Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II ☆

Tefko Saracevic
School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ, USA

"*Relevant:* having significant and demonstrable bearing on the matter at hand."

"*Relevance:* the ability (as of an information retrieval system) to retrieve material that satisfies the needs of the user."

Merriam Webster (2005)

"All is flux."

Plato on Knowledge in the *Theaetetus* (about 369 BC)

I. Prologue: How Parts I and II Are Connected Across Time and What This Work Is All About

In vol. 6, 1976, of *Advances in Librarianship*, I published a review about relevance under the same title, without, of course, "Part I" in the title (Saracevic, 1976). [A substantively similar article was published in the *Journal of the American Society for Information Science* (Saracevic, 1975)]. I did not plan then to have another related review 30 years later—but things happen. The 1976 work "attempted to trace the evolution of thinking on relevance, a key notion in information science, [and] to provide a framework

☆ A version of this review is to appear as an article subdivided in two parts in the *Journal of the American Society for Information Science and Technology.*

ADVANCES IN LIBRARIANSHIP, VOL. 30

ISSN: 0065-2830
DOI: 10.1016/S0065-2830(06)30001-3

within which the widely dissonant ideas on relevance might be interpreted and related to one another" (ibid.: 338).

Building on the examination of relevance in the preceding (1976) review, this (2006) work updates the travails of relevance in information science for the past 30 years or so. Relevance still remains a basic notion in information science, and particularly in information retrieval (IR). The aim of this work is still substantially the same: it is an attempt to trace the evolution of thinking on relevance in information science for the past three decades and to provide an updated, contemporary framework within which the still widely dissonant ideas on relevance might be interpreted and related to one another.

II. Introduction: How Information Technology Made the Study of Relevance Ever More Relevant

In human history, relevance has been around forever, or as long as humans tried to communicate and use information effectively. Computers have been around for the last 50 years or so, the Internet for some 25, the Web for about 15. In this short time, the contemporary information technology (IT), including information systems based on IT, changed or transformed a great many things in society—from education to health care, from earning a living to leisure, from physics to classics, from government to being governed, from being young to being old, … IT changed information activities dramatically, namely the way we acquire, organize, store, preserve, search, retrieve, communicate, interact with, and use information. In all of those information activities relevance plays a most significant, underlying and yet elusive role. Similarly, relevance plays a significant, underlying role when these activities are performed by information systems as well; for these systems are designed primarily to respond with information that is potentially relevant to people.

IT is not elusive; relevance is. IT is tangible; relevance is intangible. IT is relatively well understood formally; relevance is understood intuitively. IT has to be learned; relevance is tacit. IT has to be explained to people; relevance does not.

In his 1776 book *An Inquiry into the Nature and Causes of the Wealth of Nations* Adam Smith, regarded as the father of economics, set out the mechanism by which he believed economic society operated; among others, he explained market decisions as often being governed by an "invisible hand." In the same spirit, while the hand of relevance is invisible, it is governing. Somewhere, somehow, the invisible hand of relevance, under its own or other names, enters the picture in all information activities and all information

systems. As far as people are concerned, relevance is tacitly present and inescapable. Relevance is the reason people use IT in their information activities. Conversely, information systems are primarily designed to provide potentially relevant information or information objects to people. In this lies the significance of relevance.

Positioning people and IT together in this discussion is deliberate to point out basic premises, distinctions, problems, and conflicts. Information systems, through a complex set of operations based on ever-changing and improving algorithms, retrieve and offer their versions of what may be relevant. People go about their ways and assess their own version of relevance. Both treat relevance as a relation. But each may have different premises for what is involved in the relation and in assessing that relation. There are two interacting worlds: the IT world and the human world; and two basic categories of relevance: systems' and humans'. The two worlds interact with various degrees of problems and conflict, from none to a lot. Our concern here is primarily with the human world of relevance. Relevance is treated here as a human condition, which it is. While we can never get far from systems, this review does *not* cover how systems deal with relevance. Treatments of relevance in IR—in algorithms, measures, evaluation—are beyond the scope of this review.

In information science, as well as other related fields, the emergence and proliferation of IT provided an impetus for study of the notion of relevance, aimed at a better and more formal understanding of it. Lexical definitions are very important, but do not suffice; besides, we are not really able to resolve issues of relevance through definition alone (Froelich, 1994). Thus, as in all other scientific and scholarly endeavors, when faced with a basic phenomenon or notion, scholarly inquiry does not ask the naïve question: *What is relevance?* Instead, the basic question was and still is: What is the *nature* of relevance? Following are more precise questions:

- What are the manifestations of relevance?
- What is the behavior of relevance?
- What are the effects of relevance?

The organization of the present review follows this reasoning. It sets the stage with an *Introduction* and a *Historical Footnote*. Next are three sections addressing the nature of relevance: a general one synthesizing its meanings, following with more specific sections on theories and models of relevance. The review then continues with a section about various manifestations of relevance, and concludes with sections that deal with experimental and observational findings on behavior and effects of relevance. Each section ends with a summary that in effect provides an interpretation and synthesis of contemporary

thinking on the topic treated or suggests hypotheses for future research. Analyses of some of the major trends that shape relevance scholarship, together with suggestions for research agendas, are offered in the epilogue.

While knowledge for knowledge's sake in investigations of the notion of relevance is sufficient impetus, there is also pragmatic potential. The history of science and technology is full of instances where a better understanding of the basic notions or phenomena underlying a technology led to development of more effective and successful technologies and systems. A fruitful, though sometimes convoluted and arduous, translation was realized. Hopefully, a better understanding of relevance may lead to better information systems. This clearly illustrates the significance of relevance research. Considering and understanding relevance as a notion is still relevant, if not even more so, to building and operating information systems—now ever more complex in the Web environment—that effectively provide information to users pertaining to their problems at hand.

III. Historical Footnote: A Reminder of How Relevance Came into Being in IR and Affected a Lot of Things

The term "information retrieval" (IR) was coined by mathematician and physicist Calvin N. Mooers (1919–1994), a computing and IR pioneer, just as the activity started to expand from its beginnings after World War II. He posited that IR:

> embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation (Mooers, 1951, p. 25).

Over the next half century, IR evolved and expanded, but basically, it continues to concentrate on the topics Mooers described.

The key difference between IR and related methods and systems that long preceded it, such as those developed in librarianship for bibliographic description and classification, is that IR specifically included "specification for search." The others did not. From Charles Ammi Cutter (1837–1903), who postulated bibliographic requirements at the end of the 19th century to the contemporary International Federation of Library Association and Institutions' (IFLA) report on *Functional Requirements for Bibliographic Records* (FRBR), the goal was to "provide a clearly defined, structured framework for relating the data that are recorded in bibliographic records to the needs of the users of those records" (IFLA, 1998: 2.1). User needs are defined in relation

to the following generic tasks that are performed by users when searching and making use of national bibliographies and library catalogues:

- using the data to **find** materials that correspond to the user's stated search criteria
- using the data retrieved to **identify** an entity
- using the data to **select** an entity that is appropriate to the user's needs
- using the data in order to acquire or **obtain** access to the entity described (emphasis in the original) (ibid.: 2.2).

In FRBR (and all the way back to Cutter), the process of search is not specified, it is assumed that it would happen. User needs, which should be fulfilled, were specified, but how the search should be performed was not. Data in bibliographic records were then organized to fulfill the specified needs. In IR, the user's needs are assumed as well, but the search process is specified in algorithmic details and data is organized to enable the search.

The fundamental notion used in bibliographic description and in all types of classifications, ontologies, or categorizations, including those used in contemporary databases, is *aboutness.* The fundamental notion used in IR is *relevance.* It is not about any kind of information but about *relevant* information. Fundamentally, bibliographic description and classification concentrate on describing and categorizing information objects; IR is also about that *but*, *and this is a very important* "*but*," in addition IR is about searching as well, and searching is about relevance. In the realm of computer science, differences between databases and IR were often discussed in terms of differences between structured and unstructured data, which is OK, but this fails to define the fundamental difference in the basic notion used: *aboutness* in the former and *relevance* in the latter. Therein lie both similarity and difference. Relevance entered as a basic notion through the specific concentration on searching. Budd (2004, p. 449) lamented that "the preponderance of writing on relevance comes from information science" and little or none can be found in librarianship. The explanation is simple: librarianship was concerned with aboutness and thus, it produced a considerable literature about aboutness and little or none about relevance. Conversely, information science was concerned about relevance and thus, it produced a considerable literature about relevance and little or none about aboutness.

In a sense, aboutness may be considered as topical relevance, which is one manifestation of relevance discussed later. However, topical relevance in IR is construed through indexing (or some other form of representation) to be directly searchable in specified ways—and, as pointed out, searching is about relevance.

By choosing relevance as a basic, underlying notion of IR, related information systems, services and activities—and with it, the whole field of information science—went in a direction that differed from approaches taken

in librarianship, documentation, and related information services, as well as expert systems and contemporary databases in computer science. Of course, this generalization, as all generalizations, simplifies the situation, but illustrates the effect of choices.

A number of suggestions were made to use uncertainty, instead of relevance, in IR but they never took hold. If they had, we would have very different IR systems today. For example: the basis of expert systems is uncertainty (or rather reduction of uncertainty based on if-then rules). As a result, expert systems are very different from IR ones. In comparison to IR, expert systems are not as widely adopted and used. The reason may be due to the choice of the underlying notion. Relevance is a human notion, widely understood in similar ways from one end of the globe to the other. Uncertainty is not. Besides, the assumption that information decreases uncertainty does not hold universally; information may also increase uncertainty.

Historically, relevance crept in unannounced. At the start of IR, more than a half century ago, nobody made a big point about it. IR systems were constructed to do relevance, but nobody talked about it. Still, principles posited then are valid today. It was, and still is, accepted that the main objective of IR systems is to retrieve information or information objects relevant to user queries, and possibly needs. Actually, the first discussions of relevance in the early 1950s were not about relevance, but about non-relevance or "false drops"—unwanted information retrieved by IR systems. The first full recognition of relevance as an underlying notion came in 1955 with a proposal to use "recall" and "relevance" (later, because of confusion, renamed "precision") as measures of retrieval effectiveness in which relevance was the underlying criterion for these measures (Kent *et al.*, 1955). Over time, many other measures were suggested, but did not take. Precision and recall remain standard measures of effectiveness to this day, with some variations on the theme. They measure the probability of agreement between what the system did or failed to retrieve or construct as relevant (systems relevance), and what the user assessed or derived as relevant (user relevance). Relevance became and remained the underlying criterion for measuring the effectiveness of IR.

There were, still are, and always will be many problems with relevance. This is not surprising. Relevance is a human—not a system—notion and human notions are messy. Oh well, they are human. Problems led to investigations on the nature of relevance in information science. Exposition of many views and a number of experiments followed. Those before 1975 were synthesized in Part I (Saracevic, 1976), those since are in this review. However, a few of the pre-1975 works are included in this review as well in order to provide a historical context where appropriate.

IV. Meaning of Relevance: How Relevance Is Universally Well Understood, How It Is Understood in Information Science and, Nevertheless, How Problems with Relevance Are in Its Understandings

A. Intuitive Understanding

I already stressed this in Part I: Relevance does not have to be explained; it is universally understood. It is an intuitive, primitive, "y'know" notion (Saracevic, 1976, p. 90). People understood and understand relevance similarly over time, space, cultures, and domains:

> Nobody has to explain to users of IR systems what relevance is, even if they struggle (sometimes in vain) to find relevant stuff. People understand relevance intuitively (Saracevic, 1996, p. 215).

Intuitively, we understand relevance to encompass a relation—relevance always involves some version of "to" either stated explicitly or referred implicitly. This was always so. To illustrate the point: Following the etymology for "Relevant" the *Oxford English Dictionary* (2nd ed.) has this quote from awhile ago: "1646 CHAS. I *Lett. to A. Henderson* (1649) 55 "To determine our differences, or, at least, to make our Probations and Arguments Relevant." Jumping forward a few centuries to illustrate the same point, the title of an article in the *Chronicle of Higher Education* (September 30, 2005: B1) enthused: "Thoughtful Design Keeps New Libraries Relevant." In both cases "to," while implicit, was clearly there. The "to" is the context for relevance. For relevance context is *it.*

What is actually relevant may not be understood similarly, but what is relevance is. Relevance is a thoroughly human notion; this is its great strength and great weakness. As all human notions, relevance is messy and not necessarily perfectly rational. The role of research is to make relevance less messy.

B. Beyond Intuitive

On a fundamental level, relevance is understood as a relation; relevance is a tuple—a notion consisting of a number of parts that have a relation based on some property or criteria. In other words, relevance has a number of dimensions along which the parts may be related, connected, and interpreted. None of these are necessarily fixed; they may change as circumstances change.

Relevance always involves a relation between a P (or a number of Ps) and a Q (or a number of Qs) along some property R (or a number of Rs). Parts

P and Q could be intangible objects (such as ideas, concepts, information) or tangible objects (such as documents, machines, processes) or a combination of both intangible and tangible objects (such as tasks, situations, responsibilities). Properties R (such as topicality, utility) provide a base and context for establishing a relation, that is relation between Ps and Qs are considered as to relevance along properties R. These properties may be explicit or implicit, well formulated or visceral, rational or not entirely so—on a continuum.

Relevance is also considered as a measure of relatedness. If we consider communication, then our intuitive understanding is that relevance has also something to do with effectiveness of communication. Thus, the relation between objects Ps and Qs along properties Rs may also be ascertained as to some measure S (or a number of Ss), where S may be expressed along different magnitudes, such as strength, degree, or some other quantity or quality. Measures S may be explicit or implicit, well formulated or visceral, rational or not entirely—on a continuum.

Thus, relevance is considered as a *property* along which parts are related and may also be considered as a *measure* of the strength of the related connection.

C. Understanding in Information Science

Understanding of relevance in information science evolved over time and was adapted to specific circumstances. In information science, we consider relevance as a relation between information or information objects (the Ps) on the one hand and contexts, which include cognitive and affective states and situations (information need, intent, topic, problem, task) (the Qs) on the other hand, based on some property reflecting a desired manifestation of relevance (topicality, utility, cognitive match) (the Rs). As mentioned, the Ps and Qs could be tangible or intangible. In addition, we also measure the intensity of the relation on some scale (degree of relevance, or utility, or pertinence) (the Ss). Thus, in information science, relevance is a relation and a measure. If Ps are considered as external and Qs as internal then relevance reflects a relation between external and internal objects along internal and external contexts, including measure(s) that reflects strength or effectiveness of the relation. It is worth stressing that the context is formulated through a dynamic interaction between a number of external and internal aspects, from a physical situation to cognitive and affective states, to motivations and beliefs, to situations, and back to feedback and resolution. Context is complex.

This generalization corresponds with the general pattern for numerous definitions of relevance that were offered in information science as specified

in Part I (Saracevic, 1976, p. 99). The pattern is: "*Relevance is the A of a B existing between a C and a D as determined by an E*," where A may be "*measure, degree, estimate, relation …;*" B may be "*correspondence, utility, fit …;*" C may be "*document, information provided, fact …;*" D may be "*query, request, information requirement …;*" and E may be "*user, judge, information specialist.*" Almost every definition offered still fits this pattern. In Part I, relevance was also considered "as a measure of the effectiveness of contact between a source and a destination in a communication process" (ibid.: 91).

D. The Big Question and Challenge

We also understand that relevance is not given, it is established. This leads to the next question and the big challenge for information science: *How does relevance happen?* The obvious sub-questions are: *Who does it, under what circumstances, and how?* Some of the relevance theories and models, reviewed in next two sections, tried to answer these questions.

In information science, we consider relevance as an inference: it is *created* by inference, but also it is *derived* by inference. This is not an either–or proposition; rather there is a continuum from creating to deriving relevance. A simplified explanation: systems or automatons create relevance and users derive relevance. However, situations could be more complex, because people can act as automatons (fully or to some degree) to create relevance as systems do, and systems can be somewhat "intelligent" to derive some aspect of relevance. Thus, to account for such circumstances there is a need for a continuum, rather than a binary distinction between creation and derivation. It is a matter of degree. Still creation–derivation is a useful distinction, adding to our understanding of relevance in information science. The inference—creation or derivation—follows some intent. In other words, intentionality is involved along the general conception of intentional mental states discussed by Searle (1984). His concluding statement holds for relevance as well:

> Because it is just a plain fact about human beings that they do have desires, goals, intentions, purposes, aims, and plans, and these play a causal role in the production of their behavior (ibid.: 15).

IR systems create relevance—they take a query, process it by following some algorithms, and provide what they consider relevant. People derive relevance from obtained information or information objects. They relate and interpret the information or information objects to the problem at hand, their cognitive state, and other factors. IR systems match queries to objects in their belly to construct those that are relevant to the query, possibly rank order them, and regurgitate the results. Users take the results and derive what may

be relevant to them. But users can read into results a lot more than correspondence between noun phrases or some such in queries and objects, used primarily by systems for matching. Moreover, users can and do find other information objects or other information relevant to their problem that is not retrieved by a system for a variety of reasons, for example not reflected in the query to start with. Several excellent examples of how relevance is derived above and beyond that which is topically retrieved are given by Harter (1992: 607ff). Specifically, Harter provides examples of topics that interest him and then analyzes a number of articles that are not directly related to the topics as stated, but are relevant. He demonstrates through examples how relevance is derived from articles as related to the cognitive state of an individual ("psychological relevance") that is very different than topical relevance as considered by a system. "Topical relevance concerns itself only with a restricted form of language. It ignores the user" (ibid.: 613).

A similar argument about non-matching topicality was provided by Green (1995); Green and Bean (1995) present extensive examples of derived relevance using the topics of a religious thematic guide and the referred passages derived in that guide. More dramatic examples are provided by Swanson and Smalheiser (1997, 1999). In these articles they summarize a decade-long effort in which they took several areas of medicine and showed causal connections between previously unrelated phenomena to derive relevance relations where none existed before; these relations were derived from literature and later confirmed in clinical testing.

The situation is actually more complex than presented. Yes, people may and do derive relevance from ideas and clues in articles that no system could readily recognize, at least as yet. But, that depends also on domain expertise (Vakkari and Hakala, 2000). Greater expertise on a topic leads to more potent derivate powers for relevance. Lesser expertise leads to lesser powers for deriving relevance. With little expertise, one constructs relevance as an automaton. White (in press, a, b) discusses these hypotheses at great length, with examples throughout both articles, and provides essentially the same distinction between created and derived relevance.

