Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance

Tefko Saracevic
School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ 08901. E-mail: tefko@scils.rutgers.edu

All is flux.
—Plato on Knowledge in the *Theaetetus* (about 369 BC)

Relevance is a, if not even *the*, key notion in information science in general and information retrieval in particular. This two-part critical review traces and synthesizes the scholarship on relevance over the past 30 years or so and provides an updated framework within which the still widely dissonant ideas and works about relevance might be interpreted and related. It is a continuation and update of a similar review that appeared in 1975 under the same title, considered here as being Part I. The present review is organized in two parts: Part II addresses the questions related to nature and manifestations of relevance, and Part III addresses questions related to relevance behavior and effects. In Part II, the nature of relevance is discussed in terms of meaning ascribed to relevance, theories used or proposed, and models that have been developed. The manifestations of relevance are classified as to several kinds of relevance that form an interdependent system of relevancies. In Part III, relevance behavior and effects are synthesized using experimental and observational works that incorporated data. In both parts, each section concludes 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 work are offered in conclusions.

Prologue to Part III: How It Is Connected and What This Work Is All About

To provide a continuation from the preceding Part II, a few basic descriptions about this work are repeated.

As stated in the Preface to Part II, in 1975 I published a review about relevance under the same title, without, of course, "Part I" in the title (Saracevic, 1975). There was no plan then to have another related review 30 years later—but things happen. The intent of the 1975 work was "to explore the meaning of relevance as it has evolved in information science and to provide a framework within which various interpretations of relevance can be related" (Saracevic, 1975, p. 321).

Building on the examination of relevance in the preceding (1975) review, this work (2007) follows the travails of relevance in information science for the past 30 years. It is an update. 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.

The organization of the present review, offered in two parts, addresses the questions related to relevance nature, manifestations, behavior, and effects. Following the *Introduction* and a *Historical Footnote,* the preceding Part II (this issue, pp. 1915–1933) started with a general section on the nature of relevance by synthesizing its meanings, following with sections on theories and models of relevance that are, in effect, further elaborations on the nature of relevance. Part II ended with a section about various manifestations of relevance. This Part III deals with experimental and observational findings on human relevance behavior and effects of relevance. Part II is oriented toward scholarship that addressed relevance concepts, whereas Part III is oriented toward scholarship that provided tangible results based on experimentation or observation. The text of Part III is meant to continue from Part II beginning with the seventh section, Relevance Behavior and continuing through the ninth section, Epilogue. The third to the eighth sections end with summaries that provide a personal interpretation and a critical synthesis of contemporary thinking on the topic treated in the cited studies or suggest hypotheses for future research—in effect conclusions in summaries, although based on reviewed studies, should mainly be treated as hypothesis to encourage further research.

Received April 1, 2006; revised November 18, 2006, January 4, 2007, June 12, 2007; accepted June 12, 2007

© 2007 Wiley Periodicals, Inc. ● Published online 10 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20681

The rationale for summaries was to concentrate on synthesis of specific data and findings, rather than on discussions and conjectures. Analyses of some of the major trends that shape relevance work are offered in the Epilogue.

Relevance Behavior: 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, including cognitive IR, extensively deal with relevance as well (e.g., Ingwersen, & Järvelin, 2005, Spink & Cole, 2005). 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 & White, 1999), on judgment of cognitive authority and information quality (Rieh & Belkin, 2000), on users' assessments of Web pages (Tombros, Ruthven, & Jose, 2005), or on relation between search terms, index terms, and documents judged as to relevance (Kim, 2006). 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 assumed to indicate relevance implicitly. In other words, although relevance was not explicitly discussed, an action such as saving a page or document is regarded as implying relevance; relevance is not stated, but implied. Although 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 contain 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 studies, also are excluded. I probably missed some studies and did not include repetitive articles (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: "To study, among others, [*object of research*]." I classified the studies into appropriate sections according to the main object of research. Most, if not all, studies included a discussion of a framework (underlying theories, models, concepts, and the like); however, this discussion is omitted in their description that follows because it is covered in preceding sections of this review. Where appropriate, some summaries include numerical results. However, the principal results from all of the studies, with a number of caveats, are summarized, i.e., interpreted, synthesized, and generalized at the end of the section and suggested as hypotheses.

Relevance Clues

What makes information or information objects relevant? Or more specifically, what do people look for in information or information objects to infer relevance? Two distinct approaches are used in deciphering this question. In the first, the topic approach, the course of deriving topical or nontopical relation is analyzed. This approach (represented by Green & Bean, 1995; Swanson & Smalheiser, 1997, 1999) was treated in the Part II subsection *The Big Question and Challenge* (this issue, pp. 1915–1933). The second or clues approach, treated here, follows the research agenda proposed by Schamber, Eisenberg, and Nilan (1990; reviewed in Part II's subsection *Dynamic Model*, this issue, pp. 1915–1933) 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 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 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, Belkin, and Kantor (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 to identify characteristics of texts affecting relevance judgments. They identified six facets of judgment of document usefulness.

