International Perspectives on the History of Information Science and Technology

Proceedings of the ASIS&T 2012 Pre-Conference on the History of ASIS&T and Information Science and Technology

Edited by
Toni Carbo and
Trudi Bellardo Hahn

asis&t
ASIST Monograph Series

Research on Relevance in Information Science: A Historical Perspective

Tefko Saracevic

Abstract

Relevance is a fundamental concept in information science. The aim of this paper is to provide a historical perspective on two large questions: *Why did relevance become a central notion of information science?* and *What did we learn about relevance through research in information science?* As to the first question, there are no historical documents addressing it, thus conclusions are derivative. Relevance emerged as a central notion in information science because of extensive theoretical and practical concerns with and commitments to searching and not only with organization of information. In turn, searching was allowed using modern information technology. Aboutness is a fundamental notion related to *organization of information*, while relevance is a fundamental notion related to *searching and retrieval of information.* As to the second question there are more than 300+articles that contain experimental, empirical or observational data about relevance and about a dozen or so articles about relevance that are based on scholarly, primarily philosophical, argumentation. A sample of these is used to synthesize major areas of relevance research. The relation of information retrieval testing to relevance and related research is explored. In the aftermath of the first IR tests, two large projects were funded at the end of the 1960s – their effect is synthesized. Also addressed is research on dynamics and situational aspects of relevance. Conclusions concentrate on general achievements and failures of relevance research.

Relevance:

1. *relation to the matter at hand*
2. *the ability (as of an information retrieval system) to retrieve material that satisfies the needs of the user.*

Merriam-Webster Dictionary Online

Introduction

Every field has some central idea or ideas. Retrieval of *relevant information* and not just any kind of information (and there are many) is a central idea of information science. Information retrieval (IR), a major branch of information science, is about *relevant* information. Thus, the notion of relevance is fundamental to information science.

As most fundamental notions, relevance is intuitively well understood—nobody has to explain it to anybody in the world. That is its strength. That is why the systems aiming at retrieval of relevant information to users, including search engines, are so well accepted globally—differences in cultures, societies, and mores do not matter. But relevance is a human, not technical, notion. And that is its weakness. As are all human notions relevance is messy, built in with many variables that are hard to control and fathom formally. All the algorithms in all the systems in the world are trying to approximate, with various degrees of success, the human notion of relevance. That is what they are all about.

The aim of this paper is to provide a historical perspective on two large questions:

1. *Why did relevance become a central notion of information science?*
2. *What did we learn about relevance through research in information science?*

Sources of Information

The first question, why relevance, was not contemplated to any extent in information science; it seems not to be of interest, since relevance as a fundamental notion is simply taken as a given. Thus, there are no direct historical documents to contemplate and the discussion here is derivative.

The second question, what we learned, can be synthesized from about 300 or so articles that contain experimental, empirical or observational data about relevance (cited in the papers listed in the next paragraph) plus about a dozen or so articles about relevance that are based on philosophical argumentation. Because the length of the paper is restricted, for each conclusion or opinion only a representative article is cited. Articles that discuss relevance in a contemplative, sage-on-stage, manner (and there are quite a few) are ignored.

Relevance was a subject of a number of reviews that appeared over time. Further information, particularly of historical nature, can be found in these reviews. Among these are reviews by Schamber, Eisenberg & Nilan (1990), Shamber (1994), Mizzaro (1997), and Borlund (2003), and a book by Ingwersen & Järvelin (2005). My own work on relevance is represented and follows in these comprehensive articles: Saracevic (1975), (2007a), (2007b), and (2008). Parts of this article are synthesized from Saracevic (2007a and b).

IR systems offer their version of what may be relevant, based on algorithms and technological processes. People go about their ways and assess their own version of relevance, based on their problem-at-hand and context. There are two interacting worlds: the technological and the human world, and two basic categories of relevance: systems' and humans'. Our concern here is with the human world of relevance. While we can never get far from systems, I am *not* reviewing how systems deal with relevance, but how people do, fully recognizing that everything in systems is created by people. Although the literature on that topic is huge and vibrant, treatments of relevance in IR systems – algorithms, processes, technologies, evaluation - are beyond the scope of this review. As yet, no attempt has been made to unify the two very different categories of relevance; who knows if it is possible at all?

Why Did Relevance Emerge as a Central Notion in Information Science?

