The History of Information Retrieval Research

Mark Sanderson: School of Computer Science and Information Technology, RMIT University, GPO Box 2476, Melbourne 3001 Victoria, Australia. mark.sanderson@rmit.edu.au; +61 3 992 59675

W. Bruce Croft: Department of Computer Science, 140 Governors Drive, Box 9264, University of Massachusetts, Amherst, MA 01003-9264, USA. croft@cs.umass.edu; +1 413 545-0463

Abstract

This paper describes a brief history of the research and development of information retrieval systems starting with the creation of electro-mechanical searching devices, through to the early adoption of computers to search for items that are relevant to a user's query. The advances achieved by information retrieval researchers from the 1950s through to the present day are detailed next, focusing on the process of locating relevant information. The paper closes with speculation on where the future of information retrieval lies.

Keywords

Information Retrieval, History, Ranking Algorithms

Introduction

The long history of information retrieval does not begin with the internet. It is only in the last decade and a half of the IEEE's one hundred years that web search engines have become pervasive and search has become integrated into the fabric of desktop and mobile operating systems. Prior to the broad public day-to-day use of search engines, IR systems were found in commercial and intelligence applications as long ago as the 1960s. The earliest computer-based searching systems were built in the late 1940s and were inspired by pioneering innovation in the first half of the 20th century. As with many computer technologies, the capabilities of retrieval systems grew with increases in processor speed and storage capacity. The development of such systems also reflects a rapid progression away from manual library-based approaches of acquiring, indexing, and searching information to increasingly automated methods.

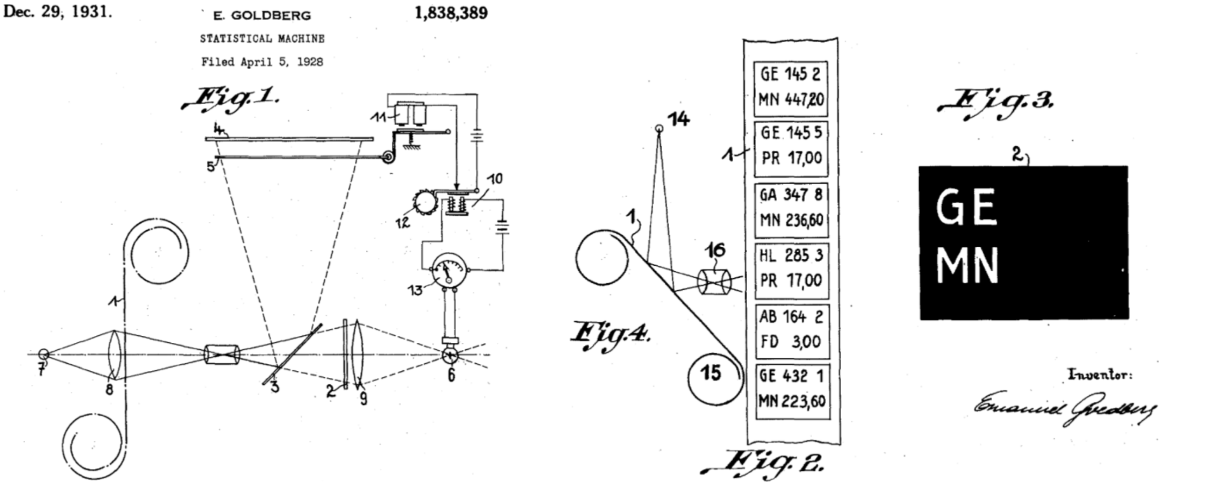
An information retrieval (IR) system locates information that is relevant to a user's query. An IR system typically searches in collections of unstructured or semi-structured data (e.g. web pages, documents, images, video, etc.). The need for an IR system occurs when a collection reaches a size where traditional cataloguing techniques can no longer cope. Similar to Moore's law of continual processor speed increase, there has been a consistent doubling in digital storage capacity every two years. The number of bits of information packed into a square inch of hard drive surface grew from 2,000 bits in 1956 to 100 billion bits in 2005[1]. With the growth of digitised unstructured information and, via high speed networks, rapid global access to enormous quantities of that information, the only viable solution to finding relevant items from these large text databases was search, and IR systems became ubiquitous.

This brief review of past work focuses on the algorithms that take a user's query and retrieve a set of relevant documents. This paper opens with a review of the early developments of electro-mechanical and computational devices that searched manually generated catalogues. This is followed by a description of how IR moved to automatic indexing of the words in text and how complex Boolean query languages gave way to simple text queries. The automatic techniques and theories that supported them have continued to be developed for more than forty years, and provided the framework for successful web search engines. This review finishes with a perspective on the future challenges for IR.