DOI: 10.1002/asi

- 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 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 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) to suggest a general model for document selection using relevance clues.
- Fidel and Crandall (1997) studied 15 engineering information 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 to derive criteria for judging the reports relevant or not relevant. They identified 13 criteria explaining why a report was relevant; 14 criteria explaining why it was not relevant.
- Barry and Schamber (1998) compared results from two of their studies (Barry, 1994; Schamber, 1991) 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 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 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, 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 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 138 retrieved documents 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 the student's own query, which were presented in different representations to derive and compare criteria for relevant, partially relevant, and nonrelevant judgments. They identified 29 criteria in six categories and compared the presence of their criteria with criteria from 10 other studies.
- Toms, O'Brien, Kopak, and Freund (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 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.

Image clues. *What makes images relevant? Are clues used in relevance inference about images similar to those for texts?* Only one study addressed these questions.

- 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 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 criteria after the search.

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, the user's cognitive state changes and 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. To observe differences in judgments at different stages (initial, final, citing) and among different cases, Smithson had the users judge a combined total of 1,406 documents for relevance at the initiation and completion stages of the case.

DOI: 10.1002/asi

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 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 to identify relevance criteria used in early and later stages of the subjects' research. They identified 11 criteria in the early stages and another 8 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 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 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 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, 9 graduate students who searched for documents to support their own research also were evaluated at Stages 1 and 2 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 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 of the groups relate to shaping boundaries to a topic.

Relevance Feedback

What factors affect the process of relevance feedback? A short explanation of relevance feedback from the human perspective: I find a relevant document, go through it and, on the basis of something in that document, go on and reformulate my search or identify something else that I should consult. In information retrieval (IR), relevance feedback (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.

- Koenemann and Belkin (1996) used 64 undergraduate students to search two topics from TREC 2 on a subset of the TREC collection using a nonfeedback IR system as a base and three systems that incorporated various types of feedback to assess the effectiveness of relevance feedback. They found that relevance feedback improves performance by at least 10% and is preferred by users.
- 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 to analyze the nature of feedback involving users, intermediaries, searches, and results. The users judged 6,225 retrieved documents as to relevance. The researchers identified 885 feedback loops grouped in five categories depicting different types of feedback.
- Jansen, Spink, and Saracevic (2000) analyzed logs of 51,423 queries posed by 18,113 users on the Excite search engine to determine a number of query characteristics, including the incidence of relevance feedback. They found that 5% of queries used RF.
- Quiroga and Mostafa (2002) studied 18 graduate students who searched a collection of 6,000 records in consumer health on a system with various feedback capabilities. The researchers provided a verbal protocol of proceedings to categorize factors that influence relevance feedback assessments. They identified 15 factors in four categories related to users and three categories of factors related to documents.
- Ruthven, Lalmas, and van Rijsbergen (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 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, as is Koenemann and Belkin (1996), but they are included here to show the human side investigated.

Summary of Relevance Behavior[1]

Caveats abound. Numerous aspects of the studies reviewed can be questioned and criticized. Criteria, language,

[1] Bulleted conclusions in the summary of this and the next section are a collective synthesis from several studies that are enumerated; the mentioned studies came to similar conclusions, usually in different words, and I derived the synthesis from their conclusions. These are suggested as hypotheses. When a single study resulted in a conclusion and the conclusion was not replicated in some form or other by other studies in the section, then that study is cited as an exception

DOI: 10.1002/asi

measures, and methods used in these studies were not standardized and they varied widely. In that sense, although no study was an island, each study was done more or less on its own. Thus, the results are only cautiously comparable. Still, it is really refreshing to see conclusions made based on data, rather than on the basis of examples, anecdotes, authorities, or contemplation. Generalizations below are derived from the studies reviewed by first examining and then synthesizing the actual data and results as presented, rather than just incorporating conclusions from the studies themselves. As mentioned, generalizations should primarily be treated as hypotheses. The language and concepts in summaries, while obtained from studies, are standardized.

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 derived from data in 16 studies reported in the earlier subsection Relevance Clues.

- 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* use *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, whereas faculty assigns a lot. *Different users, tasks, progress in tasks, classes of users* use *similar criteria*, but may apply *different weights*.
- Although 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, mental effort
    - *Affective match*: emotional responses to information, fun, frustration, uncertainty
    - *Belief match*: personal credence given to information, 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). (In a similar sense this is posited in subsection *Stratified Model*; Part II, this issue, pp. 1915–1933.)
- 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, which include topical relevance, are rated highest in importance, but interact with other criteria—they are not the sole criteria.
- 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 given criteria differs depending on the rating—e.g., weight for the same criterion on a document judged relevant differs from the weight of a document judged not relevant. *Different ratings of relevance* use *similar criteria* but may apply *different weights*.
- Similarly, although 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. (Conclusion based on a single study that was reported in Choi and Rasmussen, 2002).

