Document retrieval: shallow data, deep theories; historical reflections, potential directions

Karen Spärck Jones
Computer Laboratory, University of Cambridge
William Gates Building, JJ Thomson Avenue, Cambridge CB3 0FD, UK
sparckjones@cl.cam.ac.uk

Advances in information retrieval, 25th European Conference on IR Research, ECIR 2003
Ed. F. Sebastiani, LNCS 2633, Berlin: Springer, 2003, 1-11

Abstract

This paper reviews the development of statistically-based retrieval. Since the 1950s statistical techniques have clearly demonstrated their practical worth and statistical theories their staying power, for document or text retrieval. In the last decade the TREC programme, and the Web, have offered new retrieval challenges to which these methods have successfully risen. They are now one element in the much wider and very productive spread of statistical methods to all areas of information and language processing, in which innovative approaches to modelling their data and tasks are being applied.

Introduction

Two ideas have played a crucial role in automated information retrieval. They are not in themselves computational ideas, but computers were necessary to make them work. Specifically, computers made them so much easier to apply that the quantitative changes in information management that followed automation have become qualitative ones.

Many things about information and searching for it are quite timeless (though they may not be recognised as such in shiny current computing contexts). But the two very simple ideas that pervade modern retrieval systems have effected an information revolution. One of these ideas is taking words as they stand. The other is counting their stances. It is not merely sufficient, it is necessary for document or text retrieval to respect the actual words that people use. It is thus also essential for effective retrieval to respond to the ways words are distributed in documents: the relative frequencies with which words occur and cooccur mark topics and their significance in texts.

Stated thus, these ideas may appear banal, and just what *any* approach to indexing and searching requires. But the latter has not historically been the case, and working out the consequences of these simple ideas has been the distinguishing feature of automated retrieval since research on it began fifty years ago. These two ideas, obvious though they now seem, were important novelties in the context of traditional library classification in the 1950s. The innovative retrieval work associated with them that began in the 50s and grew in the 60s provided the base for current Web search technology, and the wholly new information world that many take this to represent. In this paper I will look at the way these ideas have evolved, and try to identify critical points in this evolution both in the past and, now, for the future.

Innovative ideas

The stimulus for new thinking about indexing and searching in the 50s came from the growth of, and increasing specialisation in, the scientific literature. Traditional subject classification schemes were too general, too rigid, and too prescriptive to support effective retrieval from collections of scientific papers. Indexing had to be more fine-grained, more flexible, and more reflective of the documents themselves.

The researchers of the 50s responded to these needs with a whole range of notions and tools. These included: using given text words and phrases, however specialised, for indexing, as opposed to independent category labels (Taube et al. 1952); defining topics at search time by postcoordination, rather than at file time by precoordination (also Taube et al.); choosing indexing vocabularies to reflect file topic distributions, not external structures (Mooers 1951); exploiting numeric data to refine topic characterisation (Luhn 1957); and developing formal models for retrieval based on statistics and probability (Maron and Kuhns 1960). The important point about all of these proposals is that they lent themselves to automation even, already, in the pre-computer form of punched card manipulation, as with Taube and Mooers, but more strikingly, using early computers, in Luhn's work (see Schultz 1968).

Much of this early work was concerned, in one way or another, with classification. The presumption was that indexing and retrieval depend on classification, whether of what there is 'out there' or of what is said about this. Though the world of information was changing with a growing and richer literature, it was still a world of information; and since manipulating information depends on classification, it seemed that what was called for was new views, and methods, of classification. The novel ideas about retrieval just mentioned were therefore seen as grist for classification mills.

The UK-based Classification Research Group (1969) was especially active in investigating ways of shaking up traditional approaches to information description and classification so as to make indexing more responsive to document realities, especially subject specialisation and rapid scientific and technological change; but their presumption was still that constructing classifications and indexing with them would be manual rather than automatic. The more radical research on classification that flourished on both sides of the Atlantic in the late 50s and early 60s focussed on automatic methods of classification, and hence of indexing. This research embraced the new ideas about derivative indexing, grounded in the texts of the documents themselves, in a much more wholehearted way, since classification was based on bottom-level distributional data about words.

