FEATURE ARTICLE

A Brief History of Search Results Ranking

Stephen Robertson
University College London

The theory and practice of search results ranking, as currently offered by most web search engines, is older than one might think. The first proposal for a system of ranking was in a JACM paper in 1960. Through the remainder of the twentieth century, extensive research was done on ranking systems – on devising methods of ranking, on the use of learning methods in association with ranking, and on the comparative evaluation of alternative methods. The web search engines of the tail end of the century made much use of this work.

It has become the norm, in web search particularly but also in other search tasks, for the system to generate a ranked list of results – this is well-established practice, and is central to the technology and business of the web search engines. It has become fertile ground for all sorts of methods under the general heading of machine learning. Ranking in the search context has a venerable history, but in contrast to some other areas of computer science, this history is little known outside specialist circles. Many computer scientists, for example, will be familiar with or at least aware of the contributions of Huffman to compression, or of Codd to databases, or of Dijkstra to algorithms; but comparable knowledge of search is unusual – despite the fact that a significant fraction of the world's population uses search engine technology on a daily basis.

The aim of this paper is to sketch some history, particularly from the last century, up to the rise to dominance of Google in web search. This is a personal view of this history, and the work specifically mentioned is a personal selection (including some of my own work); but I believe it to amount to a fair representaton of the development of ideas. As a secondary aim, I hope to debunk the myth of PageRank, which remains in widespread currency. But more of that anon. A good and more extensive history, with many more references, is provided by Sanderson and Croft[1].

TO 1960

The term *information retrieval* was first coined in the 1950s, although ideas concerning the storage, organisation and retrieval of information, generally in the context of library documents, had been around for much longer. Well before the use of computers, there were mechanical methods that allowed searches to be made for queries containing multiple concepts. Also in the 1950s, HP Luhn of IBM put forward some ideas about computer-based information retrieval.[2]

Luhn's ideas were associated with work by himself and others on automatic indexing and abstracting (see Edmundson and Wyllys' 1961 paper).[3] The primary aim of the automatic indexing work was to automatically extract from a document *index terms*, words or perhaps phrases that could be used in something like a printed index or a library card catalog. There was also the idea of autmatically generating text abstracts of papers. This work led to a number of ideas for assigning numerical weights to words, to indicate their importance in describing the topic or substance of a document. One source of information for such weights could be statistical data about the usage of words. These ideas were revisited in the context of ranking in the following decade.

Results ranking was proposed by Maron and Kuhns in 1960.[4] The abstract to this paper includes the following:

> *"The technique ... allows a computing machine, given a request for information, to make a statistical inference and derive a number ... for each document, which is a measure of the probability that the document will satisfy the given request. The result of the search is an ordered list of those documents ..."*

The basis for the Maron/Kuhns idea of a ranking function is a probabilistic argument, concerning the probability that a user, arriving with a specified request, would be satisfied with a particular document. Although this argument was largely forgotten for a decade or more, it subsequently became the basis for a substantive theory of ranking for retrieval. One of the ideas implicit in the Maron/Kuhns paper (briefly mentioned explicitly, but rejected as impractical) is the possibility of having the system learn, over a body of users and uses, what kinds of items different users might like to see.

However, despite the notion of a computing machine (and publication in JACM), much of the paper still revolves around the notion of a human librarian indexing a document – that is, the librarian provides a searchable description of each document in a standardised language. A user is supposed to present their query in terms of one of the concepts formally defined in the library system. There is no notion of a query containing multiple concepts or terms.

No practical system resulted from the Maron/Kuhns proposal. However, the field was just opening up for research and experimentation.

1960S

In the early 1960s, first at Harvard and then at Cornell, Salton began his long series of experiments with the Smart system, which did indeed generate ranked output.[5] Each document was indexed by, and therefore searchable on, the natural language words in the text (excluding some common words such as *the*, *of*, *and*). The model was a somewhat curious mathematical analogy, based on regarding documents and requests as occupying a high-dimensional vector space, but it embraced the possibility of multi-term queries, and led to many useful methods and ideas. Ranking was derived from a notion of distance in the vector space. It was natural in multi-term queries to assign weights to terms, to determine how strongly each term should influence the ranking; initially the weighting methods used were simple and ad-hoc, but see further below.