Information science came forward after the Second World War, together with a number of other fields that followed the scientific and technological triumphs of the War. In a hugely influential article that appeared just following the end of the War, Vannevar Bush, a scientist, inventor and most importantly head of the US scientific effort during the War, defined a critical problem and proposed a solution (Bush, 1945). Bush defined the problem as "the massive task of making more accessible a bewildering store of knowledge" and suggested a technological solution. In other words, Bush addressed the problem of information explosion. The problem is still with us, but now not only in science and technology, as in Bush's concerns at the time, but in all areas of human endeavors. Bush suggested a machine named Memex that should have a capability for "association of ideas," and duplication of "mental processes artificially". While Memex was never built, it still remains a goal. But the idea of technological solutions persisted and, with advent of computers and telecommunications, became a global, highly successful reality. People and most importantly granting agencies listened. Bush was also instrumental in launching the National Science Foundation (NSF); the law establishing it mandated, among others, funding for research on advances in scientific and technical information. Under ever changing NSF divisions and programs the mandate is followed to this day, but all of them have one thing in common: an emphasis on supporting research dealing with a "technological fix."

When the American Documentation Institute (ADI) was founded in 1937 by Watson Davis (1986-1967) it followed by name if not by practice the European field of documentation (Shera & Egan, 1950). A central concern of European documentation, founded toward the end of 19th century in Belgium, was organization of scientific literature. Eventually for that purpose, a Universal Decimal Classification (UDC) was developed and maintained by the International Federation for Documentation (FID) founded in 1895 and dissolved in 2002. UDC followed the principles of Dewey Decimal Classification (DDC), invented in 1876 by Melvile Dewey (1851-1931). In many ways UDC was more elaborate than DDC.

Samuel Clement Bradford (1878-1948), a British mathematician and librarian, published the first (and only) text in English on documentation (Bradford, 1948). Among others, in the book he championed UDC and described various documentary practices and services. However, UDC never took real hold in the US and ADI was never even concerned about UDC. Instead at the start and for years after, both Watson Davis and ADI were prominently promoting microfilm, the technology of the day, as a technical solution to problems of scientific communication (Schultz & Garwig, 1969). The European version of documentation was about organizing scientific information primarily through classification; the American version was about a "technological fix."

A side note: Bradford's "Documentation" became well known and highly cited not for UDC and related matter in the book, but for a chapter entitled "The documentary chaos" – from an article first published in 1934. In there, Bradford described a study of distribution or scatter of articles on a subject across journals where they appeared. He formulated a "law of distribution of papers on a given subject in scientific periodicals" that became known as Bradford's Law, motivating development of bibliometrics as an area of study in information science.

Up to 1952 ADI was an organization of institutional members only, after that date it became a scientific and professional organization of individuals–a move that completely changed the nature of the organization and even of the field. In 1950 ADI started publishing *American Documentation* (predecessor of the current *Journal of the American Society for Information Science and Technology*) as a "a quarterly review of ideas, techniques, problems, and achievements in documentation," with the purpose, among others, "for the publication of original research in the field; for reporting investigations of new techniques, mechanisms and devices for documentation and their applications both in the United States and abroad"(Tate, 1950). *American Documentation* (and successors) became the major outlet for reporting advances in information science globally, including research on relevance.

The term "information retrieval" (IR) was coined by mathematician and physicist Calvin N. Mooers (1919-1994), a computing and IR pioneer (and son-in-law of Watson Davis):

> Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. ... 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).

Mooers did not use the term "relevance," but the notion was implied by "useful" and the context of "need for information." Throughout decades that followed IR changed dramatically from Mooers' days, but the basic idea as formulated then is still valid. *Searching was added and with it relevance entered unannounced.*

Actually, Mooers' article was about a specific scheme for coding (called Zatocoding) as applied to edge-notched punch cards and the associated technology used for searching and retrieval. Microfilm and searching did not go together at all, thus other technologies of the day were explored that would allow searching. Following Bush and others, the emphasis was not only on *organization and storage* of information but also on *search and retrieval.* Short lived technologies that allowed for searching were edge-notched punch cards, Hollerith cards, optical coincidence (peek-a-boo) cards, uniterm cards and the like. The literature of the 1950s in *American Documentation* and elsewhere, is full of their descriptions, particularly including searching methods. By the end of 1950s they were all replaced by computers.

As to methods of searching Hans Peter Luhn (1896-1964) a computer scientist at IBM, inventor, a major pioneer in the field, and ADI president at the time of his death, was the first to formally describe searching using Venn diagrams (Luhn 1953). In addition, Mortimer Taube (1910–1965), an early entrepreneur in the field with a PhD in philosophy, as the inventor of coordinate indexing, was the first to describe searching in terms of Boolean algebra (Taube & Wachtel, 1953). While these were first attempts at formalization of search, neither Luhn nor Taube mentioned relevance, but searching implied it.

