Scientometrics (2015) 102:2275–2299
DOI 10.1007/s11192-014-1483-4

Co-cited author retrieval and relevance theory: examples from the humanities

Howard D. White

Received: 29 October 2014 / Published online: 26 November 2014

Abstract Given a user-selected seed author, a unique experimental system called AuthorWeb can return the 24 authors most frequently co-cited with the seed in a 10-year segment of the Arts and Humanities Citation Index. The Web-based system can then instantly display the seed and the others as a Pathfinder network, a Kohonen self-organizing map, or a pennant diagram. Each display gives a somewhat different overview of the literature cited with the seed in a specialty (e.g., Thomas Mann studies). Each is also a live interface for retrieving (1) the documents that co-cite the seed with another user-selected author, and (2) the works by the seed and the other author that are co-cited. This article describes the Pathfinder and Kohonen maps, but focuses much more on AuthorWeb pennant diagrams, exhibited here for the first time. Pennants are interesting because they unite ego-centered co-citation data from bibliometrics, the TF*IDF formula from information retrieval, and Sperber and Wilson's relevance theory (RT) from linguistic pragmatics. RT provides a cognitive interpretation of TF*IDF weighting. By making people's inferential processes a primary concern, RT also yields insights into both topical and non-topical relevance, central matters in information science. Pennants for several authors in the humanities demonstrate these insights.

Keywords Author co-citation analysis · Bibliometric visualizations · Cognitive information science · Pennant diagrams · Relevance · TF*IDF weighting

Introduction

Researchers in the humanities often characterize their interests in terms of individual creators or artists or scholars. They specialize in names around whom scholarly literatures have gathered—Melville, for example, or Dürer or Bertrand Russell or Verdi. Such names

H. D. White (✉)
College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA
e-mail: whitehd@drexel.edu

are frequently co-cited with other authors. Co-cited authors' names can address wide-ranging interests that no single topical phrase captures. This motivates a mapping service for the humanities—one that puts user-selected seed authors and their co-citees on view, one that instantly reveals the intellectual company a seed author keeps. The system now called AuthorWeb is such a service (White et al. 2001; Buzydlowski 2002; Buzydlowski et al. 2003; Lin et al. 2003). Experimentally operational on the Web for more than a decade, AuthorWeb combines bibliometric data, visualizations, and document retrieval. That is, since its maps are made from co-citation data in scholarly literatures, they are also designed to retrieve the co-citing documents from those literatures. Integrated into a digital library, they could provide leads to browsing and searching author-based specialties. The user need only *recognize* interesting co-citation ties in a map, rather than asking for them in advance—a principle well known to designers of user-friendly retrieval systems.

The AuthorWeb maps are made from a file of the Arts and Humanities Citation Index (AHCI) that contains 1.26 million bibliographic records from the period 1988–97. AHCI's publisher, Thomson Reuters, gave the file to what is now Drexel University's College of Computing and Informatics for research purposes. The file may be updated with more recent records in the future, but, as a prototype, AuthorWeb does not chiefly depend on the currency of its data. It was designed to be a proof of concept.

As far as I am aware, no system other than AuthorWeb instantly maps co-cited author data today. Current major sources of citation data—the Web of Science, Scopus, Google Scholar—are not likely to provide a service like it in the near future, nor do they provide others with the necessary data for large-scale use. But open-access digital libraries may increasingly link citation data to bibliographic records. Some version of AuthorWeb's open software is thus potentially usable in a digital library that has standardized citation data.

The present paper discusses the three kinds of co-citation maps produced by Author-Web: pathfinder networks (PFNETs), Kohonen self-organizing maps (SOMs), and pennant diagrams. However, it is mainly devoted to pennant diagrams, a type of map added to PFNETs and SOMs several years after the AuthorWeb system was first made public. Pennants were introduced in White 2007a, b and further illustrated and discussed in Schneider et al. (2007), White (2009, 2010a), and White and Mayr (2013). These earlier pennants were made from various kinds of bibliometric data (not just co-cited authors) in the databases of DialogClassic.[1] The simpler AuthorWeb pennants have not hitherto been described in a publication. Those here feature several European authors as seeds, two chosen by me and three chosen by others. They are tests of whether AuthorWeb can intelligibly map authors never mapped before, with outcomes that readers of the present article may judge.

Pennants are a way of visualizing terms, here authors' names, whose frequency counts in a database have been weighted according to the formula TF*IDF. This formula will later be discussed at length, but for now let me say that I created pennants in part to explain why it has been much used by designers of document retrieval systems. The explanation is based on relevance theory (Sperber and Wilson 1995; Wilson and Sperber 2004; Clark 2013), a major subfield of linguistic pragmatics that was introduced into information science by Harter (1992). Information scientists have discussed relevance for more than half a century, yet have never defined the concept as clearly as Sperber and Wilson's work makes possible. Relevance theory (RT) provides a *cognitive* interpretation of TF*IDF

[1] Dialog, the "Cadillac" of bibliographic information services since the 1960s, was purchased by ProQuest in 2006. In 2013 ProQuest took the decades-old DialogClassic software out of service and terminated access through it to the Thomson Reuters citation databases.

weighting that pennants can illustrate. Like White (2011), the present article therefore makes psychological claims unusual in bibliometrics. Yet Harter (1992: 612–613) linked RT not only to relevance judgments in document retrieval but to bibliometrics as well. I will therefore use AuthorWeb pennant diagrams from the humanities to further explicate relevance as a central concept in information science.

AuthorWeb and maps for Mann

To use AuthorWeb, a researcher or student simply enters the name of *one* seed author of interest (e.g., the novelist Thomas Mann)—a cognitive task deliberately kept minimal. The system then instantly retrieves and maps the 24 other authors who are most frequently co-cited with the seed. Although the maps are very shallow introductions to the study of any author, that is by design. They are meant to be light, fast overviews, which is why they present just 25 authors as labeled nodes. That number assures enough authors to suggest the complexity of an author-centered specialty, but also minimizes overlapping names, a problem that often mars visualizations of bibliometric data.[2] The goal is to make the maps just rich enough to have heuristic value for a student, teacher, literature reviewer, or intellectual historian. Perhaps their main use would be as aids to discovery for literature searchers newly caught up in the scholarship surrounding an author. As such, they can be quickly called up and quickly discarded.

In AHCI, many thousands of names—the famous and the not-so-famous, scholars as well as artists—may be used as seeds, although well-cited authors work best. In co-cited author retrieval, a seed author is implicitly *a subset of writings from the seed's oeuvre*. Co-cited authors are *subsets of writings from other authors' oeuvres*. The seed's subset is sent out to retrieve these other subsets that are cited with it in later publications. This is the "ego-centered" mode of retrieval introduced in White (2000) and elaborated in Cronin and Shaw (2001), Bar-Ilan (2006), White (2007a, 2009), McCain (2010), and Hu et al. (2012).

One must always add that AHCI draws its citation data from articles in humanities journals, as opposed to books, and that it covers articles in English from American or British journals much more fully than those in other languages and from other countries. Nevertheless, it does cover some of the latter. Recall, too, that it indexes *anything these articles cite*, which includes books and other items regardless of language. AuthorWeb maps are therefore based on cited works of any sort—for example, Franz Kafka's novels in German, when the citing articles are in English, German, or French.

The PFNETs, SOMs, and pennants of AuthorWeb all define a specialty in terms of co-cited author data from published research. Each offers somewhat different information on how given sets of 25 authors—the seed plus 24 others—are interrelated. All emerge from the referencing behavior of citers in general, as opposed to that of any one scholar. They are thus "pictures of the literature" that can reveal connections otherwise invisible and possibly new to users—connections that bear on fields such as canon formation studies, audience reception theory, the study of interdisciplinarity, the sociology of literature, and the history of ideas. Users do, of course, need domain knowledge to interpret and evaluate what they see.

[2] More than 25 nodes can be mapped in extensions of AuthorWeb software to medical databases in the Visual Concept Explorer (Zhu et al. 2005; Lin et al. 2007), but such additions arguably lead to information overload.

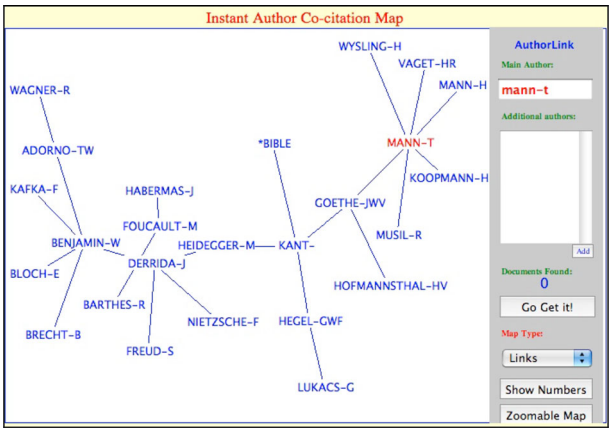

Fig. 1 Pathfinder network for Thomas Mann

The three types of AuthorWeb maps appear in Figs. 1, 2 and 3. I chose the first seed, the German novelist and Nobel laureate Thomas Mann. By default, AuthorWeb first shows him and his co-citees as a PFNET. From this beginning, one can immediately pass to a SOM or a pennant if desired.