The basic ideas of experiments with user relevance judgements were established in the two Cranfield experiments, at the then Cranfield College of Aeronautics in England.[6] The principal idea of Cranfield-type experiments lay in the notion of relevance – a much-debated idea, generally assumed to require human judgement. That is, given a query and a set of documents, an information retrieval system should offer the user those documents that the user would regard as relevant to their information need (that is, what caused the user to issue the query). In the context of a ranking system, this would mean that the relevant documents should (as far as possible) occur early in the ranking. Such experiments are concerned only with the *effectiveness* of retrieval, not with such matters as search speed.

The first Cranfield experiment (ending 1962) had no conception of ranking, but the second (1962-66) used a very simple form of ranking (actually a partial order with many ties, but nevertheless rather more sophisticated than simple yes-no set retrieval). It is worth noting that this experiment was in no way associated with computers or computer science: the prime mover was Cyril Cleverdon, the Librarian at Cranfield, and the highest form of technology used was the card catalog (together with an electromechanical calculator the size of a large typewriter, which was used to calculate the statistical tables deriving from the experiment). Nevertheless, it had a profound effect on the research field of computational information retrieval.

The work of Salton and his collaborators, over approximately 35 years (in fact reaching the era of the web search engine) was central to the development of the field. Many ideas in the areas of automatic language processing and ranking have their origins in Salton's laboratory. Rocchio, a member of Salton's group, proposed a form of automatic feedback – allowing the system to improve its ranking on the basis of user relevance judgements. Actually two forms: one relating to an individual search, where feedback by a searcher during the course of the search enables a reranking of the items the searcher has not yet seen; and one which modifies the indexing for subsequent searchers.

Salton's system used no human indexing, only the text of documents. The Cranfield experiments included both free text and human indexing – but provided some evidence that searching systems based only on free text might be effective. There are still many systems in the world that make good use of human indexers, for certain kinds of material, but the current generation of web search engines relies entirely on free text.

1970S AND 80S

Among the many ideas investigated in this period, the idea of learning from explicit user feedback, in particular from user relevance judgements, continued to stimulate some of the ongoing research. Note, however, that the use of implicit feedback, for example learning from user actions such as clickthrough or dwell time, has not yet emerged. Learning methods were mainly home-grown within the information retrieval community, but some use was made of generic pattern-matching / machine learning ideas.

In 1972, Spärck Jones (Cambridge, UK) published a form of term weighting,[7] based on the frequency of the term in the collection, which proved to be a core technique in much subsequent ranking work – it was taken up quickly by Salton's group, and added to their portfolio of ranking ideas. In 1976, Spärck Jones and the present writer published a Rocchio-like feedback method, for an individual search session, but based on a well-founded probabilistic argument; the principle was to assign weights to query terms to reflect their relative discrimination power in identifying relevant documents.[8] In 1977, using some unpublished material by Cooper (University of California Berkeley), I published an analysis of the Probability Ranking Principle, a formalisation of the principle used in the 1960 paper.[9] In 1979, Croft and Harper (Cambridge) showed how the Robertson/Spärck Jones feedback method could be used in the absence of explicit feedback.[10] In 1982, Cooper and Maron and I analysed the relationship between the two forms of feedback, within and across uses of the system.[11]

In 1989, Fuhr (Dortmund, Germany)[12] made the first serious attempt at a system which learnt across users/uses. While Rocchio had used relevance judgements to modify the indexing of the documents, and Cooper, Maron and I had analysed the structure of the problem without proposing any particular methods, Fuhr's idea was to tune the ranking algorithm itself. Over the next few years there were some more attempts at such ideas (together with much more on individual-session relevance feedback), by Fuhr and Cooper and others. However, in those days we did not have enough data to make serious improvements in ranking effectiveness – that would have to wait on the web.