Bibliographic classifications, subject headings, and indexing languages were used for organizing information or information records for a long time, some schemes and practices going back centuries. All are based on the notion of *aboutness.* Choice of a given classification code, subject heading, or index term denotes what a document or part thereof is *about.* They assume but do not address searching at all. In other words, all deal with inputs and take outputs as a given. Searching is taken for granted.

No attempt was ever made to define searching related to any classification or subject heading scheme, be it formally or pragmatically. Charles Ammi Cutter (1837-1903) a most influential American librarian, developer, and author in the 19th century , among others formulated widely influential "Rules for a dictionary catalog" (first edition 1876, fourth edition, rewritten, 1904). Right at the beginning (p.12) Cutter defined what are the "Objects [meaning objectives of a dictionary catalog]." Over a century after Cutter's first edition the International Federation of Library Association and Institutions issued a report on the Functional Requirements for Bibliographic Records (FRBR) (IFLA, 1998). Among others, "four generic user tasks have been defined for the purposes of this study. The tasks are defined in relation to the elementary uses that are made of the data by the user ... " (Ibid. section 6.1). In essence these "tasks" are directly based on Cutter's "Objects". They are assumptions about what users will be searching for. Neither Cutter nor IFLA's FRBR say anything about how is this to be accomplished. Searching just happens. A great many studies on searching and human information behavior showed that these assumptions are not warranted at all. Searching is a complex process, affected by many variables and subject to many variations. It is also a subject of never ending exploration for improvements.

Suggesting a formal definition of aboutness, Maron (1977) made a careful distinction between aboutness and relevance. Aboutness is a fundamental notion related to *organization of information,* while relevance is a fundamental notion related to *searching and retrieval of information.* While related, the two are still very different processes. Aboutness relates to subject and in a broader sense to epistemology while relevance relates to problem-at-hand and in a broader sense to context and pragmatism.

The question at the heading of this section can be answered thus: relevance emerged as a central notion in information science because of extensive theoretical and practical concerns with and commitments to searching and not only to organization of information. In turn, searching was accomplished using modern information technology. People search using technology to find information relevant to their problem-at-hand and context. Systems provide ways and means of organizing of and searching for information that attempt to provide higher probability of finding relevant information by people. Interestingly, organization of information can and is being done without recourse to specific tools (such as classification), by algorithms that concentrate on exploiting patterns in the raw data, for example, as revealed by word counts, links, page ranks, and the like, all geared toward searching. Traditionally, librarianship concentrated on organization of information and thus on aboutness, while information science concentrated on searching and thus on relevance. This describes both their relation and difference.

What Did We Learn About Relevance through Research in Information Science?

Literature about relevance in information science is extensive but not overwhelming. Part of it is about issues based in philosophy, such as the treatise by Hjørland (2010) in which he argues for treating relevance from a subject knowledge or epistemological point of view and rejects the duality of system-user relevance. While illuminating, arguments in such treatises are not elaborated here due to restrictions on length. Instead concentration is on studies containing data or observations.

Relevance and Testing

As mentioned, with an added emphasis on searching relevance came into information science unannounced at the start of the 1950s: the desired outcome was to retrieve relevant information. In a short period of time during the 1950s numerous competing IR systems and schemes were suggested. As claims and counterclaims escalated testing was advocated for a resolution—this is not surprising since many if not most of IR developers were scientists and engineers for whom testing is a mandatory part of development. Much that we learned about relevance in intervening years has a connection with testing of IR systems and techniques.

The issue with relevance came to a head with a big bang in the mid 1950s during an attempt to test the performance of two competing IR systems developed by separate groups: one developed by the Armed Services Technical Information Agency (ASTIA) using subject headings, and the other by Mortimer Taube and his company named Documentation Inc., using uniterms (keywords searched in a Boolean manner) (Gull, 1956). The study is a classic example of the law of unintended consequences, showing not only that relevance inferences differ significantly among groups of judges, but also inadvertently uncovering a whole range of issues that IR evaluation struggles with to this day. The results are worth recalling. In the test, each group searched 98 requests using the same 15,000 documents, indexed separately, in order to evaluate performance based on relevance of retrieved documents. *However, each group judged relevance separately.* Then, not the systems' performance, but their relevance judgments became contentious. The first group found that 2200 documents were relevant to the 98 requests, while the second found that 1998 were relevant. There was not much overlap between groups. The first group judged 1640 documents relevant that the second had not, and the second group judged 980 relevant that the first had not. You see where this is going. Then they had 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 on relevance assessment hovers indeed around that figure. For instance, Shaw, Wood, Wood, & Tibbo (1991) found consistency to be about 40%; and over a decade later Vakkari & Sormunen (2004) found it to be 45% - the context of each study was quite different while the results were similar.

