Effects of Inconsistent Relevance Judgments on Information Retrieval Test Results: A Historical Perspective[1]

Tefko Saracevic

> Only by continuous self-appraisal can a large information system make itself responsive to the needs of the scientific community.
>
> Concluding sentence in Lancaster (1969)

Abstract

The main objective of information retrieval (IR) systems is to retrieve information or information objects relevant to user requests and possible needs. In IR tests, retrieval effectiveness is established by comparing IR systems retrievals (systems relevance) with users' or user surrogates' assessments (user relevance), where user relevance is treated as the gold standard for performance evaluation. Relevance is a human notion, and establishing relevance by humans is fraught with a number of problems—inconsistency in judgment being one of them. The aim of this critical review is to explore the relationship between relevance on the one hand and testing of IR systems and procedures on the other. Critics of IR tests raised the issue of validity of the IR tests because they were based on relevance judgments that are inconsistent. This review traces and synthesizes experimental studies dealing with (1) inconsistency of relevance judgments by people, (2) effects of such inconsistency on results of IR tests and (3) reasons for retrieval failures. A historical context for these studies and for IR testing is provided including an assessment of Lancaster's (1969) evaluation of MEDLARS and its unique place in the history of IR evaluation.

LIBRARY TRENDS, Vol. 56, No. 4, Spring 2008 ("The Evaluation and Transformation of Information Systems: Essays Honoring the Legacy of F. W. Lancaster," edited by Lorraine J. Haricombe and Keith Russell), pp. 763–783

Introduction

Information retrieval systems came into being shortly after the Second World War addressing the problem of controlling the information explosion, primarily as related to scientific and technical information. Vannevar Bush (1890–1974) is credited with defining the problem and suggesting a solution that caught wide attention. As to the problem, he defined it this way: "The summation of human experience is being expanded at a prodigious rate" and "our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose" (Bush, 1945, p. 2). Bush suggested a technological solution in the form of a device he called *memex*—"a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory" (ibid., p. 6). As yet, memex has not been built. It was a vision. However, the idea of inadequacy of existing methods for controlling the information explosion and of providing a technological solution caught on immediately after the Second World War. Among other things, it affected the development of information retrieval (IR) by using new techniques and systems that rested on technology. Importantly, Bush's ideas were a motivation for funders, such as the National Science Foundation in the United States, to support IR development and testing.

As defined by Calvin Mooers (1919–94), a mathematician, physicist, and pioneer in the field, "information retrieval . . . embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation" (Mooers, 1951, p. 25). Of course, IR systems and techniques have undergone evolutionary and even revolutionary changes since 1951, but basically, they still concentrate on the same aspects Mooers defined.

The difference between IR and related methods and systems that long preceded it—classifications, subject headings, various indexing methods, or bibliographic descriptions, including the contemporary Functional Requirements for Bibliographic Records (IFLA, 1998)—is that IR specifically included "specification for search." The others did not include searching in their specification; searching was simply assumed. In IR, searching is specified in algorithmic detail and the algorithms keep changing and improving. This is the first key difference.

The second key difference was the choice (at the beginning more by assumption than deliberate selection) of relevance as the underlying, basic notion:

> The fundamental notion used in bibliographic description and in all types of classifications or categorizations, including those used in contemporary databases, is *aboutness*. The fundamental notion used in IR

> is *relevance*. It is not about any kind of information, and there are great many, but about *relevant* information. Fundamentally, bibliographic description and classification concentrate on describing and categorizing information objects; IR is also about that but, *and this is a very important "but,"* in addition IR is about searching as well, and searching is about relevance. (Saracevic, 2007a, p. 1917)

Retrieval of relevant information or information objects became and still is the primary objective of IR systems.

The two choices in IR, algorithms for searching and relevance as the basic notion and objective, not only affected but even governed testing that grew to be a very important activity in IR. From the outset of IR testing, which had already started by the mid-1950s, relevance served as the criterion on the basis of which performance of various IR systems or algorithms were compared. Relevance is a human notion and relevance judgments are human assessments, bringing with them all kinds of issues and problems common to many human notions and types of assessments. Well, they are human. One of the issues is that human relevance assessments (like a great many other human assessments) are not consistent, raising the obvious question on the effect of inconsistency in judgments on the results of IR testing.

The aim of this article is to review studies that contained data (as opposed to discussion only) related to questions implied above: *What are the effects of inconsistent human relevance judgments on relative performance of different IR algorithms or approaches? Does inconsistency affect test results?* In the process, I am providing a historical perspective to these questions and to the general description of IR testing that follows. In addition, I am reviewing and honoring the classic test of Wilf Lancaster (1969) that differed in significant ways from IR tests that followed. His was a unique contribution to IR testing.

Note that the present article is an enlargement of one part of the relevance study reported in Saracevic (2007a, 2007b). In that study I dealt comprehensively with relevance as the basic notion in information science while in this review I am focusing and enlarging on the part that dealt with the relation between relevance and information retrieval testing.

Testing in Information Retrieval

From the very start of practical development of IR systems dating to the late 1940s, searching was based on Boolean logic (AND, OR, NOT), even though at the start "Boolean" was not mentioned by name and computing technology was yet to be used (Mooers, 1951; Perry, 1951). Shortly thereafter, coordinate indexing, developed by Mortimer Taube and colleagues at a company named Documentation Inc., was a direct outgrowth of these ideas and it took the IR world by storm; interestingly, Taube referred to coordinate indexing, following Bush, as "association of ideas" (Taube and

Associates, 1955). It was based on *uniterms*, single terms assigned to documents to represent the content, that were later "coordinated" in searching, meaning searched in a Boolean fashion. Uniterms were predecessors of modern techniques in IR. While originally they were assigned and searched by human indexers and searchers, now computers are doing a similar job using various algorithms. In other words, uniterms were a granddaddy of IR. With a wide adoption of coordinate indexing, Boolean logic was fully recognized as the basis for searching in IR. A variety of specific, even competing, approaches and tools were developed and applied in practical realizations of coordinate indexing and IR in general.