Through the 1980s, many groups worked on ranked-output systems; prominent among these were Salton's group and Croft's group at the University of Massachussets, Amherst. Other models for ranking emerged, including van Rijsbergen's (Dublin/Glasgow) notion of retrieval as probabilistic logical inference.[13]

This work built up to an exposion of interest in the early 90s. But before I explain this development, I should talk about the commercial world.

REAL USERS, COMMERCIAL SYSTEMS

Commercial search systems, drawing from the library and publishing worlds, had begun in the late 60s with offline searching, and were flourishing by the 80s, offering online searching. Mostly they used databases of scientific abstracts. The search function was essentially Boolean (with a few extensions to standard Boolean logic), resulting in the retrieval of undifferentiated sets of results. For example, I might specify a query in a database of computer science abstracts such as "history AND (search OR (information AND retrieval))". The system might then tell me that there are 3 items that match my query (or 0 or 1087 as the case may be). I could then display these records, but they would appear in no particular order. That is, these commercial search engines paid no attention to the idea of ranking. However, they did establish the idea that in order to service a user request quickly, with a large collection of documents, you have to have first constructed an index – an idea the early web search engines took for granted. An index, in this context, is essentially a look-up list of words – each word points to a list all the documents containing it. The art of index building was very well-developed by the end of the 80s.

The idea of ranked results did find its way into a few systems on the border between the research domain and the world of real users and real collections of documents. OKAPI was a system for searching a library catalogue, experimental but serving real users with real information needs, at the then Polytechnic of Central London in the early 80s (see Venner et al.[14]). It was one of the first systems to invite unstructured queries, and to offer ranked results to real library users. Using a text-only terminal, the user would type a query and see a list in pages of 9 items each, with a single digit keystroke to go to the full record for any item. There was also a single-session relevance feedback function, allowing a user to mark items of interest, following which the system would automatically modify the search to find further items ("more like the ones you have chosen"). Around the same time there were similar system: MUSCAT (see Porter[15]) was operating at the Scott Polar research institute in Cambridge, England; in the USA, CITE (see Doszkocs[16]) was developed for the National Library of Medicine database.

TREC

Returning to the research world, the big event of the early 90s was TREC, the Text Retrieval Conference (run by Harman at the National Institute of Standards and Technology, announced 1990, and continuing to this day).[17] This is a combination of cross-group collaboration and competition, intended to drive forward the state of the art of text retrieval – but very much in the mould of the Cranfield experiments 30 years earlier. That is, the main focus has been on search system effectiveness, in the sense of retrieving items that the user would consider relevant. Each year, a number of challenge tasks are offered; effectiveness metrics are defined, but the relevance judgements on which they are based are not released until after the participating teams have submitted their responses to each task. One of the main objects of competition in the early years of TREC was exactly the ranking engine. My group from City University of London (which now included the OKAPI team) was one of the groups taking part, and over the first three years I developed a ranking function, based again on a probabilistic model, called BM25 – which proved highly successful in the competition.

Other models and other ranking ideas came from other groups, particularly Croft's group – it was a hugely productive period of research. Another form of statistical model, called the Language Model, was put forward by Ponte and Croft[18], and independently by Hiemstra[19], also very successfully. The model has many variants, but essentially takes language (the texts of both documents and queries) to be the result of a statistical generation process involving the probabilistic selection of successive words.

EARLY WEB SEARCH

We are now getting into the beginnings of the web search engine. In 1993-5, several search engines were started: JumpStation, WebCrawler, Lycos, Yahoo! for example. Probably WebCrawler was the first to offer ranked search results.

There are two innovations which really stand out from the early search engines. The first is the crawl – you have to gather data for your index file by crawling the web. The second is related to the first: in the process of crawling, you gather information about a page not only from the page itself, but also from the page that pointed you to it. The existence of the link is important, but the anchor text associated with the link is even more important. You keep this information in your index: it tells you something about the page that the page itself may not. It will figure in your ranking algorithm. It's not clear which was the first search engine to make use of anchor text in this way, but it certainly started during this period.