The corollary that IR evaluators learned from Gull: *Never, ever use more than a single judge per query.* They don't.

Measures

In mid the 1950s Allen Kent and James W. Perry, both chemists and pioneers in information science, wrote a series of articles about techniques of IR. In one of the articles they suggested measures for evaluating performance of IR systems. They were "precision" and "relevance" (later because of confusion renamed "recall") (Kent, Perry, Leuhrs, & Perry, 1955). This was the first full recognition of relevance as an underlying notion of retrieval—relevance was the criterion for these measures. Precision and recall remained standard measures of effectiveness to this day, with a number of variations on the theme. They measure the probability of agreement between what the system retrieved/not retrieved as relevant (systems relevance) on the one hand and what the user assessed as relevant (user relevance) on the other hand, where user relevance is the gold standard for comparison. Relevance became and remained the underlying criterion for measuring the effectiveness of IR. By now, it is cemented there.

Tests

From 1954 to 1966 Helen Brownson (1917-) was at NSF the Program Director for Documentation Research, Office of Scientific Information (predecessor of the current Division of Information and Intelligent Systems – IIS). The years of Brownson's leadership and initiatives would eventually prove to be significant ones for IR tests and general IR developments due to the research, publications, reports, and funds that Brownson and her staff dispersed (Jayroe, forthcoming). One of the many grants she invited was from Aslib (Association for Special Libraries and Information Bureaux), Aeronautical Group Committee, located in Cranfield, England. The invitation was for testing various indexing systems in order to measure and compare their performance. She encouraged the group to submit a grant proposal for $28,000 (about $240,000 in 2012 dollars), which the NSF subsequently funded starting in 1957 and Cyril Cleverdon (1914-1997; a British science librarian) conducted under the name of "Cranfield tests" (Cleverdon, 1967, 1991). This was a pioneering effort: the first attempt to control the application of different indexing systems to the same body of documents, to conduct test searches of the indexed material, to obtain relevance assessments of retrieved results, and to comparatively analyze the results using measures of precision and recall as formulated by Kent and colleagues, all under laboratory conditions. The study developed an IR testing methodology that became standard ("Cranfield methodology"). It is based on a model of IR systems called "traditional model" that in essence represents the process as static. It also uses a set of assumptions about relevance assessments discussed below.

Fast forward some three decades: starting in 1992 IR tests that began under the umbrella of the Text REtrieval Conference (TREC), but on a vastly larger scale, followed the Cranfield methodology. TREC, continuing to this day, is a long-term effort at the [US] National Institute for Standards and Technology (NIST), which brings various IR teams from around the globe together annually to compare results from different IR approaches under laboratory conditions (Voorhees & Harman, 2005). TREC inspired other international laboratory efforts using the same testing methodology, among them in the Europe Initiative for Evaluation of XML Retrieval (INEX), and Conference and Labs of the Evaluation Forum (CLEF) (formerly Cross-Language Evaluation Forum). The Cranfield methodology is multiplying.

The golden rod for all these tests was and still is relevance assessment by users or user surrogates (or by somebody, somehow). When it comes to relevance judgments, the central assumption in any and all IR tests using Cranfield and derivative approaches, such as TREC, INEX and CLEF, has five postulates assuming that relevance is (these assumptions were not stated in any test but are there nevertheless):

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

These simplified assumptions about relevance assessments enabled successful IR testing. Over the years, test results led to significant improvements in IR algorithms and procedures. IR extended from "bags of words" and texts to areas covering a variety of other expressions and media. Not surprisingly, these assumptions were criticized. In turn, questioning of these assumptions led to a significant amount of research on relevance. In other words, assumptions under which IR testing is conducted stimulated relevance research and led to a lot of interesting results. IR testing and relevance research got connected, at least for a while.

Even a cursory examination of these assumptions can show that none of them hold. This was confirmed in a number of experiments and studies. Let us examine some of them.

Topicality: Topicality plays an important, but not at all an exclusive, role in relevance inferences by people. By no means is topicality the sole criterion. We make a distinction among: *Cognitive relevance or pertinence; Situational relevance or utility; Affective relevance*; and others (summarized in Borlund, 2003, and Ingwersen & Järvelin, 2005).

