Why Is Relevance Still the Basic Notion in Information Science?

(Despite Great Advances in Information Technology)

Tefko Saracevic

Distinguished Professor Emeritus
School of Communication and Information, Rutgers University
New Brunswick, New Jersey, USA
tefkos@rutgers.edu

Relevance:
... relation to the matter at hand
... the ability (as of an information retrieval system) to retrieve material that satisfies the needs of the user.
Merriam-Webster Dictionary Online

Abstract

Relevance is a fundamental notion in information science. The aim of this paper is to provide a perspective on two large questions, the first historical the second contemporary: (1) *Why did relevance become a central notion of information science?* (2) *In this day and age of huge advances in information technology, why did relevance still remain a central notion?* In the 1950s, 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 connected to and made possible by many innovations in computers and computing. Contemporary advances in information technology brought about great many changes. Search engines, social media, and a myriad of new information resources emerged and transformed the world. Searching for information

In: F. Pehar/C. Schlögl/C. Wolff (Eds.). Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19th–21st May 2015. Glückstadt: Verlag Werner Hülsbusch, pp. 26–35.

is used widely all over the globe by all kinds of populations and reasons. However, searching is still based on relevance. Relevance was and still is a fundamental notion related to *searching and retrieval of information.* Conclusions emphasize that information technology and myriad applications are changing at an accelerated pace. However, no matter what, relevance is here to stay. Relevance is timeless. Concerns about relevance will always be timely.

Keywords: Relevance, Information science, Information retrieval, Searching

1 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, central notion 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 and a variety of search apps in social media, are so well accepted globally – differences in cultures, societies, and mores do not matter. However, relevance is a human, not a technical, notion. That is its weakness. As all human notions, relevance is messy. Relevance encompasses many variables that are hard to control and even fathom formally. Relevance always, repeat always, involves a context as well. All the search 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, that is why they exist.

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

1. *Why did relevance become a central notion of information science?*
2. *In this day and age of huge advances in information technology, why did relevance still remain a central notion?*

The first question 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. As mentioned, it is universally understood. Thus,

there are no direct historical documents to contemplate and the discussion here is derivative. The same applies to the second question. Advances in information technology in general and information search technology in particular, do not reflect on relevance at all. Relevance is what they are aiming at even without saying – also thought as universally understood.

Major sources

Relevance was a subject of a number of major, even outstanding, reviews that appeared over time. Among them are reviews by Schamber, Eisenberg & Nilan (1990), Shamber (1994), Mizzaro (1997), Borlund (2003), Ingwersen & Järvelin (2005), Hjørland (2010), and Huang and Soergel (2013). All these reviews contain, among others, information of historical nature. However, none of these reviews asked any of the two questions raised above.

My own work on relevance is represented in these comprehensive articles: Saracevic (1975), (2007a), (2007b), and (2008). Large parts of this article are synthesized from Saracevic (2012), where the first question was dealt with in some detail; the second question was not addressed at all.

2 Why relevance? A bit of history

Why did relevance become a central notion of 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 (1890–1974), a scientist, inventor and most importantly head of the U.S. 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 the area of 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. Bush's idea of a tech-

nological solutions persisted – with expansion of computers and telecommunications in decades that followed it became a global, highly successful reality. People and most importantly granting agencies listened. [Bush was also instrumental in launching in the U.S. the National Science Foundation (NSF); the 1950 law establishing NSF mandated, among others, funding for research on advances in scientific and technical information.] All efforts that followed to this day have one thing in common: an emphasis on supporting research, development, and applications dealing with a "technological fix" to a variety of information problems and issues. This includes information retrieval.

Information retrieval

The term "information retrieval" (IR) was coined by mathematician and physicist Calvin N. Mooers (1919–1994), a computing and IR pioneer:

> 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.*

As to methods of searching, Hans Peter Luhn (1896–1964) a computer scientist at IBM, inventor, a major pioneer in the field, and American Documentation Institute (ADI) president at the time of his death, was first to formally describe searching using Venn diagrams (Luhn 1953). In addition, Mortimer Taube (1910–1965), an early entrepreneur in the field with a Ph.D. in philosophy, 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 by that name, but through searching they implied it. They addressed searching because the technology of the day allowed and even demanded it – the technolgy was all about seraching.

To briefly connect the notion of *aboutness* and the notion of *relevance*: 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. Searching is taken for granted. In other words, all deal with inputs and take outputs as a given. No attempt was ever made to define searching related to any classification or subject heading scheme, be it formally or pragmatically.