The work done at the Cambridge Language Research Unit in the UK illustrates this development very well (Needham and Spärck Jones 1964). It sought to apply general, formal classification techniques to text vocabularies, characterised by their document occurrences, in order to obtain thesaurus classes of words with similar document distributions. The members of a class could be substituted for one another in searching to obtain query-document topic matches, while the classes themselves could be freely combined in postcoordinate fashion.

This period of research on statistically-based indexing and classification was a very exciting time, as Stevens 1965 and Stevens et al. 1965 show. The work both in this specific area and more generally in novel approaches to indexing and classification was thoroughly international, with many European contributors, for example to the Elsinore Classification Research Conference (Atherton 1965).

This was also the period when Salton's research at Harvard and then in the thirty-years of SMART work at Cornell began (Salton 1968, 1971). This too started by seeking to automate then-conventional forms of indexing and searching but, particularly through working with abstract texts, gradually moved to placing more emphasis on non-conventional techniques like term weighting. Bely et al. (1970)'s very sophisticated work on automatic controlled-language indexing using SYNTOL also indirectly encouraged simpler

techniques, since the elaborate indexing devices they used were not especially effective for retrieval. This work is more properly seen as a pioneering project on information *extraction*.

In all of the statistically-based studies, automation was critical in making it possible both to deal with volumes of data (though these were trivial by modern standards), and to apply procedures consistently and objectively: manipulating numerical data about text words is something computers are much better at than humans.

Even more importantly, automation made it possible to do very large numbers of systematic and controlled experiments to evaluate methods of indexing and searching. The Cranfield projects (Cleverdon 1967, Spärck Jones 1981) were pioneering efforts in system evaluation and, as Cleverdon noted, while essentially manual, sought to simulate machine objectivity. But even though the number of studies, e.g. of different indexing devices, was very impressive, it was evident from the subsequent SMART work that with automatic indexing and searching, far more extensive tests across different environment variables or system parameters could be done.

The work on relevance feedback done both at Cornell from the 60s and in Cambridge a little later (Salton 1971, Spärck Jones 1980) illustrates the gain from automation, both for retrieval itself and for research on retrieval. Of course human searching involves the use of relevance information; but the specific feedback methods, just like those for term weighting (or classification), are wholly unsuited to human implementation and wholly suited to machine application. At the same time, the enormous volume of experiments with different weighting strategies and formulae since the late 60s could never have been carried out without automated test rigs.

Though automation for cataloguing and catalogue searching, as done by the Library of Congress and OCLC, and for systems for the journal literature, as exemplified by Medlars or Inspec, appeared in the later 60s, this did not immediately lead to radical innovations in the nature of indexing. The scientific literature systems continued with manual indexing and subject-based description, though this was now done with specialised thesauri and controlled languages rather than the older universal classification schemes. The first move towards more modern approaches appeared with the ability to search eg actual title words in MEDLINE, rather than only descriptor fields, though with Boolean not ranking searches. In these operational systems the advantage of automation came primarily from the 'administrative' convenience it provided through access to large databases, rapid searching, and so forth.

The Science Citation Index, on the other hand, showed automation is a more original light, since citation indexing constituted a new type of indexing and could never have been done on a large scale without computers. Moreover, though the indexing did not exploit frequency data in the style adopted for term weighting, it did show the importance of quantitative data referring to documents.

Development and consolidation

Research in the 1970s and 80s continued to develop and test the new approaches of the 60s to indexing and searching using text words and statistical data. This work did not make much impact on the large-scale bibliographic search services, which very reasonably concentrated on other issues of user importance, like rapid document delivery. Moreover even where full-text files were in question, as with legal systems, queries remained in the Boolean mode. But at Cornell and e.g. Cambridge and City University in the UK, work on both theory and practice for statistically-based methods continued. This was successful in further developing appropriate formal models for retrieval, as in Salton 1975, van Rijsbergen 1979, and Robertson and Spärck Jones 1976, and in extending the range of experiments that confirmed the value of these approaches (e.g. Salton and Buckley

1988). These tests included not only ones to compare variations on statistical methods, but ones that suggested that they could compete successfully with conventional controlled indexing (Salton 1986). Research in this area was indeed now extended to enrich at least some conventional bibliographic services, as in Biebricher et al. 1988.

