Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 44 (2008) 963–970
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Looking back: On relevance, probabilistic indexing and information retrieval

Paul Thompson *

Dartmouth College, 6211 Sudikoff Laboratory, Hanover, NH 03755, USA

Received 13 January 2007; received in revised form 23 October 2007; accepted 30 October 2007
Available online 21 December 2007

Abstract

Forty-eight years ago Maron and Kuhns published their paper, "On Relevance, Probabilistic Indexing and Information Retrieval" (1960). This was the first paper to present a probabilistic approach to information retrieval, and perhaps the first paper on ranked retrieval. Although it is one of the most widely cited papers in the field of information retrieval, many researchers today may not be familiar with its influence. This paper describes the Maron and Kuhns article and the influence that it has had on the field of information retrieval.

Keyword: Probabilistic information retrieval

1. Introduction

Forty-eight years ago Maron and Kuhns published "On Relevance, Probabilistic Indexing and Information Retrieval" (1960). This was the first paper to present a probabilistic approach to information retrieval, and perhaps the first paper on ranked retrieval. Although it is one of the most widely cited papers in the field of information retrieval, many researchers today may not be familiar with the influence that it has had. This paper describes the Maron and Kuhns article and the influence that it has had on the field of information retrieval, and also suggests that there is more to be learned from that paper today.

After World War II, when general purpose computers were first developed to calculate ballistic trajectories, to aid in the design of nuclear weapons, and to decipher coded enemy messages, among the first uses that were considered for the new computers was document retrieval (Bagley, 1951). In 1945 Vannevar Bush had published his seminal, conceptual article on the Memex, a machine that would provide a scientist with personalized navigation of scientific literature (Bush, 1945). Maron and Kuhns referred to this problem of finding scientific papers as the library problem. Mooers had coined the term "information retrieval" (1950), but the problem was perhaps most often thought of as access to scientific information. In 1946 the Royal Society

* Tel.: +1 603 646 8747; fax: +1 603 646 1672.
E-mail address: Paul.Thompson@dartmouth.edu

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.10.002

of the United Kingdom had held the British Empire Scientific Conference (Brown, 1959). At this conference planning began for another Royal Society conference, the International Conference on Scientific Information in 1948, which was the first international conference on this topic. Various new retrieval systems were developed during the 1950s and early 1960s, often based on notched card technology (Mooers, 1950; Taube & Wooster, 1958), and new manual indexing schemes were studied (Cleverdon, 1962). Then in 1958 – the same year that Maron and Kuhns, along with Ray, published the first of two technical reports (1958, 1959), which ultimately led to Maron and Kuhns 1960 paper – the International Conference on Scientific Information was held in Washington, DC (National Academy of Sciences, 1959). Planning for the conference had begun in 1956, but even more importance was attached to the conference with the 1957 launch of Sputnik and the perception that the Soviet Union's handling of scientific information was more advanced than elsewhere. It was claimed that about 1000 people attended the conference, while another 1000 who wanted to attend were turned away.

It was natural to think of the automated solution to accessing scientific information, or the library problem, in terms of what had gone before, that is the practice of librarianship. In 1960 librarians cataloging a document provided metadata by which a document was represented in a card catalog. Each document had a unique identifier, or accession number, but this number was not used to provide intellectual access to the document. Rather several other entry points were provided: (1) a main entry, generally the author's name; (2) a title entry; (3) one or more subject headings. Physical cards were made for each of these entries and the cards were stored in the card catalog in alphabetical order. When asked his views on automated solutions to the library problem during this time, Norbert Wiener, the noted mathematician and cybernetician, said that he had no need for a solution to the problem as he personally knew of all work being done in any field of interest to him. Of course, all researchers were not similarly well situated with respect to their fields, and now the problem has only been worsened given the increasing rates of publications available not only in libraries, but on the Web, and in organizational repositories.