Since information science deals with creation and derivation, systems and users, we understood early on that there is not only one kind of relevance, but also several. They were even labeled differently, like "topical relevance," "user relevance," and so on, as reviewed later in the section *Manifestations of Relevance*. Of course, information science is not the only field to recognize that relevance has a number of manifestations. In information science, however, this is a very pronounced understanding, because we match various kinds of relevance and evaluate performance on that basis. Among other things, this also leads to intellectual disputes as to the primacy of one kind of relevance over others.

Here are two final points about understanding relevance in information science. First, either derived or constructed relevance usually involves a process of selection. Information or information objects are selected as relevant (or expressed on some continuum of relevance) from a number of available existing, or even competing, information objects or information. The selection is geared toward maximization of results, minimization of effort in using the results, or both. Second, the selection process involves a series of interactions of various kinds. Thus, an understanding of relevance also recognizes that a selection and interaction process is involved.

E. Summary: Attributes of Relevance in Information Science

We consider relevance as having a number of dimensions or attributes:

Relevance is a relation.
Relevance is a property.
Relevance is a measure.
Relevance has a context, external, and internal.
Relevance may change.
Relevance has a number of manifestations or kinds.
Relevance is not given.
Relevance is inferred.
Relevance is created or derived.
Relevance involves selection.
Relevance involves interaction.
Relevance follows some intentionality.

These attributes of relevance can be summarized as follows (Cosijn and Ingwersen, 2000; Saracevic, 1996):

- *Relation*: Relevance arises when expressing a relation along certain properties, frequently in communicative exchanges that involve people as well as information objects.
- *Intention*: The relation in expression of relevance involves intention(s)—objectives, roles, and expectations. Motivation is involved.
- *Context*: The intention in expression of relevance always comes from a context and is directed toward that context. Relevance cannot be considered without a context:
  - *Internal context*: Relevance involves cognitive and affective states.
  - *External context*: Relevance is directed toward a situation, tasks, problem-at-hand. Social and cultural components may be involved as well.
- *Inference*: Relevance involves assessment about a relation, and on that basis is created or derived.
- *Selection*: Inference may also involve a selection from competing sources geared toward maximization of results and/or minimization of effort in dealing with results.
- *Interaction*: Inference is accomplished as a dynamic, interacting process, in which an interpretation of other attributes may change, as context changes.
- *Measurement*: Relevance involves a graduated assessment of the effectiveness or degree of maximization of a given relation, such as assessment of some information sought, for an intention geared toward a context.

These conceptualizations reflect a general understanding of the meaning of relevance in information science. But as always, the devil is in the details. When these general understandings are translated into theories, models and practices; into systems and users; into inputs and outputs; then the general understanding, as enumerated, does not serve or guide us well—translation from a general understanding to pragmatic application is very difficult. How to actually construct or derive relevance, how to measure it, who does it, and with what effect is an entirely different matter; at times even wrought with controversy. In the same category belongs the question: *How much relevance is enough?* Still, we understand relevance better than we did 30 years ago.

V. Theories of Relevance: What Theoretical Constructs Were Borrowed From Elsewhere and How We Still Don't Have an Applicable Theory of Relevance

After all, relevance is a universal human notion and thus of scholarly interest in fields other than information science. Extensive theories on relevance appear in several fields, among them logic, philosophy, and communication. Relevance theories in logic were not used in information science, and thus are only briefly characterized here to illustrate a possible connection. Those in philosophy were used to some extent and were extensively reviewed in Part I, thus only a synthesis is provided. Finally, a theory of relevance in communication, formulated in the 1980s and 1990s, had some impact on thinking about relevance in information science, thus it is reviewed here in some detail as theory-on-loan, that is as a theory that is used and interpreted within the context of information science.

A. Relevance in Logic

For some 2000 years, logicians have been struggling with the notion of relevance, particularly in deduction of inferences. To avoid fallacies, a necessary condition for an inference from A to B is that A is relevant to B. In that sense, confirmation of conclusions from premises is based on relevance. Relevance logic is an attempt to construct logics that reject theses and arguments that commit fallacies of relevance. Several systems of relevance were developed in semantics and proof theory (Mares, 1998). The widely cited seminal work by Anderson and Belnap (1975) and Anderson *et al.* (1992) is a standard for contemporary treatment and critiques of relevance logic.

Several attempts were made to apply a formal system of logic to IR that involved consideration of relevance (e.g., starting with Cooper, 1971 and continuing with van Rijsbergen, 1986; Nie *et al*., 1995 and others as summarized by Lalmas, 1998 and Sebastiani, 1998) but they are outside the scope of this review. However, all are based on the underlying notion that there is a connection between relevance and logical consequences. No attempt has been made, so far, to apply relevance logic to the study of relevance as a notion in information science. The mentioned work by Anderson and Belnap may be a plausible borrowed theory for such an extension.

However, logic was used in the explication of relevance in artificial intelligence (AI). A special issue on relevance in the journal *Artificial Intelligence* deals with the treatment of relevance in the domain of AI (Subramanian *et al*., 1997). In two articles, logic, together with the concept of belief, was used as a basis for a formal treatment of relevance and its properties. Lakemeyer (1997) formalized relevance relations in the context of propositional logical theories from an agent's point of view and relative to his/her deductive capabilities and beliefs. Beliefs were also used in developing a set of formal axioms of casual irrelevance (Galles and Pearl, 1997). Overall, interest in relevance in AI was fleeting and faded away. However, involving beliefs with relevance makes the approach interesting, even though logic formalities, as applied in cited works, may be highly restrictive in any pragmatic sense. The notion of belief has not yet penetrated relevance theorizing in information science, even though on the face of it the idea may be of interest. Beliefs are a murky concept, but they may affect relevance.

B. Relevance in Philosophy

A number of philosophers, particularly in the area of phenomenology, were interested in relevance. Of particular interest to information science are the works by Schutz (1970) and Schutz and Luckman (1973). The latter is a summary of Alfred Schutz's lifelong ideas, posthumously completed by his collaborator Thomas Luckman. Schutz's concepts related to relevance were already summarized in Part I (Saracevic, 1976, p. 84–85), but are mentioned here again since they continue to have implication for theoretical thinking on relevance in information science; it is knowledge worth borrowing. Briefly, Schutz characterized structure and functioning of the "life-world"—situations that people face in the reality of everyday life. These situations form layers—life-world is stratified. Relevance is the principle for stratification and dynamic interplay among strata. But there is not a single relevance, but rather an interdependent system of relevances (plural). He proposed a typology of relevances with three main categories: thematic

(in the 1970 work called "topical"), interpretational, and motivational. These concepts are echoed in many later works on relevance in information science, even without reference to Schutz.

1. Application in Information Science

Schutz is cited a number of times as an appropriate framework in information science; his viewpoint is very much reflected in works on manifestations of relevance. The two following philosophical perspectives, which emanated from information science, are very different than Schutz's.

In the first, Hjørland (2002) suggests an epistemological perspective for considering relevance and other fundamental concepts at play in IR, such as interpretation of texts and information needs. In supporting this position, Hjørland demonstrates relevance criteria in four epistemological schools: empiricism, rationalism, historicism, and pragmatism. Each provides a different criterion for considering relevance. In essence, as stated in his conclusions, he rejects "the cognitive view [which] tends to psychologize the epistemological issues (to study knowledge by studying the individual)," and advocates "the socio-cognitive view, which tends to epistemologize psychological issues (to see individual knowledge in a historical, cultural, and social perspective)" (ibid.: 268). Epistemology is suggested as the proper way to approach relevance. In a similar vein, Froelich (1994) previously had suggested applying hermeneutics (study of how context makes and shapes interpretation) to the study of relevance, because relevance is an act of interpretation.

In the second perspective, taking a philosophy stance (but not Schutz's or Hjørland's), Budd (2004) reviews treatment of relevance in information science (with a lament that it is not treated in librarianship), and invokes ideas from a number of philosophers, including Wittgenstein and Habermas, as possible explanations. While Budd's review does not offer a theoretical synthesis, but only a selective enumeration, it provides a juxtaposition of a wide range of different views and concepts related to relevance, involving philosophy as well.

Relevance is also philosophical. The works reviewed, however, were not much more than proposals for what to do rather than philosophical treatises on relevance in information science.

C. Relevance in Communication

Information and communication are related, but there is also a distinction. Information is a *phenomeno*n. Communication is a *process*. A process in which

information is dispersed or exchanged. The process of communication encompasses a vast array of human activities and has many facets and manifestations. Similarly, the phenomenon of information encompasses many manifestations—there are many kinds of information—and is interpreted in many senses. Concept of "communication" could be understood and used, similarly as "information," in numerous ways. Not surprisingly then, the field of communication is also broad and expansive. The study of communication intersects with a number of other fields, including, linguistics, semantics, psychology, cognitive science, philosophy, and related areas. The study of relevance in communication also comes from an interdisciplinary tradition. Since one of the theories about relevance that emerged in the study of communication was prominently treated in information science, it is described here in some detail.

The most comprehensive and ambitious contribution to theorizing on relevance in a communication framework was made by Sperber and Wilson (1986, 1995) (abbreviated here as S&W), with the latest synthesis by Wilson and Sperber (2004) (abbreviated here as W&S). Their "Relevance Theory" has an overarching goal of explaining what must be relevant and why to an individual with a single cognitive intention of a conceptual nature. It is based on an inferential model of communication that views communication in terms of intentions, as opposed to the more traditional and widely accepted source–message–destination model (also called the classical code model since messages are coded and decoded). The inferential model considers that the critical feature of most human communication—verbal or non-verbal—is an expression and recognition of intentions. "Relevant information is information worth having" (S&W, 1995, p. 264).

Relevance Theory was originally associated with everyday speech or verbal communication, but later was extended to cover wider cognitive processes. Authors consider it a cognitive psychological theory. It has a high goal of being a theory of cognition and of communication, tying them together on the basis of relevance. However, the basic problem addressed in the theory is how relevance is created in dialogs between persons. It explains "what makes an input worth picking up from the mass of competing stimuli" (W&S, 2004, Section 1). In somewhat awkward language, they argue about ostensive behavior or "ostension," manifestations, and presumptions of relevance. Simply put, out of many stimuli, we pay attention only to information which seems relevant to us; furthermore, to communicate is to claim someone's attention, and hence to imply that the information communicated is relevant. They firmly anchor relevance in a given context and talk about contextual effects—relevance is contextual. They also consider relevance assessment as comparative, not quantitative—relevance is comparative.

At the center of their theory they postulate two principles, claiming to reflect universal tendencies:

1. *Cognitive Principle of Relevance*: The claim that human cognition tends to be geared to maximization of relevance.
2. *Communicative Principle of Relevance*: The claim that every ostensive stimulus conveys a presumption of its own relevance.

In other words, human cognition is relevance oriented, and so is human communication. The two principles lead to the specification of how relevance may be assessed in terms of two components: *cognitive effects* and *processing effort*:

Relevance to an individual:

1. Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of input to the individual at that time.
2. Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time (W&S, 2004, Section 2(1)).

This serves as an explanation as to what makes us "pick out the most relevant stimuli in {our} environment and process them so as to maximise their relevance" (W&S, 2004, Section 3). The two Principles of Relevance and the two components of assessment are at the heart of the theory, with the first being explanatory and the second predictive.

The proposition of maximization in the Cognitive Principle of Relevance evokes a similar, if not identical, explanation postulated by Zipf (1949) in the Principle of Least Effort. Furthermore, treating relevance as an underlying principle in both cognition and communication evokes the explanation of what makes the life-world tick by Schutz (1970), as mentioned above. Neither was considered in S&W's Relevance Theory.

Needless to say, Relevance Theory, as a major, comprehensive attempt to provide explanations and principles about cognition and communication anchored in relevance, attracted followers and critics. Critics voiced a number of themes, among them restriction in scope, contradictions in arguments, and the total absence of any connection to human motivations—in other words, in the theory they treated humans as perfect rational beings. Gorayska and Lindsay (1993) summarized these critiques, pointing out the theory's shortcomings from the point-of-view of the pragmatic use of the notion in everyday language—it does not fit—but also recognized the value of the theory and proposed future directions for research.

The strength of the theory lies in proposing a number of explanations and operational, predictive principles about cognition and communication in terms of relevance. A Relevance Theory at last! Two weaknesses are mentioned here, beside the ones mentioned by critics as cited above. The first

weakness concerns the nature of their proofs and grounds for generalization. They use hypothetical conversations between two protagonists, Peter and Mary, to provide both examples and proof (Peter/Mary dialogs get tiring fast). But more seriously, proof by example is no proof. The second weakness is that in the two decades since its first appearance, the theory was not tested empirically or experimentally. A theory is scientific if it is refutable (i.e., testable). While the authors proposed a number of possible tests and talked about forthcoming experiments (W&S, 2004, Section 6), such tests and experiments have not come forth as yet. Moreover, none are in sight. Relevance Theory is appealing, but it is also untested. It awaits verification and possible modification as a result. Of course, the fact that a theory is not tested is not grounds for rejection. However, an untested theory may also be untestable. In that case, it is not a scientific theory. The question is still open whether Relevance Theory is testable to start with. Nevertheless, it does provide a number of insights about relevance and its behavior.

1. Applications in Information Science

In information science, Harter (1992) provided the first attempt to apply S&W's Relevance Theory to information science in general and IR in particular. He starts with an emphatic rejection of topical relevance, that is, the notion and practice in IR where relevance is treated as to its topicality only. As a solution, he embraced the notion of relevance as being exclusively related to cognitive states that change dynamically, calling this "psychological relevance." Relevance is what causes cognitive changes in a given context. This will be further discussed in the section on *Manifestations of Relevance*, because the essence of Harter's proposal is to consider a given type or manifestation of relevance as the primary or even exclusive property.

Harter deduced a number of excellent insights into relevance behavior. The strength of Harter's notion of psychological relevance is that he has attempted to base the concept on a broader and elaborate theoretical basis, namely S&W's Relevance Theory. The weakness is that actually he has not done that, beyond borrowing some concepts and terminology. Besides, as with S&W's Relevance Theory, Harter's construct was not tested. He discussed, however, the difficulty of testing and applying it in practice. Still, the value of his attempt to gain some theoretical footing for relevance in information science is in itself groundbreaking. Unfortunately, he did not get there but he pointed the way and opened a wide-ranging and raging discussion.

A second and much more comprehensive attempt to transfer S&W's Relevance Theory into an information science framework was done recently

by White (in press a, b). In this massive work, White confines S&W's Relevance Theory to the application of the *cognitive effects* and *processing effort*; he did not use the two relevance principles. In an effort to integrate Relevance Theory, IR and bibliometrics, he proposed that cognitive effects and processing effort are also components in relevance assessments in the context of IR and can be used as predictive mechanisms for the operational assessment of relevance. Briefly, White translated the widely applied approach in IT based on terms called tf*idf (term frequencies, inverse document frequencies) into bibliometric retrieval based on citations; used this to create a new two-dimensional visual display of retrieved bibliometric results called a pennant diagram (because it looks like one); interpreted the dimensions of the diagram in terms of cognitive effects and processing effort; derived a number of practical examples; and engaged in extensive interpretation of results and discussion of reasoning behind them, in a similar vein as S&W. (Even Peter and Mary made a prominent appearance.) White has significantly extended the interpretation of S&W Relevance Theory to information science circumstances and interests, with both the strength and the weaknesses of the theory present. Its strength is that he actually put his constructs to practical work. While the proposed bibliometric retrieval and associated pennant diagram may have been done without recourse to Relevance Theory, the borrowed constructs (cognitive effects and processing effort) provided grounds for extensive abstract explanations of both processes and results. They offer insight about retrieval above and beyond the statistical nature of the process and rank listing of results. However, the weakness of the nature of proof present in S&W's work is also present here. Besides, White's work is not a test of Relevance Theory as claimed; it is structures, concepts, and terminology on loan.

Both works—Harter's and White's—are worthwhile in their efforts to adapt a theory—the field should be stimulated to think about such adaptations and think about theory, but the question remains whether the theory being adapted is worthwhile to start with.

D. Summary: Still in Search of a Theory

As yet, authors on relevance in information science have not developed any indigenous theory cum theory about the notion, nor have they successfully adapted theories from other fields, despite a few attempts. Where theories were borrowed for use, they were merely described, interpreted, and declared appropriate. They were not tested. However, and to their credit, they were conceptual and terminological borrowings used for extending our collective insight about relevance. They made us think.

We are still in search of a theory of relevance applicable to the context of information science and particularly IR. In other words, we are still in search of a conceptual basis, a set of testable principles and propositions, to explain the notion of relevance applicable to information science practice, to explain its manifestation, and to predict its behavior and effects. Of course, practice can be successfully pursued in absence of a theory. The history of technology has a great many examples, IR being just one of them. But, a great many substantial advances have been achieved based on a theory; the history of modern technology has even more such examples. As the adage says: there is nothing more practical than a good theory.

A number of authors have suggested outlines of an applicable theory of relevance. For instance, Park (1994), echoing Harter, suggested a possible framework for "a theory of user-based relevance" (title) to emerge from qualitative research using a naturalistic approach and paradigm. The attempt was interesting, but the proposal lead nowhere. Several other proposals of the same genre are not treated here for the same reason.

These attempts to borrow and adapt theories have a positive effect on clarifying empirical knowledge and understanding about relevance in information science. Schutz's reference to systems of relevances (plural) suggests a number of manifestations of relevance that are already recognized, and his reference to "horizon" suggests the inclusion of contexts as inevitable. S&W's cognitive effects and processing efforts suggest dimensions used in assessing relevance, including its dynamic nature, are also well recognized.

While we were not successful in developing or adapting a "good" theory of relevance for information science, we were certainly rich in proposing a number of models depicting elements or variables involved in relevance, as summarized in the next section. Yet, there are differences between theories and models in scientific endeavors. Theories explain and predict; models enumerate. Theories are about why and how; models are about what is involved or occurring. Theories guide; models provide structure. So, on to models.

VI. Models of Relevance: How Relevance Was Reviewed and Reviewed, and How a Few Models Came Out of Reviews

For whatever reason, relevance is an eminently suitable subject for review. Interestingly, there was a 15-year gap in relevance reviews between the one by Saracevic (1975) and those that began appearing on an almost regular basis since after 1990.