Pre-history – mechanical and electro-mechanical devices

Conventional approaches to managing large collections of information originate from the discipline of librarianship. Commonly, items such as books or papers were indexed using cataloguing schemes. Eliot and Rose claim this approach to be millennia old: declaring Callimachus, a 3rd century BC Greek poet as the first person known to create a library catalogue [2]. Facilitating faster search of these physical records was long researched, for example, Rudolph filed a US patent in 1891 for a machine composed of catalogue cards linked together, which could be wound past a viewing window enabling rapid manual scanning of the catalogue. Soper filed a patent for a device in 1918 [3], where catalogue cards with holes, related to categories, were aligned in front of each other to determine if there were entries in a collection with a particular combination of categories. If light could be seen through the arrangement of cards, a match was found.

Mechanical devices that searched a catalogue for a particular entry were also devised. The first person to build such a system appears to be Emanuel Goldberg who tackled this problem in the 1920s and '30s. A series of patents were granted to Goldberg for a machine that searched for a pattern of dots or letters across catalogue entries stored on a roll of microfilm. Goldberg's original patents can be found on the web sites of the German and US patent offices. Part of the US version filed in 1928 [4] is shown in Figure 1. Here it can be seen that catalogue entries were stored on a roll of film (No. 1 of the figure). A query (2) was also on film showing a negative image of the part of the catalogue being searched for; in this case the 1st and 6th entries on the roll. A light source (7) was shone through the catalogue roll and query film, focused onto a photocell (6). If an exact match was found, all light was blocked to the cell causing a relay to move a counter forward (12) and for an image of the match to be shown via a half silvered mirror (3), reflecting the match onto a screen or photographic plate (4 & 5).



According to a biography of Goldberg by Buckland [5], three prototypes of the machine were built, one of which was said to be integrated into Goldberg's desk. Buckland quotes a colleague of Goldberg as saying "*He was telling us that he was the only person in the world as far as he knew who had on his desk a document retrieval capability ... He would dial a number, press a button and after three seconds [a microfilmed copy of] the document would be projected*".

A number of other researchers produced devices based on similar principles: Mooers [6] described investigations by Davis and Draeger in 1935 on searching with microfilm. This work, according to Mooers was taken up by Vannevar Bush in the late 1930s, who, with his students, built a film-based searching prototype. The work undoubtedly informed Bush's well-known proposal of the Memex system in 1945 [7]. The culmination of this approach appears to be Shaw's rapid selector [8], which was reported to search through a 2,000 foot reel of film. Each half of the film's frames had a different purpose: one half for 'frames of material'; the other for 'index entries'. It is stated that 72,000 frames were stored on the film, which in total were indexed by 430,000 entries. Shaw reported that the selector was able to search at the rate of 78,000 entries per minute.

Other mechanical technologies were examined. Luhn, for example, made a selector using punch cards, light and photocells. Prototypes of this system were completed in 1950 and demonstrated in 1951. A key feature of this system was that a consecutive sequence of characters could be matched within a larger string [9]. The system searched at the rate of 600 cards per minute. At this time, the term "information retrieval" was first used. Presenting a paper at a conference in March 1950, Calvin Mooers wrote "*The problem under discussion here is machine searching and retrieval of information from storage according to a specification by subject… It should not be necessary to dwell upon the importance of information retrieval before a scientific group such as this for all of us have known frustration from the operation of our libraries – all libraries, without exception.*" [10]. Mooers' paper described IR systems using punch cards. As reported by Jahoda [11], these mechanical systems continued to be developed and used until the advent of computers when this approach to IR was surpassed.

Early use of computers for IR

To discuss means of dealing with a perceived explosion in the amounts of scientific information available, a specially convened conference was held by the UK's Royal Society in 1948. At it, Holmstrom described a "*machine called the Univac*" capable of searching for text references associated with a subject code. The code and text were stored on a magnetic steel tape [12]. Holmstrom stated that the machine could process "*at the rate of 120 words per minute*"[1]. It appears that this is the first reference to a computer being used to search for content.

Mitchell [13] described a project to model the use of a Univac computer to search 1,000,000 records indexed by up to six subject codes; it was estimated that it would take 15 hours to search that many records. Nanus [14] detailed a number of computer-based IR projects run in the 1950s, including one system from General Electric that searched over 30,000 document abstracts; see also Brownson [15] for another review of implementations of computerised IR systems in that decade, including mention of IR work conducted in the Soviet Union in the 1950s. The impact of computers in IR is highlighted when Hollywood drew public attention to the innovation with the comedy "Desk Set", which came out in 1957. It centred on a group of reference librarians who were about to be replaced by a computer.