Maron and Kuhns proposed a radically new approach to the information retrieval problem. They suggested that two-valued indexing of documents be replaced by weighted indexing, where the weights were to be interpreted as probabilities. They viewed the retrieval problem as a problem involving clues and inference – that an information retrieval system should predict, given a user's query, which documents in the collection would most probably be relevant to that user and then rank those documents in descending order by those computed values of probability of relevance. Thus the output response to a library query was not a set of documents whose indexes "matched" the query term(s), but rather a ranking of documents by their computed values of probability of relevance. Maron and Kuhns called their approach "Probabilistic Indexing" and it was, in fact, a theoretical attack which replaced traditional two-valued indexing and matching with a statistical approach involving use of library statistics and user queries as clues to make predictions about the relevance of documents in the collection. It should be noted that "relevance" has been interpreted in many ways in the field of information retrieval (Saracevic, 1975). As Maron commented later (1984), his view of relevance is equated with a user's wanting a document. Cooper and Maron also developed a utility-theoretic interpretation of relevance in terms of subjective utility (1978).

Probabilistic concepts and statistical techniques for information retrieval had been considered by others before Maron and Kuhns, but not for the purpose of ranking documents. Fairthorne (1961) describes Mooers' (1950) definition of information retrieval in these terms. "The term [information retrieval]... denotes the recovery from a given collection of documents, and with stated probability, of a set of documents that includes, possibly together with some irrelevant ones, all documents of specified content; or, a set of documents that includes nothing but documents of specified contents, but possibly not all of them". Thus the probability that was considered was that of having, on the one hand, perfect recall, or on the other hand, perfect precision. Ranking was not considered. An excellent discussion of the origins of probabilistic information retrieval is given by van Rijsbergen (2005).

2. What the Maron and Kuhns paper showed

In addition to being the first paper on probabilistic information retrieval, the Maron and Kuhns paper was the first to introduce the notions of query expansion and expansion of the set of relevant, retrieved documents.

Section 2.1 shows how probability of relevance was calculated. Section 2.2 describes how the set of relevant retrieved documents was expanded. Section 2.3 describes the general search strategy proposed by Maron and Kuhns. Section 2.4 discusses the contribution of the experimental results.

2.1. Deriving the relevance number

Maron and Kuhns first considered the probability that if a user requests information on subject or topic I_j, then that user will be satisfied with Document D_i, i.e., judge D_i to be relevant. Although Maron and Kuhns used different notation, this can be represented as $P(D_i|I_j)$, where D_i is the event of document D_i being relevant to the user's information need. By Bayes' theorem $P(D_i|I_j)$ is proportional to $P(D_i) * P(I_j|D_i)$. $P(D_i)$, the prior probability of relevance can be determined by library statistics. In other words if n is the number of times that any document is judged relevant and m is the number of times that that document being judged is D_i, then $P(D_i) = m/n$. Thought of most simply, interpreting $P(D_i)$ in this way seems to imply a static retrieval system with no new documents entering or leaving, but $P(D_i)$ could be estimated in more dynamic settings. $P(I_j|D_i)$ is the probability that if D_i is retrieved and found relevant by a searcher that I_j would be the searcher's query term. An estimate of this probability is to be provided by the probabilistic indexer. Maron and Kuhns also showed how probabilities of relevance associated with Boolean combinations of terms could be derived.

2.2. Expanding the set of relevant, retrieved documents

Next Maron and Kuhns considered ways to automatically produce the best set of documents to return for a given request. Given a request R and class C of retrieved documents whose indexing matches the Boolean logic of the request, how can one either: (a) automatically generalize the request R to R', where R' will select class C', which is larger than C, or (b) leave R as it is, but automatically generalize C to C'? To achieve either type of automatic generalization, Maron and Kuhns introduce the concept of an Index Space, where each point represents an index term and an n-dimensional Document Space wherein individual documents are represented as n-dimensional vectors. They discuss semantic and statistical similarity of terms, where statistical similarity would relate to the tendency of two terms to index the same documents. They then define three measures of similarity of terms in Index Space. First, the probability that if a term I_j is assigned to a document, that another term I_k will also be assigned to the documents, that is $P(I_k|I_j)$, Thus a query using term I_j would be expanded by adding the term I_k for which $P(I_k|I_j)$ is the highest. They refer to this as a conditional probability search. They also define $P(I_j|I_k)$, or the inverse conditional probability search, which for a query using term I_j picks the I_k, which provides the highest value for $P(I_j|I_k)$. Finally, they propose a third measure, a coefficient of association based on how far $P(I_j \cdot I_k)$ diverges from $P(I_j) * P(I_k)$. If I_j and I_k were independent, these two values would be equal. These three measures can be used to automatically modify R, the request, by adding similar terms. Maron and Kuhns also show how similarities (or dissimilarities) between documents can be determined based on these measures, so that the set C of documents returned by a request can be automatically expanded. Looking at such similarities among index terms or documents was novel in 1960. Since then many researchers have experimented with a wide variety of similarity measures among index terms and documents.