Figures 1, 2 and 3 depict Mann studies, a typical specialty. All three contain the same co-cited authors. (The surname-hyphen-initials format of their names is built into the AHCI file that Drexel received.) Mann's top co-citees, like those of countless other seeds, include scholars who specialize in him (e.g., Hans Wysling), pertinent theorists (e.g., Theodor Adorno, whose knowledge of music informed Mann's novel about a composer, *Doctor Faustus*), and comparable creative artists (e.g., Franz Kafka, Bertolt Brecht, Robert Musil, Johann Wolfgang von Goethe). The maps in Figs. 1, 2 and 3 are typical in their sweep across different implicit disciplines, genres, and periods.

The word "author" can denote either a person ("Mann was born in Lübeck.") or an oeuvre ("She has a big shelf of Mann in German"). In AuthorWeb, it refers mainly to oeuvres. That is, the co-cited *authors* in the maps are always extracted from strings of co-cited *works* (White 2007a: 552–553). To illustrate, Fehn (1988) co-cites a monograph by Freud with an undated edition of Mann's novella *Mario und der Zauberer* [*Mario and the Magician*], and the AHCI strings look like this:

FREUD-S, 1921, MASSENPSYCHOLOGIE IC,[3]
MANN-T, NULL, MARIO UND DER ZAUBER,

Besides creating maps, AuthorWeb automatically places the seed's name, here Mann's, in the search box at top right. The larger search box below it is where additional authors' names can be put by clicking on them if a co-cited author retrieval is wanted. The "Go Get It!" button initiates the retrieval. Clicking on, say, Freud's name in any of the maps will place it in the search box. A second click will then retrieve the documents (such as Fehn 1988) that cite anything by Freud with anything by Mann. A click on one of the retrieved documents will show its references in abbreviated string form, including the specific works

[3] Which expands to: Freud, Sigmund. (1921). Massenpsychologie und Ich-Analyse. Wien: Internationaler Psychoanalytischer Verlag. [Group Psychology and the Analysis of the Ego. Vienna: International Psychoanalytic Press.] The commas at the end of strings indicate truncations.

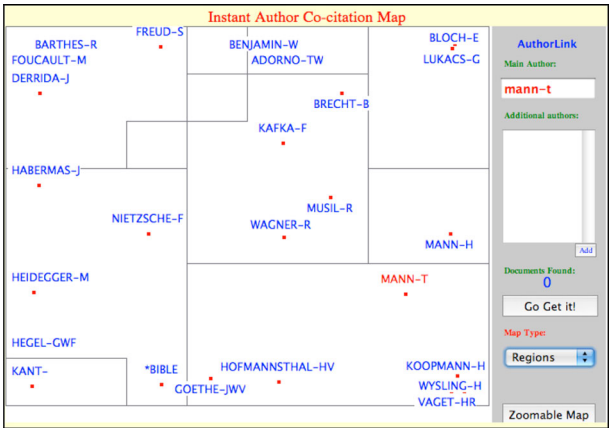

Fig. 2 Self-organizing map for Thomas Mann

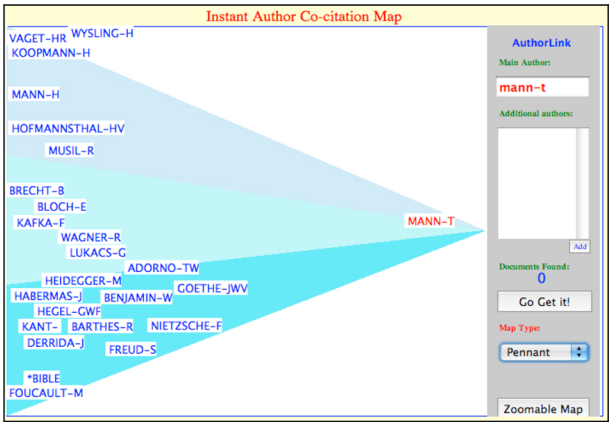

Fig. 3 Pennant diagram for Thomas Mann

by Freud and Mann that were jointly cited. In a real sense, AuthorWeb is a recommender system like Amazon's for book purchases: "People who cited your seed author's works also cited these works by…"

In Thomson Reuters databases, when a single article cites one or more works by two different authors, the co-citation count for that pair increases by one (White 2011: 3360). Thus, if an article cites Mann's novel *Doctor Faustus* and Goethe's dramas *Faust* and *Egmont*, it is not the one citation to Mann and the two to Goethe that are counted. It is the article in which Mann and Goethe are co-cited. The number of articles citing any such pair is that pair's co-citation count.

Pathfinder networks

PFNETs, as in Fig. 1, are designed to reveal the strongest links between pairs of authors in a co-citation matrix. The algorithm first retrieves Mann's co-citation counts with the other authors in the database and selects the 24 co-citees with the highest counts. It then retrieves

the co-citation count for each of the 24 with every other member of the 24. This results in a 25 × 25 symmetric matrix of counts in which 25(24)/2 = 300 pairs are unique. Finally, the PFNET algorithm displays links representing only the *highest* (or tied highest) count for any pair, eliminating all the rest. The radically simplified network is generally very intelligible, as it would not be if all the non-zero counts actually present in the data were shown as links.

Mann, for example, has non-zero counts with everyone else in the map, but only six authors have their highest counts with him. They are the novelists Heinrich Mann (his brother) and Robert Musil, Goethe, and the Mann scholars Wysling, Hans Rudolph Vaget, and Helmut Koopmann. Hugo von Hofmannsthal's highest count is with Goethe, not Mann; Wagner's is with Adorno; and so on. Goethe links Mann to a very distinguished body of European philosophers, critics, composers, and literary artists. (Also the Bible; clearly these are not co-author relations.) It is easy to pass from this map to one involving another known relationship—say, Nietzsche and Wagner independently of Mann—and then back, if desired.

Self-organizing maps

SOMs, as in Fig. 2, are also generated from the 25 × 25 matrix of counts. Instead of links, these neural networks, also known as Kohonen feature maps after their Finnish originator, make use of a distance metaphor to show relationships between terms, here authors' names. As stated in White et al. (2004: 5300), "The more frequently co-occurring terms, which presumably have greater mutual relevance, occupy more proximate regions of the map. SOMs are designed to render not just the highest co-occurrence counts between terms, but relatively high co-occurrences across groups of terms. They are a softer-focus kind of mapping than PFNETs, but they, too, suggest specific combinations of terms on which the user might want to base retrievals."

In direct comparisons of AuthorWeb PFNETs and SOMs in the humanities, Buzydlowski (2002) found that his 20 domain experts were about equally divided in their preferences for one type of map over the other. Those preferring the SOMs considered their regions more evocative than the less ambiguous links of the PFNETs. It will be seen, however, that Figs. 1 and 2 agree fairly well in the authors they place relatively close to Mann and in other groupings of highly related authors, Wagner being an exception.

Before turning to the pennant for Thomas Mann, which has its own section below, I must discuss various aspects of pennants in general. As stated above, they have their basis in Sperber and Wilson's (1995) relevance theory (RT), and so that must be sketched as well. It is worth doing so, because RT has implications for information science, and in particular for the psychology of document retrieval, that go well beyond their context here (White 2007a, b). RT can even lend a psychological cast to various bibliometric distributions (White 2009, 2010a, 2011).

Pennants and their relevance-theoretic background

The first principle of RT is that, by evolutionary adaptation, human cognition tends to maximize relevance. S&W define relevance as a property of inputs to human minds. Inputs include both perceptions of one's surroundings and communications from others. The focus here is on communications, specifically written utterances. RT has primarily been used to explain how hearers grasp what speakers actually mean from what they in fact say.

But it has also been used to explain how readers grasp writers' underlying meanings from their texts (Wilson 2011), including those on the Internet (Yus 2011). Writings qualify for relevance-theoretic analyses because, like interpersonal talk, they are overtly intentional. Writers want both to inform readers and to communicate that information, and readers must recognize these intentions if communication is to succeed (Clark 2013: 112–119). Writers therefore engage in what RT calls ostensive communication: that is, they *demonstrate* they want to claim readers' attention by expressing their thoughts in a durable medium that is then made available to audiences ranging from one person to large publics. Pennants, although a very atypical kind of writing, are similarly intended to claim attention—to be heeded as inputs from which useful meanings can be derived.

Any individual's mind, according to RT, comprises large stocks of existing assumptions, glossed in Goatly (1997: 137) as "beliefs/thoughts." Assumptions are thus *cognitive contexts* that an individual can access in response to inputs (such as bibliographic information from pennants). Among the innumerable inputs the individual might heed, the ones actually heeded are those with the greatest perceived relevance at the moment.

The relevance of an input to an individual depends on two factors that operate simultaneously. They are sometimes cast as a ratio: *relevance = cognitive effects/processing effort* (S&W 1996; Goatly 1997: 139). Given a context of assumptions, that is, the relevance of an input increases *directly* with its effects, but *inversely* with the effort needed to process it. A greater cost in effort decreases relevance. Nevertheless, an individual may be willing to expend more effort if prospective effects seem worth it. Thus the relevance of an input may vary over time.

Effects in RT are mainly matters of inference rather than of coding and decoding. RT subordinates a code model of communication, such as the famous one in Shannon and Weaver (1949), to an inferential model that descends, with considerable modifications, from Grice (1989). The inferential model more adequately explains how communication works, since what we are able to infer from utterances is almost always more than what speech or writing actually encodes. Speakers and writers routinely rely on this inferential ability in their audiences; they do not attempt to spell everything out (Clark 2013: ch. 2). To test this claim, read any novel, watch any movie.