IR as a research discipline was starting to emerge at this time with two important developments: how to index documents and how to retrieve them.

Indexing – the move towards words

In the field of librarianship, the way that items were organised in a collection was a topic that was regularly debated. The classic approach was to use a hierarchical subject classification scheme, such as the Dewey Decimal system, which assigned numerical codes to collection items. However, alternatives were proposed, most notably Taube et al's Uniterm system [16], which was essentially a proposal to index items by a list of keywords. As simple an idea as this seems today, this was at the time a radical step. A few years later, Cleverdon conducted a detailed comparison of retrieval effectiveness using Uniterms and the more classic classification techniques [17]. His conclusion that Uniterms were as good as and possibly better than other approaches caused much surprise and his work came in for extensive scrutiny [18]. However, Cleverdon's experimental results were found to be correct and as a result the use of words to index the documents of an IR system became established. Many aspects of Cleverdon's *test collection* approach to evaluation are still used in both academic research and commercial search testing today.

Ranked retrieval

The style of search used by both the electro-mechanical and computer-based IR systems was so-called Boolean retrieval. A query was a logical combination of *terms* (a synonym of word in IR literature), which resulted in a set of those documents that exactly matched the query. Luhn [19] proposed and Maron, Kuhns, and Ray [20] tested an alternative approach, where each document in the collection was assigned a score indicating its relevance to a given query. The documents were then sorted and those at the top ranks were returned to the user. The researchers manually assigned keywords to a collection of 200 documents, weighting those assignments based on the importance of the keyword to the document. The scores assigned to the documents were based on a probabilistic approach. The researchers hand tested their ranked retrieval method, showing that it outperformed Boolean search on this test collection with 39 queries. In the same year as Maron et al's work, Luhn suggested "*that the frequency of word occurrence in an article furnishes a useful measurement of word significance*" [21]; his approach later became known as *term frequency* weighting.

This ranked retrieval approach to search was taken up by IR researchers, who over the following decades refined and revised the means by which documents were sorted in relation to a query. The superior effectiveness of this approach over Boolean search was demonstrated in many experiments over those years, see p237 in Spärck Jones's book [22] for a list of these experiments. Work in the 1950s established computers as the definitive tool for search. What followed was the growth of a commercial search sector and the consolidation of IR as an increasingly important research area.

1960s

The 1960s saw a wide range of activities reflecting the move from simply asking if IR was possible on computers to determining means of improving IR systems. One of the major figures to emerge in this period was Gerard Salton, who formed and led a large IR group, first at Harvard University, then at Cornell. The group produced numerous technical reports (the ISR reports), establishing ideas and concepts that are still major areas of investigation today.

One of these areas is the formalization of algorithms to rank documents relative to a query. Of particular note was an approach where documents and queries were viewed as vectors within an N dimensional space (N being the number of unique terms in the collection being searched). This was first proposed by Switzer [23] and later,

[1] Note, the UNIVAC isn't generally thought to have come into existence until 1951, the date when the first machine was sold, Holmstrom presumably saw or was told about a pre-production version.

the similarity between a document and query vector was suggested by Salton to be measured as the cosine of the angle between the vectors using the cosine coefficient [24] (p236).

Another significant innovation at this time was the introduction of relevance feedback [25]. This was a process to support iterative search, where documents previously retrieved could be marked as relevant in an IR system. A user's query was automatically adjusted using information extracted from the relevant documents. Versions of this process are used in modern search engines, such as the "Related articles" link on Google Scholar. Relevance feedback was also the first (but not the last) use of machine learning in IR.

Other IR enhancements examined in this period included the clustering of documents with similar content; the statistical association of terms with similar semantic meaning, increasing the number of documents matched with a query by expanding the query with lexical variations (so called stems) or with semantically associated words. For coverage of this past research see Salton's book [24], the proceedings of the 1964 conference on Statistical Association Methods for Mechanized Documentation [26], and Van Rijsbergen's book [27].

In this decade, commercial search companies emerged out of the development of bespoke systems built for large companies or government organisations. See Dennis [28] for a description of one of these early systems that was searching tens of thousands of items. Bjørner states that one of the first companies dedicated to providing search was Dialog formed in 1966 from the creation of an IR system for NASA [29].

A striking aspect of this time was the low level of interaction between the commercial and IR research communities. Despite researchers' consistent demonstration that ranked retrieval was a superior technique, almost all commercial searching systems used Boolean search. This situation didn't change until the early to mid-1990s with systems such as WESTLAW's WIN system [30] and the growth of web search engines.