Very soon, the perennial questions asked of all systems were raised: *What is the effectiveness and performance of given IR approaches? How do they compare?* It is not surprising that these questions were raised in IR. At the time; most developers, funders, and users associated with IR were engineers or scientists or worked in related areas where the question of testing was natural, even obligatory. In addition, IR testing began in the late 1950s within a certain context as described by Cyril Cleverdon in his acceptance speech for the 1991 Association for Computing Machinery, Special Interest Group on Information Retrieval Gerard Salton Award:

> These new techniques generated considerable argument, not only between the proponents of the different systems, but also among the library establishment, many of whom saw these new methods as degrading their professional mystiques. . . . Controversy over the new methods was still raging, with extravagant claims on one side being countered by absurd arguments on the other side, without any firm data being available to justify either viewpoint. (Cleverdon, 1991, pp. 3, 4)

Kent et al. (1955) were first to propose measures for testing IR effectiveness; they suggested "recall" and "relevance" (later, because of confusion, renamed "precision"), where relevance was the underlying criterion for these measures. Respectively, they measure the probability of agreement between what the system retrieved or failed to retrieve as relevant (systems relevance) and what the user assessed as relevant (user relevance) where user relevance is the gold standard on the basis of which evaluations are made.[2] Other measures were suggested, but not adopted. With some variation on the theme, precision and recall remained standard measures of IR effectiveness to this day with relevance as the underlying criterion.

The first IR test on record was attempted in the early 1950s, as reported by Gull (1956) and recounted later in the section, Inconsistency in Human Relevance Assessments. In short, the test collapsed because of disagreement in relevance assessments between two competing groups. Historically, early IR tests that were most influential were collectively known as "Cranfield tests," done in the 1950s and 1960s at the (U.K.) Cranfield College of Aeronautics (to become Cranfield Institute of Technology in 1969 and Cranfield University in 1993) under the leadership of Cyril Cleverdon (1914–1997). As summarized in Cleverdon (1962, 1967, 1991), the 1962 report refers to Cranfield I and the 1966 and 1967 and in Cleverdon, Mills, & Keen (1966) reports to Cranfield II tests.[3] Cranfield tests also became controversial. For instance, Swanson (1965, 1971), among others, argued that the method of obtaining relevance judgments had influenced the results. Thus, as in the Gull (1956) test, relevance assessments entered again as a point of contention in IR testing. They remain contentious to this day.

In Cranfield I tests, four methods for representing information were compared: Universal Decimal Classification (UDC), alphabetical subject catalog, faceted classification, and uniterms. This was the first and last time that traditional library techniques (the first three) were tested together with a technique representing IR (uniterms). The results were not anticipated by proponents of each system, namely on many counts, the four systems performed pretty much the same:

> No system that has been investigated has shown itself to be so markedly superior as to justify its use in all conditions. . . . The most surprising finding was that "uniterm," as a descriptor language, can be given a high rating on many counts. It achieved the best overall figures in the test, it presented no serious difficulties for the technical searchers . . . and was notably successful with short indexing time. (Cleverdon, Mills, and Keen, 1966, p. 92)

Of course, there were numerous critiques of the tests and findings. Today, it is hard to imagine the emotionalism that followed the test—they were contrary to many firmly held beliefs. My favorite critique that Cleverdon repeated a number of times was: "You had no right to be so intelligent with the uniterm system; it is meant to be used by people of low intellect" (Cleverdon, Mills, & Keen, 1966, p.6).

Cranfield II was devoted to testing various index language devices based on natural language. Thirty-three types of index languages were investigated starting with single terms and then adding word forms and synonyms; broader, related, and narrower terms; and term phrases, hierarchies, and combinations thereof, with alterations of levels of specificity and exhaustivity of indexing (Cleverdon, 1967). Some results were surprising, even revolutionary at the time: "Neither we nor anybody else had considered it as remotely possible that an index language based on single terms in the natural language of the documents would be so effective that the performance could only be improved by confounding word forms or true synonyms" (Cleverdon 1991, p. 8). This can be done by computers. The Cranfield results paved the way.

Cranfield tests were significant for two other reasons. First, they established a model of IR, called the traditional or laboratory IR model, that was used in IR testing later by Gerard Salton (1927–95) in the famous SMART experiments (summarized in Salton, 1971 and Salton & McGill, 1983), that later morphed into the comprehensive Text Retrieval Conference (TREC) experiments conducted from 1992 to date (Voorhees & Harman 2005).[4] Unlike Cranfield tests, SMART and TREC were fully automated. The model that came out of Cranfield tests has been in continuous use in IR testing for half a century. The emphasis in the model is on processing information objects by IR systems and then matching them with queries to produce retrieved results. The processing and matching is algorithmic; the goal of the algorithms is to maximize retrieval of relevant information or information objects. In the purest form of this model, the user is represented by a query only and not considered beyond that at all; also, interaction with anything outside the system is not a consideration, as if the system is a self-contained black box. Relevance assessments are done by a user, or user surrogate, and the effectiveness of retrieved outputs, using different approaches or algorithms, is compared to these assessments. Testing is based on a number of assumptions, one of them being that human judgments of relevance are consistent (Saracevic 2007b, p. 2132). Needless to say, the evident restrictions of the model came under numerous critiques, more recently and thoroughly by Ingwersen & Järvelin (2005).