Cognitive effects in RT are of three kinds. Each is an inference in which a new input modifies an individual's existing assumptions. The information may (1) strengthen an assumption, (2) eliminate an assumption, or (3) combine with an assumption to yield a new conclusion (Clark 2013: 102). People continually seek to know more about their worlds, and when a new input actually brings this about, S&W term it a "positive cognitive effect." As relevance maximizers, people want positive cognitive effects. ("Positive" does not mean pleasing, but truthful, reality-based.) Effects are examples of cognitive change—of new inputs modifying existing contexts of assumptions, just as new information is said to modify knowledge structures in cognitive information science (White 2010b).

Since S&W define relevance as a ratio, it is natural to ask how cognitive effects and processing effort in individuals can be measured. The answer is, only ordinally. No one, that is, can measure either of the two factors exactly on ratio-level scales through introspection or by any known instrument. Nevertheless, when the implications of comparable utterances differ markedly, we can sense it, and their degrees of relevance will differ. Consider three utterances made on November 22, 1963, all requiring the same effort to process:

Aldous Huxley died today.
C. S. Lewis died today.
President Kennedy died today.

For almost everyone, including devoted admirers of Huxley and Lewis, the third piece of news had profoundly greater implications—greater effects—and was thus by far the most relevant of the three, even if the difference cannot be captured on a ratio-level scale. In like (if less dramatic) fashion, when people judge the relevance of documents to queries in information retrieval (IR), the scale values tend to be in terms of *more* or *less*. (Eisenberg 1988 records an attempt to measure relevance judgments more precisely.) In IR, moreover, judges often disagree in how they score the same document. RT would explain this by claiming that everyone judges the relevance of utterances in contexts of assumptions that vary over individuals. Information scientists have reached much the same conclusion (Saracevic 2007: 2134–2137).

To show how utterances with the same cognitive effects differ perceptibly in processing effort, RT uses artificial examples like the following from Clark (2013: 105, using his numbering scheme):

> Suppose I ask you whether it is raining and you reply:
> (35) Yes, it is.
> This response is relevant because it confirms my initial assumption, which counts as a positive cognitive effect. Now imagine you replied instead:
> (36) Yes, it is raining and it rained in Aberdeen on the second of July 1864.
> If nothing follows from knowing whether it rained in Aberdeen on 2nd of July 1864, then (36) is less relevant than (35). This follows because (36) requires more processing effort but does not lead to any extra cognitive effects.

Somewhat more realistic examples of processing effort appear in the discussion to come.

The effects-effort ratio and TF*IDF

As it happens, the effects-effort ratio provides a psychological rationale for TF*IDF. The standard textbook by Manning et al. (2008: ch. 6) discusses TF*IDF as a formula for weighting query terms and indexing terms in computerized document retrieval systems. When numbers are plugged into TF*IDF, it yields predictions of how relevant documents will be to ordinally-judging people. The predictions involve ranking documents by their TF*IDF weights. Usually the terms being weighted will be topical in nature, and relevance will depend on how well documents match the topical sense of users' queries (Green 1995).

TF is term frequency—a count of how frequently a significant query term appears in any document in the system. (Researchers often supplement TF with a factor that normalizes documents of different lengths.) DF is document frequency—a count of how many documents in the system contain a query term. IDF is inverse document frequency—the raw DF count divided into the total number of documents in the system. The IDF factor was introduced as "statistical specificity" in Sparck Jones (1972), because, being inverse, it is higher for terms that occur *less* frequently in the document collection and lower for terms that occur *more* frequently. IDF boosts less frequent terms for being presumably more specific to the searcher's interest, while penalizing more frequent terms for being less specific and hence less informative. Lastly, in a version of the formula in Manning et al. (2008: 118), both TF and IDF are damped by taking logs, so that the weight of query word_i in document_j is given by

$$\left(1 + \log\left(\mathrm{TF}_{i,j}\right)\right) * \log(N/\mathrm{DF}_i)$$

where N is the number of documents in the system.

As components of (broadly) intentional utterances by system designers, TF predicts cognitive effects on the user, and IDF predicts the user's processing effort. Multiplying TF by inverse DF is like dividing cognitive effects by processing effort. However, since greater statistical specificity is supposed to reduce processing effort, higher IDF weights predict *less* effort, and lower weights predict *greater* effort. This inverse measure is therefore clearer if it is renamed *ease of processing*, so that high weights mean *easy*, and low weights mean *hard*. But it remains a scale of processing effort as in RT.

Relevance theory suggests how designers use TF*IDF to increase the relevance of retrieved documents to a user's query. High TF weights elevate documents in which significant query terms occur relatively frequently. If the user sees these terms emerging in the top-ranked retrievals, it may strengthen his or her assumption that the retrievals match the query and so are relevant—a positive cognitive effect (often one of many). And if relevance goes up when effort goes down, higher IDF weights will elevate documents containing terms more specifically related to the query, which presumably are easier for the user to process. The two factors operate simultaneously, just like cognitive effects and processing effort in RT.

Relevance and topicality

It hardly needs adding that top-ranked retrievals are not always on topic. Yet retrieval system designers do the best they can. In the technical language of RT, they cannot routinely produce *maximal* relevance—the greatest possible effects for the least possible effort—which is what users automatically seek. But designers can try for *optimal* relevance—adequate effects for no unjustifiable effort (Higashimori and Wilson 1996). Optimal relevance for *speakers* is related to least effort for *hearers* in Carston and Powell (2008: 342):

> Quite generally, an utterance comes with a presumption of its own optimal relevance; that is, there is an implicit guarantee that the utterance is the most relevant one the speaker could have produced, given her abilities and her preferences, and that it is at least relevant enough to be worth processing. That utterances carry this presumption motivates a particular comprehension procedure, which, in successful communication, reduces the number of possible interpretations to one: in essence, it licenses a hearer to consider possible interpretations in order of their accessibility (that is, to follow a path of least effort) and to stop as soon as he reaches one that satisfies his expectation of relevance.

This specifies the main function of least effort in RT: from among the multiple interpretations any utterance *could* have, the hearer accepts the one that *does* have satisfactory cognitive effects in context and then *stops processing*. As a rule, the hearer will thus need to process only one interpretation before moving on to the next utterance.

The same account can be applied to written communications between systems designers and users. The utterances of designers are the document surrogates their systems deliver in response to users' queries. Designers have far less control over these utterances than they would over their own speech or writing, but, given present technology, that is a limit in their *abilities* (as in the quote above). They know that their utterances will be read by users

with built-in expectations of relevance. They know that, in deciding whether a surrogate is relevant to a query, users will seek the most accessible interpretation. They know that the most accessible interpretation—the one likely to cost users least effort and be stopped at—is one based on whether a surrogate *matches the sense of terms* in the query. And while term-matching can now be algorithmically expanded, refined, or supplemented, it remains at the heart of document retrieval because inferences about term-matching are easiest for users to draw. Thus originates the primacy of topical relevance in information science (White 2007b: 585–586).

We are perennially left with a large question, however: what about document surrogates that are in fact relevant to the query but do not match its terms? RT suggests a hypothesis: they are harder to process. A corollary of this hypothesis is that those with enough background knowledge to process them will be specialized researchers. As a practical matter, judgments on whether surrogates and queries agree in sense can by made by anyone who knows the topical vocabulary. But not everyone will be willing to assess the relevance of surrogates that *do not match* queries. Where do we find non-matches of this sort? In *citations*, which frequently connect documents not obviously related. This has always been a selling point for citation databases. Given a work or an oeuvre as query, they can return relevant documents that do not match the query's title terms or subject indexing.

Relevance theory, TF*IDF, and co-occurrence data

Sorting *more* relevant from *less* relevant documents is what Sparck Jones (1972) and subsequent information scientists intended TF*IDF to accomplish in literature searches. I simply claim here that the formula can be used to relevance-rank *the terms that co-occur with the seed,* as well as the documents in which they co-occur.

A seed term is a kind of query. In relevance-theoretic language, it indicates a user's context of assumptions in which the terms that co-occur with it are new inputs. When a user processes the seed and a new input jointly, both non-negligible cognitive effects and effort may result. The relevance of the new input to the seed depends in part on the cognitive effort it takes to process them together.

The skewed distributions of bibliometrics are formed when numerous terms that co-occur with a seed are ranked by their co-occurrence counts. In White (2010a) I took several such distributions and showed how terms ranked by a single TF*IDF weight differed qualitatively at the upper and lower ends. Top-ranked terms were semantically related to the seed in specific and obvious ways, while bottom-ranked terms were much more general in their implications. They frequently co-occurred with the seed, but their semantic connections to it were much less obvious. The terms thus differed in their cognitive effects and the effort they cost to process.

For example, in White (2010a) the descriptor "Information Needs" is used as a seed to retrieve the other descriptors assigned with it to documents in the ERIC database. When these co-assigned descriptors are weighted and ranked by TF*IDF, those that rise to the top include terms like "User Needs (Information)" and "Information Seeking." Their relevance to the seed costs little effort to see and thus is greater than that of terms pushed to the bottom by TF*IDF, such as "Community" and "Relationship." The relevance of the latter to the seed is harder to see because of their semantic distance from it.