In addition to reviewing the progress in relevance research or challenging a prevalent paradigm or line of thought, these reviews also provided a synthesis on the basis of which relevance models were projected. We concentrate here on several models proposed in major reviews. Models are abstractions forming general ideas from specific examples. Their importance is great because they are a basis for given standpoints that predicate given types of actions and exclude other types. Indeed, different relevance models suggest different actions.

A. Dynamic Model

For a fleeting decade, relevance had its Camelot. It was in Syracuse. From about the mid-1980s until about the mid-1990s, a series of doctoral dissertations at the School of Information Studies, Syracuse University, addressed various aspects of relevance, reflecting a vigorous research environment under the guiding spirit of Robert Taylor and Jeffrey Katzer. These dissertations resulted in a number of articles (Carol Barry, Michael Eisenberg, Myke Gluck, Joseph Janes, Linda Schamber) reviewed later in this work. The Syracuse relevance school also produced a notable and widely cited review that had an extensive impact and changed the view of what is important in relevance. When well done, critical reviews can do that.

Schamber *et al*. (1990) re-examined thinking about relevance in information science, addressed the role of relevance in human information behavior and in systems evaluation, summarized major ideas and experiments, and came to a forceful conclusion that relevance should be modeled as being dynamic and situational. The idea was echoed in Schamber (1994), in which she connected the wider area of human information behavior studies with relevance, organized along the issues of relevance behavior, measurement, and terminology. Of course, dynamic properties of relevance had been discussed in previous decades and demonstrated in experiments as readily acknowledged by the authors, but it was their insistence on the primacy of the dynamic and situational nature of relevance—all is flux—that struck a chord.

They went further and proposed a rich research agenda for the investigation of users and relevance. Research questions were asked about: criteria that users employ in assessing relevance and consistency of their application; the characteristics of documents that are included in these criteria; indicators or clues in documents reflecting these characteristics; recognition of document-based clues by users; and recognition of document-based clues by systems.

The strength of the review was that it suggested a model of relevance in terms of the dynamics of human information behavior and situations in

which this behavior occurs. Moreover, it directed attention to a connection between aspects of documents (documentary relevance clues) and human relevance assessment. It modeled document clues as to relevance. As a result, a clues-oriented research developed, as synthesized in the section *Behavior of Relevance*.

The weakness was twofold. First, stating by itself that relevance is dynamic and situation dependent is not much more than a truism recognized in one way or another since Plato when he contemplated the nature of knowledge. It falls under the category "What else is new?" or "Can it be any other way?" Second, the concept of situation really was not elaborated on, even though promised in the title. Other investigations, reviewed later, specifically addressed both the dynamic and the situational behavior of relevance. Still, this conceptual contribution attracted wide attention and set the stage for further research.

B. Dual Model

Another review with high resonance was produced by Mizzaro (1997) as a "whole history of relevance" (title). The review was a comprehensive classification of 157 studies divided over three periods: Before 1958, 1959–1976, and 1977–1997. Within each period, he classified papers as dealing with one or more of seven different aspects:

1. methodological foundations,
2. different kinds of relevance,
3. beyond-topical criteria adopted by users,
4. modes for expression of the relevance judgment,
5. dynamic nature of relevance,
6. types of document representation,
7. agreement among different judges.

In effect, the seven aspects provide a convenient model along which works, conceptualizations, and findings about relevance may be categorized and compared.

In conclusions, Mizzaro posits the orientation of works in different periods:

> The "1959–1976" period is more oriented toward relevance inherent in documents and query. In the "1977-present" period … the researchers try to understand, formalize, and measure a more subjective, dynamic, and multidimensional relevance (ibid.: 827).

This duality reflects approaches to modeling relevance to this day.

C. Split between System and User Models

Relevance is a participant in a wider battle royal that started in the 1980s and is still going on. It involves two opposing views or models of IR: systems and users. The user side vehemently criticized the system side. The systems side barely noticed that it was attacked. A few reconciliatory authors tried to resolve the differences. In effect the invisible hand of relevance is behind the battle—how to deal with relevance is really what the battle is all about. The arguments resemble those presented in the late 1950s in C. P. Snow's memorable, though dated book *The Two Cultures*, in which he discusses the failure of communication between the sciences and the humanities (the "two cultures" of the title) (Snow, reprinted 1993).

In a massive study of co-citation patterns in information science for the period 1972–1995, White and McCain (1998), among others, mapped the structure of the field showing two broad clusters calling them "domain analysis" and "information retrieval." Their conclusion: "Two subdisciplines of information science are not yet well integrated" (ibid.: 337) and, "as things turn out, information science looks rather like Australia: heavily coastal in its development, with a sparsely settled interior" (ibid.: 342). This holds for relevance—it indeed has two cultures, each with its own model; they are not integrated, and they map like Australia. Despite attempts at bridging, as reviewed below, the two cultures are mostly foreign to each other.

The systems viewpoint, obviously, considers IR from the systems' side ignoring the user. It is based on a model of IR, called traditional or laboratory IR model, in which the emphasis is on systems processing information objects and matching them with queries. The processing and matching is algorithmic; the goal of the algorithms is to create and maximize retrieval of relevant information or information objects. In the purest form of this model, the user is represented by a query and not considered in any other respect; also, interaction is not a consideration. The model has been in continuous and unchanged use since the Cranfield experiments (Cleverdon, 1967) to experiments conducted under the fold of Text REtrieval Conference (TREC) (Voorhees and Harman, 2005) (TREC, started in 1992, is a long-term effort at the [US] National Institute for Standards and Technology (NIST), that brings various IR teams together annually to compare results from different IR approaches under laboratory conditions).

The user viewpoint considers IR from the user's rather than the systems' side, taking the system as a given. The user is considered way beyond the query by seeking to incorporate a host of cognitive and social dimensions, and interaction into the model. The user viewpoint does not have a single

model that has been agreed upon, although quite a few have been proposed reflecting different perspectives (e.g., Ingwersen, 1996).

While there were rumblings long before, the frontal attack championing the user side came in a critical review by Dervin and Nilan (1986). While reviewing alternative approaches to the assessment of information needs, they issued a call for a significant paradigm shift in information needs and uses research from systems orientation to user orientation, underscoring that the systems approach is inadequate. The review, considered a turning point in user studies, was much cited, often as a sort of a manifesto. The Dervin and Nilan review did not consider relevance per se, but nevertheless relevance was predominant. Of course, studies of human information behavior (which include information seeking and user studies) can investigate aspects that do not involve relevance. However, when considering any aspect of retrieval, relevance is present either explicitly or as an invisible hand.

User studies became a burgeoning area of research with the following justification:

> By looking at all kinds of criteria users employ in evaluating information, not only can we attain a more concrete understanding of relevance, but we can also inform system design (Schamber *et al*., 1990, p. 773).

"*Informing systems design*" became a mantra not only for relevance studies, but also for all studies of human information behavior and information seeking in particular; it even concludes the introduction in this review. Seems logical. But it is not really happening. Why? The question was analyzed and lamented upon by a number of researchers and commentators about the state of affairs in information science. Researchers representing the systems viewpoint simply took a stance: "*Tell us what to do and we will do it.*" But the user side was not "*telling*" beyond the mantra. Unfortunately, "*telling*" is not that simple. Relevance is a factor of human intelligence. Human intelligence is as elusive to "algorithm-ize" for IR as it was for AI.

As it turns out, both sides in the battle are wrong. Dervin and Nilan and followers were wrong in insisting on the primacy or exclusivity of the user approach. Systems people were wrong in ignoring the user side and making the traditional IR model an exclusive foundation of their research for decades on end. Neither side got out of their box. Deep down the issue is really not a system vs. user approach. It is not system relevance *against* user relevance. The central issue and problem is: *How can we make the user and system side work together for the benefit of both?* When IR systems fail, the main reason is a failure in relevance; thus, that is the best reason for advocating the resolution of the system–user problem in an integrative manner.

A number of works have tried to reconcile the two viewpoints, suggesting integrative relevance models as a solution to the problem. Starting from the user viewpoint, Ingwersen and Järvelin (2005) produced a massive volume outlining the integration of approaches in information seeking and IR in context. The goal of the effort:

> It is time to look back and to look forward to develop a new integrated view of information seeking and retrieval: the field should turn off its separate narrow paths of research and construct a new avenue (ibid.: vii).

This they did, with relevance playing a major and explicit role. They reviewed any and all models used in IR and in information seeking research, and produced an extensive model integrating cognitive and systems aspects of IR. The Ingwersen-Järvelin integrative model, anchored in cognition, is complex, reflecting the complexity of the process and situation. The model has five central components:

> each consisting of data structures representing the cognitive structures of the actors involved in their generation, maintenance, and modification in time: 1) the IT setting; 2) the information space holding objects of potential information value to 3) information seekers via 4) interface mechanism—all set in 5) socio-organizational context (ibid.: 306).

The model is also an integrated relevance model. In addition, they defined several manifestations or kinds of relevance as discussed in the next section.

In a similar vein, Ruthven (2005) reviews various approaches to relevance, from systems to situational to cognitive, and advocates an approach that integrates IR and information seeking research. While he starts from a systems viewpoint, he also fully recognizes the limited nature of the ensuing relevance definition in that model. Among others, he reviews different kinds of relevance assessments (non-binary, consensus, completeness) and suggests that "allowing users of IR systems to make differentiated relevance assessments would seem a simple extension to the standard IR interface" (ibid.: 71). (Well, is it really "simple"?). He also deals with relevance dynamics—the issue of changing user assessments of relevance over time and comments how IR systems have responded poorly to this phenomenon. Ruthven rightly concludes:

> How we use relevance in the design of IR systems—what evidence of relevance we see as important, how we believe this evidence should be handled, what inference we draw from this evidence—define what we see as the task of retrieval systems (ibid.: 77).

D. Stratified Model

Relevance is a tangled affair involving interaction between and among a host of factors and variables. In philosophy, Schutz (as reviewed in section *Theories of Relevance*) considered people in their everyday social world ("life-world"), which is not a homogeneous affair (and is really tangled!); he suggested that the life-world is stratified into different realities, with relevance being at the root of the stratification of the life-world. Models that view a complex, intertwined object (process, structure, system, phenomenon, notion) in a stratified way were suggested in a number of fields from linguistics to medicine to meteorology to statistics and more. "*Stratified*" means that the object modeled is considered in terms of a set of interdependent, interacting layers or levels; it is decomposed and composed back in terms of layers or strata.

After reviewing and reconsidering various relevance models, I proposed a stratified model for relevance (Saracevic, 1996). It is another integrative model. I further extended the stratified model to include IR interactions in general, encompassing a number of specific processes or notions that play a crucial role in IR interaction: relevance, user modeling, selection of search terms, and feedback (Saracevic, 1997). Various elements in and derivations from the model were also elaborated on and extended by Cosijn and Ingwersen (2000). Relevance is placed within a framework of IR interaction. In the stratified model, IR interactions are depicted as involving a number of layers or strata; inferences about relevance are created or derived in interaction and interplay among these strata. Generally, the stratified model is aimed at decomposing the complex interaction between people, information, and technology on the one hand, and showing the interdependence between the elements involved on the other hand.

The stratified model starts with assumptions that: (i) users interact with IR systems in order to use information and (ii) that the use of information is connected with cognition and then situational application, that is, it is connected with relevance (Saracevic and Kantor, 1997). These assumptions also follow from relevance attributes as summarized in the section *Meaning of Relevance*. The major elements in the stratified model are user and computer, each with a host of variables of their own, having a discourse through an interface, as depicted in Fig. 1.

The user side has a number of levels. I suggest three to start with: *Cognitive*, *Affective*, and *Situational.* The suggested computer levels are *Engineering* (hardware), *Processing* (software, algorithms), and *Content* (information resources). It should be recognized that each level can be further

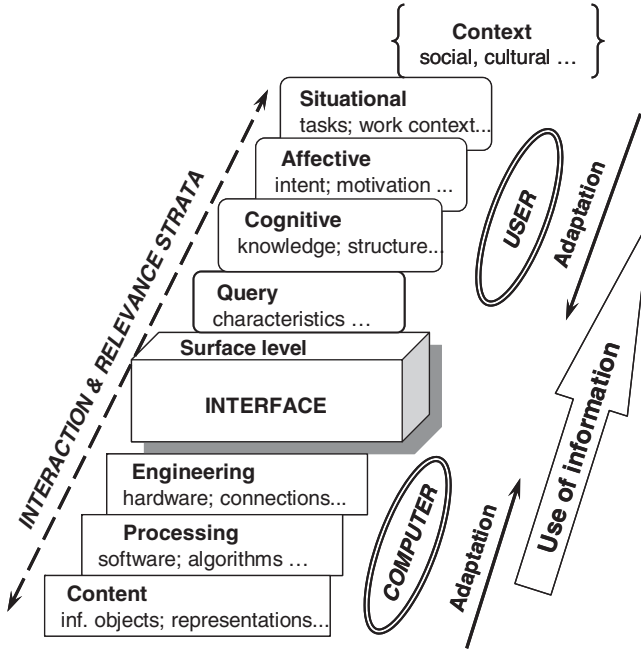

Fig. 1 Stratified model of relevance interactions.

delineated, or that others may be added, depending on the given set of conditions or emphasis in analysis.

A variety of interactions are instantiated on the interface or surface level, but the interface is not the focus of interactions despite the fact that it can in its own right effectively support or frustrate other interactions. We can think of interaction as a sequence of processes occurring in several connected levels or strata. *The IR interaction is then a dialog between the participants—elements associated with the user and with the computer—through an interface, with the main purpose being to affect the cognitive state of the user for effective use of relevant information in connection with an application at hand, including a context.* The dialog can be reiterative, incorporating among other things, various feedback types, and can exhibit a number of patterns—all of which are topics for study.

Each strata/level involves different elements and/or specific processes. On the human side, processes may be physiological, psychological, affective, and cognitive. On the computer side, they may be physical and symbolic. The

interface provides for an interaction on the *surface* level in which:

1. Users carry out a dialog by making utterances (e.g., commands) and receiving responses (computer utterances) through an interface with a computer to not only do searching and matching (as depicted in the traditional IR model), but also engage in a number of other processes or "things," beyond searching and matching, such as: understanding and eliciting the attributes of a given computer component, or information resource; browsing; navigating within and among information resources, even distributed ones; determining the state of a given process; visualizing displays and results; obtaining and providing various types of feedback; making judgments; and so on.
2. Computers interact with users with given processes and "understandings" of their own, and provide given responses in this dialog; they also may provide elicitations or requests for responses from the user in turn.

Let me elaborate on the nature of relevance from the stratified model point of view. We assume that the primary (but not the only) intent on both the user and the computer side of IR interaction deals with relevance. Given that we have a number of strata in interaction, and that in each of them may be considerations or inferences as to relevance, then relevance can also be considered in strata. In other words, in IR we have a dynamic, interdependent *system of relevances* (note plural). Similarly, this plurality was depicted by Schutz, from whom I took the term "system of relevances," and by Sperber and Wilson, who talked about principles of relevance. In IR, relevance manifests itself in different strata. While often there may be differences in relevance inferences at different strata, these inferences are still interdependent. The whole point of IR evaluation, as practiced, is to compare relevance inferences from different levels. We can typify relevance as it manifests itself at different levels, and we can then study its behavior and effects within and between strata—as treated in relevant sections.

E. Summary

All IR and information seeking models have relevance at their base either explicitly or as an invisible hand—in effect they are relevance models. A variety of integrative relevance models, above and beyond the simple traditional IR model, have been proposed. Basically, the models sought a framework within which the complexity of relevance may be analyzed, and the widely dissonant ideas on the notion may be interpreted and related to one another.

Among them, the stratified model has been suggested not only for modeling relevance but also for modeling interaction in IR, and more broadly in human–computer interaction (Saracevic, 1997). As examples, Rieh and Xie (2006) adapted it for a study of patterns of interactive

reformulation of queries posed on the Web, and Spink and Cole (2005b) for deriving a multitasking framework for cognitive IR. At its base, relevance involves interaction. Interaction is interplay between numbers of elements—so is relevance. Interaction is a tangled affair—so is relevance. The stratified model is suggested as one way to untangle them.

Proposing more complex models was an advance in relevance scholarship. However, suggesting models and applying them are two different things. Time will tell if the integrative models and approaches to IR will be successful in furthering research and practice.

IR systems chose to deal with a (if not even THE) most simplified model and manifestation of relevance (later called weak relevance). Within that model, IR is a proven success. Now that much more complex models and manifestations of relevance have been identified, together with suggestions to be incorporated in IR, the challenge to translate them into IR research and practice has increased a lot. *A LOT!*

VII. Manifestations of Relevance: How Relevance Is Not One Thing But Many And How They Are Interpreted

"*How many relevances in information retrieval?*" is a title of an article by Mizzaro (1998). Indeed, how many? Manifestation is a realization, a display of existence, nature, qualities, or presence of some thing. Like many other notions or phenomena, relevance has a number of manifestations. Think of energy: potential energy and kinetic energy are some of its manifestations. For some phenomena or notions, it is not that easy to identify the variety of manifestations and to distinguish among them. Think of manifestations of love. Or information. Or relevance.

As already pointed out, in information science, relevance was early on distinguished as comprising various kinds. It was an explicit realization that relevance has different manifestations. With time and recognition of a number of problems with relevance, a cottage industry has developed in identifying and naming different kinds or manifestations of relevance, or presenting arguments about various manifestations. Manifestations of relevance also became argumentative.

As noted, relevance, among other things, indicates a relation. Efforts to specify manifestations of relevance have concentrated on identifying what given objects are related by a given kind of relevance—the Ps and Qs discussed in the section *Meaning of Relevance*. Different manifestations are manifested by different objects being related and/or by different properties used

for a relation. Sometimes, the efforts also involved naming different manifestations—such as adding a qualifier in the form of [*adjective*] relevance, for example "*topical* relevance" or using a distinct name to denote a distinct manifestation, for example "pertinence." Relevance gained adjectives. Relevance gained names. But that which we call relevance by any other word would still be relevance. Relevance is relevance is relevance is relevance. The arguments about manifestations concentrated more on the primacy of given manifestation rather than their nature. Here is an attempt to interpret the proposed manifestations and replay the manifestation arguments.