Second, for the first time in Cranfield tests the familiar precision-recall graphs were drawn and the "law" of inverse performance between recall and precision was formulated (Cleverdon, 1962, pp. 72, 89, 90). To this day, graphing of precision-recall figures is an established way to demonstrate and compare performance, and improving on the inverse relation is a major goal of most procedures in IR tests.

SMART tests also signified a departure of IR from the original Boolean logic for searching and retrieval to more sophisticated approaches that allowed for different information organizations and subsequent outputs, such as ranking and clustering by relevance, where relevance is determined by the system, of course. A variety of approaches and algorithms were used and tested, so tests became more involved as well. TREC further extended these approaches and algorithms, even involving numerous new areas for IR, such as retrieval of recordings of speech, across multiple languages and much more, as recounted on the TREC site, http://trec.nist.gov/. Not surprisingly, IR tests became still more involved.

Determining Relevance in Information Retrieval Tests

As mentioned, IR tests are based on comparing systems relevance—responses to a query that a system deemed and retrieved as relevant following whatever procedure—and user relevance—user's (or a surrogate's) assessment as to relevance of retrieved answers or of any information or information objects in the system, even if not retrieved. User relevance is the gold standard against which system relevance, that is, system performance, is compared. Thus, performance assessment of a given system (algorithm, procedure . . .) follows from and is based on human judgment of relevance of given information or information object to a given query or need. The key issue is obtaining acceptable relevance judgments that can then be used as a standard for calculating recall and precision. Once these are obtained, calculations are straightforward. Well, almost. The assessments have to involve not only the retrieved answers, but also all potentially relevant documents in the collection (or in a representative sample, or in a pooled set of answers) so that recall can be calculated. One of the best descriptions of these and other requirements of IR testing was concisely provided by Tague-Sutcliffe (1992).

Establishing this gold standard is one of the main problems, even conundrums, of IR testing. Not surprisingly then, in many reports of IR tests, the critical step showing how relevant objects became relevant is often shrouded in mystery. Or, it is glossed over. Or, it is accepted from a previous source without further ado. Or some collective group, such as "judges" or "librarians" or "searchers" or "students" is mentioned as bearing the responsibility. Or, some such explanation. It is hard to get at it.

The objective of relevance judgments in IR tests is to get as close as possible to real-life situations so that test results would have real-life validity. This is very, very difficult to achieve. Thus, simulation methods have been developed. Basically, there are four methods by which relevance judgments have been obtained that are regarded as gold standards:

1. By the user or questioner—person who posed own question made the judgment as well;
2. By a user surrogate(s)—such as a specialist (or by consensus of a group of specialists) who perform judgments on the topic of a given question in their specialty;
3. By an information professional (or by consensus of a group of professionals) who is professionally entrusted or involved with some aspect of the process, who performs judgments on the topic of a given question that is not necessarily in their specialty, but is familiar with what is going on; and
4. By "bystanders" signifying none of the above—for example, by students asked to do a given task of judgment, including possible prescreening.

The first method involves "real users" and the others "laboratory-type users." Here are some examples. In Cranfield I, "the search questions had been obtained from several hundred individuals in 58 different organisations, mainly in England and America. Each question was based on a single document in the test collection, and a search was considered successful if that particular paper was located in the catalogue" (Cleverdon, 1991, p. 4; full report in Cleverdon, 1962, pp. 8–9, 52). This is a variation of the theme of the second method above. Questions came from an unknown number of individual specialists who were asked to pose a question(s) on the basis of a source document, and the gold standard was the document from which the question came. But additional documents were retrieved, and the issue became how to deal with them as to relevance. These "were assessed in relation to the appropriate question" (ibid., p. 52). Presumably, the project members did the additional relevance assessments, thus bringing in the third method. In Cranfield II, the procedure for getting the gold standard was changed: a number of authors of recent research papers (in aeronautics) provided a question based on the problem that led to the research, together with more questions that arose during the conduct of research; the authors also were given a set of references to judge as to their relevance to these questions (Cleverdon, Mills, & Keen, 1966, p. 16). The source documents and evaluated references comprised the gold standard for each question. This is a combination of the first and second method. However, some prescreening also was done by students, so the fourth, or bystander, method was used as well. Generously, the Cranfield collection with relevance assessments was provided as open source for sharing. Subsequently, it was used in many IR tests, including SMART. With this, Cranfield relevance assessments migrated as well.

All IR tests that followed used one or more of these methods for establishing gold standards, the first method used the least because it is the most difficult to secure. Here is a sampling: Lancaster (1969) and Saracevic et al. (1988) used the first method; SMART test collections used the second and third method; TREC uses the second method, with some derivative tests using the third and fourth method; Shaw et al. (1991) used the second and third method. Needless to say, all of these tests faced similar difficulties as the Cranfield tests in obtaining gold standards, but subsequently, all abandoned the use of a source document as the standard the way it had been used in the Cranfield tests. In some form or other, sometimes real users but mostly surrogates— specialists, information professionals, or bystanders—were the ultimate relevance judges for gold standards.

Analysis of Retrieval Failures in IR Tests

For any system or process, diagnosing the reason(s) for failure is often a key issue in testing in general. Here, we are considering IR tests where analysis of failures was done on the basis of retrieval effectiveness measures, namely precision and recall. These were: the Cranfield I test, (failure was not analyzed in Cranfield II), Lancaster (1969) test of MEDLARS, and Blair & Maron (1985) tests of a legal collection. That's it. Diagnosing failure has not become a part of major IR tests. Thus, we are dealing here with a very limited universe. Just to mention a connection: Wilfrid Lancaster was in 1963 a member of the Cranfield team.