TF*IDF has historically been geared to retrieval with topical noun phrases from natural language. Almost never has it been applied to cited or co-cited authors' names. However, it

functions with these names as if they were topical terms. Indeed, to those with the right domain knowledge, they *are* topical terms. For example, many information scientists would know that the name "Leo Egghe" implies "Mathematical Bibliometrics," just as "Mathematical Bibliometrics" implies "Leo Egghe." (This is encyclopedic, as opposed to linguistic, knowledge.) In this sense, "Leo Egghe" is a topical term, as are the names of millions of other authors.

Accordingly, White (2010a) takes the name of an author, "Katy Börner," to connote "the kinds of things Börner writes about in her oeuvre," and uses her name a seed in the INSPEC database to retrieve the descriptors assigned to her publications. When, again, these descriptors as weighted and ranked by TF*IDF, those on top are *specifically* relevant to Börner's research, such as "Data Visualization," "Citation Analysis," and "Digital Libraries." The broader ones toward the bottom, such as "Computational Complexity" and "Diagrams," are not irrelevant, but, at least on first exposure, they relate less to her work than to that of computer scientists in general. They are thus less informative, less relevant.

The upshot is that, in skewed distributions like those generated by "Information Needs" and "Katy Börner," relevance is given a new psychological interpretation. Since AuthorWeb pennants are based on similarly skewed distributions of author names in the AHCI database, they inherit this interpretation. Co-cited authors' names imply oeuvres. Oeuvres comprise works whose bibliographic descriptions, when retrieved, may or may not be obviously related to the works of the seed. If two works are frequently co-cited (which will also add to the co-citation counts of their authors), RT suggests an explanation: they produce positive cognitive effects when taken together, and the connection between them is not difficult to grasp. I argue in White (2011) that this is why citers have frequently co-cited them in the first place. For example, if "Diana Crane" as a seed retrieves "Derek de Solla Price" as a high-ranked co-citee, it is in part because both wrote on the topic "Invisible Colleges," and citers have referred to those works jointly in later articles. Someone seeing Price in the Crane distribution may already know that their paired names imply this topic, or may learn it by exploring articles that co-cite their writings.

Pennants are designed to prompt cxplorations of the latter sort. However, as two-dimensional figures, they position authors not by a single TF*IDF weight but by TF and IDF values separately. The raw data from AHCI are:

TF: The seed author's co-citation count with each of the other 24 authors (and with himself or herself).
DF: The total citation count for each of the 25 in the database. (In the seed's case, TF and DF are equal).
N: The total number of document records in the database.

These values are converted as in Manning et al. and plotted on two logarithmic axes (unlabeled in AuthorWeb) to form the pennant. The seed author's TF and IDF values are always greatest, which places the seed at or near the pennant's tip. Document length is not an issue in this kind of retrieval, because the database itself constitutes one big document from which occurrence and co-occurrence counts (DF and TF counts) are taken. Therefore the TF*IDF formula is not expanded to normalize for document length.

Pennant structure and the Mann pennant

The axes of a pennant lend themselves to fairly consistent interpretations that can be illustrated with Fig. 3, the pennant for Thomas Mann.

Horizontal axis

The log TF values on this axis predict cognitive effects, from lower at left to higher at right. The higher an author's co-citation count with the seed, the more the author is pulled toward the seed on it. The empirical citation record thus predicts that, when the more rightward authors are read with the seed, the cognitive effects on the reader will be greater. In Fig. 3, for example, works by Wysling, Adorno, Freud, Benjamin, Nietzsche, and Goethe are predicted to have greater cognitive effects than, say, Foucault's when read in connection with Mann. To see what works these are, the AuthorWeb user must move from (1) authors on the pennant to (2) articles that cite those authors, and then to (3) works cited in those articles. The predictions may of course be wrong for a given user, but they are based on evidence from citers, and they will often be intelligible to domain experts.

RT's main cognitive effects can be illustrated here. Before seeing the Mann pennant, a doctoral student named Mary may assume that Goethe will appear on it, because both he and Mann based major works on the Faust legend. Input from the pennant will *strengthen* this assumption, and the retrieval of articles in which Goethe's *Faust* is indeed co-cited with Mann's *Doctor Faustus* will strengthen it further. Or Mary may assume that Heinrich von Kleist will appear on it, because she has a paperback of Kleist's stories for which Mann wrote an admiring introduction. The pennant will *eliminate* this notion; Kleist does not make the cut among Mann's top 24 co-citees. Or Mary may assume that the playwright Bertolt Brecht will not appear on the pennant, since he and Mann seem to her dissimilar writers. But since Brecht does appear on it, she *draws a new conclusion*—that scholars may link them because both were emigrés from Hitler's Germany who lived in Los Angeles during the 1940s. In all three cases, new information from the pennant readily interacted with Mary's existing assumptions to produce non-negligible cognitive effects, and thus the pennant was relevant to her. Furthermore, pennants are *collections* of utterances, any one of which may be tested for relevance by a user. If the Mann-Goethe link in the pennant interests Mary more than the others, then it, and what she can learn by pursuing it, will be more relevant to her than the others. Relevance in RT is a matter of degree, just as in information science (Saracevic 2007: 2133).

When looking at any AuthorWeb map, decoding symbols is part of Mary's cognitive task: for example, she must be able to decode the surnames plus initials in Roman letters, and the links (edges) or proximities that code relationships among them. But she could have carried out these decoding tasks yet still found a pennant irrelevant, because it produced in her negligible cognitive effects (Harter 1992). This would happen, for instance, if she did not know or was wholly indifferent to the authors in the map. It is only because she can go beyond decoding their names to the more important step—drawing useful inferences about them within her own cognitive contexts—that the pennant attains relevance for her. (S&W originally called cognitive effects "contextual effects.") In other words, Mary's knowledge in a particular area of Mann studies has been improved, at least provisionally. If further inputs bear the improvement out, she experiences positive cognitive effects.

In the "European" pennants to come, I myself do not recognize a fair number of authors, and so they are relevant to me only as content-free examples. I lack accessible

cognitive contexts—domain knowledge—in which these authors' names can produce noteworthy cognitive effects. I might look them up in Wikipedia, but that would increase the effort of processing them, a prospect that further lessens their relevance. Of course, if I do make the effort of looking them up, I might be compensated by finding them newly relevant to me, because they activate interests hitherto latent. This mutability illustrates why information scientists frequently claim that relevance is "dynamic" (Saracevic 2007: 2128–2129). What actually changes is the cognitive context in which I now regard a previously unknown author. The change occurred as new information—the Wikipedia article—interacted with my existing assumptions to produce effects of the three kinds listed above. A good many information scientists would probably agree with S&W that experiencing relevance in this sense is the same as becoming informed (cf. Harter 1992; Furner 2004; Saracevic 2007; White 2007a, b).

Vertical axis

The log IDF values on this axis predict the ease of processing a co-citee in relation to the seed. To repeat, log IDF values are computed from an author's *total citation count* in the database. (Log TF values, again, are computed from an author's *co-citation count with the seed*.) Because log IDF is an inverse scale, the *lower* an author's total citation count, the *higher* that author is placed on the vertical axis. Moreover, as a general feature of co-cited author pennants, higher authors are predicted to be *easier* to relate to the seed than lower authors. This result may seem puzzling and requires further explanation.

The total citation counts of co-citees indicate their fame. As an inverse scale, log IDF pushes a seed's *least famous co-citees* to the top of the pennant and *the most famous co-citees* to the bottom. The least famous co-citees—those with relatively few citations—are often scholars who specialize in the seed's works and whose names thus have relatively narrow implications (like Börner's high-ranked descriptors). The most famous co-citees—those with many citations—are often artists or thinkers who do not write about the seed's works at all and whose names have implications far broader than those works (like Börner's low-ranked descriptors). These famous authors are co-cited with the seed because scholars make non-obvious connections between the seed's works and theirs. AuthorWeb pennants are divided into three sectors to highlight this structure, which exemplifies Sparck Jones's (1972) "statistical specificity" in a new way.[4]

The key assumption behind IDF or "statistical specificity" is that, in bibliographic databases, relatively rare terms tend to specify a document's content better than relatively common ones. Therefore, documents with terms that match specific terms from a query are ranked higher by IDF in retrieval displays. The ranking predicts that these documents match the sense of the query better than others, and that their retrieval-worthiness is evident. This also happens when IDF ranks co-cited authors' names. That is, when items from top-ranked oeuvres are retrieved in the form of full bibliographic references, they refer to the seed's own name or works at the global level of titles.

Examples from Fig. 3 include Koopmann's *Thomas-Mann-Handbuch* (1990), Vaget's *Thomas Mann Kommentar zu sämtlichen Erzählungen* [*Thomas Mann: Comments on the Complete Stories*] (1984), and a chapter by Wysling that is co-cited with eight of Mann's non-fiction works in Lepenies and Harshav (1988):

[4] In White (2007a, b, 2009, 2010a) I divided the pennants into sectors on the basis of my own qualitative judgments. AuthorWeb pennants are simply divided into thirds mechanically, and so their qualitative implications are even more approximate.

