Programming
5.39	Virtual Reality Languages	2.59	Digital Simulation
5.35	Rendering (Computer Graphics)	2.57	Graphs
5.24	Graphical User Interfaces	2.55	Computational Complexity
5.13	Solid Modelling	2.27	Diagrams
5.07	Architectural CAD	2.2	CAD

Table 3 continues this line of analysis by presenting the descriptors and identifiers assigned in INSPEC to publications by the computer scientist Katy Börner. She is known for her work in many areas of information visualization, including the visualization of bibliometric data. In her case, the terms in the two columns may seem to contrast less than those in Table 2. This may be because the terms at right are not the bottommost in her distribution, which was truncated to exclude terms assigned less than three times. Nevertheless, the left-column terms provide a more individualized portrait of Börner's interests than do those at right. Note the emphasis on aspects of visualization, computer graphics, and interface design; note also the prominence of "Citation Analysis" and "Digital Libraries," which are not typical interests in mainstream computer science. Most of the terms at right, by contrast, are more generic and characterize the interests of thousands of researchers. "Art" is distinctive to Börner, but a person must know a lot about her career to say why that descriptor is used with her.

If such a person were given a descriptor from each column—e.g., "Data Visualization" and "Object-Oriented Programming"—and asked to choose which of them better represents what Börner is all about, it seems highly likely that the former would be chosen. This is not because "Data Visualization" is more specific than "Object-Oriented Programming" as a general proposition. It is because it is *more specific to Börner*; it is easier to see how it fits her.

Again, this is the kind of discrimination that tf*idf is supposed to make (although the computer scientists who use tf*idf in their programs never discuss successes such as these and are probably unaware of them). From the standpoint of RT, one would presumably have to *think less* to say that "Data Visualization" is relevant to Börner. This predicted faster response could be empirically tested for consistency across judges. (There is a new field in linguistics called experimental pragmatics in which certain predictions from RT are being empirically tested. Sperber is active in it.)

Table 4. Authors co-cited with Blaise Cronin in Social Scisearch

tf*idf	Top 15 terms	tf*idf	Bottom 15 terms
12.33	Björneborn L	7.3	Latour B
12.15	Almind TC	7.24	Ruggie JG
12.06	Thelwall M	6.98	Waltz KN
11.97	Bar-Ilan J	6.86	Abbott A
11.84	Davenport E	6.84	Shannon CE
11.59	Rousseau R	6.75	Long JS
11.58	MacRoberts MH	6.63	Granovetter MS
11.55	Brooks TA	6.48	Rogers EM
11.35	Cano V	6.36	Kuhn TS
11.33	Baird LM	6.23	Drucker PF
11.33	McCain KW	6.15	Porter ME
11.31	Vinkler P	5.85	Wittgenstein L
11.28	Goodrum AA	5.74	Bell D
11.19	Peritz BC	5.55	Bourdieu P
11.17	White HD	5.52	Giddens A

Tables 4 and 5 are based on another kind of data—counts of the journal articles in which authors are co-cited with a seed author. The seeds in these tables are Blaise Cronin and Concepción S. Wilson. Like Katy Börner, they are established figures in IS and have been appropriated here because many readers will be able to interpret data on them. Knowledge of authors and writings is highly specialized within domains and exemplifies the subjectivity of certain kinds of relevance judgments. What insiders immediately see as relevant will usually be altogether lost on outsiders. The "parochialism" of the examples is thus necessary because, unlike descriptors and identifiers, authors' names indicate subject matter only implicitly; everything they convey must be read into them.

The use of tf*idf to rank co-citation data on authors and publications was first tried in White (2007a). The results in Tables 4, 5, and 6 replicate parts of the analyses there and are consistent, on a smaller scale, with the earlier findings. The co-citation record indicates that both Cronin and Wilson are more closely identified with authors in the broad subfield of informetrics than in other subfields of IS, such as experimental information retrieval or user studies. Clues in the lists of top co-citees shade Cronin toward citation analysis (e.g., MacRoberts, Brooks, Cano, Baird, McCain, Peritz, White) and Wilson toward mathematical bibliometrics (e.g., Egghe, Wolfram, Rousseau, Tague-Sutcliffe, Glanzel, Van Leeuwen, Bradford). But names in the top 15 of each seed author overlap, and the intellectual boundaries between them are clearly not hard and fast.