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. Here are some generalizations derived from data in eight studies reported in the above subsection Relevance Dynamics:

- For a given task, it seems that the users' inferences about specific information or information object are dependent on the stage of the task.
- However, users' 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 users' 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 selections* are made in *different stages using similar criteria*, but possibly with *different weights*.
- As time progresses and a task becomes more focused, it seems that the discriminatory power for relevance selection increases. *Increased focus* results in *increased discrimination* and *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. Here are some generalizations derived from data in the five studies reported in the above subsection Relevance Feedback:

DOI: 10.1002/asi

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

Effects of Relevance: 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 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.

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 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, 3 with associate's degrees, 41 with or working on bachelor's degrees, 19 with or working on master's degrees, and 6 with or working on PhD degrees) to (a) respond to an unspecified set of geography-related questions using two packets of geographic materials, and (b) recall their recent experience where geographic questions were raised with responses coded by two coders on a 5-point relevance scale to study the effects of geographic competence and experience on relevance inferences (1995 study) and compare user relevance and satisfaction ratings (1996 study).
- Dong, Loh, and Mondry (2005) asked a physician (whose assessment was considered the gold standard), 6 evaluators with biology or medical backgrounds, and 6 without such backgrounds to assess for relevance 132 Web documents retrieved by a metacrawler in relation to specific medical topics to measure variation in relevance assessments due to their domain knowledge and develop a measure of relevance similarity.
- Hansen and Karlgren (2005) used 8 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 to investigate how judges assess the relevance of retrieved documents in a foreign language, and how different scenarios affect assessments.

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 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 (1988) used five professional searchers each to search 40 questions, posed by 40 users (19 faculty, 15 graduate students, and 6 professionals from industry) with real information needs. Their pooled results were presented to the users for relevance assessment 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%.

Further studies that show the degree of agreement on relevance assessments between different groups of judges are reviewed in the below subsections *Beyond consistency* and *But does it matter?*—they relate to individual difference studies presented here and are considered in summary below.

Relevance Judgments

What factors affect relevance judgments? There are 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

DOI: 10.1002/asi

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 factors 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 nonrelevant ones are equally nonrelevant.
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 intravariation in relevance assessments among judges. Even if there are, it does not matter; there is no appreciable effect in ranking performance.

A sixth, *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, because this is really a question for IR evaluation methods rather than relevance judgments, the completeness postulate is not addressed 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 subsection *Issue of primacy—weak and strong relevance* (Part II, this issue, pp. 1915–1933). The postulates are stringent laboratory assumptions, easily challenged. In most, if not all laboratory investigations in science, things are idealized and simplified 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. Information retrieval evaluation, with or despite of its weak view of relevance, played a significant role in that achievement. Clearly, advances in technology played a major role as well.

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 and user 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.

Beyond topical. Do people infer relevance based on topicality only? This question was treated in the subsection *Topical relevances* (Part II, this issue, pp. 1915–1933) and the above subsection *Relevance Clues*, thus not rehashed again. It is brought up here because it is one of the postulates in the central assumption for IR evaluation. The short conclusion is that it 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 above subsection *Summary of Relevance Behavior*, 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 1,288 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 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, 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 the analysis included 264 evaluated documents. The study was done 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.
- Xu (2007) asked 113 undergraduate students to search the Web for documents of personal interest for self-education or relaxation and then choose and evaluate two documents that they have browsed; thus, analyses included 226 evaluated documents. The study was done to test five hypotheses related to criteria for informative relevance (resulting from "epistemic information searches"—desire for knowledge)

and affective relevance (resulting from "hedonic information searches"—information for fun or affective stimulation) as opposed to situational relevance (resulting from problem-oriented searches). He found that topicality, novelty, and reliability significantly contribute to informative relevance, but scope and understandability do not, and topicality and understandability significantly contribute to affective relevance, but novelty does not. This and the preceding work are the only studies that did hypothesis testing as to relevance criteria; others provided either frequency counts or description only.