Wysling, Hans (1967). "Geist und Kunst": Thomas Manns Notizen zu einem "Literatur-Essay." Thomas-Mann-Studien 1: 123–233. ["Intellect and Art": Thomas Mann's Notes to a "Literature Essay." Thomas Mann Studies 1: 123–233.]

Anyone who sees such examples can stop processing them immediately; their relevance to the seed is clear. Besides textual ties of this sort, authors in the top sector often have *social* ties with the seed that informed users can recognize. For instance, at top left in Fig. 3 is Mann's brother Heinrich, with whom he voluminously corresponded. Social ties are less common in the lower sectors (White 2007a, 2009, 2010a).

The vertical axis predicts that authors in the midsector will be harder to associate with Mann than the authors above them, but easier to associate than the authors below them, because their breath of implication is somewhat less. Brecht, Kafka, and Wagner are roughly Mann's peers in terms of total citation counts, and like him they are artists, implying the comparability of their creative works with his. (Comparable artists in the top sector—Heinrich Mann, Hofmannsthal, and Musil—have lower citation counts in AHCI.) Also opposite Mann one sees the thinkers Ernst Bloch and Georg Lukács, who have strong ties to imaginative literature and, along with Brecht, Adorno, and Walter Benjamin, a Marxist orientation. Mann was a public intellectual as well as a novelist, and many of his cited works are essays (e.g., literary criticism or discussions of politics) rather than fiction. Hence, he has a natural place in the world of ideas, as represented by the philosophers and theorists in the pennant.

The bottom sector displays authors whose citation counts in AHCI exceed Mann's. Far from specializing in his work, these luminaries may not even be from his period (e.g., Goethe, Kant, Hegel), and their writings, while relevant to the German novelist's, have vastly wider implications than "Mann studies." Unlike, say, Koopmann's *Thomas-Mann-Handbuch*, their titles tend not to evoke his work directly, which increases the effort of processing them for topical relevance. (Exceptions do occur, and some will be seen here, such as Goethe's explicit link with Mann through the Faust legend.)

Whether full titles are easy or hard to relate to the seed once they have been obtained, it costs effort even to get them because of AHCI's severe abbreviations (e.g., Freud-S, 1921, Massenpsychologie Ic,). Only when bibliographic references have been spelled out in full (and perhaps translated) can one tell what is really going on with the ease-of-processing scale, and that requires digging outside AHCI.

Why TF*IDF has been popular

The differences in pennant sectors occur through blind IDF weighting of author's names. Authors whose works can be related to the seed's at a glance are supposedly more relevant. This belief explains why TF*IDF ranking has long been popular with designers of information retrieval systems. They tend to work by the principle, *Give people what they say they want and make it obvious*. So a designer reasons: "Your say your interest is Thomas Mann. Is this document relevant to Mann-T?"

Koopmann-H, 1990, T Mann Hdb

Since the *Thomas-Mann-Handbuch* is easy to relate to the seed, designers want it to be among the first items the user looks at. Users will presumably be pleased to see whatever lets them stop processing more documents. The same algorithm puts Koopmann high in the pennant.

Designers like IDF because it has a history of improving their scores in evaluation trials. But one can also understand why Harter (1992) adopted RT to use against them. He had come to reject the notion, pervasive in document retrieval experiments, that relevance means simply *on topic*. Following RT, he wanted a more general meaning—*productive of cognitive effects*. He knew from experience that effects go well beyond inferring whether a document matches a query in sense; for instance, they can include creative leaps in which a new document combines with existing assumptions to yield a wholly new conclusion (as we saw with Mary). Harter's argument is persuasive: people routinely find uses for documents that do not match their stated topics of interest. (They also decline to use, for one reason or another, documents that *do* match their topics.) Co-cited author pennants subsume ties of *both* kinds—those based on exact or near term-matches and those requiring inferences in which term-matching is absent.[5]

Suppose, then, a designer asks: "Is this document relevant to Mann-T?"

Goethe-JWV, Null, Wilhelm Meisters Wan,

The string refers to an undated edition of Goethe's *Wilhelm Meisters Wanderjahre* [*Wilhelm Meister's Journeyman Years*], which Dowden (1994) co-cites with another novel, Mann's *Der Zauberberg* [*The Magic Mountain*]. By pushing Goethe down in the pennant, IDF also pushes down his Meister novel, whose relation to Mann's is not immediately apparent. Yet it is highly doubtful that humanities researchers would be disturbed by this. What probably *would* disturb them is a claim that an academic like Koopmann is more relevant to Mann than a titan like Goethe. That is not claimed here. The relevance predicted by IDF on the pennant's vertical axis is very superficial. It speaks to the immediate evaluation of documents by topic, rather than to hard-won literary interpretation. Someone who has made the effort to read works by both Goethe and Mann may well have experienced cognitive effects that are correspondingly great, and, to that person, Goethe's relevance to Mann is much greater than, say, Koopmann's. And indeed the scholarly consensus pulls Goethe closer to Mann on the horizontal axis than any other author. His position in the bottom sector simply predicts that relatively few of his associations with Mann will be self-evident.

To sum up, if citers repeatedly co-cite authors, and more particularly certain works, there are probably thematic reasons for it. (For instance, Goethe's two Wilhelm Meister novels and Mann's *The Magic Mountain* might be studied together as bildungsromans.) Bibliometric visualizations can at best only hint at these reasons; like bibliographies and library catalogs, they are ancillary to reading, not substitutes for it. But implicit connections among writings are humanists' stock in trade, and so even names in the bottom sector of a pennant may be easy enough for them to associate with the seed. They might know, for example, why Mann and Nietzsche have been frequently co-cited in past scholarship, or be able to guess why correctly.

A small case in point (White 2002): on first seeing an AuthorWeb map for the British novelist Kingsley Amis, I guessed that he was co-cited with George Orwell and Ray

[5] Harter et al. (1993) sampled pairs of *citing-cited* documents (not co-cited documents) and analyzed their subject indexing for presumed semantic closeness. They found that such closeness could not be taken for granted, because the descriptors assigned to the pairs rarely matched exactly. But perfect descriptor match is a highly restrictive standard. It misses partial matches, leaves out terms from titles and abstracts altogether, and ignores semantic ties that occur in body text, rather than at the level of the entire works. More generally, it discounts the human ability to infer semantic relations that fall outside exact term-matching. However, since Harter (1992) had already advocated Sperber and Wilson's subjective approach to relevance, the inadequacy of an "objective" descriptor-based approach seems to be the real point of the 1993 article.

Bradbury because of his study of science fiction dystopias, *New Maps of Hell*, and that, in contrast, he was co-cited with John Wain, John Osborne, and John Braine, his fellow "angry young men" of the 1950s, because of his novel *Lucky Jim*. Both inferences proved correct.

Processing effort in information science

We can transfer this line of reasoning to information science in a thought experiment. Imagine you are asked to guess why certain indexing terms might frequently co-occur in the literature—which is also why they might be productively ANDed together in an online search. Your first pair is Children AND Handguns. Chances are, you could quickly and easily give an explanation that retrieved documents would prove correct. Your second pair is Cotton Batting AND Water. Here, because these terms do not jointly suggest a subject matter to most people, it would presumably be much harder to explain why they might co-occur (cf. White 2002). They do retrieve documents in at least one database, but the retrievals are incoherent.

Now extend this framework to the author-pairs of a pennant diagram. Suppose you have considerable domain knowledge in information science and are interested in the work of Marcia J. Bates. Given her name as seed, you are asked to say why other names might be co-cited with her. This amounts to predicting the broad subject matter of documents, as yet unseen, that Bates and a co-citee would retrieve as an ANDed pair. Your first input is Nicholas J. Belkin. The Bates-Belkin pair would seem easy to process because, singly and jointly, the two names suggest topical areas such as information-seeking behavior or interactive document retrieval. In AuthorWeb you could test your inference by retrieving documents that co-cite this pair and judging them for relevance to your prediction—which might also be your own topical interest.

Your next input is Wallace Stevens. Taking this to be the American modernist poet, it is difficult if not impossible to say why his work and Marcia Bates's would be frequently co-cited. In the absence of a clearly overarching topic, joining his name with hers increases processing effort without any compensating cognitive effects, just as happened when the statement about rain in Aberdeen in 1864 was joined to the statement about rain in the here and now. The difficulty cannot be measured absolutely, but we can intuit that the pair Bates-MJ AND Stevens-W is much harder to process than the pair Bates-MJ AND Belkin-NJ. Assumptions about Bates as an initial context are quickly strengthened by seeing Belkin as new information; expectations of relevance are satisfied, and one *stops processing*. With Stevens, no effect occurs, expectations are not satisfied, and one merely gives up.

This brings us to Fig. 4, the pennant that AuthorWeb actually produces with Bates-MJ as seed. (AuthorWeb can map information scientists who are cited in journals covered by AHCI.) Belkin appears in the top sector, and so do other information scientists such as Salton-G and Matthews-JR. Pulled closer to Bates are the information scientists Fidel-R, Fenichel-C, Saracevic-T, and Borgman-CL, predicting greater cognitive effects for anyone who reads their works jointly with hers. The prediction makes sense, because both she and they write in relatively qualitative ways about behavioral aspects of online information retrieval systems.