Binary: Users do not use only binary relevance assessments, but infer relevance of information or information objects on a continuum and comparatively. Numerous experimental studies confirmed this, among the first being Eisenberg & Hue (1987). However, even though relevance assessments are not binary they seem to be bi-modal: high peaks at end points of the range (not relevant, relevant) with smaller peaks in the middle range (somewhat not relevant or relevant) (Spink & Greisdorf, 2001).

Independence: The order in which documents are presented to users seems to have an effect. It seems that documents presented early have a higher probability of being inferred as relevant. However, when a small number of documents is presented, order does not matter (Eisenberg & Barry, 1988; Huang & Wang, 2004).

Stable: 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 (first observed experimentally by Rees & Schultz, 1967, and more recently by Taylor, 2012). Relevance is dynamic and situational (Schamber, 1994).

Consistent: Human judgments about anything related to information are not consistent in general, and relevance judgments are no exception. Why should they be? Individual differences are a, if not *the*, most prominent feature and factor in relevance inferences (three studies about degree of consistency of relevance assessments among judges were already mentioned). Early on it was demonstrated that individual differences involve a number of variables affecting judgments (Cuadra et al. 1967).

Aftermath of First Tests

Cranfield tests produced some interesting, even unexpected results. The most startling one was that simple Uniterms (keywords from texts) performed as well and even slightly better than three much more complex (and highly promoted) systems—based on UDC, alphabetical subject index, and faceted classification. In reminiscences Cleverdon (1991) remarked: "The publication of the final report [in 1962] attracted wide interest, caused considerable annoyance to the advocates of the different systems, and received some praise but much criticism. Most of this could be ignored, such as the comment, "You had no right to be so intelligent with the Uniterm system; it is meant to be used by persons of low intellect." Cranfield tests did not use computers, but computer searching was simulated using paper strips.[1] Uniterms were equivalent to keywords extracted by a computer algorithm. Their good showing was a boost for later developers of computerized IR systems, such as by Gerard Salton (1927-1995), computer scientist and IR pioneer who did long term development and tests of a variety of new IR algorithms and methods in his SMART IR System from 1958 on (Salton, 1989). Tests in SMART used the Cranfield methodology, including the traditional IR model and assumptions listed.

A different kind of Cranfield criticism addressed means and ways of relevance assessments, e.g. Swanson (1965). Relevance came to the forefront as an issue, widely debated.

Partly because issues that rose in the aftermath of Cranfield tests, in October 1964 NSF sponsored a Study Conference on Evaluation of Document Searching Systems and Procedures (Brownson, 1965). One of the major conclusions was that "lack of sufficient knowledge regarding the character and variability of human assessment of relevance" was a major obstacle to progress in the area of evaluation methodology and to test and evaluation activities. Two general classes of factors (variables) were hypothesized to have a definite and important effect over the relevance assessment process:

a) Subject variables dealing with the education, involvement, and performance of the relevance assessor, and

b) Situational variables dealing with the environment within which the relevance assessment is made and with the object of assessment (titles, abstracts, etc).

Subsequently, in 1965 NSF issued a Request for Proposal (NSF 65-111) entitled "Statement of requirements for an empirical study of relevance assessments in relation to document searching" with the basic objective to determine the patterns, variability, and basis of relevance assessments. (as quoted in Rees & Schultz, 1967). From proposals, NSF selected two two-year projects for support, one at System Development Corporation (SDC), with Carlos Cuadra and Robert Katter as principal investigators and the other at Western Reserve University (later Case Western Reserve U) with Alan Rees and Douglas Schultz as principal investigators. Final reports by Cuadra et al. (1967) and Rees & Schultz (1967) summarize the experiments and results. The two projects cooperated closely during the investigations. (Disclosure: I was a participant in the WRU project, at the time I started my PhD studies).

These were the first and only projects on relevance ever funded by NSF or any other large US federal agency, the first and only relevance projects ever that had a relatively large and dedicated team for some length of time, and also the largest relevance research projects ever. In these respects they were unique in history of relevance research. Three of the four principal investigators (Cuadra, Katter, & Schultz) were psychologists, Rees was a literature major. The projects were very different. Despite that, collectively they brought exploration of psychological and cognitive variables and related methods of investigations into relevance research. These stayed with relevance research to this day.

[1] "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).