1970s

One of the key developments of this period was that Luhn's term frequency (*tf*) weights (based on the occurrence of words within a document), were complemented with Spärck Jones's work on word occurrence across the documents of a collection. Her paper on inverse document frequency (*idf*) introduced the idea that the frequency of occurrence of a word in a document collection was inversely proportional to its significance in retrieval: less common words tended to refer to more specific concepts, which were more important in retrieval [31]. The idea of combining these two weights (*tf•idf*) was quickly adopted; see Salton and Yang [32] for an early exploration of such ideas.

A number of researchers worked to formalize the retrieval process. Salton synthesised the outputs of his group's work on vectors to produce the vector space model [33]. This approach to describing the retrieval process underpinned many research retrieval systems and much research for the coming two decades. Nowadays, the ranking formulas proposed by Salton are rarely used, however, viewing documents and queries as vectors in a large dimensional space is still common.

An alternative means of modelling IR systems involved extending Maron, Kuhns and Ray's idea of using probability theory. Robertson defined the probability ranking principle [34], which determined how to optimally rank documents based on probabilistic measures with respect to defined evaluation measures. A further paper from Robertson and Spärck Jones [35] along with a derivation of the probabilistic model in Van Rijsbergen's book [27] stimulated much research on this form of modelling. Van Rijsbergen showed that the basic probabilistic model assumed that words in a document occurred independently of each other, which is a somewhat unrealistic assumption. Incorporating term dependency into ranked retrieval started to be examined, which lead to a wide range of research in later years.

1980s – mid 1990s

Building on the developments of the 1970s, variations of *tf•idf* weighting schemes were produced (Salton and Buckley [36] reviewed an extensive range) and the formal models of retrieval were extended. The original probabilistic model did not include *tf* weights and a number of researchers worked to incorporate them in an effective and principled way. Amongst other achievements, this work ultimately led to the ranking function BM25 (Robertson et al, 199?), which, although not as principled an approach as some researchers would have liked, has proven to be a highly effective ranking function and is still commonly used.

Advances on the basic vector space model were also developed and probably the most well-known is Latent Semantic Indexing (LSI), where the dimensionality of the vector space of a document collection was reduced though singular-value decomposition [37]. Queries were mapped into the reduced space. Deerwester and his colleagues claimed the reduction caused words with common semantic meaning to be merged resulting in queries matching a wider range of relevant documents. Tests in the original paper were described as only "*modestly encouraging*"; nevertheless, the paper has been highly influential.

Unlike the purely numerical approach of LSI for extending the range of documents a query could match, others explored computational linguistics approaches considering the syntax of words, their semantics; addressing anaphora, ambiguity, and named entities. A great deal of this work led to little or no improvement in the effectiveness of retrieval systems. One technique that was found to produce a level of improvement was stemming, the process of matching words to their lexical variants. Although stemming algorithms date back to the 1960s, Porter in the late 1970s developed a compact set of English language stemming rules; his Porter stemmer [38] continues to influence stemming design today.

TREC

One concern in the academic community in the late 1980s and early 1990s was that the size of document collections being used for testing was small compared to the collections that some commercial search companies were working with at the time. Donna Harman and colleagues formed TREC (Text REtrieval Conference), an annual exercise where a large number of international research groups collaborated to build test collections several orders of magnitude larger than had been in existence before [39]. Working with these new data sets showed that the existing weighting and ranking functions were not ideally suited for these different collections. It was also becoming clear that different collections required different ranking and weighting approaches. This realisation was to be further confirmed as web search engines started to be developed in the late 1990s.

Learning to rank

Up to this point, the ranking functions used in search engines were manually devised and tuned by hand through experimentation. Fuhr [40] described work where the retrieval function was learned based on relevant documents identified for an existing set of queries. Whereas Rocchio's relevance feedback tuned the query for a particular search, Fuhr's idea was to tune the ranking function for all queries for a particular document collection. The idea was followed up soon after, [41], [42], but only became truly effective when more training data became available in web query logs in the 2000s, along with better learning methods that are able to handle large numbers of features.

Mid 1990s – present

Although Berners-Lee created the World Wide Web in late 1990, the number of web sites and quantity of pages was relatively small until 1993. In those initial years, conventional manual cataloguing of content sufficed. In the middle of 1993, as recorded by Gray's survey[2], there were around 100 web sites; six months later, there were over four times that number, six months after that, the number had increased fourfold again. Web search engines started to emerge in late 1993 to cope with this growth. The arrival of the web initiated the study of new problems in IR. This point also marked a time when the interaction between the commercial and research oriented IR communities was much stronger than it had been before. Ideas developed in earlier years were pushed further and implemented in the commercial search sector.