Beyond binary. Are relevance inferences binary, i.e., 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 to study the distribution of judgments and observe whether the participants perceived the break point between relevant and nonrelevant 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) 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 to determine the distribution of judgments on a continuous scale.
- Su (1992) used 30 graduate students, 9 faculty, and 1 staff member 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 to determine whether a single measure or a group of measures reflecting various relevance criteria is/are the best indicator of successful retrieval. She identified the "value of search results as a whole" as the best measure reflecting IR performance.
- Janes (1993) rearranged relevance judgment data from two older studies (Cuadra, Katter, Holmes, & Wallace, 1967; Rees & 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 to investigate the distribution of relevance judgments.
- Spink, Greisdorf, and Bateman (1998) used data from four studies involving a total of 55 users (37 graduate students, 18 not identified as to academic status), who, in 55 searches related to their information needs, retrieved 4,926 documents in response. Users were grouped on the basis of a number of variables and also asked to provide a rationale for their relevance judgments to establish and compare criteria used for judging documents as relevant, partially relevant, and not relevant.
- Greisdorf and Spink (2001) used 36 graduate students in three studies, who in 57 searches related to their personal or academic information need, retrieved 1,295 documents. The students were asked to indicate relevance assessments using various scales and criteria 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 1,059 documents. The students were asked to indicate relevance assessments using various scales and criteria 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 1,432 documents in response. The students were asked to assess their results using a number of relevance criteria on a continuous relevance scale to study the users' evaluation as related to different regions of relevance.

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 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 articles discussing the issue followed, but only at the end of the 1980s did the question start receiving experimental treatment:

- Eisenberg and Barry (1988) conducted two experiments, the first experiment with 42 and the second with 32 graduate students. 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 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 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) 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 to examine the influence of the order and size of document presentation on relevance judgments.

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 questions are treated in the above subsection *Relevance Dynamics*, thus not reviewed here. However, briefly, 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, the criteria for judging relevance are fairly stable.

Beyond consistency. Are relevance judgments consistent among judges or a 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. Information retrieval 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, 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 2,200 documents were relevant to the 98 requests, whereas the second found that 1,998 were relevant. There was not much overlap between groups. The first group judged 1,640 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 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.

Consistency in relevance judgments was derived from or addressed in the following studies:

- 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 5,307 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 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 1,525 relevant articles (3.5%) were retrieved by all three search groups. Expert searchers retrieved twice as many relevant documents as novice searchers, but novice searchers retrieved some that expert searchers did not. This is the only real-life study on the question.
- Shaw, Wood, Wood, and Tibbo (1991) used four judges to assess the relevance of 1,239 documents in a 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, 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 rejudge the two sets to compare changes in judgments due to an 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 rejudge 20 documents in each of two topics that were previously judged as to relevance by users to compare users' versus nonusers' 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 5,271 documents already judged on relevance in 38 topics in TREC 7 and 8 on a graded 4-point scale (as opposed to a binary scale used in TREC) 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 on a system that provided interactive relevance feedback capabilities. The results had preassigned relevance ratings by TREC assessors on a binary relevance scale with additional reassessment by two assessors on a 4-point scale—these were called *official assessors*. This was done to study the consistency between user (student) relevance assessment and those by official assessors and the difference in identification of relevant and highly relevant documents. They found that the student users identified 45% of items judged relevant by TREC assessors and 83% of items judged highly relevant by additional official assessors.
- Lee, Belkin, and Krovitz (2006) used 10 experienced searchers (not indicated as to status) to compare two lists of 30 documents each for 10 TREC topics. The documents were judged as to relevance by three judges beforehand; then the lists were ordered so that precision level varied from 30% to 70%. Subjects indicated their preference between two lists of various precision levels for each topic. The study was done to examine the ability of subjects to recognize lists that have a higher precision level, called *right lists* as they contain more relevant documents. The range of recognition of right lists varied from 14.6% to 31.2%. Agreement in relevance judgments was 24%.

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): "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" (p. 43).

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 six 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. Five of the six 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 1,268 abstracts in the field of library and information science, to assess the relevance of those 1,268 documents to their queries (called the A judgments). Then a second, independent set of relevance judgments (B judgments) was obtained by asking eight new judges to assess for relevance six additional queries not of his or her own 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 2,600 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 to compare the performance of different representations in relation to different judgment sets. He found that the Lesk and Salton hypothesis is confirmed: The relative ranking of the seven different representations remained the same over two sets of judgments. There was one exception, however, where ranking changed.
- Burgin (1992) used a collection of 1,239 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 3,204 computer science documents (titles and in most cases, abstracts) that already had relevance judgments by SMART judges to compare two retrieval techniques. Then two judges (paper authors, called Judge 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 Judge 1 and 2). They found that the overall agreement between original assessors (SMART) and two new assessors (Judge 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 & Harman, 2005, pp. 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 rejudged a pool of up to 200 relevant and 200 nonrelevant 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

DOI: 10.1002/asi

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 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.