The SDC project specified a host of variables of interest and conducted a large number of experiments, involving groups of variables related to: (1) documents, (2) information requirement statements, (3) the judge, (4) the judgment condition, and (5) the available mode of expression. Variations in results involving these variables were observed in experiments where subjects were students and librarians from near-by institutions (SDC was located in Santa Monica, California). To illustrate, an example of results for a study on inter-judge agreement is in the footnote.[2]

The WRU project also involved a considerable number of variables, but concentrated the effort on one large experiment derived from real-life circumstances. The project took a funded research study in diabetes from the WRU medical school, distinguished different stages of the project as it progressed from inception to completion, provided a set of documents for judgment, and assembled different classes of potential users to conduct experiments on relevance assessments. Again, for illustration some results on the same variable on inter-judge agreement are in the footnote.[3]

Both projects presented great amounts of data to share in elaborate final reports. (In addition to the final report from the SDC project came one journal article (Cuadra & Katter, 1967), but none out of the WRU project). On a most general level the projects showed that relevance is indeed measurable and that numerous, but final, numbers of variables are involved. The projects also introduced rigorous statistical analyses, as applied in other social sciences, to relevance research.

The projects ended up with reams and reams of data. Many results, particularly related to individual differences and dynamic changes, were original and even surprising. However and disappointingly, neither project came up with any generalizations from the results. We cannot see the forest for the trees. In retrospect this is possible because both project were overdesigned, trying to cover at one time way too much in terms of variables and sub-variables. But it is also possible because of the very nature of relevance – variability is such that it makes too hard to generalize.

One can only speculate why relevance research was funded only through this one NSF initiative and never again by NSF or any granting agency. Were the results from these studies so complex or so unconvincing? Was the research on the topic too human-oriented without technological connections and implications? Was the phenomenon deemed hopeless or unimportant for research? Or not promising for translation into anything practical—such as improved or new IR systems? Were research policies changed, and that was that? Because the reason for funding relevance research at that time was to address problems with IR testing, was it thought that the problems had been resolved and thus there was no need to address relevance anymore? No documents exist to answer these questions.

Whatever the reasons, this lack of funding had a great effect on future relevance research. All projects were small; many were the result of solitary work on PhD dissertations; there were no research teams; and there was poverty in resources. Relevance became poor, but even on a small scale relevance research still persisted.

Back to Testing

A lot of relevance research was and still is connected to testing of IR algorithms and procedures inspired by the fact that testing methods use human relevance assessments as a golden rod for comparison. But in turn, results from relevance research that showed individual and group differences in relevance judgments and variations in consistency of judgments came back to haunt

[2] 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 8 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 field. 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.

[3] Rees & Schultz (1967) used a total of 153 judges divided in 7 groups that were given 16 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 21 medical librarians - searchers was 44%, 21 medical librarians - non-searchers was 40%, 14 medical experts - researchers was 58%, 14 medical experts - non-researchers was 56%, 29 scientists was 55%, 25 residents was 51% and 29 medical students was 50%.

IR testing. A loop between IR testing and relevance research was completed.

The issue is: *How can IR testing results be considered as reliable if they are based on human relevance assessments that are shown to be inconsistent?* The question resulted in experiments about reliability of testing results but it was changed to: *Does ranking of different algorithms or procedures in test results change due to inconsistency of human relevance assessments?* Of course, this is a critical question for all IR testing using Cranfield methods to this day.

There were seven or so experimental studies conducted to date trying to answer these questions; first was Lesk & Salton (1968) and the latest was Voorhees (2001); all are summarized in Saracevic (2008, Table 2). 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. Conclusions drawn based on results from these experiments are as follows:

> " ...[studies found that] disagreement among judges seems not to affect or affects minimally the results of relative performance among different systems [ranking] 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. ... 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. ... 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" (Saracevic, 2008, p. 779, 780).

In contrast, finding that relevance assessments are inconsistent stopped being of concern to testing of IR systems. The current stance in various test laboratories, such as TREC, INEX, CLEF, is that inconsistency in relevance assessment does not matter. That may be right indeed, but then it may not. No matter, research on the issue stopped.

Relevance Dynamics and Context

For a fleeting decade, relevance had a place where research on a variety of relevance topics flourished. From about the mid- 1980s until about the mid-1990s, a series of doctoral dissertations at the School of Information Studies, Syracuse University, addressed various aspects of relevance, reflecting a vigorous research environment under the guiding spirits of Robert Taylor and Jeffrey Katzer. These dissertations produced a number of articles (Michael Eisenberg, Linda Schamber, Carol Berry, Myke Gluck, reviewed in Saracevic, 2007a). The Syracuse relevance school also produced a notable and widely cited critical review that had an extensive impact and changed the view of what is important in relevance (Schamber, Eisenberg, & Nilan, 1990). When well done, critical reviews can do that. They re-examined thinking about relevance in information science, addressed the role of relevance in human information behavior and in systems evaluation, summarized major ideas and experiments, and came to a forceful conclusion that relevance should be modeled and studied as being dynamic and situational. Of course, dynamic properties of relevance had been discussed for decades before and demonstrated in experiments as readily acknowledged by authors, for instance numerous critics pointed out that Cranfield-like tests take relevance as static while in reality relevance is highly dynamic. Their insistence on the primacy of the dynamic and situational nature of relevance struck a chord and changed the outlook and direction of relevance research. Note: One has to realize that dynamic aspects of relevance are quite different than situational aspects—former relate to the process and later to the context. Thus, two different streams of research are involved. A sample of studies is provided.