Analysis of failures was one of the objectives of the Cranfield I test. By failure it was meant "analysis of all cases . . . where source document was not retrieved" (Cleverdon, 1962, p. 38). The reasons for failure were classified as to (1) question (six reasons), (2) indexing (ten reasons), (3) searching (six reasons), and (4) system (six reasons). The analysis to determine causes of failure proved to be time consuming, from one to two hours per case, and complex, often involving consultation. The results indicated that the following percentages of failures were due to factors related to: question, 17 percent; indexing, 60 percent; searching, 17 percent; and system, 6 percent. Human decisions were most often causes for failure, particularly as to how questions were handled and interpreted, how indexing was done, and how searching was conducted.

Lancaster (1969) conducted a large and comprehensive evaluation of MEDLARS (Medical Literature Analysis and Retrieval System) operated by the U.S. National Library of Medicine. At the time it was a computerized system for retrospective searching on demand and had some 800,000 citations. When MEDLARS moved online it became Medline, the most widely used biomedical resource in the world that annually adds some 600,000 articles. Lancaster's was not a laboratory evaluation. It involved 299 regular, real questions posed over a twelve-month period by MEDLINE users who agreed to be part of the study. Users received a random sample of 25 to 30 retrieved articles plus additional articles found by means outside of MEDLARS (known by requesters as relevant searches outside MEDLARS) and evaluated these articles as to relevance to their request. (Additional articles were supplied in order to create a base for calculation of recall.) The average precision was 50 percent and recall was 58 percent—these figures were later widely used as general indicators of performance for IR systems. But Lancaster cautioned that averages can be misleading—some searches operated with high precision and recall at the same time, while others with very low recall.

Lancaster analyzed two types of failures: recall failures (relevant documents that were not retrieved) and precision failures (retrieved documents that were not relevant). There were 797 recall failures and 3,038 precision failures. As to recall failures 10 percent were due to index language, 35 percent due to searching, 37 percent due to indexing, and 25 percent due to inadequate user-system interaction. (A document can be missed due to more than one cause, thus the percentages add to more than 100.) As to precision failures 36 percent were due to index language, 32 percent due to searching, 13 percent due to indexing, 17 percent due to inadequate user-system interaction, and 2 percent due to value judgment. A large number of failures were due to inadequate searching and user-computer interaction; Lancaster made a number of suggestions on how to improve them. These suggestions are still relevant today. In practice, searching and human-computer interactions still involve a great many human decisions, no matter how automated and sophisticated the systems may be.

Here follows a summary of another large study involving failure analysis. It is also the last study of this kind. Blair & Maron (1985) conducted a study that involved retrieval from a system named STAIRS (Storage and Information Retrieval System) developed by IBM that automatically indexed full texts of documents. Like Lancaster's, the test was not laboratory but real-life based. The collection involved 40,000 documents (about 350,000 pages of text) that were assembled and used in the defense of a large corporate lawsuit. Two lawyers, principal defense attorneys in the suit, generated fifty-one information requests that were searched by paralegals who were also information professionals. The searches were repeated until lawyers (requestors) indicated that they had enough relevant information to defend the lawsuit on that issue or question. Lawyers indicated the relevance of answers. Precision, as always, was easily calculated. To establish a recall base, Blair and Maron also included answers from "sample frames consisting of subsets of the unretrieved database that we believed to be rich in relevant documents" and took random samples from these subsets—these were also provided to lawyers for judging. Precision was 79 percent but recall was 20 percent—which they considered a surprisingly low figure. They gave reasons for "deterioration of recall" (i.e., the system retrieving only one in five relevant documents) as being due to the large file size, restrictions of natural language indexing, and failures in searching. They did not provide figures for each reason, only examples. Test results became controversial, as were all test results from IR testing. Salton (1986) provided a critique of the test by showing examples from the other test and concluded at the outset: "that not only is this level of performance typical of what is achievable in existing, operational retrieval environments, but that it actually represents a high order of retrieval effectiveness" (ibid., p. 649). Blair & Maron (1990) answered and clarified the results. In essence, Salton defended full-text indexing vigorously by questioning Blair & Maron's conclusion about the ineffectiveness of automatic full-text indexing. Today, the controversy is forgotten. Full-text indexing is fully accepted, but failure analyses, a la Lancaster and Blair & Maron are no longer conducted.

A lot can be learned from failure analyses, particularly about human performance. Regrettably, failure tests are no longer conducted, mostly because they are complex, very time consuming, and CANNOT be done by a computer. This type of testing is now relegated to history. Lancaster is the major contributor to that history. His explanation of difficulties also provides the reasons why we have not seen more failure tests:

> The "hindsight" analysis of a search failure is the most challenging aspect of the evaluation process. It involves, for each "failure," an examination of the full text of the document; the indexing record for this document (i.e., the index terms assigned . . .); the request statement; the search formulation upon which the search was conducted; the requester's completed assessment forms, particularly the reasons for articles being judged "of no value"; and any other information supplied by the requester. On the basis of all these records, a decision is made

> as to the prime cause or causes of the particular failure under review. (Lancaster, 1969, p. 123)

Inconsistency in Human Relevance Assessments

People differ, sometimes considerably, in decisions related to a variety of information processes, such as indexing, classification, searching, and yes, relevance as well. Measured are individual or group differences in terms of a degree of agreement/disagreement, overlap, or inter- or intraconsistency. For illustration here are some results from studies of individual differences in information processes other than relevance:

- In a recent study of inter-indexer consistency, Medelyan & Witten (2006) found an average consistency of 38 percent according to one measure and 49.5 percent with another measure, while in an older study Zunde & Dexter (1969) found indexing consistency of 24 percent according to one and 41 percent according to another measure (averages differ depending on what measure is used—measures are not standardized).
- In studies of selection of search terms for the same questions by different searchers, Iivonen (1995) found 40.3 percent consistency for specific and 24.4 percent for general searches, and *Saracevic, Chamis, & Trivison Kantor* (1988) found that the mean overlap was 27 percent.