A. Starting from Duality

In 1959, Brian Vickery was first to recognize that relevance has different manifestations (Vickery, 1959a, b). Inadvertently, the way in which he did it also precipitated a pattern of discussion about relevance manifestations that continues to this day. In a paper in the *Proceedings of the International Conference on Scientific Information* (a highly influential conference and publication), Vickery defined: the "controlling [criterion] in deciding on the optimum level of discrimination, we may call *user relevance*" (italics his; 1959a, p. 863). In another paper about terminology and retrieval, he discussed what is meant by "relevant to a particular sought subject" (1959b, p. 1227). He identified a duality of relevance manifestations *and* he treated each separately.

User relevance on the one hand and *subject (topic, system) relevance* on the other. These represent the basic relevance manifestations. Each involves different relations. Each can and was further refined and interpreted; each can be thought as a broad class with subclasses. In retrieval they dance together, sometimes in intricate patterns and with various levels of success. This is the nature of any and all retrievals of information. This is why we consider relevance as interaction. The interplay between the two manifestations cannot be avoided; however, the effectiveness may differ greatly depending on how the interplay is accomplished. The two should be complementary, but at times they are in conflict. The duality was explicit in reviews discussed in the preceding section.

In a paper with the shortest title in the history of relevance writings, Bookstein (1979) pursues the formalization of an operational interpretation of relevance for IR "[to help] the reader disentangle at least part of the web of notions surrounding one of the most basic concepts of our discipline" (ibid.: 272). In discussing what people intend when they use the term "relevant" (quotes his) and what the basic functions of an IR system are, Bookstein explicitly recognizes a "duality of viewpoints," and concludes that it "accounts for much of the confusion surrounding the notion of relevance" (ibid.: 269).

Relevance is confusing. Yes it is, but the duality cannot be avoided despite the confusion such duality creates. It can only be made less confusing.

In a different way, the tension within the relevance duality was expressed as "objective versus subjective relevance" (Swanson, 1986: title). As to the two types of relevance, Swanson equates them to Popper's "Worlds" and opines:

> Whatever the requester says is relevant is taken to be relevant; the requester is the final arbiter … Relevance so defined is subjective; it is a mental experience." "A possibility exists that such a request is logically related to some document. … That relationship is then the basis that the document is objectively related to the request (Ibid.: 391, 392).

Swanson's argument about objective relevance is based on logical relations between requests and documents, and a possible degree of confirmation. Pessimistically, Swanson concludes: "[For the purpose of an IR search] I believe that the problem of accounting for or describing subjective relevance is essentially intractable" (ibid.: 395). This pessimism is in stark contrast to the optimistic mantra of user studies, described in the preceding section, stating that such studies have a potential of contributing to better designs. Thus, we have another duality: optimistic and pessimistic relevance. To date, the pessimistic kind is pragmatically ahead.

In a similar vein, Ingwersen and Järvelin (2005) considered algorithmic relevance as an "objective assessment made by a retrieval algorithm," and topical relevance, pertinence, situational relevance, and socio-cognitive relevance as being "of higher order due to their subjectivity" (ibid.: 381, 385).

While it is hard to think about anything in relevance as fully objective, considering relevance in these terms follows ideas of Karl Popper (thought of as the greatest philosopher of science in the 20th century) and his three interacting "Worlds":

> World One, is the phenomenal world, or the world of direct experience. World Two is the world of mind, or mental states, ideas, and perceptions. World Three is the body of human knowledge expressed in its manifold forms, or the products of the second world made manifest in the materials of the first world (e.g., books, papers, symphonies, and all the products of the human mind). World Three, he argued, was the product of individual human beings. The influence of World Three, in his view, on the individual human mind (World Two) is at least as strong as the influence of World One (Popper, 1972).

At its base, relevance is dual, perhaps a product of interaction between different Worlds, but from that dualism grows a pluralistic system of relevancies.

B. Beyond Duality

Relevance manifestations are cornucopian. There is much more to relevance manifestations than duality. A number of works suggested other or additional relevance manifestations. For instance, Cooper (1971) introduced "logical relevance," and improving on this, Wilson (1973) introduced "situational relevance." Harter (1992) championed "psychological relevance." "Topical relevance" was a perennial topic of discussion. "Pertinence" and "utility" were used by a number of authors. And we also have "system relevance," "documentary relevance," and so on, as discussed below.

In the previously cited article, Mizzaro (1998) tried to create order and clarify the issue of relevance manifestations by suggesting a classification that accommodates all of them. He proposed that relevance manifestations can be classified in a four-dimensional space: (i) *information resources* (documents, surrogates, information); (ii) *representation of user problem* (real information need, perceived information need, request, query); (iii) *time* (interaction between other dimensions as changed over time); and (iv) *components* (topic, task, context). Accordingly, Mizzaro suggests that each manifestation of relevance can be represented by where it fits into this four-dimensional space as a partial order. For example:

> *rel (Surrogate, Query, t(q_0), [Topic]* stands for the relevance of a surrogate to the query at time *t(q_0)*, with respect to the topic component (the relevance judged by an IR system (ibid.: 311).

If we agree to these four dimensions, including the given definitions of what they contain, then Mizzaro is right: various manifestations can indeed be consigned to various dimensions. But is it a space? Hardly, for the expression of some logical placements of and distances between different manifestations cannot be derived.

Starting from the cognitive viewpoint and the idea that relevance judgments evolve during the process of IR interaction, Borlund (2003) developed a framework for viewing relevance that also can be considered a classification of various manifestations of relevance. She analyzed three instances of relevance relations, also enumerating aspects or variables involved: (i) the types of relevance relationships employed in traditional non-interactive IR during an IR session; (ii) the types of relevance relationships involved in a given instance of an IR session, which includes situational relevance as viewed by Wilson (1973); and (iii) the types of relevance relationships which include the interrelationship between judgment of situational relevance and the development of the information need during a dynamic and interactive IR session. Depicted manifestations are from non-interactive to situational to interactive.

The three instances build on each other, and the third, as expected, is most comprehensive. Topical, situational, and cognitive relevances are modeled.

Both of these works provide a framework for conceptualizing various attributes of relevance and a classification for relevance manifestations. But as the first sentence of this section ponders: We still do not know "how many relevances" there are.

1. User Relevances

User relevances follow from user context. And user context was a main consideration in a number of relevance models already discussed. But what does that mean? What manifestations are involved? One way is to classify them is as internal and external.

Internally, the most prominent variable in which relevance plays a role is in changes in the cognitive state. This prompted Harter (1992) to introduce *psychological relevance*; more often labeled *cognitive relevance*—meaning a relation between *information objects or information* and the *user's cognitive state*. But this is not the only internal aspect. Carol Kuhlthau studied extensively and longitudinally the process of information seeking (her work, beginning in the 1980s, is synthesized in Kuhlthau, 2004). While she did not study relevance per se, she derived a model of information seeking ("Kuhlthau's model") that involved not only cognitive but also affective aspects of users. Following this, Saracevic (1996) added to other relevance manifestations also *motivational or affective relevance*—a relation between *information or information objects* and *intents, goals, motivations, frustrations of a user*. Cosijn and Ingwersen (2000) elaborated further on Saracevic (1996), and defined five manifestations of relevance: *algorithmic*, *topical*, *cognitive*, *situational*, and *socio-cognitive.* However, they made a distinction between motivation and intention or intentionality, and even "placed affective relevance not as a manifestation nor as an attribute, but as a dimension in line with time" (ibid.: 546). They considered that affective relevance is time dependent over all manifestations except algorithmic relevance. Affective relevance is also contentious.

Externally, we consider that a user is faced with something in relation to which relevance is inferred. This introduced *situational relevance*—a relation between *information or information objects* and *situation, task, or problem at hand facing the user*. However, "external" really is not wholly external—it also involves user interpretation of that externality. Cosijn and Ingwersen (2000, p. 547) made a further distinction: in addition to situational relevance, they introduced *socio-cognitive relevance*, a relation between *information or information objects* and *situation, task or problem at hand as perceived in socio-cultural context.* The context of relevance has a further context.

2. Topical Relevances

Vickery (1959b) labeled it *subject relevance*, but, more often than not, we call it *topical relevance.* Both terms denote the same relation: between *information or information objects* and *the topic or subject under consideration*. Topical relevance may be inferred from the output of an IR system, or completely independent of any system—from any set of information objects, for example from the pile of documents in my office that I gathered there over the years. Topical relevance may or may not involve an IR system.

Documentary relevance also denotes topical relevance, but is restricted to documents as texts, rather than a whole class of information objects that include not only texts, but other informational artifacts, such as images, music, speech, or multimedia. Ingwersen and Järvelin (2005) introduced *bibliographic relevance*—a relation between *representations of metadata* (e.g., as found in a catalog) and *the topic or subject under consideration*.

To narrow relevance manifestation down to systems, we have *system relevance*—a relation between *information or information objects retrieved by the system* and *the query*. Sometimes it is also called *algorithmic relevance* to denote the method of inference. It has been argued that in a narrow sense system relevance is always perfect; the system retrieved that which the query asked for. Not so. The whole point of the evaluation of different algorithms is that they produce different outputs for the same query and from the same set of documents in the system's belly.

3. Issue of Primacy: Weak and Strong Relevance

Does topical relevance underlie all others? Do all other manifestations of relevance follow from topical relevance and does it have primacy among relevance manifestations? As we can imagine, there are two schools of thought: yes and no.

In the first, topicality is basic. For example, in summarizing definitions of topical relevance, pertinence, and utility, Soergel (1994) suggests that an entity—information object—is topically relevant if it can help to answer a user's question; it is pertinent if topically relevant and appropriate for the user—a user can understand and use the information obtained—and it has utility if pertinent and gives the user new information. In this view, pertinence and utility follow from topicality.

In the second school of thought, topicality is not basic, there is relevance beyond topicality. Non-topical relevance can be derived, as discussed in the section *Meaning of Relevance*, from information objects that are not directly topically related.

The issue boils down to the query and request on one hand and information interpretation and derivation on the other. In a strict correspondence between a query and request, topical relevance is basic. But, if we approach the issue with human intellect and imagination, then many interpretations can be made. Topical relevance is but one manifestation, while others may be of even more interest—relevance is not necessarily restricted to a single, direct correspondence. Topical relevance by itself may be labeled as weak relevance: the second interpretation, relevance beyond topicality, includes derivate powers of the intellect and is more argumentative. This may be labeled as strong relevance: there is weak relevance and strong relevance. Weak relevance goes with systems, strong with people. Duality strikes again.

Topical relevance is certainly the basis for system or algorithmic relevance (Borlund, 2003). A simple line of reasoning is: systems retrieve that which is asked for in a query; a query represents a topic of interest. As is practiced today, the overwhelming majority of IR systems organize information objects around words; queries are expressed in words, and matching is based on words or derivative connections. These words are mostly noun phrases. Even when documents are matched based on similarity, matching is based on words. More sophisticated handling involves patterns, such as in music or image retrieval, or links, such as in citation retrieval or Google's page-rank retrieval. But word-based retrieval is still on the throne. In turn, word-based retrieval is based on trying to establish topical relevance. In this sense, it is also the simplest kind of relevance, no matter the sophistication of algorithms and procedures involved. Systems construct weak relevance. This does not mean that the task is simple, for words, arranged in language, are by no means a simple proposition to handle. They are human creation, complex and messy. It is very hard to deal even with the simplest, weakest kind of relevance.

C. Summary

Relevance is like a tree of knowledge. The basic structure of the system of relevances in information science is a duality. The tree of relevance has two main branches, system and human, each with a number of twigs, but it is still the same tree. The roots of the branches and the fruits (results) are a matter for exploration.

Here is a summary of the manifestations of relevance in information science, mainly following Saracevic (1997), Cosijn and Ingwersen (2000),

and Borlund (2003):

- *System or algorithmic relevance*: Relation between a query and information or information objects in the file of a system as retrieved or as failed to be retrieved, by a given procedure or algorithm. Each system has ways and means by which given objects are represented, organized, and matched to a query. They encompass an assumption of relevance, in that the intent is to retrieve a set of objects that the system inferred (constructed) as being relevant to a query. Comparative effectiveness in inferring relevance is the criterion for system relevance.
- *Topical or subject relevance*: Relation between the subject or topic expressed in a query and topic or subject covered by information or information objects (retrieved or in the systems file, or even in existence). It is assumed that both queries and objects can be identified as being about a topic or subject. Aboutness is the criterion by which topicality is inferred.
- *Cognitive relevance or pertinence*: Relation between the cognitive state of knowledge of a user and information or information objects (retrieved or in the systems file, or even in existence). Cognitive correspondence, informativeness, novelty, information quality, and the like are criteria by which cognitive relevance is inferred.
- *Situational relevance or utility*: Relation between the situation, task, or problem at hand and information objects (retrieved or in the systems file, or even in existence). Usefulness in decision-making, appropriateness of information in resolution of a problem, reduction of uncertainty, and the like are criteria by which situational relevance is inferred. This may be extended to involve general social and cultural factors as well.
- *Affective relevance*: Relation between the intents, goals, emotions, and motivations of a user and information (retrieved or in the systems file, or even in existence). Satisfaction, success, accomplishment, and the like are criteria for inferring motivational relevance.

VIII. Behavior of Relevance: Or Rather, How People Behave Around Relevance and How It Was Studied

Strictly speaking, relevance does not behave. People behave. A number of studies examined a variety of factors that play a role in how humans determine relevance of information or information objects. Relevance behavior studies are closely related to information seeking studies and to the broad area of human information behavior studies. Not surprisingly then, texts that deal with human information behavior or cognitive IR, which is based on human information behavior, extensively deal with relevance as well (e.g., Ingwersen and Järvelin, 2005; Spink and Cole, 2005a). Many studies on various aspects of human information behavior are related to relevance behavior but are not included here for space reasons. Examples include studies on decisions about documents in reading and citing (Wang and White, 1999), on judgment of cognitive authority and information quality (Rieh and Belkin, 2000), or on users' assessments of Web pages (Tombros *et al*., 2005). Kelly (2005) reviewed a host of studies about human decisions during

interaction with the Web (or other information resources); the focus was on decisions as to what to examine, retain (save, print), reference, annotate, and the like. Such decisions are treated as implicitly indicating relevance. In other words, although relevance was not overly indicated, an action such as saving a page or document is regarded as implying relevance; relevance is not stated but assumed. While related to relevance by assumption, studies on implicit or secondary relevance are also not included here.

In this and the next section, I concentrate *exclusively* on observational, empirical, or experimental studies, that is, on works that contained some kind of data directly addressing relevance. Works that discuss or review the same topics but do not contain data are *not* included, with a few exceptions to provide a context. Works that are related but do not directly treat relevance, as the aforementioned human information studies, also are excluded. I probably missed some studies and did not include repetitive papers (where the same study is reported again), but I believe this coverage of relevance studies with data for the last three decades is a fairly complete representation. This seems to be it. A few studies before that period are included for context. Relevance experimental and observational studies were very much alive in the 1960s; they had a hiatus from the mid-1970s until the late 1980s, and started on a revival path in the early 1990s.

Studies are briefly summarized following this pattern:

- [*author*] used [*subjects*] to do [*tasks*] in order to study [*object of research*].

If the authors had several objects of research, only those related to relevance are mentioned, thus the full statement should actually be read as: "in order to study, among others, [*object of research*]." While most, if not all, studies included a discussion of a framework (underlying theories, models, concepts, and the like), this discussion is omitted in their description that follows because it is covered in preceding sections of this article. Where appropriate, some summaries include numerical results. However, the principal results from all of the studies, with a number of caveats, are synthesized and generalized at the end of the section.

A. Relevance Clues

What makes information or information objects relevant? Or more specifically: What do people look for in information or information objects in order to infer relevance? Two distinct approaches are used in deciphering this question. In the first, or topic, approach, the course of deriving topical or non-topical relation is analyzed. This approach (represented by Green and Bean, 1995 and Swanson and Smalheiser, 1997, 1999) was treated in the section Meaning of

Relevance. The second or clues approach, treated here, follows the research agenda proposed by Schamber *et al*. (1990) (reviewed in the section Models of Relevance) to study criteria or clues found in given information or information objects (usually documents) that people use in assessments of relevance. The first approach deals with topical relevance only; the second includes cognitive, situational, and affective relevance as well.

Specifically, clues research aims to uncover and classify attributes or criteria that users concentrate on while making relevance inferences. The focus is on criteria users employ while contemplating what is or is not relevant, and to what degree it may be relevant. A wide range of clues or criteria were investigated. Different observational studies came up with different lists and classifications. Here are summaries of various studies:

- Schamber (1991) interviewed 30 users of weather information using different sources, from oral reports to documents and maps in order to derive and categorize their relevance criteria. She identified 22 categories in 10 groups.
- Park (1993) interviewed four faculty and six graduate students who received an online search related to their real need in order to study the thought processes of users evaluating retrieved bibliographic citations. She identified three major categories that included 22 subcategories of variables affecting relevance inferences.
- Cool *et al*. (1993) report on two studies. In the first, they asked approximately 300 freshmen in a computer science course, who were assigned to write an essay on a topic and had selected at least five sources on the topic, to indicate reasons for their selections. In the second study, they interviewed an unspecified number of humanities scholars on their use of information sources for a variety of tasks from teaching to research. Both studies were done in order to identify characteristics of texts affecting relevance judgments. They identified six facets of judgment of document usefulness.
- Barry (1994) interviewed 18 academic users (not specified as being students or faculty) who had requested an information search for documents related to their work in order to categorize their relevance criteria. She identified 23 categories in seven groups.
- Howard (1994) studied nine graduate students who had selected five to seven documents for a class assignment, and identified the relevance criteria for their selections in order to determine and compare personal constructs (criteria) used in relevance assessments. She identified 32 personal constructs grouped in two groups (topicality and informativeness).
- Wang (1997) compared 11 relevance criteria derived from a study in her doctoral dissertation with criteria from four other studies (Barry, 1994; Cool *et al*., 1993; Park, 1993; Schamber, 1991) in order to suggest a general model for document selection using relevance clues.
- Fidel and Crandall (1997) studied 15 engineering users and observed 34 sessions in which they received technical reports, asking them to think aloud about their decisions of deleting or retaining given reports in order to derive criteria for judging the reports relevant or not relevant. They identified 13 criteria explaining why a report was relevant, and 14 explaining why was not relevant.
- Barry and Schamber (1998) compared results from two of their studies (Barry, 1994; Schamber, 1991) in order to study similarities and differences in derived criteria. They identified 10 criteria in common and concluded that there is a high degree of overlap in criteria from both studies despite the difference in users and sources. This is the only study that attempted a badly needed generalization about relevance clues and criteria with a detailed analysis of data.