Research on dynamic aspects of relevance focused on interactions during IR processes. One stream of research concentrated on clues or criteria: *What do people look for in information or information objects in order to infer relevance?* Involved were a wide range of studies from the one by Cool, Belkin & Kantor (1993) interviewing computer science students and separately humanities scholars in order to identify characteristics of texts affecting their relevance judgments, to the one by Toms et al. (2005) observing users of the Web in accomplishing given tasks in order to identify and categorize a set of criteria for relevance. General conclusion: 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.

Another stream of research concentrated on changes: *Do relevance inferences and criteria change over time for the same user and task, and if so, how?* Again a wide range of studies addressed the question from the one by Smithson (1994), who in a case study approach, studied a set graduate students with a semester-long assignment in order to observe differences in judgments at different stages of the project, to the one by Anderson (2005) who observed two academics involved in scholarly research over a period of two years in order to explore relevance assessments as part of the decision-making process of individuals doing research over time. General conclusion: For a given task, it seems that the user's inferences about specific information or information object are dependent on the stage of the task. However, a user's criteria for inferences are fairly stable. As the time and the work on the task progress, users change criteria for relevance inferences, but not that much. Also, the weight given to different criteria may change over stages of work.

Concerns with situational aspects of relevance concentrated on context. Relevance is always dependent on context. There is no such thing as a search by a user that is context-free. One cannot not have a context. A classic example: An answer to: "Where is Taj Mahal?" could be: "In India" or "Down the street"—depends on the context of the question. However, beyond discussions there were no experiments on relevance as related to situation or context. But another thing happened: concerns with relevance morphed into broader aspects of interaction in IR. The problem of context is determining the conditions and circumstances that are relevant to a given information interaction; with this, the notion of information interaction context is connected with the notion of relevance in information science. Numerous experiments have shown that searching tasks influence individuals' information behavior in general and searching in particular (e.g., Byström & Järvelin, 1995; Vakkari, 2003; Li & Belkin 2008).

The interest and literature on interaction in IR in relation to context has grown to the point that the topic has its own biennial symposiums Information Interaction in Context (IIiX), held since 2008 in different locations. A paper by Balatsoukas & Ruthven (2010) is an example where study of interaction, context, and user relevance criteria were joined.[4] In a way, context is swapping relevance in IR as a focus of interests, if not as a basic notion. Relevance has broadened to information interaction.

Conclusions

The aim of this paper is to examine two questions: *Why did relevance become a fundamental notion in information science?* and *What did we learn about relevance through research in information science?*

As to the first question: information science turned toward searching using a variety of technologies even before advent of computers. Information retrieval was embraced not only for organization and storage of information but even more for search and retrieval. Searching is about retrieving relevant information—information that will be related to a problem-at-hand. The fundamental notion in organization of information is *aboutness*, while the fundamental notion in searching is *relevance*. With orientation toward searching and technology for searching, relevance became a central notion in information science.

As to the second question—what did we learn—the answer is a mixed bag. From a relatively small number of experimental and philosophical works done over a period of over a half a century we obtained evidence that relevance is a multidimensional notion, encompassing multidimensional variables, and facing multidimensional problems. Still the number of dimensions and variables is finite.

From the start of IR different and competing schemes were proposed and promoted and explored. Testing became imperative. Relevance became the criterion for measures of performance of IR systems; user assessment of relevant output became the gold standard

[4] "Participants were asked to search for real information needs that represented different search contexts (e.g. from searches about personal interest to academic related searches). This permitted the identification of several relevance criteria that naturally occur across different search contexts and the emergence of some fixation patterns, not observed before, associated to the use of these criteria. A user study was conducted that involved the completion of questionnaires, use of eye tracking technology, talk aloud protocols and post-search interviews."

for tests. Concerns with relevance and ensuing variability emerged after troubles with methodologies used in testing of retrieval systems and schemes. Relevance research became enmeshed with IR testing. Answers were sought for relevance issues in testing first and foremost. For a long time testing drove the agenda for relevance research. However, by the start of the millennium inconsistency of relevance assessment stopped being an issue for test laboratories and projects. Interest switched to a broader area of context in human searching.