Web search

Until the rise of the web, the collections that people searched were selected and edited from authoritative sources by the search service providers. On the web, however, the situation was different. Search engine developers quickly realised that they could use the links between web pages to construct a crawler or robot to traverse and gather most web pages on the internet; thereby automating acquisition of content[3]. However, this approach did nothing to ensure a crawled collection contained only authoritative material. Unscrupulous authors discovered that by manipulating the content of a page, they could alter its rank on a search engine. Methods to combat such manipulation (i.e., the various types of spam) and to also identify the best pages on the web were needed.

Two important developments to achieving these goals were link analysis and searching of anchor text – i.e. searching both the content of a web page and the text of links pointing (anchoring) to that page. Both developments were related to earlier work on the use of citation data for bibliometric analysis and search, and using "spreading activation" search in hypertext networks. The anchor text, almost always a brief summary of the page, was recognised early on as a valuable source of information (e.g. McBryan's work in 1994 [43]). The anchor texts were generally written by a number of people, making manipulation of that text harder to achieve. Using anchor text was a key feature of the Google search engine from its early development [44], along with the more famous use of link analysis methods: PageRank developed by the creators of Google and HITS which was developed at the same time by Kleinberg [45].

[2] http://www.mit.edu/~mkgray/net/web-growth-summary.html

[3] The first crawler for web search was developed for the JumpStation search engine built by Fletcher in late 1993. The search engine offered only basic search of part of the web pages gathered. The first full text search engine using a crawler was WebCrawler released in 1994.

Adding link analysis and multiple text representations of documents to existing document ranking functions meant that the internal algorithm of an IR system was becoming complex. Correctly setting parameters for these different features was a challenge, which caused a revisiting of the learning to rank approaches started by Fuhr. He was hampered by a lack of training data, however, as search engines became popular, it was realised that logs of user interactions could be exploited for this purpose. The log data is very noisy, but solutions for extracting valuable information were found; see for example, Joachims's use of logs to train a rank function [46].

Exploiting query logs

Automated exploitation of information extracted from the logs of search engines was also examined in this time. Although logs were stored and examined for many years [47], the most they had been used for was to inform subsequent manual adjustment of a searching system [48]. The true potential of extracting valuable information from these logs was only realised when large volumes of people started to use web search engines. Examining users' queries, click patterns, and reformulations of queries enabled researchers to develop more effective query processing techniques based on understanding the user's "intent", such as automated spell correction [49]; automated query expansion [50], and more accurate stemming [51].

Other advances

In the same way that query log analysis had long been known about, but was only researched in detail more recently, it had long been recognised that different users with different information needs might search for that need using the same query [52]. IR systems should be able to serve such diverse needs, by finding "differently relevant" documents to rank. Only since the late 1990s, has there been a concerted effort to tackle this problem. Carbonell and Goldstein's [53] description of their MMR diversity system was a key paper in generating interest in this area.

The retrieval models that are the basis of the core ranking function of IR systems continued to be developed in this period. Of particular note was the introduction of a probabilistic approach using language models, described by Ponte and Croft [54] and by Hiemstra [55]. By taking a new view of the matching process between documents and queries, the language model approach provided new understanding of a wide range of IR processes, such as relevance feedback, forming clusters of documents, and term dependence. Metzler and Croft [56] showed, for example, that incorporating term dependence in the form of proximity operators in a ranking function significantly outperforms term independence models.

New areas of search

The applications of search and the field of information retrieval continue to evolve as the computing environment changes. The most obvious recent example of this type of change is the rapid growth of mobile devices and social media. One response from the IR community has been the development of *social search*, which deals with search involving communities of users and informal information exchange. New research in a variety of areas such as user tagging, conversation retrieval, filtering and recommendation, and collaborative search is starting to provide effective new tools for managing personal and social information. An important early paper in this area dealt with desktop search [57], which has many similar characteristics to current search applications in the mobile world.

Much research on web IR has focused on short queries, which have little linguistic structure (typically a single noun compound). Another development has been supporting users who issue longer, more natural questions. Much of this work started with the question answering task in TREC [58], that dealt with finding simple answers in text to a limited range of questions (such as the "wh" questions "who" and "when"). This then progressed into the more detailed questions found in large community-based question-answering archives. Researchers have also been developing techniques that provide more focused answers for more detailed questions. The success of applications such as Apple's Siri, IBM's Watson, and Yahoo! Answers is in part due to this research.