Of greater note here is what tf*idf does in this context. To start with, it automatically makes a strong partition that, in effect, eases relevance judgments. For both seeds, the top 15 co-citees are all information scientists, while none of the bottom 15 are. (A possible exception is "O'Connor J" in Table 5, but that may be a homonym for several people.) Many of the names at right in both tables will be familiar to readers in informetrics as famous social scientists,

philosophers, and quantitative methodologists. Their writings are certainly relevant to those of Cronin and Wilson, but not as relevant, says tf*idf, as writings by the top co-citees.

As we have seen, tf*idf tends to associate high relevance with *obvious* connections, the ones that spring to mind most easily. Thus, in Tables 4 and 5, it has automatically brought to the top authors who are the seed's co-authors and/or doctoral students (Davenport with Cronin; Hood and Osareh with Wilson; see White 2007a for other examples of this). It has moreover pushed to the bottom authors from other disciplines who might require effortful reading and

Table 5. Authors co-cited with Concepción S. Wilson in Social Scisearch

tf*idf	Top 15 terms	tf*idf	Bottom 15 terms
9.36	Hood WW	4.59	Krippendorff K
8.66	Egghe L	4.54	Zuckerman H
8.66	Wolfram D	4.53	King J
8.5	Bossy MJ	4.48	Shadish WR
8.39	Rousseau R	4.45	Miller C
8.27	Blackert L	4.4	Crane D
8.27	White HD	4.4	Merton RK
8.21	Osareh F	4.19	O'Connor J
8.18	Tague-Sutcliffe J	4.08	Kruskal JB
8.17	Noyons ECM	4.05	Milgram S
8.07	Glanzel W	3.79	Scott J
8.05	Ingwersen P	3.7	Latour B
7.97	Deogan MS	3.67	Douglas M
7.92	Van Leeuwen TN	3.3	Kuhn TS
7.9	Bradford SC	3.21	Simon HA

creative thought to relate to the seeds, such as Wittgenstein and Bourdieu in Cronin's case or Kuhn and Latour in Wilson's.

A nearer example for both Cronin and Wilson is Eugene Garfield. He has the highest raw co-citation count—the highest tf—with either of them, but he has written voluminously and been cited thousands of times, which signals to idf that his influence is diffuse and hard to pin down. So, like other famous figures, idf relocates him far down the lists of Cronin's and Wilson's co-citees. He is of course relevant to both of them, but only generically so, as it were.

The authors that tf*idf finds most specifically relevant to Cronin are interesting. Their names, compared to Garfield's, are still rare, bibliometrically speaking. But plainly Björneborn, Almind, Thelwall, and Bar-Ilan are not more specific than Garfield (or anyone else) *as persons*. Their names here stand for their works, and the works of theirs that intersect Cronin's have a very specific focus: the conjunction of citation analysis with webometrics. A key example from Cronin would be his *Journal of Information Science* article, "Bibliometrics and beyond: Some thoughts on web-based citation analysis." Citation analysis itself has been around for decades and has a sprawling, heterogeneous literature to show for it. But its links to webometrics are relatively new, and the associated literature is relatively small, which is why tf*idf brings its representatives to the fore with Cronin in Table 5.

Table 6, the final one, displays journals and books co-cited five or more times with a famous book, Thomas S. Kuhn's *The Structure of Scientific Revolutions*, since 2000. The counts were drawn from the more recent segment of the Scisearch file on Dialog. Scisearch notoriously abbreviates book and journal titles in cryptic ways. To compensate, the titles of the top 15 items have been spelled out (but not always their subtitles), and the surnames of the book authors have been inserted (when possible). The bottom items consist entirely of journals,

whose full titles have been restored.