In case readers are unfamiliar with the notion of anchor text, it is the text associated with a clickable link from one page to another – which commonly appears in blue on web pages. Thus if I have a web page which is a recipe for pumpkin pie, and someone else has a clickable link to it which appears like this: a good recipe for pumpkin pie, then as far as a web search engine is concerned, *my* page has been tagged with the word "good", even if I never used that word on the page itself. Although many links are completely useless in this regard (for example click here), there are enough useful links around to have a considerable effect on the ranking effectiveness of the web search engines.

The first good ranker was the Altavista one. This at least is the opinion of Batelle, exressed in *The Search* ten years later[20]; the present author agrees (at the time there was precious little proper evaluation of the ranking effectiveness of web search engines). Altavista began in 1995; its system was based in part on work done in TREC in general, and on a 1994 Technical Report by Spärck Jones and myself in particular.[21] When Google came along a couple of years later, there were quite a few other search engines around, also with good rankers. Google itself had good ranking, probably one of the best – and included a new feature, PageRank (more on this below). However, Google's rise to dominance over the next few years was due to a range of different factors. For one thing, it had a wonderfully clean interface (most of the others loaded their front pages with far too much stuff). Additional factors were the size and up-to-date nature of the crawl, and the speed of search. Altavista was already languishing, the result of a number of poor commercial decisions.

RANKING SINCE 1998

We have come quite a long way in the last twenty years. One major development in web search has been the use of a whole variety of techniques that now come under the heading of machine learning. Web search engines today learn from many kinds of data, but most particularly from use and users – from implicit cues, such as clickthrough and dwell time. That is, users do not click in order to provide the search engine with feedback – they click in order to visit a new page. But a search engine can use clickthrough as an indication of which retrieved items the user likes the look of. This is a kind of learning data that was simply not available to the research systems of the last century – nor is there much scope for this in the many inhouse search systems that exist within organisations. It requires the kind of volume of data that only web-scale operations can call upon. Search engines make use of learning methods in many ways, including the ranking algorithm, but also in areas such as spelling correction and identifying spam web pages. As well as using observed user behaviour, they create extensive data sets for off-line learning from, including large numbers of queries and relevance judgements made by human judges (not the actual users).

Nevertheless, the idea of results ranking, and the associated idea of learning to rank well, have been around for almost 60 years. It is very clear that the entire field of web search owes a great deal to prior research over an extended period. The early pioneers, including Luhn and Maron and Kuhns, deserve some recognition.

Postscript: PageRank

In the early days of Google, a great deal was made of the PageRank algorithm, due to Page and Brin. The idea was assiduously promoted that PageRank *was* the ranker, and it took hold thoroughly. In 2012, for example, McCormick published a book called *Nine Algorithms that Changed the Future*.[22] In chapters devoted to such matters as data compression, cryptography, databases, computability, he pays due homage to such pioneers as Lempel and Ziv, Diffie and Hellman, Codd, Turing. But the chapter on search engine results ranking mentions nothing at all prior to 1998, and barely mentions that the Google ranker uses anything other than PageRank. He has bought into the myth, hook line and sinker. Batelle's book, mentioned earlier, is slightly more nuanced – he mentions Salton, for example – but even he seems overawed by PageRank. It remains a common misconception that Google's major innovation, and the main reason for its success, was PageRank.

So how important was PageRank in the Google ranker? It was one of a large number of features, and contributed something to the overall effectiveness of the ranker. But in my view, it was much less important than doing all the other things well. And in particular, the advantage claimed for PageRank (that it quantifies the authority of a page as viewed by other web page authors) can already be obtained from matching the query against anchor text. Anchor text is a little less subtle than PageRank as a quantitative measure, but on the other hand it is query-specific, which PageRank is not. This fact has been recognised in the information retrieval research community, supported by some evaluation work by Hawking and colleagues, for example[23] -- work which was published nine years before McCormick's book.