In information science, observations of relevance inconsistency started with IR tests. As mentioned, Gull (1956) reported on the first study aimed at IR evaluation. The study is worth recounting because inadvertently it showed that relevance assessments differ significantly among groups of judges.[5] Actually, consistency of relevance judgments was not the purpose of the study at all. IR evaluation was. The original 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 Documentation Inc. using coordinate indexing uniterms, that is, index terms searched in Boolean manner. In the test, each group indexed separately the same 15,000 documents, searched 98 requests, and then *separately* judged retrieved answers as to relevance. Then, not the performance of different systems, but the relevance judgments became contentious. The first group found that 2,200 documents were relevant to the 98 requests, while 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. Then they tried to reconcile and considered each others' relevant documents and again compared judgments. Each group accepted some more as relevant, but in the end, they still disagreed; their rate of agreement, at the end was 30.9 percent. The first-ever IR test did not continue.

Cleverdon was very much aware of this study and discussed it and the associated relevance problems at some length in both the 1962 and 1966 reports. The collapse of Gull's study influenced Cleverdon's selection of the method for obtaining relevance judgments, as it did every IR test done since then. The lesson was learned: Never, ever use more than a single judge (or a single object, such as source document) for establishing the gold standard for comparison. No test ever does.

With the test fiasco reported by Gull (1956), the whole field of information retrieval became very conscious of the fact that human relevance judgments are not consistent. It was a rude awakening. Not unexpectedly, researchers started asking: *How consistent, or rather how inconsistent are relevance judgments*? and *What factors affect consistency?*

Consistency or rather inconsistency of relevance judgments became an object of study in a number of experiments. For some studies, this was one of a number of objectives (e.g., Rees & Schultz, 1967), for others this was the main objective (e.g., Sormunen, 2002), while still for others, like in the Gull study, this was not an objective at all, but data on relevance judgment consistency can be derived (e.g., Haynes et al., 1990).

Table 1 provides a list of studies with relevance consistency data—this is not just a representative sample, but almost the total universe of such studies. Other consistency data can be derived from studies presented in Table 2 in the next section, where all the studies were of the third category mentioned above (objective different, but consistency data derivable).

Studies are summarized following the pattern: "[*author*] used [*subjects*] to do [*tasks*] in order to study [*object of research*]." In this way, the sample, method, research question, and results are put together for direct familiarization and for observation of considerable differences between various studies, which make generalizations difficult and hypothetical. Note that seven of the ten studies in the table were also reviewed in Saracevic (2007b); three older studies (Resnick & Savage, 1964, Rees & Schultz, 1967, and Cuadra et al., 1967) were added here to provide a longer historical perspective.

Table 1. Studies Reporting on Consistency of Relevance Judgments.

Resnick & Savage (1964) in the first relevance consistency study on record, used forty-six technical professionals to assess relevance of thirty-four technical reports and patent disclosures to indicate which of these are relevant to their interest in order to observe intra-consistency of relevance judgments. The judges were divided into four groups each receiving a different representation—full text, citation, abstract, including citation, and title. The experiment was repeated after one month. Respectively, intra-relevance agreements on judgments were for full documents 54%, for citations 70%, for abstracts 61%, and for titles 63%.

Rees & Schultz (1967) used a total of 153 judges divided in seven groups (as listed below) that were given sixteen documents in diabetes related to a real research project to judge the relevance of the documents to each of three research stages in order to, among others, observe the inter-consistency of relevance judgments by each group. Respectively, inter-relevance agreement for twenty-one medical librarians—searchers was 44%, twenty-one

Table 1. *continued.*

medical librarians—non-searchers was 40%, fourteen medical experts—researchers was 58%, fourteen medical experts—non-researchers was 56%, twenty-nine scientists was 55%, twenty-five residents was 51% and twenty-nine medical students was 50%.

Cuadra & Katter (1967) used 230 seniors and graduate students in psychology (with different levels of experience) to rate relevance of each of nine psychology journal abstracts against each of eight short information requirement statements in order, among others, to observe the degree of inter-judge agreement in relevance ratings as related to the level of training of the judges in the filed. Four levels of experience were established. The inter-judge correlations for the four experience levels from lowest to highest were .41, .41, .49, and .44.

Haynes et al. (1990) studied MEDLINE use in a clinical setting and not relevance consistency. However, their report does include data from which consistency rates can be derived. They used forty-seven 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 thirteen physicians experienced in searching and three librarians to replicate 78 of those searches where relevance was judged by a physician with clinical expertise in the topic area in order to compare retrieval of relevant citations according to expertise. For the replicated searches, all searcher groups retrieved some relevant articles, but only 53 of the 1,525 relevant articles (3.5%) were retrieved by all three search groups. This is the only real-life study on the question.