Problematically, however, the name closest of all to Bates-MJ is Stevens-W in the bottom sector. Is this Wallace Stevens the poet? It must be, because other famous poets—Eliot-TS and Williams-WC—are also there. Our doctoral student Mary would recognize

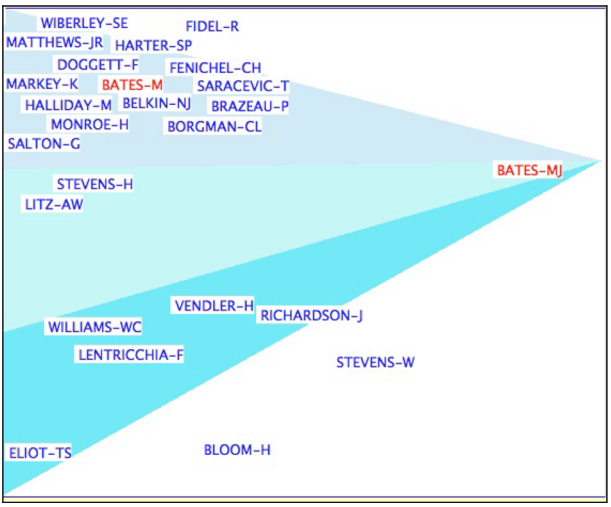

Fig. 4 Pennant diagram for Bates-MJ

some literary critics as well—Bloom-H, Vendler-H, Richardson-J, Lentricchia-F. It is very implausible that information scientists would repeatedly cite such names with Bates's. So why do they appear? As might be guessed, it is because of the software's inability to disambiguate—more precisely, to infer what *you* meant by Bates-MJ. AuthorWeb has conflated Marcia J. Bates's co-citees with those of Milton J. Bates, a prominent writer on Wallace Stevens.[6] Consequently, 13 names on the map belong to Wallace Stevens studies. Table 1 lists the respective camps. Mix-ups of this sort are by no means the rule in AuthorWeb, but, given how AHCI abbreviates authors' names, they do happen.

In a real experiment, a researcher could present subjects with "Bates-MJ" and various co-citees from Table 1. Almost surely information scientists would classify the "Marcia" pairs faster and more accurately than the "Milton" pairs, and with greater consistency across judges. For specialists in modern American poetry, who bring very different cognitive contexts (i.e., domain knowledge) to the task, it would be just the opposite.

RT naturally makes disambiguation of terms a key part of understanding utterances. Depending on which Bates is meant, all the co-citees in the *other* column of Table 1 would be failures of understanding on the system's part. For users interested in Marcia Bates, entire oeuvres under Milton Bates are false positives: retrieved but not relevant. The RT ratio tells us why: negligible cognitive effects/high processing effort. The rejection of any document on grounds of insufficient relevance can be explained in this way, which sharpens our understanding of *precision* as a measure in document retrieval.[7]

[6] The form Bates-M appears because the same article sometimes cites Marcia or Milton both with and without their middle initial.

[7] T. S. Eliot is not *completely* irrelevant to Marcia Bates's work. Information scientists occasionally quote his lines in *Choruses from "The Rock"*: "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?" Since a few of them have also cited her in the same article, her co-citation count with Eliot is not zero (cf. Bates 2010).

Table 1 Two M. J. Bateses and their co-citees

Marcia J. Bates	Milton J. Bates
Gerard Salton	Wallace Stevens
Tefko Saracevic	T. S. Eliot
Nicholas J. Belkin	William Carlos Williams
Christine L. Borgman	Harold Bloom
Karen Markey	Frank Lentricchia
Stephen P. Harter	Helen Vendler
John Richardson	Joan Richardson
Raya Fidel	Holly Stevens
Carol H. Fenichel	Harriet Monroe
Joseph R. Matthews	Peter Brazeau
Stephen E. Wiberley	Arthur Walton Litz
	Mark Halliday
	Frank Doggett

Pennants for four European authors

Four European seed authors are presented in Figs. 5, 6, 7 and 8. Two are contemporary: *Mieke Bal*, a Dutch literary and cultural theorist at the University of Amsterdam, and *Anthony Milton*, a British historian of seventeenth century England at the University of Sheffield. Two are historical: *Ludvig Holberg*, an eighteenth century Danish-Norwegian playwright, and *Karl Kraus*, an Austrian satirist and journalist of the late nineteenth and early twentieth centuries. The comments in this section are conjectures about why these seeds and certain other authors were co-cited—thematic predictions, as it were. While not based on extensive analyses of data, they could be tested by actual retrievals, along lines sketched in the Kingsley Amis example above.

Mieke Bal (suggested by Rens Bod, University of Amsterdam). A feature of Bal's pennant is that, unlike Mann's, it consists almost entirely of scholars and theorists, as opposed to artists, and their titles tend to concentrate subject matter better than artistic

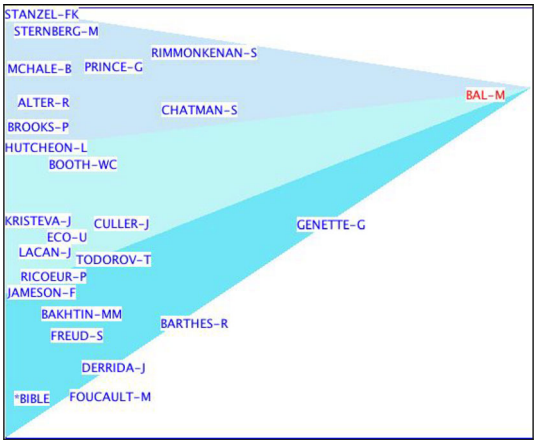

Fig. 5 Pennant diagram for Mieke Bal

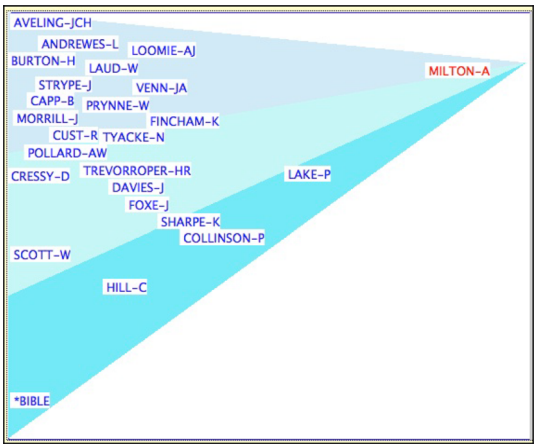

Fig. 6 Pennant diagram for Anthony Milton

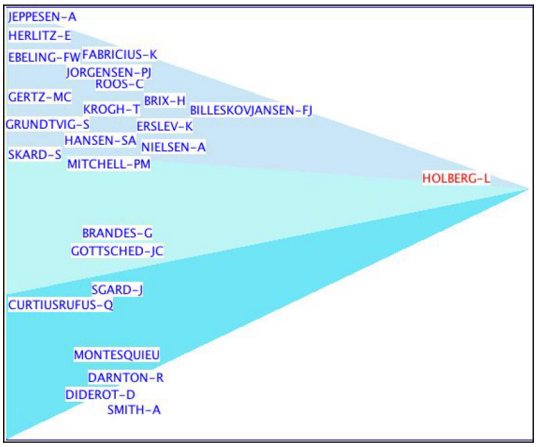

Fig. 7 Pennant diagram for Ludvig Holberg

titles. She herself has written about many artists, but none are among her top co-citees here. As a seed, she is pulled upward in Fig. 5 because she is less cited in AHCI than the powerhouse names in the middle and lower sectors. (This imbalance also appears in other pennants here.) Her co-citees exhibit the familiar transition from relatively obscure at top to very famous at bottom. During the period AuthorWeb covers, 1988 through 1997, her fields of literary theory and cultural studies were dominated by the French post-structuralists Michel Foucault, Jacques Derrida, and Roland Barthes. They are in fact godlike in AHCI, and so their presence in the bottom cluster, along with that of Freud and other renowned theorists, is unsurprising.

Bal has published feminist readings of Biblical stories, which partially explains why she is co-cited with the Bible (and with Robert Alter in the top sector). But the major theme underlying her pennant—the one that explains why certain names are pulled toward hers on the horizontal axis—is narratology, the analysis of properties of narratives. One of her co-citees, Gerald Prince, wrote *A Dictionary of Narratology*, and he could interpret much of the map simply by pointing to his bibliography, where writings by Bal and most of the

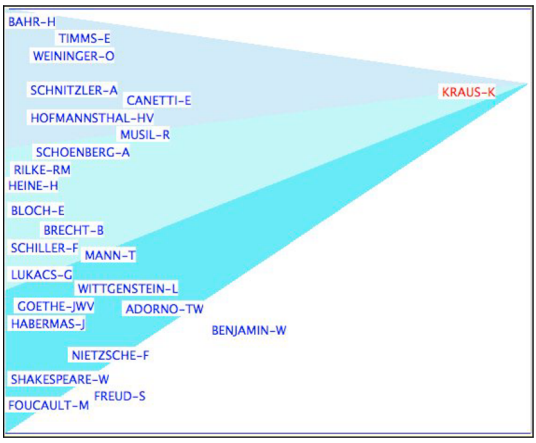

Fig. 8 Pennant diagram for Karl Kraus

other mapped authors appear. Tzvetan Todorov named the field. Bal herself published *Narratologie* in French in 1977, and her most frequent co-citee is the famous French theorist Gérard Genette (one of whose books is, in translation, *Narrative Discourse: An Essay in Method*). Other authors pulled rightward toward Bal are also narratologists—Shlomith Rimmon-Kenan (*Narrative Fiction: Contemporary Poetics*), Seymour Chatman (*Story and Discourse: Narrative Structure in Fiction and Film*) and Roland Barthes ("An introduction to the structural analysis of narrative").