- Vorhees (2001) used 50 topics created for the TREC-9 Web track and asked assessors to judge retrieved pages on a 3-point scale: relevant, highly relevant, not relevant (as opposed to general TREC assessments that use a binary relevance scale—relevant and not relevant). The assessments were done by a primary judge and then the relevant and highly relevant documents were reassessed by two other secondary assessors. All assessors were also asked to identify the best page or pages for a topic. The study was done to examine the effect of highly relevant documents on the performance ranking of the different IR techniques tested. She found that "different retrieval systems are better at finding the highly relevant documents than those that are better at finding generally relevant documents." (p. 76) This conclusion contradicts the finding of the previous (Vorhees, 2000) study, which concluded that relative effectiveness of retrieval systems is stable despite differences in relevance judgment sets. "The ability to separate highly relevant documents from generally relevant documents evidently is correlated with systems functionality, and thus differences among systems are reflected in the average score" (Vorhees, 2001, p. 77). The agreement among three assessors as to the best pages for a topic was 34%.

Summary of Effects of Relevance

Caveats abound again and for the same reasons mentioned in the summary of the previous section. Although similar or even identical research questions were asked in a number of studies, the criteria and methodologies differed so widely (they were all over the place) that general conclusions offered below are no more than possible hypotheses. As in the summaries for the preceding section, generalizations below are derived from the studies reviewed by first examining and then synthesizing the actual data and results as presented, rather than just incorporating conclusions from the studies themselves. Language, while derived from the studies, is standardized.

Judges. A very limited number of factors related to relevance judges were studied. This is in sharp contrast to a much large number of factors studied in various studies of indexers and searchers (e.g., Saracevic & Kantor, 1988). Here are some generalizations derived from data in four studies reported in subsection *Relevance Judges*:

- Subject expertise seems to be one variable that accounts strongly for differences in relevance inferences by group of judges—*higher expertise* results in *higher agreement*, *less differences*.
- Lesser subject expertise seems to lead to more lenient and relatively higher relevance ratings—*lesser expertise* results in *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. (Conclusion based on a single study only; Hansen & Karlgren, 2005).
- Different search requests or what authors call *task scenarios* make a difference in the relevance assessment process as to time, but seem not to affect the degree of agreement. *Longer task scenarios* result in *more time spent in assessment*; *all task scenarios* have a *similar degree of agreement among judges*. (Same single study as above)

Individual differences. Most often, studies of individual differences in relevance inferences involved observing plain statistics of differences or degrees of overlap, with little or no diagnostics of underlying factors. Here are some generalizations derived from data in two studies reported in subsection *Individual differences*, as well as from related data in eight studies reported in subsection *Beyond consistency* and six studies in subsection *But does it matter?*:

- A relatively large variability can be expected in relevance inferences by individuals or groups of individuals with similar backgrounds. Individual differences are *a*, 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 as previously reviewed (Saracevic, 1991).
- 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.

Judgments. A number of factors affect relevance judgments; for instance and as mentioned, Schamber (1994) listed 80 factors grouped in six categories and Harter (1996) 24 factors grouped in four categories, both in tables. Instead of creating another table, I grouped studies along questions related to assumptions made in IR evaluations. Here are some generalizations derived from data in 29 studies reported in the subsection *Relevance Judgments* as a whole including all the subsections:

- Relevance is measurable—this is probably the most important general conclusion from all the studies containing data.

- Not surprisingly, none of the five postulates in the central assumption of IR laboratory testing holds.
  - However, using these postulates (representing a simplified or weak view of relevance) in a laboratory evaluation produced significant improvements in IR techniques.

Topicality. A perennial research question and hot discussion topic revolved along the issue of whether topicality is the only or the most important attribute in relevance inferences by people. Here are some generalizations derived from data in 3 studies in subsections *Beyond topical* and 15 studies reported in subsection *Relevance Clues*:

- Topicality of information or information objects is not at all an exclusive criterion or attribute in relevance inferences by people.
- A number of other relevance attributes play a role and are used in conjunction and interaction with topicality, as enumerated in the subsection *Summary of Relevance Behavior*.
  - However, in great many situations topicality plays a primary role in inferences of relevance of information or information objects.

Measures. Another perennial question for investigation relates to measures and measuring of relevance inferences, including distribution of relevance judgments along some gradation from relevant to not relevant. Here are some generalizations derived from data in nine studies reported in the subsection *Beyond binary*:

- 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, relevant) with smaller peaks in the middle range (somewhat not relevant or relevant). The highest peak is on the not 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.
  - Another however—when assessing the use of search outputs considering a complete set of retrieved information objects, the value of search results as a whole seems to be the critical criterion that users apply in making relevance inferences. (Based on a single study; Su, 1992).
- 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.