Shaw, Wood, Wood, & Tibbo (1991) used four judges to assess the relevance of 1,239 documents in the cystic fibrosis test collection to 100 queries. Judged documents were divided into four sets: A from query author/researcher on the subject, B from 9 other researchers, C from four postdoctoral fellows, and D from one medical bibliographer, in order to enable performance evaluations of different IR representations and techniques using any or all of the judgment sets. The overall agreement between judgment sets was 40%.

Janes & McKinney (1992) used four students as users with information requests to judge as to relevance two sets of retrieved documents that differed in the amount of information presented (primary judges) and then used four undergraduate students without and four graduate students with searching expertise (secondary judges) to re-judge the two sets in order to compare changes in judgments due to increase in provided information between primary and secondary judges. The overlap in judgment of relevant documents (calculated here as sensitivity) between all secondary judges and primary judges was 68%.

Janes (1994) used thirteen students inexperienced in searching, twenty experienced student searchers and fifteen librarians to re-judge twenty documents in each of two topics that were previously judged as to relevance by users in order to compare users' versus non-users' relevance judgments. The overall agreement in ratings between original users' judgments and judgments of the three groups was 57% and 72% for the respective document sets.

Sormunen (2002) used nine master's students to reassess 5,271 documents already judged on relevance in thirty-eight topics in TREC-7 and 8 on a graded four-point scale (as opposed to a binary scale used in TREC) in order to compare the distribution of agreement on relevance judgment between original TREC and newly reassessed documents and seek resolution in cases of disagreement. He found that 25% of documents rated relevant in TREC were rated not relevant by the new assessors; 36% of those relevant in TREC were marginally relevant; and 1% of documents rated not relevant in TREC were rated relevant.

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

Table 1. *continued.*

Lee, Belkin, & Krovitz (2006) used ten experienced searchers (not indicated as to status) to compare two lists of thirty documents each for ten TREC topics. The documents were beforehand judged as to relevance by three judges; 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 in order 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%

Before making conclusions, here is a note of caution. As was mentioned in Saracevic (2007b, p. 2129), for synthesizing findings caveat abound:

> Numerous aspects of the studies reviewed can be questioned and criticized. Easily! Criteria, measures, and methods used in these studies are not standardized. While no study was an island, each study was done more or less on its own. . . . Thus, the results are hardly comparable. 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. Summary conclusions . . . derived from the studies reviewed should be really treated as hypotheses.

From the nine studies in Table 1 and from data in seven studies in Table 2 reported in the next section, we can draw some hypothetical generalizations (Saracevic, 2007b, p. 2137):

> The inter- and intra-consistency or overlap in relevance judgments varies widely from population to population and even from experiment to experiment, making generalizations particularly difficult and tentative.

- 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 percent.
- *Higher expertise* results in a *larger overlap. Lower expertise* results in *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.*

Tests of Using Human Relevance Judgments in IR Tests

Cranfield and SMART tests and later TREC tests as well, stirred a wide debate and generated a considerable amount of harsh criticism. Critics concentrated especially on relevance judgments used as gold standards—on methods by which they were obtained, on their inadequacy, shortcomings, and so on (e.g,. Swanson, 1965, 1971). The critiques are succinctly summarized by Harter (1996, pp. 37, 38, 43, 45):

> Relevance judgments form the bedrock on which traditional experimental evaluation model is constructed. . . . Relevance assessments are anything but stable and they vary significantly depending on the variable being investigated. . . . That variations in relevance judgments are likely to change the values of recall and precision is obvious. . . . We can no longer rest the evaluation of information retrieval systems on the assumption that such variations do not significantly affect the measurement of information retrieval performance. . . . On the other hand, the reaction to this research [showing variations in relevance judgments] and criticism from experimental researchers who use relevance assessment to conduct Cranfield-like experiments on information retrieval systems has been mostly silence . . . with very few exceptions [As exceptions, Harter discusses studies by Lesk & Salton, 1968; Cleverdon, 1970; Kazhdan, 1979; and Burgin, 1992 included in Table 2; mostly, he dismisses them because of "their lack of involvement with the variables associated with real users."]

Despite sometimes emotional criticism, Harter (and others in the same vein) raises serious and even critical questions: *Given that relevance judgments are inconsistent, which they are to various degrees as amply demonstrated, how does this affect results of IR evaluation? Because of that, are IR test results valid, reliable and to be trusted in a scientific sense?* Answers need to be decisive for accepting results of such tests.

There were seven experimental studies conducted to date trying to answer these questions—I believe this is the whole universe of such studies. Considering hundreds of IR tests done over the years since Cranfield, this is a small universe; nevertheless, I do not believe they can be dismissed as Harter (1996) did. Table 2 presents descriptions of and conclusions from these seven studies.

Table 2. Studies Reporting on the Effect of Inconsistency of Relevance Judgments on IR Test Results

Lesk & Salton (1968) used eight students or librarians (not specified as to which) who posed forty-eight 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 each of the eight judges to assess for relevance six additional queries not of his/her own in order to rank system performance obtained using four different judgments sets (A, B, their intersection and union). They found that the overall agreement between original assessors (A) and eight new assessors (B) was 30% and concluded after testing three different IR techniques that all sets of relevance judgments produce stable performance ranking of the three techniques.

Table 2. *continued.*

Cleverdon (1970) used three subject experts in aerodynamics (the field of the collection) to separately judge relevance of documents retrieved for forty-two questions in Cranfield II tests for which known relevance scores were originally established by users in order to observe "whether the new sets of relevance decisions made any significant difference in the order of merit, as determined by the normalized recall of the indexing language" (ibid., p. 11). Nineteen indexing languages were tested. Rank correlation showed that relevance decisions by different judges did not significantly affect the comparative results of original rankings for these languages—the rank correlation between original results and three new sets was .92, .92, and .94 respectively. Overall agreement in relevance decisions was not given, although it could be calculated from data in appendices.