2.3. A search strategy

Having defined relevance numbers and measures of similarity, or dissimilarity, in Index Space and Document Space, Maron and Kuhns show how all of these ideas can be used in a search strategy. The components of the search strategy include:

- Input, including the request, possibly weighted.
- A probabilistic matrix, including dissimilarities among documents, significance measures for index terms (based on inverse document frequency, anticipating *idf* weighting), and measures of closeness between index terms.
- The a priori probability of relevance for each index term.

- Output, i.e., the method of selecting documents, based on Boolean matching and ranking by probability of relevance.
- Control numbers, e.g., parameters such as the total number of relevant documents desired or relevance thresholds.
- Operations, including the basic selection process and elaborations of the retrieved set based on measures of index term or document closeness.

2.4. Experimental results from the Maron and Kuhns paper

Readers interested in details of the results of Maron and Kuhns' experiments should refer to the 1960 paper. The specific results are not so much important intrinsically, as they are important because they provided an empirical validation of the first probabilistic theory of information retrieval. Moreover, Maron and Kuhn's empirical approach was one of the early empirical studies published in the scientific literature. The paper showed that experiments could be conducted to test and evaluate alternative theories and techniques. Other larger scale empirical evaluations were already underway even at the time of Maron and Kuhn's earlier technical reports. Cleverdon (1959) in his Area 4 paper at the International Conference on Scientific Information mentioned above gives a good characterization of the state-of-affairs in 1958.

> Recent years have seen a two-pronged attack to deal with the problems which have been caused by the immense growth in the amount of recorded information, the greater complexity of the subject matter and the increasing interrelationship between subjects. First, there have been many attempts to devise new indexing systems which will be an improvement on the conventional methods and, on the other hand, a great deal of work has been done developing the mechanics which can be used, from the simpler kinds of hand-sorted punched cards to high-speed computing machines.
> Several theoretical evaluations have been made of the various systems, but it appears that the position has been reached where it is necessary to make a practical assessment of the merits and demerits of information retrieval systems. A project which will attempt to do this has been started under the direction of the author with the aid of a grant from the National Science Foundation to the Association of Special Libraries and Information Bureaux (Aslib).

Maron and Kuhns not only devised a new indexing system, but they also evaluated it both theoretically and empirically. The project to which Cleverdon referred was the Cranfield project, which has become the model for large scale experimental evaluation of retrieval systems. Other empirical work was already underway, e.g., Luhn (1961), or about to begin, e.g., Salton and Lesk (1968), but much of this work was either proprietary or only available in technical reports. I.J. Good had also briefly discussed probabilistic information retrieval in an IBM technical report published shortly after the 1958 conference on scientific information (1958). He mentions having heard of Maron's ongoing work at Ramo Woolridge at the 1958 conference.

Although Maron and Kuhns did not use the term relevance feedback, the notion of relevance feedback was implicit in their discussion of techniques for expanding the set of relevant retrieved documents. It was not until the work of Rocchio (1966) that relevance feedback was treated explicitly.

3. Information Retrieval since the Maron and Kuhns paper

Over the next several years another approach came to dominate academic research in information retrieval in the United States. This was the vector space model developed by Salton and his students from around 1960 until the present (Salton & McGill, 1983). The vector space model is based on the idea of representing all documents in a collection, and queries against the collection as n-dimensional vectors, where each component represented one, generally normalized, word in the collection. (Again, the notion of representing documents as weighted (as opposed to binary) vectors was used in the 1960 Maron and Kuhns paper.)