Anthony Milton (suggested by John Rigby, University of Manchester). In general, Rigby (2013, personal communication) thought AuthorWeb's bibliometric methods "look promising." He writes: "[Milton's] contributions are I think to the second reformation in England, as he calls it, in which the English church underwent further significant development in doctrine and organisation during the seventeenth century beyond what had occurred during the era of Hooker…" Milton's Web page indeed gives his specialties as "Early Modern England, 17th c. Anglo-Dutch relations; royalism; Church of England 1603–1700." The author drawn closest to him in Fig. 6 is Peter Lake, who is likewise a specialist in Tudor and Stuart England and also a co-editor of books to which Milton has contributed chapters. Rigby's gloss on Christopher Hill (whose citation count—i.e., fame—in the pennant is second only to the Bible's) is that he "provided a great provocation to at least two generations of scholars keen to respond to what was termed his 'lumping' tendency, which served his Marxist interpretation. Hill was a very successful historian producing a great deal of work and an extensive synthesis. It is difficult to ignore him perhaps, even if his work has been extensively and energetically responded to, qualified, and of course much rebutted."

Perhaps confusingly, Milton's pennant merges contemporary historians (e.g., John Morrill, Nicholas Tyacke, Hugh R. Trevor-Roper) with historical personages (Lancelot Andrewes, Henry Burton, John Foxe, William Laud, William Prynne). This occurs in AuthorWeb when authors from distinct eras have similar citation counts. As it happens, Milton's self-organizing map (SOM), which is not shown, is better at separating present-day writers from their centuries-old sources—an argument for systems that can generate different kinds of maps from the same seed.

Ludvig Holberg (suggested by Gunnar Sivertsen, Nordic Institute for Studies in Innovation, Research and Education). The co-citation counts underlying Fig. 7 are low, ranging

from one to three for this Scandinavian playwright and man of letters. Sivertsen (2013, personal communication) expresses the usual reservations about AHCI: its exclusion of citations from books and numerous journals, and its meager coverage of German scholarship on Holberg, which exceeds that of scholars writing in English. Even so, the AuthorWeb maps made sense to him. Below, his notes on Holberg's co-citees show an expert doing by return e-mail what would take another scholar much longer. Ability to find relevance, in other words, is greatly affected by differences in people's cognitive contexts. The cognitive effects in this case are Sivertsen's inferences, given verbatim, about why the bulleted authors appear with Holberg:

> F. J. Billeskov-Jansen, professor at Copenhagen University, was the most influential Holberg scholar in Denmark (and the world) throughout six decades 1940–2000. Most of his publications were in Danish, but he wrote a monograph on Holberg in Twayne's World Author Series as well as the article about Holberg in *Encyclopedia Britannica*. In 1946, he founded the international journal *Orbis Litterarum*, one of the few journals in the humanities from Scandinavia that has been covered by AHCI for a long time.
> Montesquieu and Diderot: Holberg is often compared to the two as the most important Enlightenment figure in Scandinavia, but Holberg also wrote a critical essay on *L'Ésprit des Lois* upon its publication and had the essay translated into French and German.
> Gottsched introduced Holberg's comedies to Germany in the 1740s and had them translated and staged there. He also included a discussion of Holberg in his works on drama theory.
> Mitchell was the translator of Holberg's essays into English in the twentieth century.
> Brix, Krogh and Roos are other important Holberg scholars in Denmark in the twentieth century.
> Brandes is the well known Danish critic who also wrote about Strindberg and Ibsen. He published a monograph on Holberg in 1884.
> Skard (Sigmund) was a Norwegian professor of American literature who also studied classical literature (and in this connection studied Holberg). But his main field of research explains why he is visible here.
> Grundtvig is the well known Danish nineteenth century poet, probably connected to Holberg here because of their general importance in Danish literature.
> Holberg was also a historian; that is perhaps why Quintus Curtius Rufus appears. But he had stronger influences from other classical writers.

Karl Kraus (my selection). The satirist, essayist, and playwright Kraus (Fig. 8), longtime editor and then sole producer of the journal *Die Fackel* (*The Torch*), has recently been re-introduced to the English-speaking world by the American novelist Jonathan Franzen (2013). However, I chose him because, as a Viennese author, he seemed appropriate to map for a presentation on AuthorWeb in Vienna (GESIS 2013). Kraus's pennant comprises mostly artists and philosophers; the one scholar at top is his biographer and critic Edward Timms, whose works of course name him in their titles. In the nicely evocative top sector, citers have connected Kraus to figures prominent in Viennese intellectual life during his lifetime (1874–1936). The sector implicitly picks up his social ties with them (which were often hostile). The lower sectors connect him to even more famous names in the German-speaking world, including the Viennese giants Wittgenstein and Freud, also his approximate contemporaries. According to Wikipedia, Kraus used to read from Shakespeare,

Goethe, and Brecht in his very popular public lectures, but that, of course, is not necessarily why they are co-cited with him.

The co-citee whom the pennant predicts to have the greatest effects when read with Kraus is Walter Benjamin. Since he is another theorist with towering stature among citers during 1988–97, he is in the bottom sector, implying non-obvious connections to the seed. It is therefore noteworthy that in 1931 he published a long essay entitled *Karl Kraus*. On examination, however, one finds that this essay is by no means the only work of his that citers link to Kraus, and these other works fall in the non-obvious category. The same is true of Goethe and Mann, who are linked not only through explicit matches on Faust; and of Genette and Bal, who are linked not only through explicit matches on narratology.

Ways forward

Considering both the strengths and weaknesses of AuthorWeb, how might co-cited author mapping and retrieval be developed in the future? I will offer a few ideas, still somewhat utopian.

The relevance of documents to a seed can be tested at either global or local levels. The *global* level is that of surrogates of entire works—bibliographic representations such as titles, abstracts, and subject indexing. One seeks relevance at this level by inferring how well these representations respond to the seed-query, which in turn usually represents one's topical interests. However, AHCI does not return full surrogates of co-cited documents; it identifies them only by the strings of abbreviated noun phrases seen in AuthorWeb. Some of these strings will be meaningful to users with the right domain knowledge, but no user can begin to interpret all of them. Retrieval systems of the future need to provide surrogates that fully identify *cited* works (not just *citing* works). Gradually, this form of bibliographic control might also be extended to important publications that citation databases do not now cover.

To determine relevance at the *local* level, one reads the sentences in which citations are embedded in body text. Services such as Google Scholar, Google Books, and CiteSeer already retrieve citation-bearing sentences, although these "citances" (Nakov et al. 2004) may not be easy to compare. Capability for comparing them seems likely to grow, since users will often find them more helpful than title phrases and other global indexing for inferring degrees of relevance. Advancing this capability, Jeong et al. (2014) have integrated "citance" content into author co-citation analysis and have shown how such content enriches maps of co-cited authors. The addition of content also produces author maps that differ in important respects from those based solely on co-citation counts, like AuthorWeb's.

Another key piece of *local* information is the proximity of citations to any two works within the same document. Bibliometricians have begun to analyze not only who is frequently co-cited with whom, but the width of the textual window in which the co-citations occur (Hu et al. 2013; Jeong et al. 2014: 199). In a previous example, we saw that Fehn (1988) co-cites Freud's *Group Psychology and the Analysis of the Ego* with Mann's *Mario and the Magician* (titles translated). Without knowing Fehn's article, few readers could guess how—or even whether—it relates these two works. In other words, users intrigued by the two titles cannot simply stop processing; they must move to passages of body text. And when that is done (costing further effort), the relevance of Freud's monograph to Mann's novella is not assured. Fehn may cite the two works in such different contexts that little or no connection between them can be inferred (nullifying effect).

RT provides a language for discussing this outcome, as it also does when items co-cited with the seed are perceived as highly relevant to it. For example, Mieke Bal's pennant indicated the importance of Genette and Rimmon-Kenan to her research. Note how a one-sentence window from a book not covered by AuthorWeb (Malina 2002: 145, endnote 2) *strengthens* the impression left by the pennant (or possibly leads to *new conclusions*):

> 2. The plot level is what Mieke Bal calls "fabula" (as vs. "story" and "text") and what Shlomith Rimmon-Kenan calls "story" (as vs. "text" and "narration"), a schema that corresponds to Genette's "*histoire*," "*récit*," and "*narration*" (see Bal, *Narratology*, 5–8; Rimmon-Kenan, *Narrative*, 3).

We thus need systems that quickly display citation-bearing sentences from the same source (or different sources), so that users can compare their implications. Authors cited in the same sentence or paragraph would generally seem to be ideationally closer than authors cited far apart.

Within scholarly literatures, knowledge claims are made in sentences, and an important part of research on document or passage retrieval is attention to sentences that contain citations (Ritchie 2008; Ritchie et al. 2008). Presumably humanities scholars want tools that let them quickly retrieve sentences of this sort, and those tools are emerging. But sentences and passages cannot represent whole specialties in the humanities; they are too fine-grained. AuthorWeb, in contrast, depicts specialties in ways not available elsewhere. It anticipates systems in which broad overviews of specialty literatures are the first thing a user sees. Through these overviews, particular works of the literatures, and then individual passages in them, could be reached.