Further reading

Beyond ranking functions, a wide range of IR research was extensively studied in areas such as information seeking behaviour, interface design, implementation of search engines, evaluation, and specialisations for particular collection types (e.g. social media, multimedia, etc.). Books from Manning et al [59], Hearst [60], Croft et al [61] and Baeza-Yates and Ribeiro-Neto [62] provide excellent coverage of these other research areas. Recent detailed reviews of specific research areas in IR can be found in the journal Foundations and Trends in Information Retrieval and the Morgan Claypool Synthesis Lectures on Information Concepts, Retrieval, and Services.

Conclusions and future directions

The 20th and early 21st centuries were transformational in the way people accessed information. In 1912, a person with an information need would probably go to a local library and, using a card catalogue, locate books or documents that hopefully answered that need. Because of the relative inconvenience of accessing information in that way, that person would most likely only seek to answer a small number of questions. The scope of information available to people would be limited by the size of their library; for a small number of very important needs, a loan across libraries might have been arranged. Because of the ubiquity of web-based search, it need hardly be said what the current state of the art is: for those with an internet connection, one can instantaneously access hundreds of terabytes of web pages, video clips, news, images, social media, scanned books, academic papers, music, television programs, and films; almost always through search engines. In the last few years, the access has been also possible from a mobile phone. Just about the only thing in common between the situation today and 100 years ago is that both services are generally free at the point of use.

Because the systems that are accessible today are so easy to use, it is tempting to think the technology behind them is similarly straightforward to build. This review has shown that the route to creating successful IR systems required much innovation and thought over a long period of time.

When considering possible future directions, Apple's 1987 Knowledge Navigator vision of IR is still a strong exemplar of how search systems might develop. The short film showed a college professor pulling together a lecture presentation at the last minute. The professor used a form of tablet computer running an IR system presented as an agent capable of impeccable speech recognition, natural dialogue management, a high level of semantic understanding of the searcher's information needs, as well as unbounded access to documents and federated databases.

The Knowledge Navigator identified and connected the professor to a colleague who helped him with the lecture. The broader implications of finding people (rather than documents) to aid with information needs that we see facilitated in the vast growth of social media was not really addressed in the Apple vision. What it also did not encompass was the portability of computer devices opening the possibility of serving information needs pertinent to the particular local context of location, location type, route, the company one is in, or a combination of all these factors.

Today's web search engines seem a simple tool compared to such visions of the future: there are still many opportunities to improve.

References

[1] C. Walter, 'Insights: Kryder's Law', *Scientific American*, 01-2005.

[2] S. Eliot and J. Rose, *A Companion to the History of the Book*. John Wiley and Sons, 2009.

[3] H. E. Soper, 'Means for compiling tabular and statistical data', U.S. Patent US00135169231-1920.

[4] E. Goldberg, 'Statistical Machine', U.S. Patent 183838929-1931.

[5] M. K. Buckland, *Emanuel Goldberg and his knowledge machine: information, invention, and political forces*. Greenwood Publishing Group, 2006.

[6] C. N. Mooers, 'The next twenty years in information retrieval: some goals and predictions', in *Papers presented at the western joint computer conference*, 1959, pp. 81-86.

[7] V. Bush, 'As we may think', *The Atlantic Monthly*, vol. 176, no. 1, pp. 101-108, 1945.

[8] R. R. Shaw, 'The Rapid Selector', *Journal of Documentation*, vol. 5, no. 3, pp. 164 - 171, 1949.

[9] 'New Tools for the Resurrection Of Knowledge', *Chemical and Engineering News*, vol. 32, no. 9, pp. 866-869,891, 01-1954.

[10] C. N. Mooers, 'The theory of digital handling of non-numerical information and its implications to machine economics', in *Association for Computing Machinery Conference*, Rutger University, 1950.

[11] G. Jahoda, 'Electronic searching', in *The state of the library art, (Volume 4),* Graduate School of Library Service, Rutgers in New Brunswick, N.J ., 1961, pp. 139-320.

[12] J. E. Holmstrom, 'Section III. Opening Plenary Session', in *The Royal Society Scientific Information Conference, 21 June-2 July 1948 : report and papers submitted*, London: Royal Society, 1948.

[13] H. F. Mitchell, 'The Use of the Univ AC FAC-Tronic System in the Library Reference Field', *American Documentation*, vol. 4, no. 1, pp. 16-17, 1953.

[14] B. Nanus, 'The Use of Electronic Computers for Information Retrieval', *Bull Med Libr Assoc*, vol. 48, no. 3, pp. 278-291, Jul. 1960.

[15] H. L. Brownson, 'Research on Handling Scientific Information', *Science*, vol. 132, pp. 1922-1931, 30 1960.