Independence. Although the question of whether information objects are judged dependently or independently of each other has a long history of concern, only a few studies addressed it. Here are some generalizations derived from data in four studies reported in the subsection *Beyond independence*:

- The order in which documents are presented to users does have an effect on relevance inferences by people.
  - Information objects presented early have a higher probability of being inferred as relevant.
  - However, when a small number of documents is presented, order does not matter.
- 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 are not as important as abstracts and full texts.

Consistency. For a long time it was known that relevance judgments related to the same topic or query on the same set of information objects are not consistent among individual or group of judges. But the research question was not whether the judgments are consistent, but to what degree do they overlap and how inconsistent are they. Here are some generalizations derived from data in nine studies reported in the subsection *Beyond binary* as well as from related data in six studies in subsection *But does it matter?*:

- The inter- and intraconsistency or overlap in relevance judgments varies widely from population to population and even from experiment to experiment, making generalizations particularly difficult and tentative.
  - 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.
  - In general, it seems that the overlap using different populations hovers around 30%.
  - *Higher expertise* results in a *larger overlap*. *Lower expertise* results in a *smaller overlap*.
  - 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.
  - *More judges* result in *less overlap*.
    - The lowest overlap reported was 3.5% when three search groups were used (Haynes et al., 1990).
- Subject expertise affects consistency of relevance judgments. *Higher expertise* results in *higher consistency* and *stringency*. *Lower expertise* results *in lower consistency* and *more inclusion*.

Effect on IR evaluation. Given that relevance judgments are inconsistent, which they are to various degrees, how does this effect results of IR evaluation? This is a serious question for acceptance of results of such evaluations. Here are some generalizations derived from data in six studies in subsection *But does it matter?*:

DOI: 10.1002/asi

- 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 of no effect is counter-intuitive, but a small number of experiments bears it out. However, note that the use of averaging performance affects or even explains this conclusion.
  - *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.
  - Another however—rank order of different IR techniques does change when only *highly relevant* documents are considered—this is another (and significant) exception to the overall conclusion of no effect.
  - Still another however—performance ranking over *individual* queries or topics differs significantly depending on the query.

Reflections

Reflection on approach. The pattern used in this and the previous section to synthesize studies ([author] used [subjects] to do [tasks] to study [object of research]) comes from the studies themselves. For a great many studies, this means that certain stimuli were given to subjects 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 and assumptions, including the interpretation as to the nature of higher mental processes, behaviorism fell out of favor. Methodologically, behaviorism does not include diagnostics beyond responses to given stimuli. Modified behaviorism methodologies were absorbed into 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—Skinner and colleagues were never cited. However, in a great many studies, behaviorism was there with all of its strengths and shortcomings. And in many instances, it still is.

Reflection on population. An overwhelming number of studies on relevance behavior and effects 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.

Reflection on individual differences. It has been noted (in quite a number of studies) that the overlap or degree of agreement in assessment of relevant documents differs: the more assessors (or group of assessors), the lower the overlap. In other words:

- Given the same topic or query
- for which documents were retrieved from the same collection
- and assessed by different relevance assessors (or group of assessors)
  - results in differing relevance assessments
    - i.e., differing (not different) sets of documents are assessed as relevant.

The differences may not be actually all attributable to individual (or group) factors and interpretation. Here is a hypothesis of multiple relevances:

- For the same topic or query, the same collection contains multiple sets of information objects that are relevant.

It is not only that people differ in relevance assessments, but there are indeed several, if not many, sets of relevant answers (information objects) relevant for the same topic or query in the same collection. So people select the relevant set among many relevant sets.

Given that relevance is assessed on a continuum from highly relevant to less relevant to not relevant, here is a hypothesis of high relevance:

- Given a number of relevance assessors for the same topic or query, those information objects that are most often assessed as relevant (i.e. with highest degree of agreement) are also information objects that have the highest relevance rating in general.

Reflections on information technology. As mentioned, IR algorithms and processes have improved over time. However, all is not in IR algorithms alone. Clearly, the advances in IR systems are also based on advances in information technology (IT). Searches are faster, databases larger, interfaces more flexible, reaches are global. . . all closely connected to developments in IT. In many respects, improvements in IR algorithms and processes were related to improvements in IT. But it is not clear, as yet, how these technological and algorithmic improvements have affected relevance inferences by people. Is there a correlated, complementary change in relevance behavior and effects, as far as people are concerned? Logically, it seems so. But evidence is lacking, so far.

Epilogue: What is the Matter With Relevance and What Are Some Implications for the Future?

Information retrieval (IR) came into being right after the Second World War, 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 information technology. However, relevance is also a thoroughly human notion and as all human notions, it is somewhat messy. As stated, the role of research is to make relevance complexity more comprehensible formally and possibly even more predictable.