Considering knowledge for knowledge sake, relevance research extended our knowledge about the notion—as a matter of fact it was quite successful in that respect. However, relevance research did not contribute a lot, if anything, to pragmatic knowledge: to practical issues of improving or innovating IR systems and procedures. Almost every relevance research report ends with: "The findings have implication for system design." This became a mantra. But how to do this in real life and who will do it was never addressed. Mantras do not make improvements or innovations. Thus relevance research proved to be good at extending knowledge and poor on translation into practice.

With the exception of two large projects at the end of the 1960s relevance research was not funded by granting agencies. It did not attract outside interest, it did not attract funding. Under these circumstances it is a testimony to the challenge of an interesting problem and persistence of researchers that research on relevance is continuing.

Information technology and information systems 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.

References

Anderson, T.D. (2005) Relevance as process: Judgements in the context of scholarly research. *Information Research, 10*(2) paper 226. Retrieved 6-7-2012 from http://InformationR.net/ir/10-2/paper226.html

Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts & technology behind search.* 2nd ed. Boston: Addison-Wesley.

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

Balatsoukas, P. & Ruthven, I. (2010). What eyes can tell about the use of relevance criteria during predictive relevance judgment? *IIiX - Proceedings of the 2010 Information Interaction in Context Symposium*, 389-392.

Bradford, S. C. (1948). *Documentation*. London: Crosby Lockwood.

Bush, V. (1945) As we may think. *Atlantic Monthly, 176*(11), 101-108.

Brownson, H. L. (1965). Evaluation of document searching systems and procedures. *Journal of Documentation, 21*(4), 261-266.

Byström, K. & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing and Management, 31*(2), 191-213.

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

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.*

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

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.

Cuadra, C.A & Katter, R.V. (1967). Opening the black box of 'relevance.' *Journal of Documentation, 23*(4), 291-303.

Cutter, C.A. (1904). *Rules for a Dictionary Catalog.*4th ed. rewritten. Washington, DC. Government Printing Office. Retrieved 6-2-2012 from http://digital.library.unt.edu/ark:/67531/metadc1048/m1/1/

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

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

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.

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology, 61*(2), 217-237.

Huang, M. & Wang, H. (2004). The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of the 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.* New York: Springer.

International Federation of Library Association and Institutions (IFLA) (1998). *Functional Requirements for Bibliographic Records – Final Report.* Retrieved 6-2-2012 from: http://www.ifla.org/VII/s13/frbr/frbr1.htm#2.1

Jayroe, T.J. (Forthcoming). A humble servant: the work of Helen L. Brownson and the early years of information science research. *Journal of the American Society for Information Science and Techology.* Retrieved 6-4-2012 from https://pantherfile.uwm.edu/tjjayroe/www/H.Brownson%20by%20T.Jayroe.pdf

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.

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

Li, Y. & Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing and Management, 44*(6), 1822-1837.

Luhn, H.P. (1953). A new method of recording and searching information. *American Documentation, 4*(1), 14–16.

Maron, M.E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science, 28*(1), 38-43.

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

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

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.

Salton, G. (1989). *Automatic text processing*. Addison-Wesley.

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. (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. (2008). Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends, 56*(4), 763-783.

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.

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

Schultz, C. K. & Garwig, P. L. (1969). History of the American Documentation Institute-A Sketch. *American Documentation, 20*(2), 150-162.

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.

Shera, J.H. & Egan, M. E. (1950). Documentation in the United States. *American Documentation, 1*(1), 8-12.

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

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.

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

Tate, V, (1950). Introducing American Documentation, a quarterly review of ideas, techniques, problems, and achievements in documentation. *American Documentation, 1*(1), 3-6.

Taube, M. & Wachtel, I.S. (1953). The logical structure of coordinate indexing. *American Documentation, 4*(2), 67–68.

Taylor, A. (2012). A study of the information search behaviour of the millennial generation. *Information Research, 17*(1). Retrieved 6-4-2012 from http://informationr.net/ir/17-1/paper508.html

Toms, E.G., O'Brien, H.L., Kopak, R., & Freund, L. (2005). Searching for relevance in the relevance of search. *Proceedings of Fourth International Conference on Conceptions of Library and Information Science, (CoLIS 2005).* Springer: 59-78.

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology*, 37, 413-464.

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. (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.