Other studies that addressed the issue compared different criteria with a checklist or in a brief discussion.

- Barry (1998) looked at 18 students and faculty (not differentiated as to how many in each category) who submitted a request for an online search and were presented with 15 retrieved documents. The documents were organized in four document representations in order to identify the extent to which various document representations contain clues that allow users to determine the presence, or absence, of traits, and/or qualities that determine the relevance of the document.
- Tombros and Sanderson (1998) asked two groups of 10 graduate students each to judge the relevance of a list of the 50 highest-ranked documents from 50 TREC queries in order to investigate the impact of different document clues on the effectiveness of judgments. Each subject judged relevance for five queries; one group judged documents with, and the other without, summaries, and judgment time was limited to 5 minutes.
- Schamber and Bateman (1999) used a total of 304 graduate students in five studies over several (unspecified) years to sort and rank a number of relevance criteria they used while seeking information, starting with 119 relevance criteria concepts/terms from previous studies, in order to interpret and rank user-determined relevance criteria while making relevance inferences.
- Hirsh (1999) interviewed 10 fifth-grade children, who searched various electronic sources for a class assignment, about their ways of searching and making decisions. The interviews were done during the first and third week of the project in order to examine how children make relevance decisions on information related to a school assignment. She identified nine categories of relevance criteria for textual materials and five categories for graphical materials.
- Fitzgerald and Galloway (2001) observed 10 undergraduate students using a digital library for their projects in assessing a total of 138 retrieved documents in order to derive relevance- and evaluation-related reasoning. They identified 11 relevance and 11 evaluation categories of reasoning, both entering in relevance decisions.
- Maglaughlin and Sonnenwald (2002) asked 12 graduate students with real information needs to judge the relevance of the 20 most recent documents retrieved in response to a query that were presented in different representations in order to derive and compare criteria for relevant, partially relevant, and non-relevant judgments. They identified 29 criteria in six categories and compared the presence of their criteria with criteria from 10 other studies.
- Toms *et al*. (2005) recruited 48 subjects from the general public to search the Web for answers to 16 tasks (topics) in four domains. The subjects were asked to indicate in a verbal protocol their assessment of and satisfaction with the results in order to identify and categorize a set of measures (criteria) for relevance along five relevance manifestations as formulated by Saracevic (1996). They identified 11 measures of relevance.

1. Image Clues

What makes images relevant? Are clues used in relevance inference about images similar to those for texts?

- Choi and Rasmussen (2002) interviewed 38 faculty and graduate students of American history (not differentiated as to faculty and students) on the retrieval of images using the Library of Congress *American Memory* photo archive in order to study the users' relevance criteria and dynamic changes in relevance criteria as expressed before and after the search. They used nine criteria before and identified an additional eight after the search.

B. Relevance Dynamics

Do relevance inferences and criteria change over time for the same user and task, and if so, how? The basic approach used to answer this question starts with two assumptions. As a user progresses through various stages of a task: (i) the user's cognitive state changes and (ii) the task changes as well. Thus, something about relevance also is changing. The idea of studying such dynamic changes in relevance has a long history. Rees and Schultz (1967) pioneered this line of inquiry by studying changes in relevance assessments over three stages of a given research project in diabetes. Since then, studies of relevance dynamics follow the same ideas and assumptions. Here is a representative sample of studies on this topic:

- Smithson (1994), in a case study approach, studied 22 graduate students with a semester-long assignment to produce a report on a given management information systems topic. Searches for information on the topic were performed by an unspecified number of intermediaries using online databases. In order to observe differences in judgments at different stages (initial, final, citing) and among different cases Smithson had the users judge a combined total of 1406 documents for relevance at the initiation and completion stages of the case. He found that 82% of the documents relevant in the initial stage were relevant in the final stage; 12% of the initially relevant documents were cited, but there was a large individual difference among cases.
- Bruce (1994) observed an unreported number of graduate students during three stages of search and retrieval (before, during, after) in relation to their coursework in order to study cognitive changes that occur during IR interaction.
- Wang and White (1995) interviewed 25 faculty and graduate students (not distinguished as to number) about relevance decisions they made concerning documents in the course of their research in order to identify relevance criteria used in early and later stages of the subjects' research. They identified 11 criteria in the early stages and another eight in the later stages of research.
- Tang and Solomon (1998) observed one graduate student in two sessions during the process of retrieving information for a term paper in order to study the evolution of relevance judgments.
- Bateman (1998) studied 35 graduate students during six different information seeking stages in respect to a research paper for their class. The students were asked to rate the importance of 40 relevance criteria in different stages in order to determine whether the criteria change at different stages. She found the criteria were fairly stable across stages.
- Vakkari and Hakala (2000) and Vakkari (2001) studied 11 students over a term taking a course on preparing a research proposal for a master's thesis. They observed the students' search results and relevance judgments at the beginning, middle, and final phases of their work in order to study changes in their relevance assessment. The share of relevant references declined from 23% in the initial phase to 11% in the middle and 13% in the final phase. They identified 26 criteria in six groups. They found that the distribution of criteria changed only slightly across phases.
- Tang and Solomon (2001) report on two studies: In the first, 90 undergraduate students who were given an assignment and 20 documents first as a bibliographic citation (called stage 1) and then full text (called stage 2) were asked to evaluate their relevance for the assignment; in the second study, nine graduate students who searched for documents to support their own

research also were evaluated at stages 1 and 2 in order to identify patterns in change in their use of criteria in the two studies and at different stages (i.e., from representations to full text). They found that there were dynamic changes in users' mental model (criteria) of what constitutes a relevant document across stages.

- Anderson (2005) observed two academics involved in scholarly research over a period of 2 years in order to explore relevance assessments as part of the decision-making process of individuals doing research over time. She identified 20 categories in 10 groups that users focused on in making relevance judgments. Three of the groups relate to determining the appropriateness of information and seven to shaping boundaries to a topic.

C. Relevance Feedback

What factors affect the process of relevance feedback? A short explanation of relevance feedback (RF) from the human perspective: I find a relevant document, go through it and, on the basis of something in that document, go on and re-formulate my search or identify something else that I should consult. In IR, RF is a technique aiming at improving the query being searched using terms from documents that have been assessed as relevant by users (manual RF), or by some algorithm, such as using terms from top-ranked retrieved documents (automatic RF). Manual RF has a long history in search practices by professionals and users, while automatic RF has a long history in IR evaluation. Of interest here are not the means and ways of either manual or automatic RF in IR, but the behavior of people when involved in RF.

- Spink and Saracevic (1997) used search logs and interaction transcripts from a study that involved 40 mediated searches done by four professional intermediaries on DIALOG databases in response to real information needs in order to analyze the nature of feedback involving users, intermediaries, searches, and results. The users judged 6225 retrieved documents as to relevance. The researchers identified 885 feedback loops grouped in five categories depicting different types of feedback.
- Jansen *et al*. (2000) analyzed logs of 51,423 queries posed by 18,113 users on the Excite search engine in order to determine a number of query characteristics, including the incidence of RF. They found that 5% of queries used RF.
- Quiroga and Mostafa (2002) studied 18 graduate students who searched a collection of 6000 records in consumer health on a system with various feedback capabilities. The researchers provided a verbal protocol of proceedings in order to categorize factors that influence RF assessments. They identified 15 factors in four categories related to users and three categories of factors related to documents.
- Ruthven *et al*. (2003) used 15 undergraduate and 15 graduate students to search six simulated search topics on an experimental and a control system in five experiments in which they assessed retrieved documents as to relevance in order to examine the searchers' overall search behavior for possibilities of incorporating manual RF into automatic RF. They found, among other things, that users are more satisfied when RF was available, and that their search was more effective. This is really an IR systems study, but it is included here to show the human side investigated.

D. Summary

Caveats abound. Numerous aspects of the studies reviewed can be questioned and criticized. Easily! Criteria, measures, and methods used in these studies are not standardized. While no study was an island, each study was done more or less on its own. As to the population of users, students were the primary target and studied ad nausea—but more about that later. Thus, the results are hardly generalizable. Still, it is really refreshing to see conclusions made on basis of data, rather than on basis of examples, anecdotes, authorities, or contemplation. Summary conclusions below are derived from the studies reviewed and should be really treated as hypotheses.

Relevance clues. Clues studies inevitably involved classification; their results were categories of criteria used by users or factors affecting users in inferences about relevance, including different characteristics of information objects. Classification schemes and category labels more or less differed from study to study. However, the most important aspect of the results is that the studies independently observed a remarkably similar or equivalent set of relevance criteria and clues. With all the caveats, here are some generalizations to be treated as hypotheses:

- Criteria used by a variety of users in inferring relevance of information or information objects are finite in number and the number is not large; in general, criteria are quite similar despite differences in users. *Different users = similar criteria.*
- However, the weight (importance) different users assign to given criteria differs as to tasks, progress in task over time, and class of users. For instance, children assign little or no importance to authority, while faculty assigns a lot. *Different users, tasks, progress in tasks, classes of users = similar criteria = different weights.*
- While there is no wide consensus, on a general level, clues and associated criteria on which basis users make relevance inferences may be grouped as to:
  - *Content*: Topic, quality, depth, scope, currency, treatment, clarity.
  - *Object*: Characteristics of information objects (e.g., type, organization, representation, format, availability, accessibility, costs).
  - *Validity*: Accuracy of information provided, authority, trustworthiness of sources, verifiability.
  - *Use or situational match*: Appropriateness to situation, or tasks, usability, urgency; value in use.
  - *Cognitive match*: Understanding, novelty, effort.
  - *Affective match*: Emotional responses to information, fun, frustration, uncertainty.
  - *Belief match*: Credence given to information, acceptance as to truth, reality, confidence.
- These groups of criteria are *not* independent of each other. People apply multiple criteria in relevance inferences and they are used interactively.
- The interaction is between information (or object) characteristics (top three above) and individual (or human) characteristics (bottom four). This is posited in the stratified model presented in the *Models of Relevance* section.
- Content-oriented criteria seem to be most important for users. However, as pointed out, they interact with others. In other words, criteria related to content, including topical relevance, are

rated highest in importance, but interact with other criteria—they are not the sole criteria:
  - However, when assessing the use of search outputs, *the value of search results as a whole* seems to be the critical criterion that users apply in making relevance inferences on retrieved information objects.
- Criteria used for assigning different ratings (e.g., relevant, partially relevant, not relevant) are substantially (but not completely) similar. However, the weight (could be positive or negative) assigned to a given criterion differs depending on the rating—for example weight for the same criterion on a document judged relevant differs from the weight of a document judged not relevant. *Different ratings of relevance = similar criteria = different weights.*
- Similarly, while the criteria are similar, the importance of criteria changes from the presentation of document representations to the presentation of full text. Some become more important, some less—no clear pattern has emerged.
- Of all document representations (excluding full text), titles and abstracts seem to produce the most clues.
- Visual information provides clues that make for a faster inference than textual information does.

Dynamics. Ultimately, dynamic studies involved observing changes over time, even though time itself was not involved directly in any of the studies as a variable. Some things indeed change over time, while others stay relatively constant.

- For a given task, it seems that the user's inferences about specific information or information object are dependent on the stage of the task.
- However, the user's criteria for inferences are fairly stable. As the time and the work on the task progress, users change criteria for relevance inferences, but not that much. The user's selection of given information or information objects changes—there is a difference. Also, the weight given to different criteria may change over stages of work. *Different stages = differing selections* but *different stages = similar criteria = different weights*.
- As time progresses and a task becomes more focused, it seems that the discriminatory power for relevance selection increases. *Increased focus = increased discrimination = more stringent relevance inferences*.
- As to criteria, user perception of topicality seems still to be the major criterion, but clearly not the only one in relevance inferences. However, what is topical changes with progress in time and task.

Relevance feedback. Human feedback studies reported here inevitably involved IR systems and search results; however, concentration was on how people behaved in relation to feedback:

- Human RF involves several manifestations in addition to commonly used search term feedback, it includes content, magnitude, and tactics feedback.
- Users seem to be more satisfied with systems in which they can incorporate their RF; when they use RF, retrieval performance increases. This is valid for laboratory systems and conditions. *Use of RF = increase in performance*:
  - However, when RF is available in real-life systems and conditions, users tend to use RF very sparingly—RF is not used that much.
- Searching behavior using RF is significantly different than when not using it as reflected in relevance assessments, selection of documents, time used, and ways of interaction:
  - However, criteria used in RF are similar to (or even a subset of) criteria used in relevance inferences in general.

IX. Effects of Relevance: Or Rather, What Influences Are Related To Relevance Judges and Judgments

It works both ways: Relevance is affected by a host of factors and, in turn, it affects a host of factors as well. A number of studies addressed questions about effects or variables concerning relevance judges and judgments. The synthesis below is organized along these questions. Of course, factors in these categories are interdependent, as is everything with relevance.

As in the preceding section, I will concentrate *exclusively* on observational, empirical, or experimental studies, that is, on works that contained some kind of data directly addressing relevance. Works that discuss or review the same topics but do not contain data are *not* included, with a few exceptions in order to provide context. Where appropriate, some summaries include numerical results. Main results from all studies, with a number of caveats, are synthesized and generalized at the end of the section.

A. Relevance Judges

What factors inherent in relevance judges make a difference in relevance inferences? A similar question was investigated in relation to a number of information-related activities, such as indexing and searching. Not many studies addressed the question in relation to relevance, and those that did concentrated on a limited number of factors, mostly involving the effects of expertise:

- Regazzi (1988) asked 32 judges, researchers, and students (but numbers for each group are not given), to rate as to relevance 16 documents in alcohol studies to a given topic in order to compare differences in relevance ratings, perceived utility and importance of document attributes and also to ascertain effects of various factors, such as learning during the process.
- Gluck (1995, 1996) used 82 subjects (13 high school students, three with associate's degrees, 41 with or working on bachelor's degrees, 19 with or working on master's degrees and six with or working on Ph.D. degrees) to: (i) respond to an unspecified set of geography-related questions using two packets of geographic materials and (ii) recall their recent experience where geographic questions were raised with responses coded by two coders on a five-point relevance scale in order to study the effects of geographic competence and experience on relevance inferences (1995 study) and compare user relevance and satisfaction ratings (1996 study).
- Dong *et al*. (2005) asked a physician (whose assessment was considered the "gold standard"), six evaluators with biology or medical backgrounds, and six without such backgrounds to assess for relevance 132 Web documents retrieved by a meta-crawler in relation to specific medical topics in order to measure variation in relevance assessments due to their domain knowledge and develop a measure of relevance similarity.
- Hansen and Karlgren (2005) used eight students and 20 professionals with a variety of academic backgrounds whose first language was Swedish and were fluent in English to search a newspaper database according to several simulated scenarios serving as queries with results presented in Swedish and English in order to investigate how judges assess the relevance of retrieved documents in a foreign language, and how different scenarios affect assessments.

1. Individual Differences

How large are and what affects individual differences in relevance inferences? Individually (and not at all surprisingly), people differ in relevance inferences, just as they differ in all other cognitive processes in general, and involving information in particular:

- Davidson (1977) presented 25 engineering and 23 social sciences students with a given question in their area and asked them to assess the relevance of 400 documents in order to study individual differences related to variables of expertise and information openness—the individual's cognitive repertoire as indicated by various scales—open-mindedness, control, rigidity, width.
- Saracevic and Kantor (1988a, b) used five professional searchers each to search 40 questions, posed by 40 users (19 faculty, 15 graduate students and six from industry) with real information needs. Their pooled results were presented to the users for relevance assessment in order to observe the overlap in retrieval of relevant documents among different searchers. They found that the overlap in retrieval of relevant documents among the five searchers was 18%.

B. Relevance Judgments

What factors affect relevance judgments? A short answer: a lot of them. In a comprehensive review of relevance literature, Schamber (1994) extracted 80 relevance factors grouped into six categories, as identified in various studies. She displayed them in a table. In another table, Harter (1996) extracted 24 factors from a study by Park (1993) and grouped them in four categories. A different approach is taken here. Rather than extracting still another table, I summarize various studies that tried to pinpoint some or other factor affecting relevance judgments organized on the basis of assumptions made in IR evaluations. The goal is not to prove or disprove the assumptions, but to systematize a wide variety of research questions for which some data has been obtained.

When it comes to relevance judgments, the central assumption in any and all IR evaluations using Cranfield and derivative approaches, such as TREC, has five postulates assuming that relevance is:

1. *Topical*: The relation between a query and an information object is based solely on a topicality match.
2. *Binary*: Retrieved objects are dichotomous, either relevant or not relevant—even if there was a finer gradation, relevance judgments can be collapsed into a dichotomy. It implies that all relevant objects are equally relevant and all non-relevant ones are equally non-relevant.
3. *Independent*: Each object can be judged independently of any other; documents can be judged independently of other documents or of the order of presentations.
4. *Stable*: Relevance judgments do not change over time; they are not dynamic. They do not change as cognitive, situational, or other factors change.
5. *Consistent*: Relevance judgments are consistent; there is no inter- or intra-variation in relevance assessments among judges. Even if there are, it does not matter; there is no appreciable effect in ranking performance.

A sixth, or *completeness*, postulate can be added for cases where only a sample of the collection (rather than the whole collection) is evaluated as to relevance (such as when only pooled retrievals are evaluated). This postulate assumes that the sample represents all relevant objects in the collection—no relevant objects are left behind. Zobel (1998) investigated the issue of completeness in relation to the TREC pooling method; however, since this is really a question for IR evaluation methods, rather than relevance judgments, the completeness postulate is not treated further here.