During the 1970s Cooper and Maron published a paper on utility-theoretic retrieval (1978), and Cooper published a paper on Gedanken experimentation indexing (1978). The Cooper and Maron paper on utility-

theoretic indexing proposed that indexers not assign terms interpreted as probabilities in the Maron and Kuhns sense, but rather that they estimate the utility which a searcher using that term as a search term would derive from retrieving the document. In the Gedanken indexing paper Cooper reflected on how an indexer might make such utility estimates. During this time Cooper was also the first to formally state the probability ranking principle, though the use of the principle goes back to the Maron and Kuhns paper (1977). Meanwhile in England a new approach to probabilistic information retrieval was being developed (Miller, 1971). This has been considered the initial paper on what was later called the Model 2 approach to probabilistic information retrieval (van Rijsbergen, personal communication). Model 1, then, was the approach to probabilistic information retrieval begun by Maron and Kuhns (Robertson, Maron, & Cooper, 1982). It was Miller's work, and that of Barkla (1969), which influenced work on Model 2, rather than the work of Maron and Kuhns (van Rijsbergen, 2005). In the early 1970s, the concept of term frequency inverse document frequency weighting $tf * idf$ was developed by Sparck Jones (1972), in the context of probabilistic retrieval and by Salton and Yang (1973), in the context of the vector space model. This was shortly followed by a paper by Robertson and Sparck Jones, which was the first to formally define Model 2 (1976).

The direct influence of the Maron and Kuhns paper continued in two ways during the 1970s, 1980s, and beyond. First, there were developments of Model 1 itself. Second, Model 1 and Model 2 were combined to form a unified model, Model 3 (Robertson et al.). While, after some initial work (Maron, 1984; Maron, Curry, & Thompson, 1986; Robertson, Maron, & Cooper, 1983; Thompson, 1990a, 1990b) Model 3 was not developed further in the 1980s, the indirect influence of Model 1, through Model 3, has continued in some more recent re-examinations of the unified model (Bodoff, 1999; Bodoff & Robertson, 2004; Robertson, 2003). As was mentioned above, Cooper and Maron formalized Model 1 and generalized the theory to include a utility-theoretic interpretation (1978). Maron and Cooper considered relevance to be a binary relation between a user with an information need and a document. They did not consider relevance to have degrees. On the other hand, they acknowledged that some documents might be more useful than others. This led to their theory of utility-theoretic indexing, where the task of the indexer would be to predict the utility of the document to a user with a given query, rather than the probability of the document's being relevant. Fuhr (1986, 1989) also made further developments with Model 1. According to Fuhr, Model 1 had not been implemented over the years because of the problems in estimating its probabilistic parameters. Fuhr showed how these problems in estimation could be overcome by deriving the parameters from manual indexing.

In the world of commercial document retrieval full text document retrieval products were being developed to complement the existing bibliographic retrieval systems. Now systems could retrieve not only based on indexer assigned metadata, but on any word appearing in the document. Many saw this as the ultimate solution of the library, or document retrieval, problem. In a paper, even more widely cited than the 1960 Maron and Kuhns paper, reporting a large scale study of full text retrieval using the IBM STAIRS system for litigation support, Blair and Maron showed that full text retrieval did not live up to these expectations (1985). Attorneys and paralegals using the full text system, who believed that they were retrieving at least 75% of all relevant documents using the system, were shown to be retrieving at best only 20%. The legal information retrieval companies, Lexis–Nexis and West Publishing, initially tried to dispute these findings (Dabney, 1986), but a few years later, in the early 1990s, each of these companies provided its users with a ranked retrieval mode, implicitly acknowledging the possibility of improvement upon retrieval based on human subject indexing. West's ranked retrieval mode was provided by an implementation of the Inquery system (Turtle & Croft, 1991) from the University of Massachusetts, Amherst, suitably tailored for the legal domain, while that of Lexis–Nexis was based on the vector space model. Thus, ranked document retrieval seemed to have entered the mainstream of commercial retrieval systems in 1992, when West, soon to be followed by Lexis–Nexis and Dialog, offered its users a ranked retrieval search mode. At about the same time the Tipster program and its more widely accessible TREC program showed strong U.S. government research support for ranked retrieval. Then, an unexpected new factor entered the picture.