Kazhdan (1979) took the findings from the Lesk & Salton (1968) study as a hypothesis and used a collection of 2,600 documents in electrical engineering that had sixty queries with two sets of relevance judgments—one from a single expert and the other from a group of thirty experts—in evaluating seven different document representations in order to compare the performance of different representations in relation to different judgment sets. He found that Lesk & Salton hypothesis is confirmed: the relative ranking of the seven different representations remained the same over two sets of judgments; however, there was one exception where ranking changed.

Burgin (1992) used a collection of 1,239 documents in the cystic fibrosis collection (Shaw et al., 1991) that had one hundred 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. 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 & 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 in order 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 did 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 re-judged 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 study, an unspecified number of assessors from a different and independent institution, Waterloo University, judged more than 13,000 documents for relevance related to fifty TREC-6 topics; next, the performance of seven-four IR techniques was evaluated using three sets of judgments (primary, Waterloo union and intersection). Both studies were done in order to look at the effect of relevance assessments by different judges on the performance ranking of the different IR techniques tested. She found that in the first study, the mean overlap between all assessors (primary and secondary) was 30%, and in the second study, 33%. After testing thirty-three different IR techniques in the first and seventy-four 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.

Table 2. *continued.*

Voorhees (2001) used fifty topics created for the TREC-9 Web track and asked assessors to judge retrieved pages on a three 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 re-assessed by two other secondary assessors. All assessors were also asked to identify the best page or pages for a topic. The study was done in order 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." (ibid., p. 76) This conclusion contradicts the finding of the previous (Voorhees, 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" (ibid., p. 77). The agreement among three assessors as to the best pages for a topic was 34%.

Before making conclusions, note that the same caveats mentioned above apply to these studies as well. Here are some hypothetical generalizations derived from data in seven studies in Table 2 and summarized in Saracevic (2007b, p. 2138):

> 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 bear it out. However, note that the use of average 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.

Conclusions

The basic aim of IR systems is to provide information that is relevant to user questions and possible needs. Thus, relevance became the criterion for measures of the effectiveness of performance for IR systems and procedures. IR tests are based on comparing systems relevance with user relevance, where user relevance assessments serve as the gold standard for comparison and evaluation. Relevance is a human notion, and establishing relevance by humans is fraught with a number of problems, inconsistency in judgment being one of them. The aim of this review is to explore the relation between relevance on the one hand and testing of information retrieval systems and procedures on the other. In the process, a historical perspective is provided on the testing of IR systems, and on studies that addressed the inconsistency of relevance judgments and the effect of that inconsistency on results of IR tests.

Conclusions from these studies are provided as hypothetical generalizations (with proper caveats) at the end of the last two sections. Thus, they are not repeated here. Instead, some general observations about IR tests are made here in conclusion.

Information retrieval has a proud history. It started right at the conclusion of the Second World War by addressing the problem of information explosion, particularly in science and technology, and applying modern information technology as a solution. Over the ensuing decades, IR systems and techniques spread worldwide and are successfully used in a great many endeavors, including the contemporary search engines. In part, this is due to advances in information technology—databases are larger and enable inclusion of full texts, not just representations as when IR started—searches are faster, interfaces more elaborate and flexible, and so on. And in part, this is also due to improvements in IR algorithms and procedures. But again, in many respects, these were predicated on advances in technology. The two are intertwined.

It is true that human relevance judgments are affected by a host of factors that produce significant individual and group disagreements. Tests and pragmatic experiences, as well as common sense, have shown that. Concluding that there are no effects of inconsistent relevance judgments on rank order of tested IR procedures, as optimistically proclaimed in early tests, may not be completely warranted. Averaging has an effect; rank switches do occur at times, and the issue needs a lot of further research.

But it is also easily observable that significant advances were made over decades in IR. By many pragmatic ways of figuring, contemporary IR systems and processes are better than those of a few decades ago. Along with technology, testing played a major role in improvements of IR algorithms and processes. In other words, despite observed relevance problems from the human side, IR systems improved from the systems side.

On the historical side, it is quite interesting, if not amazing, to note that the basic methodological principles and model for testing laid down a half century ago are still governing IR testing today. IR testing is like a river that became broader and deeper but never changed its course. The course seems to be cemented.

IR systems, as conceptualized, will never get away from relevance. For people, relevance is here to stay. Thus, it is here to stay with all associated problems for IR systems as well.

Acknowledgments

First of all, I wish to thank researchers who reported on experiments reviewed here. They did the hard work and provided data. Moreover, they were the inspiration for this review.

Thanks to Yuelin Li and Ying Zhang, my assistants at the time, who tirelessly searched the literature for sources about relevance and then organized them. As with my previous work, Sandra Lanman's editing and thoughtful suggestions were really relevant.

I also wish to thank Keith Russell and Lorraine Haricombe, editors for this issue, for their invitation to contribute to this Festschrift in honor of Professor F. W. Lancaster. Wilf is a valued friend of many years. The invitation was an honor.

Notes

1. Parts of this paper were reported in Saracevic (2007a and 2007b). Verbatim quotes are clearly indicated.
2. Recall can be defined as probability that a relevant information object will be retrieved and precision that a retrieved object will be relevant.
3. Interestingly enough, Cranfield tests did not use a computer but simulated computer searching: "At that time there was no program which was remotely capable of doing what was required but fortunately a member of my staff, Michael Keen, came up with an ingenious idea which allowed us to simulate computer searching, albeit with considerable clerical effort." (Cleverdon, 1991, p. 8)

4 TREC is a long-term effort at the [US] National Institute for Standards and Technology (NIST), that brings various IR teams together annually to compare results from different IR approaches under laboratory conditions.