These are very restrictive postulates, based on a highly simplified view of relevance—it is a variation on the theme of *weak relevance*, as defined in section *Manifestation of Relevance*. The postulates are stringent laboratory assumptions, easily challenged. In most, if not all laboratory investigations in science, things are idealized and simplified in order to be controlled; IR evaluation followed that path. In a scathing criticism of such assumptions about relevance in IR evaluation, supported by empirical data from a number of studies, Harter (1996) pointed out that this view of relevance does not take into account a host of situational and cognitive factors that enter into relevance assessments and that, in turn, produce significant individual and group disagreements. However, using this weak view of relevance over decades, IR tests were highly successful in a sense that they produced numerous advanced IR procedures and systems. By any measure, IR systems today are much, much better and diverse than those of some decades ago. IR evaluation, with or despite of its weak view of relevance, played a significant role in that achievement.

Harter was not the only critic; the debate has a long history. These postulates produced no end of criticism or questioning of the application of relevance in IR tests from both the system's and the user's point of view, starting with Swanson (1971) and Harter (1971) and continuing with Robertson and Hancock-Beauleiu (1992), Ellis (1996), Harter (1996), Zobel (1998), and others. This review is not concerned with IR systems, including their evaluation, thus the arguments are not revisited here. But the postulates also served as research questions for a number of experimental or observational studies that investigated a variety of related aspects. These are synthesized here, organized along the postulates.

1. Beyond Topicality

Do people infer relevance based on topicality only? This question was treated in the preceding sections at length, thus, it is rehashed only briefly here. The question is one of the postulates in the central assumption for IR evaluation. Short conclusion: seems not. Topicality plays an important, but not at all an exclusive, role in relevance inferences by people.

A number of other relevance clues or attributes, as enumerated in the summary of *Behavior of Relevance*, enter into relevance inferences. They interact with topicality as judgments are made.

Only a few observational studies directly addressed the question, among them:

- Wang and Soergel (1998) provided 11 faculty and 14 graduate students with printouts of search results from DIALOG containing a total of 1288 documents retrieved in response to the information needs related to their projects (with no indication as who did the searches) and asked them to select documents relevant to their need in order to assess and compare user criteria for document selection. They identified 11 criteria for selection, with topicality being the top criterion followed by orientation, quality, and novelty as most frequently mentioned criteria.
- Xu and Chen (2006) asked 132 students (97% undergraduate and 3% graduate) to search the Web for documents related to one of the four prescribed search topics or a search topic of their interest, and then choose and evaluate two retrieved Web documents, thus analysis included 264 evaluated documents. The study was done in order to test five hypotheses, each specifying that a given criterion has a positive association with relevance. They found that topicality and novelty were the two most significant criteria associated with relevance, while reliability and understandability were significant to a smaller degree and scope was not significant. This is the only study that did hypothesis testing as to relevance criteria; others provided either frequency counts or description only.

2. Beyond Binary

Are relevance inferences binary, that is relevant—not relevant? If not, what gradation do people use in inferences about relevance of information or information objects? The binary premise was immediately dismissed on the basis of everyday experience. Thus, investigators went on to study the distribution of relevance inferences and the possibility of classifying inferences along some regions of relevance:

- Eisenberg and Hue (1987) used 78 graduate and undergraduate students to judge 15 documents in relation to a stated information problem on a continuous 100-mm line in order to study the distribution of judgments and observe whether the participants perceived the break point between relevant and non-relevant at the midpoint of the scale.
- Eisenberg (1988) used 12 academic subjects (unnamed whether students or faculty) with "real" information needs to judge the relevance of retrieved "document descriptions" to that need (quotes in the original) in order to examine the application of magnitude estimation (an open-ended scaling technique) for measuring relevance and to compare the use of magnitude scales with the use of category scales.
- Janes (1991a) replicated the Eisenberg and Hue (1987) study by using 35 faculty, staff, and doctoral students (not distinguished as to numbers) to judge the relevance of retrieved document sets in response to their real information need in order to determine the distribution of judgments on a continuous scale.
- Su (1992) used 30 graduate students, nine faculty and one staff as end users with real questions for which online searches were done by six intermediaries. She had the users indicate the success of retrieval using 20 measures in four groups in order to determine whether a single

measure or a group of measures reflecting various relevance criteria is/are the best indicator of successful retrieval.

- Janes (1993) rearranged relevance judgment data from two older studies (Cuadra *et al.*, 1967; Rees and Schultz, 1967) and from two of his own studies with 39 faculty and doctoral students used in the first study and 33 students and 15 librarians in the second, along the scales they used in the studies in order to investigate the distribution of relevance judgments.
- Greisdorf and Spink (2001) used 36 graduate students in three studies, who in 57 searches related to their personal or academic information need, retrieved 1295 documents. The students were asked to indicate relevance assessments using various scales and criteria in order to investigate the frequency distribution of relevance assessments when more than binary judgment is used.
- Spink and Greisdorf (2001) used 21 graduate students who, in 43 searches related to their academic information need, retrieved 1059 documents. The students were asked to indicate relevance assessments using various scales and criteria in order to investigate the distribution of relevance assessments along various regions of relevance—low, middle, and high end of judgments as to relevance.
- Greisdorf (2003) used 32 graduate students who, in 54 searches related to their personal or academic information needs, retrieved 1432 documents in response. The students were asked to assess their results using a number of relevance criteria on a continuous relevance scale in order to study the users' evaluation as related to different regions of relevance.

3. Beyond Independence

When presented for relevance judging, are information objects assessed independently of each other? Does the order or size of the presentation affect relevance judgments? The independence question also has a long history of concern in relevance scholarship. In a theoretical, mathematical treatment of relevance as a measure, Goffman (1964) postulated that relevance assessments of documents depend on what was seen and judged previously, showing that, in order for relevance to satisfy mathematical properties of a measure, the relationship between a document and a query is necessary but not sufficient to determine relevance; the documents' relationship to each other has to be considered as well. Several papers discussing the issue followed, but only at the end of 1980s did the question start receiving experimental treatment:

- Eisenberg and Barry (1988) conducted two experiments, first with 42 graduate students, and then with 32. The subjects were provided with a query and 15 document descriptions as answers ranked in two orders: either high to low relevance or low to high relevance. Each subject was given one of the orders, using in the first experiment a category rating scale, and in the second, a magnitude rating in order to study whether the order of document presentation influences relevance scores assigned to these documents.
- Purgaillis and Johnson (1990) provided approximately (their description) 40 computer science students who had queries related to class assignments with retrieved document citations that were randomly "shuffled" for relevance evaluation in order to study whether there is an order presentation bias.
- Janes (1991b) asked 40 faculty and doctoral students (numbers for each group not given) with real information requests to judge the relevance of answers after online searches by intermediaries. Answers were given in different formats (title, abstract, indexing) in order to

examine how users' relevance judgments of document representation change as more information about documents is revealed to them.

- Huang and Wang (2004) asked 19 undergraduate and 29 graduate students to rate the relevance of a set of 80 documents to a topic presented in a random order in the first phase and then sets of 5 to 75 documents presented from high to low and low to high relevance in the second phase in order to examine the influence of the order and size of document presentation on relevance judgments.

4. Beyond Stability

Are relevance judgments stable as tasks and other aspects change? Do relevance inferences and criteria change over time for the same user and task, and if so how? The question is treated in preceding section under relevance dynamics, thus not rehashed again. Short answer: Relevance judgments are not completely stable; they change over time as tasks progress from one stage to another and as learning advances. What was relevant then may not be necessarily relevant now and vice versa. In that respect Plato was right: Everything is flux. However, criteria for judging relevance are fairly stable.

5. Beyond Consistency

Are relevance judgments consistent among judges or group of judges? Many critics of IR evaluation or of any relevance application had a ball with this question, pointing out easily observed inconsistencies. However, human judgments about anything related to information are not consistent in general, and relevance judgments are no exception. Why should they be?

The great-granddaddy of all studies that put some data to the question and opened a Pandora's box was done at the very dawn of IR development in the 1950s. Gull (1956), in a study that is also a classic example of the law of unintended consequences, showed not only that relevance inferences differ significantly among groups of judges, but also inadvertently uncovered a whole range of issues that IR evaluation struggles with to this day. Actually, consistency of relevance judgments was not the purpose of the study at all. But IR evaluation was. The results are worth recalling. Gull reported on a study whose goal was to compare two different and competing indexing systems—one developed by the Armed Services Technical Information Agency (ASTIA) using subject headings, and the other by a company named Documentation Inc., using uniterms (index terms searched in a Boolean manner). In the test, each group searched 98 requests using the same 15,000 documents, indexed separately, in order to evaluate performance based on relevance of retrieved documents. *However, each group judged relevance separately.* Then, not the system's performance, but their relevance judgments became contentious. The first group found that 2200 documents were

relevant to the 98 requests, while the second found that 1998 were relevant. There was not much overlap between groups. The first group judged 1640 documents relevant that the second had not, and the second group judged 980 relevant that the first had not. You see where this is going. Then they had a reconciliation and considered each others' relevant documents and again compared judgments. Each group accepted some more as relevant, but at the end, they still disagreed; their rate of agreement, even after peace talks, was 30.9%. That did it. The first ever IR evaluation did not continue. It collapsed. And it seems that the rate of agreement hovers indeed around that figure. The corollary that IR evaluators learned: *Never, ever use more than a single judge per query.* They don't.

Only a few consistency studies were done:

- Haynes *et al*. (1990) did not intend to study consistency, but rather to assess MEDLINE use in a clinical setting. However, their report does include data from which consistency rates can be derived. They used 47 attending physicians and 110 trainees who retrieved 5307 citations for 280 searches related to their clinical problem, and assessed the relevance of the retrieved citations. Authors then used two other search groups of 13 physicians experienced in searching and three librarians to replicate 78 of those searches where relevance was judged by a physician with clinical expertise in the topic area in order to compare retrieval of relevant citations according to expertise. For the replicated searches, all searcher groups retrieved some relevant articles, but only 53 of the 1525 relevant articles (3.5%) were retrieved by all three search groups. This is the only real-life study on the question.
- Shaw *et al*. (1991) used four judges to assess the relevance of 1239 documents in the cystic fibrosis collection to 100 queries. Judged documents were divided into four sets: A from query author/researcher on the subject, B from nine other researchers, C from four postdoctoral fellows, and D from one medical bibliographer, in order to enable performance evaluations of different IR representations and techniques using any or all of the judgment sets. The overall agreement between judgment sets was 40%.
- Janes and McKinney (1992) used a previous study (Janes, 1991b) from which they selected relevance assessments by four students as users with information requests. The students judged two sets of retrieved documents that differed in the amount of information presented (primary judges) and then used four undergraduate students without and four graduate students with searching expertise (secondary judges) to re-judge the two sets in order to compare changes in judgments due to increase in provided information between primary and secondary judges. The overlap in judgment of relevant documents (calculated here as sensitivity) between all secondary judges and primary judges was 68%.
- Janes (1994) used 13 students inexperienced in searching, 20 experienced student searchers and 15 librarians to re-judge 20 documents in each of two topics that were previously judged as to relevance by users in order to compare users' vs. non-users' relevance judgments. The overall agreement in ratings between original users' judgments and judgments of the three groups was 57% and 72% for the respective document sets.
- Sormunen (2002) used nine master's students to reassess 5271 documents already judged on relevance in 38 topics in TREC-7 and -8 on a graded four-point scale (as opposed to a binary scale used in TREC) in order to compare the distribution of agreement on relevance judgment between original TREC and newly reassessed documents and seek resolution in cases of disagreement. He found that 25% of documents rated relevant in TREC were rated not

relevant by the new assessors; 36% of those relevant in TREC were marginally relevant; and 1% of documents rated not relevant in TREC were rated relevant.

- Vakkari and Sormunen (2004) used 26 students to search four TREC-9 topics that already had pre-assigned relevance ratings by TREC assessors on a system that provided interactive RF capabilities, in order to study the consistency of user identification of relevant documents as predefined by TREC and possible differences in retrieval of relevant and non-relevant documents. They found that the student users identified 45% of items judged relevant by TREC assessors.

6. But Does It Matter?

How does inconsistency in human relevance judgments affect results of IR evaluation? Aforementioned critics of IR evaluation posited, among other things, that because of inconsistency in human relevance judgments, the results of IR evaluations dependent on stated judgments are suspect. Again Harter (1996, p. 43):

> Researchers conducting experimental work in information retrieval using test collections and relevance assessments *assume* that Cranfield-like evaluation models produce meaningful results. But there is massive evidence that suggest the likelihood of the contrary conclusion.

How do you evaluate something solely on the basis of human judgments that are not stable and consistent? This is a perennial question, even a conundrum, for any and all evaluations based on human decisions that by nature are inconsistent, way above and beyond IR evaluation.

As far as I can determine there are only five studies in some four decades that addressed the issue. They are modeled on the first and often cited Lesk and Salton (1968), study that had actual data on the complaint voiced by critics. Four of the five studies had also data that show the magnitude of agreements/disagreements on relevance judgments, thus can also be used as consistency studies:

- Lesk and Salton (1968) used eight students or librarians (not specified as to which) who posed 48 different queries to the SMART system containing a collection of 1268 abstracts in the field of library and information science, to assess the relevance of those 1268 documents to their queries (called the A judgments). Then a second, independent set of relevance judgments (B judgments) was obtained by asking each of the eight judges to assess for relevance six additional queries not of his/her own in order to rank system performance obtained using four different judgments sets (A, B, their intersection and union). They found that the overall agreement between original assessors (A) and eight new assessors (B) was 30% and concluded after testing three different IR techniques that all sets of relevance judgments produce stable performance ranking of the three techniques.
- Kazhdan (1979) took the findings from the Lesk and Salton (1968) study as a hypothesis and used a collection of 2600 documents in electrical engineering that had 60 queries with two sets of relevance judgments—one from a single expert and the other from a group of 13 experts—in evaluating seven different document representations in order to compare the

performance of different representations in relation to different judgment sets. He found that Lesk and Salton hypothesis is confirmed: the relative ranking of the seven different representations remained the same over two sets of judgments; however, there was one exception where ranking changed.

- Burgin (1992) used a collection of 1239 documents in the cystic fibrosis collection (Shaw *et al*. 1991, synthesized above) that had 100 queries with four sets of relevance judgments in the evaluation of six different document representations in order to compare performance as a function of different document representations and different judgment sets (as mentioned, the overall agreement between judgment sets was 40%). He found that there were no noticeable differences in overall performance averaged over all queries for the four judgment sets; however, there were many noticeable differences for individual queries.
- Wallis and Thom (1996) used seven queries from the SMART CACM collection of 3204 computer science documents (titles and in most cases, abstracts) that already had relevance judgments by SMART judges in order to compare two retrieval techniques. Then two judges (paper authors, called judges 1 and 2) assessed separately 80 pooled top-ranked retrieved documents for each of seven queries in order to rank system performance using three different judgments sets (SMART, intersection, and union of judges 1 and 2). They found that the overall agreement between original assessors (SMART) and two new assessors (judges 1 and 2) on relevant documents was 48%. After testing two different IR techniques they concluded that the three sets of relevance judgments do not produce the same performance ranking of the two techniques, but the performance figures for each technique are close to each other in all three-judgment sets.
- Voorhees (2000) (also in Voorhees and Harman, 2005, p. 44, 68–70) reports on two studies involving TREC data. (*Reminder*: A pool of retrieved documents for each topic in TREC is assessed for relevance by a single assessor, the author of the topic, called here the primary assessor.) In the first study, two additional (or secondary) assessors independently re-judged a pool of up to 200 relevant and 200 non-relevant documents as judged so by the primary assessor for each of the 49 topics in TREC-4; then the performance of 33 retrieval techniques was evaluated using three sets of judgments (primary, secondary union, and intersection). In the second study, an unspecified number of assessors from a different and independent institution, Waterloo University, judged more than 13,000 documents for relevance related to 50 TREC-6 topics; next, the performance of 74 IR techniques was evaluated using three sets of judgments (primary, Waterloo union, and intersection). Both studies were done in order to look at the effect of relevance assessments by different judges on the performance ranking of the different IR techniques tested. She found that in the first study, the mean overlap between all assessors (primary and secondary) was 30%, and in the second study, 33%. After testing 33 different IR techniques in the first and 74 in the second test, she concluded: "The relative performance of different retrieval strategies is stable despite marked differences in the relevance judgments used to define perfect retrieval" (Voorhees, 2000, p. 714). Swaps in ranking did occur but the probability of the swap was relatively small.

C. Summary

Caveats abound again and for the same reasons mentioned in the summary of the previous section. While similar or even identical research questions were asked in a number of studies, the criteria and methodologies differed so widely that general conclusions offered below are no more than possible hypotheses.

Judges. A very limited number of factors related to relevance judges were studied. This is in sharp contrasts to a much large number of factors studied in various studies of indexers and searchers (e.g., Saracevic and Kantor, 1988):

- Subject expertise seems to be one variable that accounts strongly for differences in relevance inferences by group of judges—*higher expertise = higher agreement, less differences.*
- Lesser subject expertise seems to lead to more lenient and relatively higher relevance ratings—*lesser expertise = more leniency in judgment.*
- Relevance assessment of documents in a foreign language (for native speakers who are fluent in that language) is more time consuming and taxing. Assessment agreement among judges across languages differs; it is lower when assessing foreign language documents.

1. Individual Differences

- A relatively large variability can be expected in relevance inferences by individuals. Individual differences are the, if not THE, most prominent feature and factor in relevance inferences.
- However, the differences are comparable to individual differences in other cognitive processes involving information processing, such as in indexing, classifying, searching, feedback, and so on (Saracevic, 1991).

2. Judgments

- Relevance is measurable—this is probably the most important general conclusion from all the studies containing data.
- None of the five postulates in the central assumption of IR testing holds:
  - However, using these postulates (representing a simplified or weak view of relevance) in a laboratory evaluation over the years produced significant improvements in IR techniques.
- What is relevant depends on a number of factors, but the artifact of relevance inferences can be expressed by users on a variety of measures.
- Users do not use only binary relevance assessments, but infer relevance of information or information objects on a continuum and comparatively:
  - However, even though relevance assessments are not binary they seem to be bimodal: high peaks at end points of the range (not relevant and relevant) with smaller peaks in the middle range (somewhat not relevant or relevant). The highest peak is on the non-relevant end.
  - Following that, relevance judgments may be subdivided into regions of low, middle, and high relevance assessments, with middle being the flattest part of the distribution.
- Different document formats (title, abstract, index terms, full text) have an effect on relevance inferences. Relevance judgments do change as information is added, such as from titles, to abstracts, to additional representations. Titles seem to be not as important as abstracts and full texts.
- The order in which documents are presented to users seems to have an effect:
  - It seems that documents presented early have a higher probability of being inferred as relevant.
  - However, when a small number of documents are presented, order does not matter.