[16] M. Taube, C. D. Gull, and I. S. Wachtel, ‘Unit terms in coordinate indexing’, *American Documentation*, vol. 3, no. 4, pp. 213-218, Jan. 1952.

[17] C. W. Cleverdon, ‘The Evaluation of Systems Used in Information Retrieval (1958: Washington)’, in *Proceedings of the International Conference on Scientific Information -- Two Volumes*, 1959, pp. 687-698.

[18] C. W. Cleverdon, ‘The significance of the Cranfield tests on index languages’, in *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, Chicago, Illinois, United States, 1991, pp. 3-12.

[19] H. P. Luhn, ‘A statistical approach to mechanized encoding and searching of literary information’, *IBM Journal of research and development*, vol. 1, no. 4, pp. 309–317, 1957.

[20] M. E. Maron, J. L. Kuhns, and L. C. Ray, ‘Probabilistic indexing: A statistical technique for document identification and retrieval’, Thompson Ramo Wooldridge Inc, Los Angeles, California, Data Systems Project Office, Technical Memorandum 3, Jun. 1959.

[21] H. P. Luhn, ‘The automatic creation of literature abstracts’, *IBM Journal of research and development*, vol. 2, no. 2, pp. 159-165, 1958.

[22] K. Spärck Jones, Ed., *Information Retrieval Experiment*. Butterworth-Heinemann, 1981.

[23] P. Switzer, ‘Vector Images in Document Retrieval’, Harvard University, ISR-4, 01 1963.

[24] G. Salton, *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

[25] J. J. Rocchio, ‘Relevance Feedback in Information Retrieval’, Harvard University, ISR-9, 1965.

[26] M. E. Stevens, V. E. Giuliano, and L. B. Heilprin, *Statistical association methods for mechanized documentation: symposium proceedings*. Washington, DC: G.P.O., 1964.

[27] C. J. van Rijsbergen, *Information Retrieval*, 2Rev Ed. Butterworth-Heinemann Ltd, 1979.

[28] B. K. Dennis, J. J. Brady, and J. A. Dovel, ‘Five Operational Years of Index Manipulation and Abstract Retrieval by Computer.’, *J. Chem. Doc.*, vol. 2, no. 4, pp. 234-242, Oct. 1962.

[29] S. Bjørner and S. C. Ardito, ‘Online Before the Internet, Part 1: Early Pioneers Tell Their Stories’, *Searcher: The Magazine for Database Professionals*, vol. 11, no. 6, Jun-2003.

[30] H. Turtle, ‘Natural language vs. Boolean query evaluation: A comparison of retrieval performance’, in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 1994, pp. 212-220.

[31] K. Spärck Jones, ‘A statistical interpretation of term specificity and its application in retrieval’, *Journal of documentation*, vol. 28, no. 1, pp. 11-21, 1972.

[32] G. Salton and C. S. Yang, ‘On the Specification of Term Values in Autoatic Indexing’, Department of Computer Science, Cornell University, Ithaca, New York, 14850, USA, TR 73-173, 1973.

[33] G. Salton, A. Wong, and C. S. Yang, ‘A vector space model for automatic indexing’, *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, 1975.

[34] S. E. Robertson, ‘The probability ranking principle in IR’, *Journal of documentation*, vol. 33, no. 4, pp. 294-304, 1977.

[35] S. E. Robertson and K. Spärck Jones, ‘Relevance weighting of search terms’, *Journal of the American Society for Information science*, vol. 27, no. 3, pp. 129-146, 1976.

[36] G. Salton and C. Buckley, ‘Term-weighting approaches in automatic text retrieval’, *Information processing & management*, vol. 24, no. 5, pp. 513-523, 1988.

[37] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, ‘Indexing by latent semantic analysis’, *Journal of the American society for information science*, vol. 41, no. 6, pp. 391-407, 1990.

[38] M. F. Porter, ‘An algorithm for suffix stripping’, *Program: electronic library and information systems*, vol. 14, no. 3, pp. 130-137, 1980.

[39] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*, illustrated ed. The MIT Press, 2005.

[40] N. Fuhr, ‘Optimum polynomial retrieval functions based on the probability ranking principle’, *ACM Transactions on Information Systems*, vol. 7, no. 3, pp. 183-204, 1989.

[41] N. Fuhr and C. Buckley, ‘A probabilistic learning approach for document indexing’, *ACM Transactions on Information Systems (TOIS)*, vol. 9, no. 3, pp. 223-248, 1991.

[42] W. S. Cooper, F. C. Gey, and D. P. Dabney, ‘Probabilistic retrieval based on staged logistic regression’, in *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, 1992, pp. 198-210.