5 This study and studies that follow are reported and commented upon in Saracevic (2007b, pp. 2134ff.).

References

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM,* 28(3), 289–291.

Blair, D. C., & Maron, M. E. (1990). Full-text information retrieval: Further analysis and clarification. *Information Processing & Management, 26*(3), 437–447.

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

Bush, V. (1945). As we may think. *Atlantic Monthly, 176*(11), 101–108. Retrieved Nov. 7, 2007, from http://www.theatlantic.com/doc/194507/bush.

Cleverdon, C. W. (1962). *Report on the testing and analysis of an investigation into the comparative efficiency of indexing.* Cranfield, UK: ASLIB Cranfield Research Project. Retrieved Nov. 17, 2007, from http://hdl.handle.net/1826/836.

Cleverdon, C. W. (1967). The Cranfield tests on index language devices. *Aslib Proceedings, 19*(6), 173–194.

Cleverdon, C. W. (1970). *The effect of variations in relevance assessments in comparative experimental tests of indexing languages.* Cranfield, UK: Cranfield Library Report no.3. Retrieved Nov. 16, 2007, from http://hdl.handle.net/1826/967.

Cleverdon, C. W. (1991).The significance of the Cranfield tests on index languages. *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 1–3.

Cleverdon, Cyril W., Mills, Jack, & Keen, Michael (1966). *Factors determining the performance of indexing systems; Volume 1, Design; Part 1, Text.* Retrieved Nov. 17, 2007, from http://hdl.handle.net/1826/861.

Cuadra, C. A., Katter, R. V., Holmes, E. H., & Wallace, E. M. (1967). *Experimental Studies of*

Relevance Judgments: Final Report. 3 vols. Santa Monica, CA: System Development Corporation. NTIS: PB-175 518/XAB, PB-175 517/XAB, PB-175 567/XAB.

Gull, C. D. (1956). Seven years of work on the organization of materials in special library. *American Documentation, 7*(4), 320–329.

Harter, S. P. (1971). The Cranfield II relevance assessments: A critical evaluation. *Library Quarterly, 41*(3), 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.

Iivonen, M. (1995). Consistency in the selection of search concepts and search terms. *Information Processing & Management, 31*(2), 173–190.

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

International Federation of Library Association and Institutions (IFLA) (1998). *Functional Requirements for Bibliographic Records—Final Report.* Retrieved Nov. 15, 2007 from: http://www.ifla.org/VII/s13/frbr/frbr1.htm#2.1.

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.

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

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.

Lancaster, F. W. (1969). MEDLARS: Report on the evaluation of its operating efficiency. *American Documentation, 20*(2), 119–142.

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.

Medelyan, O., & Witten, I. H. (2006). Measuring inter-indexer consistency using a thesaurus. *Proceedings of the 2006 ACM/IEEE Joint Conference on Digital Libraries.* 274–275

Mooers, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *American Documentation, 2*(1), 20–32

Perry, J. W. (1951). Superimposed punching of numerical codes on handsorted punched cards. *American Documentation, 2*(4), 205–212.

Rees, A. M., & Schultz, D. G. (1967) *A field experimental approach to the study of relevance assessments in relation to document searching.* 2 vols. Cleveland, OH: Western Reserve University, School of Library Science, Center for Documentation and Communication Research. NTIS: PB-176 080/XAB, PB-176 079/XAB. ERIC: ED027909, ED027910.

Resnick, A., & Savage, T. R. (1964). The consistence of human judgments of relevance. *American Documentation, 15*(2), 93–95.

Salton, G. (Ed.). (1971). *The SMART retrieval system: Experiments in automatic document processing.* Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 29*(7), 648–656.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw Hill.

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

Saracevic, T. (2007a). 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*(3), 1915–1933.

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology, 58*(13), 2126–2144.

Saracevic, T., Kantor. P., Chamis, A. Y., & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science, 39*(3), 161–176.

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.

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

Swanson, D. R. (1965) Evidence underlying the Cranfield results. *Library Quarterly, 35*(1), 1–20

Swanson, D. R. (1971). Some unexplained aspects of the Cranfield tests of indexing performance factors. *Library Quarterly, 41*(3), 223–228.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation revisited. *Information Processing & Management, 28*(4): 467–490.

Taube, M. and Associates. (1955). Storage and retrieval of information by means of the association of ideas. *American Documentation, 6*(1), 1–17.

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.

Voorhees, E. M. (2001). Evaluation by highly relevant documents. *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*, 74–82.

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.

Zunde, P., & Dexter, M. E. (1969). Indexing consistency and quality. *American Documentation, 20*(3), 259–267.

Tefko Saracevic is Professor II at School of Communication, Information and Library Studies, Rutgers University in New Brunswick, New Jersey. He was the president of the American Society for Information Science and received the Society's Award of Merit (the highest award given by the society). He also received the Gerard Salton Award for Excellence in Research, by the Special Interest Group on Information Retrieval, Association for Computing Machinery (also the highest award given by the group). In a histogram of citations from papers in the *Journal of the American Society for Information Science and Technology* (JASIST & predecessor names), done by Eugene Garfield from the Web of Science for years 1956–2004 and involving 3,575 authors, Tefko Saracevic ranked first in citations to his work both in articles in the Journal (Total Local Citation Score), as well in articles globally from that Journal (Total Global Citation Score).