This table provides an abundantly clear illustration of what tf*idf does when applied to cited works (CW) in any of the Thomson Reuters databases. The only outcome that is hard to interpret is the presence of the *Journal of Physiology–London* and *Marine Ecology–Progress Series* at the head of the top 15 co-cited items. Their connection to Kuhn is certainly not obvious. It appears that articles in them were simply co-cited with his book in historical studies particular to their disciplines. Because the df counts of these journals in Scisearch are

Table 6. Works co-cited with Kuhn's *The Structure of Scientific Revolutions* in Scisearch

tf*idf	Top 15 terms	tf*idf	Bottom 15 terms
13.9	Journal of Physiology–London	2.48	Journal of Organic Chemistry
13.4	Marine Ecology–Progress Series	2.41	European Journal of Biochemistry
12.9	Criticism and the Growth of Knowledge (Lakatos & Musgrave)	2.41	FEBS Letters
12.34	Thomas Kuhn (diverse authors)	2.38	EMBO Journal
12.07	The Methodology of Scientific Research Programmes (Lakatos)	2.34	Tetrahedron Letters
12.06	The Logic of Scientific Discovery (Popper)	2.33	Analytical Chemistry
12.04	The Essential Tension: Selected Studies in Scientific Tradition and Change (Kuhn)	2.22	Biochemical and Biophysical Research Communications
12.03	World Changes: Thomas Kuhn and the Nature of Science (Horwich)	2.14	Physical Review B
12.01	Reconstructing Scientific Revolutions (Hoyningen-Huene, Levine, & Kuhn)	2.11	Journal of the American Chemical Society
11.83	Against Method (Feyerabend)	2.09	Applied Physics Letters
11.82	Thomas Kuhn: A Philosophical History for Our Times (Fuller)	2.04	Journal of Applied Physics
11.82	Genesis and Development of a Scientific Fact (Fleck)	1.92	Science
11.82	The Road Since Structure (Kuhn, Conant, & Haugeland)	1.86	Proceedings of the National Academy of Sciences USA
11.7	Sociology of Science (Merton & Storer)	1.84	Journal of Biological Chemistry
11.66	Conjectures and Refutations (Popper)	1.75	Nature

very low compared to those of the powerhouse journals at right, tf*idf singled them out for elevation to the top.

The most telling features in the table are (1) the ease with which the books at left can be related to *The Structure of Scientific Revolutions*, and (2) the difficulty of saying how the

journals at right relate to it. The latter are, of course, relevant to Kuhn's book once one passes to the level of individual articles in them, but at the level of journal titles themselves, all is obscure. As we have seen before, idf moves items with very high df counts sharply downward, even when they have large tf counts with the seed. Journals at right such as *Science*, the *Proceedings of the National Academy of Sciences USA*, and *Nature* have among the highest df counts in Scisearch, and so they are penalized for it. At the same time, tf*idf does a very notable job of foregrounding books that almost any expert (and even non-experts) would say are highly relevant to Kuhn's *Structure*. The titles at left, generated automatically, make a focused reading list for a graduate seminar in contemporary philosophy and sociology of science. Besides other books by Kuhn himself, there are books by several of his rivals in much-studied philosophical debates—Popper, Lakatos, and Feyerabend. There are critical studies of Kuhn by Fuller, Horwich, and other writers. (The latter wrote books with the abbreviated title "T Kuhn" in Scisearch that are not disambiguated here.) There are classics from the sociology of science by Merton, Storer, and Fleck. It appears that tf*idf is making a statement: "Given the context set by Kuhn's *Structure*, these are the books predicted to have high cognitive impact in relation to it and whose relation to it is easy to see." (That does not mean, of course, that they are "easy reads.")

Discussion

Deirdre Wilson was quoted above on "Relevance to an individual." The data in the present study did not come from experimental trials involving the judgments of lone individuals. They are not direct psychological tests of RT such as have been conducted by Van der Henst and Sperber (2004). Rather, in an unconventional way, they follow the investigative method of much of linguistics, which uses the reader's own ability to intuit communicative effects as empirical data: one tests the persuasiveness of the evidence by experiencing it in reading.