- Subject expertise affects consistency of relevance judgments. Higher expertise = higher consistency = more stringent. Lower expertise = lower consistency = more encompassing.
- Different search request scenarios make a difference in the relevance assessment process as to time but seem not to affect the degree of agreement. *Longer scenarios = more time spent in assessment; all scenarios = similar degree of agreement among judges.*
- A complex set of individual cognitive, affective, situational, and related variables is involved in individual differences. As of now, we know little about them and can only barely account (beyond hypotheses) for sources of variability.

3. Consistency

- The inter- and intra-consistency or overlap in relevance judgments varies widely from population to population and even from experiment to experiment, making generalizations particularly difficult and tentative:
  - In general, it seems that the overlap using different populations hovers around 30%.
  - However, it seems that higher expertise and laboratory conditions can produce an overlap in judgments up to 80% or even more. The intersection is large.
  - With lower expertise the overlap drops dramatically. The intersection is small.
  - Whatever the overlap between two judges, when a third judge is added it falls, and with each addition of a judge it starts falling dramatically. Each addition of a judge or a group of judges reduces the intersection dramatically:
    - For instance, it seems that the overlap in retrieval of relevant documents by five different professional searchers when searching the same question drops to under 20%, where pair-wise comparisons were much higher.
  - *Higher expertise = larger overlap. Lower expertise = smaller overlap. More judges = less overlap.*
- In evaluating different IR systems under laboratory conditions, disagreement among judges seems not to affect or affects minimally, the results of relative performance among different systems when using *average* performance over topics or queries. The conclusion is counter intuitive, but a small number of experiments bear it out. So far, evaluators seem right and critics wrong:
  - Rank order of different IR techniques seems to change minimally, if at all, when relevance judgments of different judges, averaged over topics or queries, are applied as test standards.
  - However, swaps—changes in ranking—do occur with a relatively low probability. The conclusion of no effect is not universal.
  - *Different judges = same relative performance (on the average).*
  - However, performance ranking over *individual* topics or queries differs significantly depending on the topic and not on the IR technique tested:
    - In that respect, note the use of averaging performance in rankings (or even using averages of averages) that in itself has an affect on results.

4. Measures

- Users are capable of using a variety of scales, from categorical to interval, to indicate their inferences.
- However, the type of scales or measures used for recording relevance inferences seems to have an effect on the results of measurement. There is no one "best" scale or measure.
- It seems that magnitude estimation scales are appropriate for judging relevance; they may be less influenced by potential bias than category scales. However, they are difficult to explain and analyze.

5. Reflection on Approach

The pattern used in this and the previous section to synthesize studies ([author] used [subjects] to do [tasks] in order to study [object of research]) comes from the studies themselves. For a great many studies, this means that certain stimuli were given to subjects in order to study resulting responses. Stimulus-response studies were the hallmark of behaviorism, an approach in psychology, championed by B. F. Skinner (1904–1990) that dominated psychology from the 1930s until the 1960s. It is based on a notion that human behavior can be studied experimentally without recourse to consideration of mental states, from the theory that there is a predictable pattern between stimulus and response in the human brain. Various schools of behaviorism developed and numerous stimulus-response studies did and still do provide valuable insight into human behavior. However, because of many shortcomings in underlying notions, assumptions and methodological approaches, behaviorism fell out of favor. The basic problem is that behaviorism does not include diagnostics beyond responses. Modified behaviorism methodologies were absorbed in cognitive psychology.

Many relevance behavior and effect studies were and still are based on behaviorism. Not all, but a great many. These produced black-box experiments where systems and users are treated as a whole, inputs controlled, and outputs observed and evaluated. In the ultimate black-box experiment, only inputs and outputs are visible and relevance is inferred on the basis of some action on the part of a user or simulated user. How come? Behaviorism and related methods were imported to relevance studies through experiments carried by the hallmark relevance studies of Rees and Schultz (1967) and Cuadra *et al*. (1967). Of the four principal investigators in those studies, three were psychologists (Douglas Schultz, Carlos Cuadra, and Robert Katter); the background of the fourth, Alan Rees, was English literature. Following behaviorism as the major approach in psychology at the time, they applied related stimulus-response methodologies, including underlying assumptions, to the study of relevance. Others followed. In all fairness, in no study can we find a reference to a work in behaviorism proper. But in a great many studies, behaviorism was there with all of its strengths and shortcomings. And in many instances, it still is.

6. Reflection on Population

An overwhelming number of studies on behavior and the effects of relevance used students as the population studied. (Well, we are not alone—in psychology, a large number of studies use students as well.) The reasons are

simple: they are readily available, the cost to involve them is minimal, and so is the effort. In a way, what was studied is *student relevance*. This is not a critique and even less a condemnation of using students as the population in relevance studies. There is nothing wrong in studying student relevance, but it is an open question whether conclusions and generalizations can be extended to other populations in real life. This is another reason why the results of studies should be treated as hypotheses. But even though students predominate as a population, let me repeat: still, it is really refreshing to see conclusions made on the basis of data, rather than on the basis of examples, anecdotes, authorities, or contemplation alone.

X. Epilogue: A Backward and a Forward Look on Relevance Scholarship With Some Suggestions for Research Agenda

IR emerged after the World War II, addressing the problem of the information explosion by using technology as a solution. Many things have changed since, but the basic problem and solution are still with us. The fundamental idea was and still is to retrieve *relevant information* with the help of technology. Thus, relevance became the central notion in information science. As treated in practice, relevance is thoroughly entangled with IT. However, relevance is also a thoroughly human notion and as all human notions, it is somewhat messy. The goal of scholarship on relevance is to make it more understandable and less messy.

Some 30 years ago, I wrote a critical review that synthesized the thinking on the notion of relevance in information science during the preceding decades. This current review is an update; together Parts I and II or the previous and the current review cover the evolution of thinking on relevance since the emergence of information science some six decades ago. The purpose of this review is to trace the evolution of thinking on relevance in information science for the past three decades and to provide an updated, contemporary framework within which the still widely dissonant ideas on relevance may be interpreted and related to one another. I concentrated on scholarship about relevance and did *not* include works dealing with applications in information systems that are geared toward retrieval of relevant information or information objects. Literature on this area is huge, but outside of the scope of this review. This work is about the notion of relevance, not about relevance in information systems.

The framework for organizing this review was derived from the way phenomena and notions are studied in science in general. In science,

phenomena are studied as to their nature, manifestations, behavior and effects. As to the nature of relevance, there has been a marked progress in past decades in the elaboration of its meaning, less marked progress in developing or adapting theories, and considerable diversity in the development of models. I suggested a stratified model as an integrative framework for viewing relevance interactions between users and computers. As regarding manifestations of relevance, a consensus seems to be emerging that there are several kinds of relevance, grouped in a half dozen or so well distinguished classes. They are interdependent when it comes to interaction between people, information, and technology. As to the behavior and effects of relevance, we have seen a number of experimental and observational studies that lifted the discourse about relevance from opinions and insights (as valuable as they are) to interpretation of data and facts. These studies addressed a number of facets of relevance, however and regrettably, generalizations must be taken as hypotheses only, because experimental and observational criteria, standards and methods were all over the place.

Each of the sections concluded with a summary—my own synthesis and interpretation of what was discussed and found. Thus, here I am not providing further summaries from the literature as conclusions. Instead, I am looking at the big picture by analyzing several critical issues and manifested trends that have impacted relevance scholarship in general in the past and, in my opinion, will continue to do so in the near future. I am also suggesting, in broad brushstrokes, problems to be considered for a relevance research agenda.

A. Research Funding

Relevance is poor. Relevance research was funded much better in the past than it is today. Whatever funding exists now is spotty and without an agenda or direction. In the United States, the National Science Foundation (NSF) funded such research way back in the 1960s, but no longer. At that time, NSF funding for relevance research produced, among others, classic experimental studies with results and conclusions that standup to this day (Cuadra *et al*., 1967; Rees and Schultz, 1967). The research agenda of funding agencies concerned with information became completely oriented toward *computers and information* to the exclusion of almost any other issue that has to do with *humans and information*—despite support for periodic workshops that talk about social and human aspects of information systems design. I checked the acknowledgements in the papers on experimental and observational studies reviewed in the preceding two sections. Less than 17% mentioned support by an external granting agency, and of those, about half are from outside the United States.

Over the past three decades, most relevance research has been funded locally, meaning individually at academic institutions, in an old-fashioned way of basement and attic research. Ph.D. students do it in the time tested, solitary way of producing a dissertation, with very limited or no funding. Assistant professors do it on their own on the way to tenure-valued publications. Most of the more comprehensive relevance projects were without budgets—funded as a part of work at local institutions. Relevance is definitively small science in comparison to the big science of information systems.

Because of poor and spotty funding, scholarship on relevance has not progressed in a meaningful, comprehensive and organized manner. As the result, the conclusion that experimental and observational studies were all over the place is not surprising. It seems to me that in the absence of some meaningful funding, progress in relevance scholarship will still be all over the place. The desired merging of the two streams, reflecting users and systems relevance, can hardly produce significant results without funding for relevance research. This does not mean that coming up with bright ideas depends *only* on funding, but it does mean that further exploration nd expansion of bright ideas in today's research environment must be funded.

B. Globalization of IR: Globalization of Relevance

As IR went global, relevance went global. Relevance went to the masses. From the very start of information science in the 1950s, scholarship on relevance was concerned primarily, if not even exclusively, with problems associated with scientific, technical, professional, business, and related information. In a significant way it still is. But things in the real world changed dramatically—new populations, new concerns entered. With the development of the Web and massive search engines starting in the mid-1990s, the public also became increasingly concerned with information in every facet of life in a very similar way. *Relevant* information is desired. The rapid, global spread of information searching is nothing short of astonishing. Millions of users perform untold millions of searches every day all over the globe, seeking the elusive, relevant information. The thirst for relevant information is global, massive, and unquenchable.

As relevance went global and public, a number of questions emerged. To what extent are the results of relevance scholarship—primarily concerned with a restricted and relatively well-defined population and information—applicable to the broad public and every conceivable type of information? A great many fascinating questions worthy of research could be asked. Here are

but a few:

- Are relevance clues similar, different?
- Is relevance behavior similar, different?
- Can the broad public be defined at all as to relevance effects?

It seems that the globalization of relevance also has exposed a need for an additional and different agenda and approach for relevance scholarship.

C. Proprietary IR: Proprietary Relevance

Increasingly, relevance is becoming proprietary because major search engines are proprietary. IR techniques used by a majority of larger search engines are well known in principle, but proprietary and thus unknown in execution and detail.

From anecdotal evidence, we know that proprietary IR systems are very much interested in relevance and that they conduct their own relevance studies. Results are not disseminated in open literature. There may have been (or not) some major advances in understanding relevance behavior and effects from studies done at proprietary systems. After all, they have developed or are trying to develop a number of innovations that include user-in-the-loop technique. For that, they must have studied users. For the most part, we do not know the results of the studies, even though we may observe the innovations themselves.

Relevance research may be developing into a public branch where results are shared freely and widely, and a proprietary branch in which research results, if any, remain secret. One cannot escape the irony of the situation. The Internet and the Web are hailed as free, universal, and democratic, and their very success is directly derived from the fact that they were indeed free, universal, and democratic. Yet, proprietary relevance research is anything but.

D. Research Agenda: Beyond

In several respects, relevance research should go beyond. Here are a few suggested "beyonds".

1. Beyond Behaviorism and Black Box

As described in some detail in the summary of the preceding section, many (not all) relevance studies followed, directly or accidentally, approaches to experimentation used in behaviorism. That is, stimulus and responses were studied, while for the most part people and/or systems were black boxes. We

can gain some understanding this way, but such understanding is generally limited and may easily be biased as well.

Other theoretical bases, assumptions and methods should be explored and implemented more fully. The black-box approach is especially limited and potentially even misleading in results, particularly when systems involved in studying human behavior and effects are a complete black box. Research that is more imaginative involves diagnostics and other non-stimuli variables. It is much harder to do, but more can be learned.

2. Beyond Mantra

Practically every study that dealt with relevance behavior and effects either began or ended (or both) with a statement to the effect that *results have implications for information systems design.* A similar sentiment is repeated in many other relevance papers that vehemently argue that the user viewpoint should be predominant. The overwhelming majority of studies did not go beyond that statement, so the statement became a mantra.

Very little was ever done to actually translate results from user studies into system design. In a way, this is not surprising. The problem is exceedingly difficult theoretically and pragmatically, as demonstrated through the interactive track of TREC, which ran over the course of nine years and conducted experiments with human participation, finding, among other things, that a number of issues need a resolution (Dumais and Belkin, 2005).

However, is the problem of incorporating to a sufficient degree users concerns, characteristics and the like into systems essentially intractable? In other words, is the pessimistic relevance a la Swanson (1986) based on reality? Alternatively, is the optimistic relevance as suggested by the mantra warranted?

I believe that the sentiment beyond the mantra is warranted, but it cannot be realized by the underlying hope that somebody, somehow, somewhere, sometime will actually do it. I believe that systems designs and operations on the one hand, and users on the other, could and should be connected in a much more consequential, involved and direct way than they are now, where the connection is from minimal to none. The interactive track of TREC was on the right track. Among the key items on the agenda is conduct of studies in tandem with system design, such as:

- study of relevance interactions in a variety of manifestations and processes in and beyond retrieval;
- study of cognitive, affective, and situational factors as they dynamically affect relevance and are affected in turn;

- study of human tendencies of least effort for maximum gain as reflected in relevance;
- study of connections between secondary or implied relevance (e.g., as in a decision to retain an information object in some way) and primary or explicit relevance where relevance is actually inferred.

The beyond mantra agenda also means that IR research itself has to go beyond the classical IR model (TREC like), and thus go beyond TREC-like evaluations as done so far, with the one exception I mentioned. Proposals for cognitive IR as advocated, among others, by Ingwersen and Järvelin (2005) are an effort in laying the groundwork for that direction. Relevance research and IR research should at least get engaged, if not married. However, this is highly unlikely to happen without a dowry—without substantial redirection of funding. Namely, the availability of funding has the marvelous ability to change and redirect mindsets and efforts.

However, a word of caution is in order. The problem of building more responsive, complex, and dynamic user-oriented processes and more complex relevance manifestations into IR systems is by no means simple. As Dumais and Belkin (2005) and cohorts discovered, it is hard, tough and consuming, requiring new mindsets, directions, approaches, measures, and methods.

3. Beyond Students

As mentioned, students were endlessly used as experimental subjects for relevance experimentation and observation. Some 70% of studies reviewed in the preceding two sections included students as population studied. Again, this is not surprising. With little or no funding, other populations are much more difficult to reach—actually, the effort is unaffordable. As a result, we are really getting a good understanding of student relevance. Regrettably, we are not getting a good understanding of relevance related to real users, in real situations, dealing with real issues of relevance. If we are to gain a better understanding of relevance behavior and effects in diverse populations, other populations should (or even must) be studied as well. Maybe student relevance is a norm and results could be generalized to other populations, but we do not know.

With relevance going global and reaching a wide diversity of populations the problem becomes more urgent and expansive. We have learned quite a bit about student relevance but, beyond anecdotal evidence and pronouncements of gurus, we really know little about mass relevance (or relevance of, by, and for the people). Relevance research should extend to those populations. However, without funding for such research, students will remain the primary population.

E. In Conclusion

IT and information systems will change in ways that we cannot even imagine, not only in the long run, but even in short term. They are changing and expanding at an accelerated pace. But no matter what, relevance is here to stay.

Acknowledgments

Under its search box, Google Scholar has a cryptic command: "*Stand on the shoulders of giants.*" A few centuries ago, Isaac Newton, referring to Galileo and Kepler, said it better: "*If I have seen further {than certain other men} it is by standing upon the shoulders of giants.*" (Letter to Robert Hooke, February 5, 1675.) And a few centuries before that, in the 12th century, Bernard of Chartres said (as reported by John of Salisbury) it even better: "*We are like dwarfs sitting on the shoulders of giants; we see more things and more distant things than did they, not because our sight is keener nor because we are taller than they, but because they lift us up and add their giant stature to our height*" (Metalogicon, III, 4).

In that spirit I wish to thank the authors synthesized in this review. I stood on their shoulders and saw further.

Thanks to Yuelin Li and Ying Zhang, my assistants, who tirelessly searched the literature for sources about relevance and then organized them. Sandra Lanman's editing and thoughtful suggestions were really relevant. Students in my Spring 2006 Ph.D. class on *Human Information Behavior* were first to read and react to the finished manuscript; their discussion stimulated a number of clarifications and Jeanette de Richemond in particular, provided further, valuable editorial corrections. Their help was most valuable and much appreciated.

I also wish to thank Eileen Abels and Danuta Nitecki, co-editors, *Advances in Librarianship*, who suggested this updated review and coaxed me into doing it.

References

Anderson, T. D. (2005). Relevance as process: Judgements in the context of scholarly research. *Information Research* **10**(2) paper 226. Retrieved Feb. 8, 2006 from http://InformationR.net/ir/10-2/paper226.html

Anderson, A. R., and Belnap, N. D., Jr. (1975). *Entailment: The Logic of Relevance and Necessity*, vol. I. Princeton University Press, Princeton.

Anderson, A. R., Belnap, N. D., Jr., and Dunn, J. M. (1992). *Entailment: The Logic of Relevance and Necessity*, vol. II. Princeton University Press, Princeton.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of American Society for Information Science* **45**(3), 149–159.

Barry, C. L. (1998). Document representations and clues to document relevance. *Journal of American Society for Information Science* **49**(14), 1293–1303.

Barry, C. L., and Schamber, L. (1998). User criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management* **34**(2–3), 219–236.

Bateman, J. (1998). Changes in relevance criteria: A longitudinal study. *Proceedings of the American Society for Information Science* **35**, 23–32.

Bookstein, A. (1979). Relevance. *Journal of the American Society for Information Science* **30**(5), 269–273.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology* **54**(10), 913–925.

Bruce, H. W. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science* **45**(5), 142–148.

Budd, J. M. (2004). Relevance: Language, semantics, philosophy. *Library Trends* **52**(3), 447–462.