Matching anchor text well is vital for a good web search engine; using PageRank is useful, but nothing more.

ACKNOWLEDGEMENTS

I am very grateful to Donna Harman and to two referees for excellent comments on versions of this paper.

REFERENCES

1. Sanderson, M. and Croft, B. "The History of Information Retrieval Research." *Proceedings of the IEEE* 100 (Special Centennial Issue), 2012, 1444-1451.
2. Luhn, H.P. "A statistical approach to mechanized encoding and searching of literary information." *IBM Journal of Research and Development* 1 (1957), 309-317.
3. Edmundson, H.P. and Wyllis, R.E. "Automatic Abstracting and Indexing Survey and Recommendations." *CACM* 4 (1961), 226-234.
4. Maron, M.E. and Kuhns, J.L. "On relevance, probabilistic indexing and information retrieval." *JACM* 7 (1960), 216-244.
5. Salton, G. *Automatic Information Organization and Retrieval*, McGraw-Hill, 1968.
6. Cleverdon, C. "The Cranfield tests on index language devices." *Aslib Proc* 19 (1967).
7. Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1972), 11-21.
8. Robertson, S.E. and Spärck Jones, K. "Relevance weighting of search terms." *Journal of the American Society for Information Science* 27 (1976), 129-46.
9. Robertson, S.E. "The probability ranking principle in IR." *Journal of Documentation* 33 (1977), 294-304.
10. Croft, W. and Harper, D. "Using probabilistic models of information retrieval without relevance information." *Journal of Documentation* 35 (1979), 285-295.
11. Robertson, S.E., Maron, M.E. and Cooper, W.S. "Probability of relevance: a unification of two competing models for information retrieval." *Information Technology - Research and Development* 1 (1982), 1-21.

12. Fuhr, N. “Optimum polynomial retrieval functions based on the probability ranking principle”, *ACM Transactions on Information Systems* 7 (1989), 183-204
13. Van Rijsbergen, C.J. “A non-classical logic for information retrieval”, *The Computer Journal* 29 (1986), 481-485.
14. Venner G., Walker S. and Mitev N.N. “Okapi: a prototype online catalogue.” *Vine* 15 (2) (1985), 3-13.
15. Porter, M. "Relevance feedback in a public access catalogue for a research library: Muscat at the Scott Polar Research Institute." *Program* 22 (1988), 1-20.
16. Doszkocs, T. “From Research to Application: The Cite Natural Language Information System.” In: *Research and Development in Information Retrieval, Proceedings, Berlin, 1982*, Springer, 1983, 251-62.
17. Voorhees E.M. and Harman D.K. (eds.), *TREC: Experiments and Evaluation in Information Retrieval,* The MIT Press, 2005.
18. Ponte J.M. and Croft W.B. “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, 275-281.
19. Hiemstra D. “A linguistically motivated probabilistic model of information retrieval.”, *Research and Advanced Technology for Digital Libraries*, 1998, 569-584.
20. Batelle, J. *The Search*, Nicholas Brealey, 2006.
21. Robertson, S.E. and Spärck Jones, K. *Simple, proven approaches to text retrieval*, U Cambridge Comp Lab Tech Report no. 356 (1994, updated 1996, 1997, 2006).
22. McCormick, J. *Nine Algorithms that Changed the Future*, Princeton, 2012.
23. Upstill, T., Craswell, N. and Hawking, D. “Predicting Fame and Fortune: PageRank or Indegree?” *Proceedings of the 8th Australasian Document Computing Symposium, Canberra, Australia*, 2003, 31-40.

ABOUT THE AUTHOR

Stephen Robertson is now retired from employment, but remains a visiting professor at University College London, Professor Emeritus at City University London, and a Fellow of Girton College Cambridge. He worked for some years at Microsoft Research Cambridge. He has been researching and publishing in the field of information retrieval for about 50 years; his 1975 PhD was from University College London. He has received the Strix Award and the Salton award. Contact him at stepheneroberts on@hotmail.co.uk