[43] O. A. McBryan, ‘GENVL and WWWW: Tools for Taming the Web’, in *Proceedings of the First International World Wide Web Conference*, 1994.

[44] S. Brin and L. Page, ‘The anatomy of a large-scale hypertextual Web search engine’, *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107-117, 1998.

[45] J. M. Kleinberg, ‘Authoritative sources in a hyperlinked environment’, *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604-632, 1999.

[46] T. Joachims, ‘Optimizing search engines using clickthrough data’, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 133-142.

[47] D. Meister and D. J. Sullivan, ‘Evaluation of user reactions to a prototype on-line information retrieval system’, NASA, Prepared under Contract No. NASw-1369 by BUNKER-RAM0 CORPORATION Canoga Park, Calif. NASA CR-918, Oct. 1967.

[48] S. E. Robertson and M. M. Hancock-Beaulieu, ‘On the evaluation of IR systems’, *Information Processing & Management*, vol. 28, no. 4, pp. 457-466, 1992.

[49] S. Cucerzan and E. Brill, ‘Spelling correction as an iterative process that exploits the collective knowledge of web users’, in *Proceedings of EMNLP*, 2004, pp. 293-300.

[50] F. Radlinski and T. Joachims, ‘Query chains: learning to rank from implicit feedback’, in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 239-248.

[51] F. Peng, N. Ahmed, X. Li, and Y. Lu, ‘Context sensitive stemming for web search’, in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 639-646.

[52] J. Verhoeff, W. Goffman, and J. Belzer, ‘Inefficiency of the use of Boolean functions for information retrieval systems’, *Communications of the ACM*, vol. 4, no. 12, pp. 557-558, 1961.

[53] J. Carbonell and J. Goldstein, ‘The use of MMR, diversity-based reranking for reordering documents and producing summaries’, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 335-336.

[54] J. M. Ponte and W. B. Croft, ‘A language modeling approach to information retrieval’, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 275-281.

[55] D. Hiemstra, ‘A linguistically motivated probabilistic model of information retrieval’, *Research and Advanced Technology for Digital Libraries*, pp. 569-584, 1998.

[56] D. Metzler and W. B. Croft, ‘A Markov random field model for term dependencies’, in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 472-479.

[57] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, ‘Stuff I’ve seen: a system for personal information retrieval and re-use’, in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003, pp. 72-79.

[58] E. M. Voorhees, ‘The TREC question answering track’, *Natural Language Engineering*, vol. 7, no. 4, pp. 361-378, 2001.

[59] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[60] M. A. Hearst, *Search User Interfaces*, 1st ed. Cambridge University Press, 2009.

[61] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*, 1st ed. Addison Wesley, 2009.

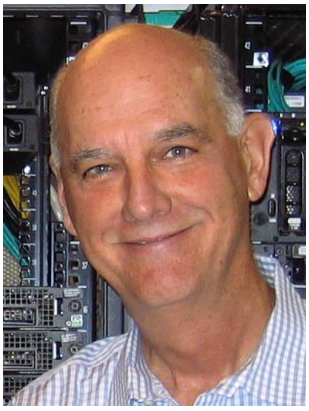
[62] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)*, 2nd ed. Addison-Wesley Professional, 2011.

About the authors

Mark Sanderson received a B.Sc. (Hons.) degree in computer science from the University of Glasgow, Glasgow, U.K. in 1988; and a Ph.D. degree in computer science also from the University of Glasgow in 1997.

From 1998 to 1999, he was a Post Doc at UMass Amherst. Then, he worked as a faculty member at the Information School in the University of Sheffield until 2010. He is now a Professor at the School of Computer Science and Information Technology at RMIT University, Melbourne, Australia.

Prof. Sanderson is associate editor of Information Processing and Management and ACM Transactions on the Web. He is Co-PC Chair of ACM SIGIR in 2012.



W. Bruce Croft received a B.Sc. (Hons) degree in 1973, and an M.Sc. in Computer Science in 1974 from Monash University in Melbourne, Australia. His Ph.D. in Computer Science was from the University of Cambridge, England in 1979.

In 1979 he joined the faculty of the Department of Computer Science at the University of Massachusetts, Amherst, USA, where he currently holds the position of Distinguished Professor.

Prof. Croft was a member of the National Research Council Computer Science and Telecommunications Board, 2000-2003, and Editor-in-Chief of ACM Transactions on Information Systems, 1995-2002. Prof. Croft was elected a Fellow of ACM in 1997, received the Research Award from the American Society for Information Science and Technology in 2000, and received the Gerard Salton Award from the ACM Special Interest Group in Information Retrieval (SIGIR) in 2003.