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 (presented in two parts) is an update; together the two reviews 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. A stratified model was suggested as an integrative framework for viewing relevance interactions between users and computers. As to 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 relevance behavior and effects, we have seen a number of experimental and observational studies that lifted the discourse about relevance from opinions, conjectures, 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 more or less as hypotheses because experimental and observational criteria, standards, and methods varied indiscriminately.

Each of the sections concluded with a summary—a personal interpretation and synthesis of contemporary thinking on the topic treated in the cited studies or suggestion of hypotheses for future research. Thus, here in the concluding section, I am not providing further summaries from the literature. Instead, I am dealing with several current issues and manifested trends that have impacted relevance scholarship in general and, in my opinion, will continue to do so in the near future.

Research Funding

Relevance is poor. Relevance research was funded much better in the past than it is today. Whatever relevance-related funding exists now, it 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 stand up to this day (Cuadra et al., 1967, Rees & Schultz, 1967). Presently at NSF, research on topics related to information is primarily funded and led by the Division of Information and Intelligent Systems (IIS), Directorate for Computer and Information Science and Engineering (CISE). Over time, the agenda, which occasionally is carefully reviewed and set by senior researchers, became completely oriented toward *computers and information* to the exclusion of almost everything or anything that has to do with *humans and information*. This is despite of the support for periodic workshops on social and human aspects of information systems design and more significantly, the recent establishment of Human–Centered Computing (HCC) as one of the three core technical areas in IIS; the

DOI: 10.1002/asi

orientation is reflected in the current solicitation for research proposals.[2]

Why such a state? In his keynote address to the Association for Computing Machinery (ACM) Digital Libraries '99 conference, David Levy (2000) concluded that "the current digital library agenda has largely been set by the computer science community, and clearly bears the imprint of this community's interests and vision. But there are other constituencies whose voices need to be heard." I am suggesting that the same conclusion can be extended currently to human-centered computing research in general and relevance research in particular.

I checked the acknowledgements in 64 articles 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. PhD 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 a result, the conclusion that experimental and observational studies varied indiscriminately 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 and expansion of bright ideas in today's research environment must be funded.

Globalization of Information Retrieval—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 human 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.

Proprietary Information Retrieval—Proprietary Relevance

Increasingly, relevance is becoming proprietary because major search engines are proprietary. Information retrieval 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 the 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- or context-in-the-loop techniques. 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.

[2] "It is anticipated that Human-Centered Computing will support computer scientists as well as social and behavioral scientists and economists whose work contributes to the design and understanding of novel information technologies. However, HCC research should primarily advance the computer and information sciences rather than the social, behavioral, or economic sciences." Program solicitation, NSF-06-572. Retrieved November 15, 2006, from http://www.nsf.gov/pubs/2006/nsf06572/nsf06572.htm

Research Agenda: Beyond

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

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, whereas 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. It should be mentioned, that many, if not most, human information behavior studies, beyond relevance studies, do not use a black box approach.

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 nonstimuli variables, as applied in a number of clues studies (reviewed in subsection *Relevance Clues*) or suggested, among others, by Ruthven (2005). It is much harder to do, but more can be learned.

Beyond mantra. Beyond TREC. 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 articles that vehemently argue that the user viewpoint should be predominant. The majority of studies did not go beyond that statement, so the statement became a mantra. Even where a specific list of implications may have been given, the statement is still a mantra.

Very little was ever done to actually translate results from user studies into system design, as discussed in detail by Ingwersen and Järvelin (2005). 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 9 years and conducted experiments with human participation, finding, among other things, that a number of issues need a resolution (Dumais & 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 à 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 the conduct of studies in tandem with system design, such as:

- The study of relevance interactions in a variety of manifestations and processes in and beyond retrieval
- The study of cognitive, affective, and situational factors as they dynamically affect relevance and are affected in turn
- The study of human tendencies of least effort for maximum gain as reflected in relevance
- The study of information and relevance contexts and ways to reflect them
- The 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 of the interactive track 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. It should be noted that one cause of the lack of translating results from user studies into system design is the lack of funding mentioned earlier in the article.

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.

Beyond students. As mentioned, students were endlessly used as experimental subjects for relevance experimentation and observation. 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 relevance 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. Of course, there is a vast amount of work on information needs, seeking and use in human information behavior studies that goes beyond students and the classical IR model. Relevance studies should follow.

In Conclusion

Information technology, information systems, and information retrieval will change in ways that we cannot even imagine, not only in the long run, but even in the short term. They are changing at an accelerated pace. But no matter what, relevance is here to stay. Relevance is timeless. Concerns about relevance will always be timely.

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 PhD class, 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

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., & Schamber, L. (1998). User criteria for relevance evaluation: A cross-situational comparison. Information Processing & 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.

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.