Burgin, R. (1992). Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing and Management* **28**(5), 619–627.

Choi, Y., and Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing and Management* **38**(5), 695–726.

Cleverdon, C. W. (1967). The Cranfield tests on index language devices. *Aslib Proceedings* **19**, 173–194.

Cool, C., Belkin, N., and Kantor, P. (1993). Characteristics of texts reflecting relevance judgments. *Proceedings of the 14th Annual National Online Meeting*, pp. 77–84. Learned Information, Medford, NJ.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval* 7(1), 19–37.

Cosijn, E., and Ingwersen, P. (2000). Dimensions of relevance. *Information Processing and Management* **36**(4), 533–550.

Cuadra, C. A., Katter, R. V., Holmes, E. H., and Wallace, E. M. (1967). *Experimental Studies of Relevance Judgments: Final Report*, 3 vols. System Development Corporation, Santa Monica, CA. NTIS: PB-175 518/XAB, PB-175 517/XAB, PB-175 567/XAB.

Davidson, D. (1977). The effect of individual differences of cognitive style on judgments of document relevance. *Journal of the American Society for Information Science* **28**(5), 273–284.

Dervin, B., and Nilan, M. S. (1986). Information needs and uses: A conceptual and methodological review. *Annual Review of Information Science and Technology* **21**, 3–33.

Dong, P., Loh, M., and Mondry, R. (2005). Relevance similarity: An alternative means to monitor information retrieval systems. *Biomedical Digital Libraries* 2(6). Retrieved Jan. 30, 2006 from http://www.bio-diglib.com/content/2/1/6

Dumais, S. T., and Belkin, N. J. (2005). The TREC interactive tracks: Putting the user into search. In *TREC. Experiment and Evaluation in Information Retrieval* (E. M. Voorhees and D. K. Harman, eds.), Chapter 6, pp. 123–145. MIT Press, Cambridge, MA.

Eisenberg, M. B. (1988). Measuring relevance judgments. *Information Processing and Management* **24**(4), 373–389.

Eisenberg, M. B., and Barry, C. (1988). Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science* **39**(5), 293–300.

Eisenberg, M. B., and Hue, X. (1987). Dichotomous relevance judgments and the evaluation of information systems. *Proceedings of the American Society for Information Science* **24**, 66–69.

Ellis, D. (1996). The dilemma of measurement in information retrieval research. *Journal of the American Society for the Information Science* **47**(1), 23–36.

Fidel, R. and Crandall, M. (1997). Users' perception of the performance of a filtering system. *Proceedings of the 20th Annual International Conference on Research and Development in Information Retrieval of the Special Interest Group on Information Retrieval, Association for Computing Machinery (SIGIR 97)*, pp. 198–205.

Fitzgerald, M. A., and Galloway, C. (2001). Relevance judging, evaluation, and decision making in virtual libraries: A descriptive study. *Journal of the American Society for Information Science and Technology* **52**(12), 989–1010.

Froelich, T. J. (1994). Relevance reconsidered—toward an agenda for the 21st century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science* **45**(3), 124–134.

Galles, D., and Pearl, J. (1997). Axioms of casual relevance. *Artificial Intelligence* **97**(1–2), 9–43.

Gluck, M. (1995). Understanding performance in information systems: Blending relevance and competence. *Journal of the American Society for Information Science* **46**(6), 446–460.

Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing and Management* **32**(1), 89–104.

Goffman, W. (1964). On relevance as a measure. *Information Storage and Retrieval* **2**(3), 201–203.

Gorayska, B., and Lindsay, R. O. (1993). The roots of relevance. *Journal of Pragmatics* **19**(4), 301–323.

Green, R. (1995). Topical relevance relationships. I. Why topic matching fails. *Journal of the American Society for Information Science* **46**(9), 646–653.

Green, R., and Bean, C. A. (1995). Topical relevance relationships. II. An exploratory study and preliminary typology. *Journal of the American Society for Information Science* **46**(9), 654–662.

Greisdorf, H. (2003). Relevance thresholds: A multi-stage predictive model of how users evaluate information. *Information Processing and Management* **39**(3), 403–423.

Greisdorf, H., and Spink, A. (2001). Median measure: An approach to IR systems evaluation. *Information Processing and Management* **37**(6), 843–857.

Gull, C. D. (1956). Seven years of work on the organization of materials in special library. *American Documentation* **7**, 320–329.

Hansen, P., and Karlgren, J. (2005). Effects of foreign language and task scenario on relevance assessment. *Journal of Documentation* **61**(5), 623–639.

Harter, S. P. (1971). The Cranfield II relevance assessments: A critical evaluation. *Library Quarterly* **41**, 229–243.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science* **53**(4), 257–270.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science* **47**(1), 37–49.

Haynes, B. R., McKibbon, A., Walker, C. Y., Ryan, N., Fitzgerald, D., and Ramsden, M. F. (1990). Online access to MEDLINE in clinical setting. *Annals of Internal Medicine* **112**(1), 78–84.

Hirsh, S. G. (1999). Children's relevance criteria and information seeking on electronic resources. *Journal of the American Society for Information Science* **50**(14), 1265–1283.

Hjørland, B. (2002). Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information Science and Technology* **53**(4), 257–270.

Howard, D. L. (1994). Pertinence as reflected in personal constructs. *Journal of the American Society for Information Science* **45**(3), 172–185.

Huang, M., and Wang, H. (2004). The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of American Society for Information Science and Technology* **55**(11), 970–979.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation* **52**(1), 3–50.

Ingwersen, P., and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.

International Federation of Library Association and Institutions (IFLA) (1998). *Functional Requirements for Bibliographic Records—Final Report.* Retrieved 10-9-2005 from: http://www.ifla.org/VII/s13/frbr/frbr1.htm#2.1

Janes, J. W. (1991a). The binary nature of continuous relevance judgments: A study of users' perceptions. *Journal of the American Society for Information Science* **42**(10), 754–756.

Janes, J. W. (1991b). Relevance judgments and the incremental presentation of document representation. *Information Processing and Management* **27**(6), 629–646.

Janes, J. W. (1993). On the distribution of relevance judgments. *Proceedings of the American Society for Information Science* **30**, 104–114.

Janes, J. W. (1994). Other people's judgments: A comparison of users' and others' judgments of document relevance, topicality, and utility. *Journal of the American Society for Information Science* **45**(3), 160–171.

Janes, J. W., and McKinney, R. (1992). Relevance judgments of actual users and secondary users: A comparative study. *Library Quarterly* **62**(2), 150–168.

Jansen, B. J., Spink, A., and Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management* **36**(2), 207–227.

Kazhdan, T. V. (1979). Effects of subjective expert evaluation of relevance on the performance parameters of document-based information retrieval system. *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2* **13**, 21–24.

Kelly, D. (2005). Implicit feedback: Using behavior to infer relevance. In *New Directions in Cognitive Information Retrieval* (A. Spink and C. Cole, eds.), pp. 169–186, Springer, Amsterdam, The Netherlands.

Kent, A., Berry, M., Leuhrs, F. U., and Perry, J. W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation* **6**(2), 93–101.

Kuhlthau, C. C. (2004). *Seeking Meaning: A Process Approach to Library and Information Services*, 2nd ed. Greenwood, Westport, CT.

Lakemeyer, G. (1997). Relevance from an epistemic perspective. *Artificial Intelligence* **97**(1–2), 137–167.

Lalmas, M. (1998). Logical models in information retrieval: Introduction and overview. *Information Processing and Management* **34**(1), 19–33.

Lesk, M. E., and Salton, G. (1968). Relevance assessment and retrieval system evaluation. *Information Processing and Management* 4(4), 343–359.

Maglaughlin, K. L., and Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of American Society for Information Science and Technology* **53**(5), ;327–342.

Mares, E. (1998). Relevance logic. In *Stanford Encyclopedia of Philosophy*. Retrieved Oct. 17, 2005 from http://plato.stanford.edu/entries/logic-relevance/#Bib

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science* **48**(9), 810–832.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers* **10**(3), 303–320.

Mooers, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *American Documentation* **2**, 20–32.

Nie, J. -Y., Brisebois, M., and Lepage, F. (1995). Information retrieval as counterfactual. *Computer Journal* **38**(8), 643–657.

Park, T. K. (1993). The nature of relevance in information retrieval: An empirical study. *Library Quarterly* **63**(3), 318–351.

Park, T. K. (1994). Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *Journal of the American Society for Information Science* **45**(3), 135–141.

Popper, K. (1972). *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford.

Purgaillis, P. L. M., and Johnson, R. E. (1990). Does order of presentation affect users' judgment of documents? *Journal of the American Society for Information Science* **41**(7), 493–494.

Quiroga, L. M., and Mostafa, J. (2002). An experiment in building profiles in information filtering: the role of context of user relevance feedback. *Information Processing and Management* **38**(5), 671–694.

Rees, A. M., and Schultz, D. G. (1967). *A Field Experimental Approach to the Study of Relevance Assessments in* Relation to *Document Searching*, 2 vols.Western Reserve University, School of Library Science, Center for Documentation and Communication Research, Cleveland, OH. NTIS: PB-176 080/XAB, PB-176 079/XAB. ERIC: ED027909, ED027910.

Regazzi, J. J. (1988). Performance measures for information retrieval systems: An experimental approach. *Journal of the American Society for Information Science* **3**(4), 235–251.

Rieh, S. Y., and Belkin, N. J. (2000). Interaction on the Web: Scholars judgment of information quality and cognitive authority. *Proceedings of the American Society for Information Science* **37**, 25–36.

Rieh, S. Y., and Xie, H. I. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing and Management* **42**(3), 751–768.

Robertson, S. E., and Hancock-Beauleiu, M. M. (1992). On the evaluation of IR systems. *Information Processing and Management* **28**(4), 457–466.

Ruthven, I. (2005). Integrating approaches to relevance. In *New Directions in Cognitive Information Retrieval* (A. Spink and C. Cole, eds.), pp. 61–80. Springer, Amsterdam, The Netherlands.

Ruthven, I., Lalmas, M., and Van Rijsbergen, K. (2003). Incorporating user search behavior into relevance feedback. *Journal of the American Society for Information Science and Technology* **54**(6), 529–549.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion of information science. *Journal of American Society for Information Science* **26**(6), 321–343.

Saracevic, T. (1976). Relevance: A review of and a framework for the thinking on the notion of information science. In , *Advances in Librarianship* (M. J. Voigt and M. H. Harris, eds.), vol. 6, pp. 81–138.

Saracevic, T. (1991). Individual differences in organizing, searching and retrieving information. *Proceedings of the American Society for Information Science* **28**, 82–86.

Saracevic, T. (1996). Relevance reconsidered '96. In *Information Science: Integration in Perspective. Proceedings of Second International Conference on Conceptions of Library and Information Science (CoLIS 1996)* (P. Ingewersen and N. O. Pors, eds.), pp. 201–218. The Royal School of Librarianship, Copenhagen.

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. *Proceedings of the American Society for Information Science* **34**, 313–327.

Saracevic, T., and Kantor, P. (1988a). A study of information seeking and retrieving: III. Searchers, searches, and overlap. *Journal of the American Society for Information Science* **39**(3), 197–216.

Saracevic, T., and Kantor, P. (1988b). A study of information seeking and retrieving: II. Users, questions, and effectiveness. *Journal of the American Society for Information Science* **39**(3), 177–196.

Saracevic, T., and Kantor, P. (1997). Studying the value of library and information services. I. Establishing a theoretical framework. *Journal of the American Society for Information Science* **48**(6), 527–542.

Schamber, L. (1991). User's criteria for evaluation in a multimedia environment. *Proceedings of the American Society for Information Science* **28**, 126–133.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology* **29**, 3–48.

Schamber, L., and Bateman, J. (1999). Relevance criteria uses and importance: Progress in development of a measurement scale. *Proceedings of the American Society for Information Science* **33**, 381–389.

Schamber, L., Eisenberg, M. B., and Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management* **26**(6), 755–776.

Schutz, A. (1970). *Reflections on the Problem of Relevance*. Yale University Press, New Haven.

Schutz, A., and Luckman, T. (1973). *The Structures of the Life-World*. Northwestern University Press, Evanston, Il.

Searle, J. R. (1984). Intentionality and its place in nature. *Synthese* **61**(1), 3–16.

Sebastiani, F. (1998). On the role of logic in information retrieval. *Information Processing and Management* **34**(1), 1–18.

Shaw, W. M., Jr., Wood, J. B., Wood, R. E., and Tibbo, H. R. (1991). The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research* **13**(4), 347–366.

Smithson, S. (1994). Information retrieval evaluation in practice: A case study approach. *Information Processing and Management* **30**(2), 205–221.

Snow, C. P. (1993). *The Two Cultures*, Canto edition. Cambridge University Press, Cambridge, UK.

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science* **45**(8), 589–599.

Sormunen, E. (2002). Liberal relevance criteria of TREC: Counting on neglible documents? Proceedings of the 25st Annual International Conference on Research and Development in Information Retrieval of the Special Interest Group on Information Retrieval, Association for Computing Machinery (SIGIR 00), pp. 324–330.

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition*. Blackwell, Cambridge MA.

Sperber, D., and Wilson, D. (1995). *Relevance: Communication and Cognition*, 2nd ed. Blackwell, Cambridge MA.

Spink, A., and Cole, C. (eds.) (2005a). *New Directions in Cognitive Information Retrieval*. Springer, Amsterdam, The Netherlands.

Spink, A., and Cole, C. (2005b). A multitasking framework for cognitive information retrieval. In *New Directions in Cognitive Information Retrieval* (A. Spink and C. Cole, eds.), pp. 99–112. Springer, Amsterdam, The Netherlands.

Spink, A., and Greisdorf, H. (2001). Regions and levels: Measuring and mapping users' relevance judgment. *Journal of the American Society for Information Science* **52**(2), 161–173.

Spink, A., and Saracevic, T. (1997). Human–computer interaction in information retrieval: Nature and manifestations of feedback. *Interacting with Computers* **10**(3), 249–267.

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing and Management* **28**(4), 503–516.

Subramanian, D., Greiner, R., and Pearl, J. (1997). The relevance of relevance. Special issue on relevance. *Artificial Intelligence* **97**(1–2), 1–5.

Swanson, D. R. (1971). Some unexplained aspects of the Cranfield tests of indexing performance factors. *Library Quarterly* **41**, 223–228.

Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *The Library Quarterly* **56**(4), 389–398.

Swanson, D. R., and Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence* **91**(2), 183–203.

Swanson, D. R., and Smalheiser, N. R. (1999). Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. *Library Trends* **48**(1), 48–59.

Tang, R., and Solomon, P. (1998). Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Processing and Management* **34**(2–3), 237–256.

Tang, R., and Solomon, P. (2001). Use of relevance criteria across stages of document evaluation: On the complementarity of experimental and naturalistic studies. *Journal of the American Society for Information Science* **52**(8), 676–685.

Tombros, A., and Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval of the Special Interest Group on Information Retrieval, Association for Computing Machinery (SIGIR 98)*, pp. 2–10.

Tombros, A., Ruthven, I., and Jose, J. M. (2005). How users assess web pages for information seeking. *Journal of the American Society for Information Science* **56**(4), 327–344.

Toms, E. G., O'Brien, H. L., Kopak, R., and Freund, L. (2005). Searching for relevance in the relevance of search. In *Proceedings of Fourth International Conference on Conceptions of Library and Information Science (CoLIS 2005)* (F. Crestani and I. Ruthven, eds.), pp. 59–78. Springer, Amsterdam, The Netherlands.

Vakkari, P. (2001). Changes in search tactics and relevance judgments when preparing a research proposal: A summary of findings of a longitudinal study. *Information Retrieval* 4(3), 295–310.

Vakkari, P., and Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation* **56**(5), 540–562.

Vakkari, P., and Sormunen, E. (2004). The influence of relevance levels on the effectiveness of interactive information retrieval. *Journal of the American Society for Information Science and Technology* **55**(11), 963–969.

van Rijsbergen, C. J. (1986). Non-classical logic for information retrieval. *Computer Journal* **30**(6), 481–485.

Vickery, B. C. (1959a). The structure of information retrieval systems. In *Proceedings of the International Conference on Scientific Information*, vol. 2, pp. 1275–1290. National Academy of Sciences, Washington, DC.Retrieved Jan. 8, 2006 from http://www.nap.edu/books/NI000518/html/1275.html

Vickery, B. C. (1959b). Subject analysis for information retrieval. In *Proceedings of the International Conference on Scientific Information*, vol. 2, pp. 855–866. National Academy of Sciences, Washington, DC.Retrieved Jan. 8, 2006 from:http://www.nap.edu/books/NI000518/html/855.html

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management* **36**(5), 697–716.

Voorhees, E. M., and Harman, D. K. (eds.) (2005). *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA.

Wallis, P., and Thom, J. A. (1996). Relevance judgments for assessing recall. *Information Processing and Management* **32**(3), 273–286.

Wang, P. (1997). The design of document retrieval systems for academic users: Implications of studies on users' relevance criteria. *Proceedings of American Society for Information Science* **34**, 162–173.

Wang, P., and Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science* **49**(2), 115–133.

Wang, P., and White, M. D. (1995). Document use during a research project: A longitudinal study. *Proceedings of American Society for Information Science* **32**, 181–188.

Wang, P., and White, M. D. (1999). A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages. *Journal of the American Society for Information Science* **50**(2), 98–114.

White, H. D. (in press a). Combining bibliometrics, information retrieval, and relevance theory: First examples of a synthesis. *Journal of the American Society for Information Science and Technology.*

White, H. D. (in press b). Combining bibliometrics, information retrieval, and Relevance Theory: Some implications for information science. *Journal of the American Society for Information Science.*

White, H., and McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science 1972–1995. *Journal of the American Society for Information Science* 49(4), 327–355.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval* 9(8), 457–471.

Wilson, D., and Sperber, D. (2004). Relevance theory. In *Handbook of Pragmatics* (G. Ward and L. Horn, eds.), pp. 607–632. Blackwell, Oxford. Also: retrieved Oct. 8, 2005 from http://www.dan.sperber.com/relevance_theory.htm

Xu, Y. C., and Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality. *Journal of the American Society for Information Science and Technology*, Published online 16 March 2006.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval of the Special Interest Group on Information Retrieval, Association for Computing Machinery (SIGIR 98)*, pp. 307–314.