In 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 start 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, based on a variety of algorithms, 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 a classification or thesaurus; it is done by computer algorithms that concentrate on exploiting patterns in the raw data, for example, as revealed by word association, 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.

3 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 during the early 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.

Relevance announced itself to the forefront of concerns in the field in a dramatic fashion because of the first IR test ever. Circumstances are worth repeating – they represent an enduring lesson. In the mid 1950s there was 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 did not, and the second group judged 980 relevant that the first did not. You see where this is going. Then they had reconciliation, considered each other's 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. *Because of relevance assessments.*

Moreover, it seems that the rate of human agreement on relevance assessment hovers indeed around that figure (Saracevic, 2007b).

Measures of performance

In mid the 1950s Allen Kent (1922–2014) and James W. Perry (1907–1971), 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 et al., 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 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.

IR testing continued to this day. Precision and recall remained standard measures of effectiveness, with a number of variations on the theme. Lesson from the first test was learned, although today hardly anybody knows its source. Tests include a single judge (or sometimes a group with a consensus) that provide a golden rod of relevance of documents (information sources) against which a system or algorithm performance is assessed.

4 Relevance and information technology

In this day and age of huge advances in information technology, why relevance still remains a central notion?

People were searching and looking for relevant information for a variety of information needs forever – long, long before the advent of information technology. However, information technology brought relevance to the fore.

It is commonly accepted that information technology involves "the development, maintenance, and use of computer systems, software, and networks for the processing and distribution of data" (Merriam-Webster Online Dictionary; term "information technology" first known use is in 1978).

However, from a human perspective information technology refers to use of that technology as tools for a variety of tasks such as storing and manipulating text and data, performing calculations, and communication. Communication aspects of information technology became most important for relevance. Actually, it is really the other way around. Relevance is the main reason why information technology is used in communication. Of course, the goal may also be communication of irrelevance, false information, propaganda and the like – all distortions – still the main concern is with the notion of relevance.

Information technology became ubiquitous – it is everywhere, omnipresent, all over the globe, even though a huge slice of humanity has no connection with it.

Contemporary advances in information technology brought about great many changes. Search engines, social media, and a myriad of new information resources emerged and transformed the world. A large literature, both popular and scientific, deals with these transformations. Every field has changed – health, government, business, sciences, professions, publishing ... you name it. Communications have changed. Politics and policies are impacted. Revolutions are assisted So are oppressions.

There is also another significant change. Populations using these contemporary information technologies have broadened considerably. The population using original information retrieval systems (mentioned in the section on history) was restricted to scientists, professionals, managers, officials, and the like. The population using contemporary systems based on information technology (variety of social media, search engines, digital information resources ..., as mentioned) is now everybody and everyone with access. This includes not only those that are dependent on or connected to knowledge economy, but also each person no matter what, with access to technology. Access has become the key.

Variety of social media, search engines, digital information resources ..., have capability for searching (browsing included). In turn, searching for all kinds of information is used widely all over the globe by all kinds of populations and for all kinds of reasons. Masses search. However, searching is still based on relevance. People search in order to find something that is relevant for them at that moment, in a given context Relevance was and still is a fundamental notion related to *searching and retrieval of information.*

The question at the start of this section can be answered thus:

*Relevance remained as a central notion in information science in contempoary applications of information technology because of extensive commitments to **searching** and not only to organization of information.*

5 In conclusion

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

References

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

Bush, V. (1945) As We May Think. *Atlantic Monthly*, 176 (11), 101–108. http://www.theatlantic.com/doc/194507/bush <23.01.2015>.

Gull, C. D. (1956). Seven Years of Work on the Organization of Materials in Special Library. *American Documentation*, 7 (4), 320–329.

Ingwersen, P. & Järvelin, K. (2005). *The Turn: Integration Of Information Seeking And Retrieval In Context*. New York: Springer.

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, X. & Soergel, D. (2013). Relevance: An Improved Framework for Explicating the Notion. *Journal of the American Society for Information Science and Technology*, 64 (1), 18–35.

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.

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.

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.

Saracevic T. (2012) Research on Relevance in Information Science: A Historical Perspective. In: Carbo, T. & Bellardo Hahn, T. (Eds.). *International Perspectives on the History of Information Science and Technology. Proceedings of the ASIS&T 2012 Pre-Conference on the History of Information Science and Technology*. pp. 49–60.

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

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.

Taube, M. & Wachtel, I. S. (1953). The Logical Structure of Coordinate Indexing. *American Documentation*, 4 (2), 67–68.