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

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

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

Cuadra, C.A., Katter, R.V., Holmes, E.H. & Wallace, E.M. (1967). Experimental studies of relevance judgments: Final report (Vols. 1–3). Santa Monica, CA: System Development Corporation.

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.

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

Dumais, S.T., & Belkin, N.J. (2005). The TREC interactive tracks: Putting the user into search. In E.M. Voorhees & D.K. Harman (Eds.), TREC. Experiment and evaluation in information retrieval (pp. 123–145). Cambridge, MA: MIT Press.

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

Eisenberg, M.B., & 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., & 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., & Crandall, M. (1997). Users' perception of the performance of a filtering system. In N.J. Belkin et al. (Eds.), 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). New York: ACM.

Fitzgerald, M.A., & 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.

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.

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

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

DOI: 10.1002/asi

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

Hansen, P., & 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. (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., & 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.

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

Huang, M., & 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., & Järvelin, K. (2005). The turn: Integration of information seeking and retrieval in context. Amsterdam: Springer.

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 & 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., & 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., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. Information Processing & 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 A. Spink & C. Cole (Eds.), New directions in cognitive information retrieval (pp. 169–186). Amsterdam: Springer.

Kim, G. (2006). Relationship between index term specificity and relevance judgment. Information Processing & Management, 42(5), 1218–1229.

Koenemann, J., & Belkin, N.J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In H-P. Frei et al. (Eds.), Proceedings of the 1996 Annual Conference of the Special Interest Group on Computer–Human Interaction of the, Association for Computing Machinery (CHI 96) (pp. 205–212). New York: ACM.

Lee, H., Belkin, N.J., & Krovitz, B. (2006). Rutgers information retrieval evaluation project on IR performance on different precision levels. Journal of the Korean Society for Information Management, 23(2), 97–111.

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

Levy, D.A. (2000, August/September). Digital libraries and the problem of purpose. Bulletin of the American Society for Information Science, pp. 22–26. D-Lib Magazine, 6(1). Retrieved November 15, 2006, from http://www.dlib.org/dlib/january00/01levy.html

Maglaughlin, K.L., & 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.

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

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

Rees, A.M., & Schultz, D.G. (1967). A field experimental approach to the study of relevance assessments in relation to document searching (Vols. 1–2). Cleveland, OH: Western Reserve University, School of Library Science, Center for Documentation and Communication Research.

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., & 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.

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

Ruthven, I. (2005). Integrating approaches to relevance. In A. Spink & C. Cole (Eds.), New directions in cognitive information retrieval (pp. 61–80). Amsterdam: Springer.

Ruthven, I., Lalmas, M., & 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. (1991). Individual differences in organizing, searching and retrieving information. Proceedings of the American Society for Information Science, 28, 82–86.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. Journal of the American Society for Information Science and Technology, 58, 1915–1933.

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

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

Shaw, W.M., Jr., Wood, J.B., Wood, R.E., & Tibbo, H.R. (1991). The cystic fibrosis database: Content and research opportunities. Library & 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.

Sormunen, E. (2002). Liberal relevance criteria of TREC: Counting on neglible documents? In K. Järvelin et al. (Eds.), 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 02) (pp. 324–330). New York: ACM.

Spink, A., & Cole, C. (Eds.). (2005). New directions in cognitive information retrieval. Amsterdam: Springer.

Spink, A., & 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., Greisdorf, R., & Bateman, J. (1998). From highly relevant to non-relevant: Examining different regions of relevance. Information Processing & Management, 34(5), 599–621.

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

DOI: 10.1002/asi

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

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

Tang, R., & 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., Ruthven, I., & Jose, J.M. (2005). How users assess Web pages for information seeking. Journal of the American Society for Information Science, 56(4), 327–344.

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In W.B. Croft et al. (Eds.), 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). New York: ACM.

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

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

Vakkari, P., & 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.

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

Vorhees, E.M. (2001). Evaluation by highly relevant documents. In D.H. Kraft et al. (Eds.), Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval of the Special Interest Group on Information Retrieval, Association for Computing Machinery (SIGIR 01) (pp. 74–82). New York: ACM.

Voorhees, E.M., & Harman, D.K. (Eds.). (2005). TREC. Experiment and evaluation in information retrieval. Cambridge, MA: MIT Press.

Wallis, P., & Thom, J.A. (1996). Relevance judgments for assessing recall. Information Processing & 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., & Soergel, D. (1999). 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., & 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., & 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.

Xu, Y. (2007) Relevance judgment in epistemic and hedonic information searches. Journal of the American Society for Information Science and Technology, 58(2), 178–189.

Xu, Y.C., & Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality. Journal of the American Society for Information Science and Technology, 57(7), 961–973

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In W.B. Croft et al. (Eds.), 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). New York: ACM.

DOI: 10.1002/asi