The 1990s brought the World Wide Web and Web search engines. Early Web search engines, especially as the Web grew to contain large numbers of documents, tended to provide unsatisfactory performance. Eventually graph structure algorithms were developed such as Kleinberg's HITS (1998) and Brin and Page's PageRank (1998). PageRank, when implemented commercially in the Google search engine along with other algorithms, such as those based on the use of anchor text, i.e., the textual context of the hyperlink in the citing

document, provided noticeably better Web retrieval performance. Basically these algorithms calculate the rank of a page based on the patterns of links among Web pages. Since these links are generally made by the human creators of Web pages, these algorithms incorporate human judgment of relevance of one Web page to another. However, since these judgments are not in the context of any particular search, graph linking structures can at best be seen as providing a Bayesian prior for relevance. On the other hand, because anchor text is considered to be part of the representation of the document to which the link is made, the human relevance judgment implicit in the link can be associated with particular queries.

While some academic research has been influenced by developments with Web search algorithms, other researchers continued to develop more general new probabilistic algorithms, in particular those based on probabilistic logic (van Rijsbergen, 1986) and on language modeling (Croft & Lafferty, 2003). Meanwhile, proprietary online retrieval services such as Westlaw, Lexis–Nexis, and Dialog, which in the early 1990s had given their users the choice of Boolean or ranked retrieval, found that the vast majority of their users preferred to stay with Boolean retrieval. The option of ranked retrieval is still offered searchers using Westlaw, Lexis–Nexis, or Dialog, but there has been little motivation for proprietary search services, such as Westlaw, Lexis–Nexis, and Dialog to conduct further research on ranking algorithms. On the other hand, there are many other commercial search sectors other than Web search, e.g., enterprise search, where new ranked retrieval algorithms are still being pursued.

4. Discussion

One might draw the conclusion from the success of ranked retrieval algorithms on the Web, and the rejection of ranked retrieval by users of the proprietary online systems mentioned above, that fully automated processing is most appropriate for the large, diverse content of the Web, while for premium content requiring manual classification and indexing, such as in legal, patent, or biomedical domains, that binary indexing and Boolean querying, practiced substantially today as in the 1960s, are still the preferred techniques. And yet, if these human indexers were to assign terms with weights, whether interpreted as probabilities, as suggested by Maron and Kuhns in 1960, or as utilities, as later suggested by Cooper and Maron (1978), improved results might be possible, as indicated by Maron and Kuhns' experiments reported in 1960. In this light, it is of interest to consider this statement from Hodge and Milstead's *Computer Support to Indexing* (1998).

> Aboutness is the inherent subject of the document and is distinguished from the "meaning" of the document, i.e., the reason for which the user is seeking the document or the purpose which it may serve for the user. The aboutness of the document is more stable than the meaning, which varies from user to user and from search request to search request for the same user. As long as a pragmatic approach is followed, the aboutness of a document can usually be determined – the uses that might be made of the document – is to a large extent unpredictable.

Maron (1977) discussed aboutness in this context as one factor contributing to relevance. In Robertson, Maron, and Cooper's unified model of probabilistic retrieval it was shown how an indexer could on the one hand assign binary index terms to a document from one thesaurus of document features, while at the same time making probabilistic estimates for terms taken from a second thesaurus of searcher features (1982). While the future uses made of a document might be difficult to predict, a well-trained indexer might be expected to make a reasonable estimate for the immediate future (Thompson, 1988). Relevance judgments provided by searchers who retrieved the document could update these initial probabilities (Thompson, 1990a, 1990b).