In linguistics, it is customary to offer examples of words, phrases, and sentences that test the reader's sense of semantics or grammar. S&W make extensive use of short dialogs between imaginary persons that test the reader's ability to infer meanings that are only implicit in what is actually said. Here, the data come from bibliographic sources that have already created relevance relations among various terms. These relations accrue through the judgments of authors, editors, and indexers and can be expressed, in part, through counts and other numeric data. Because such relations involve the verbal perceptions of many contributors, Ingwersen & Järvelin (2005: 240) see them as examples of what they call socio-cognitive relevance (italics theirs): "*Socio-cognitive* relevance assessments are tangible, e.g., by means of the citations (or inlinks) given to the [information] objects. The citations by scholarly colleagues imply commonly a certain degree of recognition, acceptance and use, and degrees of cognitive authority."

The present study has examples of socio-cognitively relevant data that have been reconfigured by a formula, tf*idf, ordinarily used in document retrieval. The examples serve to probe individual readers' intuitions. Presumably, readers will agree that the applications of tf*idf seen here are quite consistent with predictions made by RT (and more rigorous tests of the underlying hypothesis are not hard to imagine). This is one step in drawing RT and IS closer together.

But why should this convergence be desirable? The reason for wanting it is that, although RT developed independently of IS, it can explain a wide range of IS phenomena. The small study reported here shows that RT can explain what happens when tf*idf is applied to bibliometric distributions. But RT can also explain why tf*idf has succeeded fairly well over the years in real-world document retrieval and why computer scientists like it. It is because tf*idf ranks documents whose relevance to a query is easy to see higher than documents whose relevance to a query is harder to see (cf. White 2007a on *Moby Dick*). What kind of relevance is

generally preferred? *Topical* relevance, which is manifested when document terms duplicate the sense of query terms exactly or nearly or through familiar semantic ties. Above, Table 2 on "Information Needs" and Table 6 on Kuhn's book show some topical matches of this kind. Topical relevance is the stock in trade of retrieval systems designers; it is what their users expect and judge favorably when they get it—judgment that reflects credit on the designers. Harter (1992), however, argued cogently that there are important relevance relations among documents other than topical match. He imported RT into IS because of his conviction, gained after years of reading retrieval system evaluations, that *cognitive effects* in judging documents are not limited to the effects of topical matches. Non-topical effects may merely cost the right person more *processing effort*. The present study hints at that, too. The low-ranked terms in Tables 2 through 6 might seem difficult to connect to the seed term for most of us, but, in the right person, any one of them could strike sparks.

RT has explanatory power in many other areas of IS than this, and information scientists will be repaid by learning more about its potential.

References

Baird, L. M., & Oppenheim, C. (1993). Do citations matter? *Journal of Information Science*, 20, 2-15.

Goatly, A. (1997). *The language of metaphors*. London and New York: Routledge.

Harter, S.P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43, 602-615.

Ingwersen, P., & Järvelin, K. (2005). *The turn; Integration of information seeking and retrieval in context.* Dordrecht: The Netherlands.

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.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Saracevic, T. (1975). Relevance: A review of and framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26, 321-343.

Schneider, J. W., Larsen, B., & Ingwersen, P. (2007). Pennant diagrams, what is it [sic], what are the possibilities and are they useful? Presentation at the 12th Nordic Workshop in Bibliometrics and Research Policy, Copenhagen, Denmark, September 13-14, 2007. Retrieved January 20, 2009 from www.db.dk/nbw2007/files/2c_Peter_Ingwersen.pdf.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, 28, 11-21.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition.* (2d ed.) Oxford, UK, and Cambridge, MA: Blackwell.

Van der Henst, J.-B., & Sperber, D. (2004). Testing the cognitive and communicative principles of relevance. In I. A. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp. 141-171). Houndmills, Basingstoke, UK, and New York: Palgrave Macmillan.

White, H.D. (2007a). Combining bibliometrics, information retrieval, and relevance theory. Part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology*, 58, 536-559.

White, H.D. (2007b). 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, 583-605.

Wilson, D. (2007). Relevance: the cognitive principle. Lecture 3 of Pragmatic Theory (PLIN2002) 2007-08. Retrieved January 25, 2009 from http://www.phon.ucl.ac.uk/home/nick/content/pragtheory/PRAG3.doc