One approach to overviews of scholarly literatures is to map them topically as groups of linked subject headings. This sensible but dull idea addresses writings only at the global level and depends on the perceptions of indexers rather than scholars. The idea of representing literatures through their co-cited authors is still virtually unknown in the humanities, but it can lead to retrievals of both works and passages. It is also based on the behavior of scholars who create the literatures, and this arguably makes for much more absorbing maps than ordinary subject indexing. Co-cited author pennants would seem to be maps of a particularly interesting kind, not least because of their grounding in Sperber and Wilson's relevance theory. However, it remains an open question how well they fit with traditional patterns of humanities scholarship.

Acknowledgments GESIS, the Leibniz Institute for the Social Sciences in Cologne, Germany, generously supported preliminary work on this article during summer 2013. I am grateful to several GESIS colleagues for stimulating discussions: Andreas Strotmann, Philipp Mayr, Philipp Schaer, and Maria Zens. Ideas in the article were also presented in a talk at the University of Amsterdam, and I thank Alesia Zuccala (of the University's Center for Digital Humanities) and Andrea Scharnhorst (Royal Netherlands Academy of Arts and Sciences) for the invitation. Rens Bod, John Rigby, and Gunnar Sivertsen kindly suggested seed authors for me to map. I asked the latter two for reactions, and they provided them. I am pleased to quote from them here. I also thank my anonymous referees for instructive comments.

References

Bar-Ilan, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing and Management, 42*, 1553–1566.

Bates, M. J. (2010). *Information. Encyclopedia of library and information sciences* (3rd ed., pp. 2347–2360). New York: CRC Press.

Buzydlowski, J. (2002). *A comparison of self-organizing maps and pathfinder networks for the mapping of co-cited authors*. PhD dissertation. Drexel University, Philadelphia, PA. http://faculty.cis.drexel.edu/~jbuzydlo/papers/thesis.pdf.

Buzydlowski, J. W., White, H. D., & Lin, X. (2003). Term co-occurrence analysis as an interface for digital libraries. *Lecture Notes in Computer Science, 2539*, 133–144. http://web3.holyfamily.edu/jbuz/papers/lncs.pdf.

Carston, R., & Powell, G. (2008). *Relevance theory: New directions and developments. The Oxford handbook of philosophy of language* (pp. 341–360). Oxford: Oxford University Press. [Also in Oxford handbooks online and at http://www.phon.ucl.ac.uk/home/robyn/Carston-Powell-PhilHandbook-28July05%5B2%5D.pdf.

Clark, B. (2013). *Relevance theory*. Cambridge: Cambridge University Press.

Cronin, B., & Shaw, D. (2001). Identity-creators and image-makers: Using citation analysis and thick description to put authors in their place. *Proceedings of the 8th international conference on scientometrics & informetrics*, Vol. 1, 127–138.

Dowden, S. (1994). Irony and ethical autonomy in Wilhelm Meisters Wanderjahre. *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte, 68*, 134–154.

Eisenberg, M. B. (1988). Measuring relevance judgments. *Information Processing and Management, 24*, 373–389.

Fehn, A. C. (1988). Concepts of the masses and German drama in the Weimar Republic. *Seminar: A Journal of Germanic Studies, 24*, 31–57.

Franzen, J. (2013). *The Kraus project: Essays by Karl Kraus*. New York: Farrar, Straus and Giroux.

Furner, J. (2004). Information studies without information. *Library Trends, 52*, 427–445.

GESIS. (2013). *Combining bibliometrics and information retrieval. Pre-conference workshop of the International Society for Scientometrics and Bibliometrics*. Vienna, Austria, 2013. http://www.gesis.org/en/events/conferences/issiworkshop2013/.

Goatly, 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.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

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

Harter, S. P., Nisonger, T. E., & Weng, A. (1993). Semantic relationships between cited and citing articles in library and information science journals. *Journal of the American Society for Information Science, 44*, 543–552.

Higashimori, I., & Wilson, D. 1996. Questions on relevance. *University College London Working Papers in Linguistics, 8*, 111–124. http://www.phon.ucl.ac.uk/home/PUB/WPL/96papers/higashi.pdf.

Hu, X., Rousseau, R., & Chen, J. (2012). Structural indicators in citation networks. *Scientometrics, 91*, 451–460.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics, 7*, 887–896.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics, 8*, 197–211.

Lepenies, W., & Harshav, B. (1988). Between social science and poetry in Germany. *Poetics Today, 9*, 117–143.

Lin, X., Bui, Y., & Zhang, D. (2007). Visualization of knowledge structure. *Proceedings of the 11th International Conference of Information Visualization (IV2007)* (pp. 476–481).

Lin, X., White, H. D., & Buzydlowski, J. (2003). Real-time author co-citation mapping for online searching. *Information Processing and Management, 39*, 689–706. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.4125&rep=rep1&type=pdf.

Malina, D. (2002). *Breaking the frame: Metalepsis and the construction of the subject*. Columbus, OH: Ohio State University Press. http://ohiostatepress.org/index.htm?/books/book%20pages/malina%20breaking.html.

Manning, C., Raghavan, P., & Schütze, H. (2008). *An introduction to information retrieval*. Cambridge: Cambridge University Press. Draft of 2009 ed.: http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf.

McCain, K. W. (2010). The view from Garfield's shoulders: Tri-citation mapping of Eugene Garfield's citation image over three successive decades. *Annals of Library and Information Studies, 57*, 261–270.

Nakov, P. I., Schwartz, A. S., & Hearst, M. A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. *Proceedings, SIGIR'04 workshop on search and discovery in bioinformatics*, Sheffield, UK, 2004. http://biotext.berkeley.edu/papers/citances-nlpbio04.pdf.

Rigby, J. (2013). E-mail on AuthorWeb maps for Anthony Milton.

Ritchie, A. (2008). *Citation context analysis for information retrieval*. Technical Report 744, Computer Laboratory, University of Cambridge. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-744.html.

Ritchie, A., Teufel, S., & Robertson, S. (2008). *Using terms from citations for IR: Some first results*. Paper presented at the European conference on information retrieval, Glasgow, UK, 2008. http://www.cl.cam.ac.uk/~sht25/papers/ecir2008_ritchie.pdf.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology, 58*, 2126–2144.

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, 2007. http://yunus.hacettepe.edu.tr/~tonta/courses/spring2011/bby704/pennant-diagrams-2c_Peter_Ingwersen.pdf.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Sivertsen, G. (2013). E-mail on AuthorWeb maps for Ludvig Holberg.

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* (2nd ed.). Oxford: Blackwell. [1st ed., 1986.]

Sperber, D., & Wilson, D. (1996). Fodor's frame problem and relevance theory (reply to Chiappe & Kukla). *Behavioral and Brain Sciences, 19*, 530–532. http://cogprints.org/2029/1/frame.htm.

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. (2002). *Cross-textual cohesion and coherence*. Paper presented at discourse architectures: Designing and visualizing computer mediated conversation, a workshop of CHI (Computer Human Interface), Minneapolis, MN, 2002. http://pliant.org/personal/Tom_Erickson/DA_White.pdf.

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. Implications for information science. *Journal of the American Society for Information Science and Technology, 58*, 583–605.

White, H. D. (2009). Pennants for Strindberg and Persson. In *Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th birthday*. Special Volume of the E-Newsletter, International Society for Scientometrics and Informetrics (pp. 71–83). http://www.issi-society.info/ollepersson60/ollepersson60.pdf.

White, H. D. (2010a). Some new tests of relevance theory in information science. *Scientometrics, 83*, 653–667.

White, H. D. (2010b). Relevance in theory. *Encyclopedia of library and information sciences* (3rd ed., pp. 4498–4511). New York: CRC Press.

White, H. D. (2011). Relevance theory and citations. *Journal of Pragmatics, 43*, 3345–3361.

White, H. D., Lin, X., & Buzydlowski, J. W. (2001). The endless gallery: Visualizing authors' citation images in the humanities. *Proceedings of the American Society of Information Science and Technology, 38*, 182–189.

White, H. D., Lin, X., Buzydlowski, J. W., & Chen, C. (2004). User-controlled mapping of significant literatures. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Supp. 1), 5297–5302.

White, H. D., & Mayr, P. (2013). *Pennants for descriptors*. Paper presented at Networked Knowledge Organization Systems (NKOS) workshop, Valetta, Malta, 2013. http://arxiv.org/abs/1310.3808.

Wilson, D. (2011). Relevance and the interpretation of literary works. *University College London Working Papers in Linguistics, 23*, 69–80. http://www.ucl.ac.uk/psychlangsci/research/linguistics/publications/wpl/11papers/Wilson2011.

Wilson, D., & Sperber, D. (2004). Relevance theory. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics*. Oxford, England: Blackwell.

Yus, F. (2011). *Cyberpragmatics: Internet-mediated communication in context*. Amsterdam: John Benjamins.

Zhu, W., Lin, X., Hu, X., & Sokhansanj, B. A. (2005). Visualization of protein-protein interaction network for knowledge discovery. *2005 IEEE: International Conference on Granular Computing*, *1*, 373–377.