It has long been recognized that multiple sources of evidence can be combined to provide improved retrieval. This is inherent in the probability ranking principle (Robertson, 1977) and has been shown in empirical (Saracevic & Kantor, 1988a, 1988b; Saracevic, Kantor, Chamis, & Trivison, 1988) and theoretical studies (Croft, 2000, chapter 1). In the graph structure of the Web contemporary Web search engines have found one effective source of evidence. Probabilistic term assignment by human indexers, i.e., the assignment of weighted index terms, is another source of evidence, identified 48 years ago, but still completely unexploited.

The approach to document retrieval pioneered by Maron and Kuhns can be applied more broadly. Document retrieval is one example of an inductive search problem. Maron and his colleagues later developed a

system called HelpNet, which applied the probabilistic document retrieval theory which began with the Maron and Kuhns paper to the problem of finding people as sources of information, what is now called the problem of expert finding (Maron et al., 1986). More generally, the same probabilistic searching technology can be used to find a wide variety of objects, e.g., jobs, homes, or products of any kind.

5. Conclusion

Probabilistic retrieval, or more generally ranked retrieval, had its origins in the 1960 Maron and Kuhns paper. This paper provided the initial theory and empirical research for ranked retrieval. The specific approach which Maron and Kuhns advocated, based on probabilistic human indexing has not been followed, but ranked retrieval has had widespread application in academic research and in commercial systems. In the 48 years since the paper was published the field has grown enormously in at least three dimensions: depth of theory; new ramifications, techniques, and directions; and new and powerful applications.

Finally, despite the many changes in ranking algorithms, hardware, and network infrastructure that have come about since 1960, the library problem is still far from being fully solved. The suggestion of the Maron and Kuhns paper that human indexers might assign term as probabilities, rather than binarily, has not been seriously explored in the intervening years. Although there are many search environments where the expense of human indexing cannot be supported, there are others where adopting this suggestion might yet lead to improved retrieval.

Acknowledgements

The author thanks M.E. Maron, I.J. Good, and R. Solomonoff for their comments on the state of information retrieval research around 1960. The author also thanks the two anonymous reviewers whose comments helped improve the paper.

References

Bagley, P. (1951). Electronic digital machines for high-speed information searching. M.S. thesis. Cambridge, MA: Massachusetts Institute of Technology.

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), 289–299.

Bodoff, D. (1999). A re-unification of two competing models for document retrieval. *Journal of the American Society for Information Science, 50*(1), 49–64.

Bodoff, D., & Robertson, S. E. (2004). new unified probabilistic model. *Journal of the American Society for Information Science and Technology, 55*(6), 471–487.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th world wide web conference*.

Brown, Sir L. (1959). Opening session address. In *Proceedings of the international conference on scientific information I* (pp. 3–8). Washington, DC: National Academies of Sciences – National Research Council. November 16–21, 1958.

Bush, Vannevar (1945). As we may think. *Atlantic Monthly*, July.

Cleverdon, C. W. (1959). The evaluation of systems used in information retrieval. In *Proceedings of the international conference on scientific information* (Vol. 1, p. 687). Washington, DC: National Academy of Sciences.

Cleverdon, C. W. (1962). Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems, College of Aeronautics, Cranfield.

Cooper, W. S. (1978). *Journal of the American Society for Information Science, 29*(3), 107–119.

Cooper, W. S., & Maron, M. E. (1978). Foundations of probabilistic and utility-theoretic indexing. *Journal of the Association for Computing Machinery, 25*(1), 67–80.

Croft, W. B. (2000). Combining approaches to information retrieval. In W. B. Croft (Ed.), *Advances in information retrieval: Recent research from the center for intelligent information retrieval*. Boston: Kluwer.

Croft, W. B., & Lafferty, J. (Eds.). (2003). *Language modeling for information retrieval*. Boston: Kluwer Academic.

Dabney, D. (1986). A reply to West Publishing Company and mead data central on the curse of Thamus. *Law Library Journal, 78*, 349.

Fairthorne, R. (1961). *Towards information retrieval*. London: Butterworth.

Fuhr, N. (1986). Two models of retrieval with probabilistic indexing. In F. Rabitti (Ed.), *1986 – ACM conference on research and development in information retrieval* (pp. 249–257), 8–10 September 1986, Pisa, Italy.

Fuhr, N. (1989). Models for retrieval with probabilistic indexing. *Information Processing and Management, 22*(1), 55–72.

Good, I. J. (1958). Speculations concerning information retrieval. IBM Research Report RC-78, December 10, 1958.

Hodge, G., & Milstead, J. (1998). *Computer support to indexing*. Philadelphia: National Federation of Abstracting and Information Services.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM symposium on discrete algorithms*.

Luhn, H. P. (1961). The automatic derivation of information retrieval encodements from machine-readable texts. In A. Kent (Ed.). *Information retrieval and machine translation* (Vol. 3, part 2, pp. 1021–1028). New York: Interscience Publication.

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

Maron, M. E. (1984). Probabilistic retrieval models. In B. Dervin, & M. J. Voigt (Eds.), *Progress in communication sciences* (Vol. 5, pp. 145–176).

Maron, M. E., Curry, S., & Thompson, P. (1986). An inductive search system: Theory, design and implementation. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-16*(1), 21–28.

Maron, M. E., & Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery, 7*(3), 216–244.

Maron, M. E., Kuhns, J., & Ray, L. (1958). Some experiments with probabilistic indexing, RAMO-WOOLDRIDGE, December 1958, 45 pp.

Maron, M. E. Kuhns, J., & Ray, L. (1959). Probabilistic indexing: A statistical technique for document identification and retrieval, Technical Memorandum No. 3, Data Systems Project Office, RAMO-WOOLDRIDGE, June 1959, 91 pp.

Miller (1971). A probabilistic search strategy for Medlars. *Journal of Documentation, 27*, 254–266.

Mooers, C. N. (1950). The theory of digital handling of non-numerical information and its implications to machine economics. In *Association of computing machinery meeting*. March1950.

National Academy of Sciences (1959). In *Proceedings of the international conference on scientific information*, Washington, DC.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation, 33*(4), 294–304.

Robertson, S. E. (2003). The unified model revisited. In *Presented at the workshop on the mathematical foundations of information retrieval*, SIGIR 2003, Toronto, Canada.

Robertson, S. E., Maron, M. E., & Cooper, W. S. (1982). Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development, 1*, 1–21.

Robertson, S. E., Maron, M. E., & Cooper, W. S. (1983). The unified probabilistic model for IR. In G. Salton & H.-J. Schneider (Eds.), *Research and development in information retrieval* (pp. 108–117). Berlin: Springer-Verlag.

Robertson, S. E. & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 129-46. Reprinted in: P. Willett (Ed.), Document retrieval systems. Taylor Graham, 1988 (pp. 143–160).

Rocchio, J. J., Jr. (1966). Document retrieval systems – Optimization and evaluation. Harvard University Doctoral Thesis, Report No. ISR-10 to the National Science Foundation, Harvard Computation Laboratory, March.

Salton, G., & Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM, 15*, 8–36.

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

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

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

Saracevic, T., 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.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 11–21.

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

Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation, 29*, 351–372.

Taube, M., & Wooster, H. (Eds.). (1958). *Information storage and retrieval: Theory, systems, and devices*. New York: Columbia University Press.

Thompson, P. (1988). Subjective probability and information retrieval: A review of the psychological literature. *Journal of Documentation, 44*(2), 119–143.

Thompson, P. (1990a). A combination of expert opinion approach to probabilistic information retrieval, Part 1: The conceptual model. *Information Processing and Management, 26*(3), 371–382.

Thompson, P. (1990b). A combination of expert opinion approach to probabilistic information retrieval, Part 2: Mathematical treatment of CEO model 3. *Information Processing and Management, 26*(3), 383–394.

Turtle, H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems, 9*(3), 187–222.

van Rijsbergen, C. J. (2005). The emergence of probabilistic accounts of information retrieval. In J. Tait (Ed.), *Charting a new course: Natural language processing and information retrieval Essays in honour of Karen Sparck Jones* (pp. 23–38). Dordrecht, The Netherlands: Springer.

van Rijsbergen, C. J. (1986). A new theoretical framework for information retrieval. In *SIGIR'86, Proceedings of the 9th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 194–200).