Moreover, though the 70s were the heyday of AI claims for the superior merit of symbolic approaches to information and language processing, as Salton pointed out (Salton 1995), whatever AI might offer other tasks, it had never been demonstrated superior to statistical approaches for the general retrieval case. Similarly, though continued attempts were made to show that 'proper' language processing, i.e. syntactic and possibly also semantic parsing, was required for better retrieval performance, this was not supported by the test results. Rather, insofar as compound index terms, as opposed to single words, were of use, so-called statistical phrases defined by repeated word tuples were just as effective as ones obtained by explicit parsing (Salton et al. 1990, Croft et al. 1991). This is the analogue, for complex concepts, of frequency as an indicator of concept significance for simple terms. Even where some explicit language analysis was involved, as in stemming, this could be much simpler than the full-scale lemmatisation needed for other information processing tasks (Porter 1980).

The one area that remained surprisingly intractable was full-scale automatic classification, whether of terms or documents. Quite apart from the problems of identifying appropriate class definitions and viable classification algorithms, straightforward attempts to group terms to obtain a thesaurus, or to cluster documents to focus searching, could not be shown to deliver significant general improvements in retrieval effectiveness. Thus document grouping enhanced precision but with serious damage to recall (Croft 1980). Similarly, it had earlier appeared that term classes were only of any value at all when they were confined to very strongly related terms and were applied to promote extra, rather than substitute, term matches (Spärck Jones and Barber 1971).

It instead became more clear, during the 80s, that in classification as in other aspects of indexing and searching, the desired effects could be achieved by indirect rather than direct means. Thus the aim of classification, whether of terms or documents, was to bring objects with similar distributional behaviour together since the groupings obtained would, when tied to query terms, be correlated with document relevance. The whole effect of relevance feedback, especially when used to expand rather than just reweight index descriptions, is to pick up classes of terms that are motivated by shared relevant document distributions. It is true that in using known relevant documents for feedback a system has more pertinent information to exploit than in the original classification case; but the performance gains that have been consistently demonstrated in the 90s for so-called blind relevance feedback, where documents are only assumed, not known, to be relevant, illustrate a form of indirect indexing, albeit one more focussed than the earlier ones.

From one point of view, the 1980s marked time. The mass of experiments done showed that the initial ideas about the value of statistical facts for retrieval had justified staying power, but had barely affected operational systems apart from some relatively tentative initiatives (Doszkocs 1983). This was partly for the same reasons as before, namely that operational services had many other goals than just ratcheting up precision and recall, but also because, though research experiments became bigger, the service databases grew very rapidly and it was far from obvious that the research methods or results would scale up.

On the other hand, better tools for other applications, like natural language processing for text editing and database query interpretation, made it possible to conduct more far-reaching tests of language processing for retrieval, as in phrasal indexing (Salton et al. 1990), even if the results were negative.

But the 80s were significant from a rather different point of view for retrieval. This was the period when the computing community in general concentrated on user interaction and

the form of human/computer interfaces, and when established literature services began the shift from professional intermediary to end-user searching. This naturally stimulated the retrieval community to address the implications for the user's search skills, or rather lack of them. But while this most obviously led to proposals for expert system interfaces (Belkin et al. 1983), it also provided a rationale for search devices which minimise user effort while maximising the payoff from information the user is uniquely qualified to provide, as in relevance feedback exploiting statistical data.

More generally, it is clear that for retrieval, the 80s were the period before a major earthquake. The underlying plates were moving and changing shape. End-user computing was growing and taking a different form; computing power was rapidly increasing; the internet was giving remote access to files and processes a quite new convenience and utility; machine-readable full text was coming on stream; related areas like natural language processing were moving to corpus-based data extraction both for resources like lexicons and in tasks like message interpretation; AI was recognising the legitimacy of statistical approaches to knowledge characterisation and capture, and developing machine learning techniques. Thus while the retrieval innovations of the previous decades were being consolidated, the larger information world was being remade. The question is thus how these innovations have fared in this new world.

New situations

The innovations of the 60s and 70s have in fact fared very well. The 1990s have been payoff time for statistically-based retrieval. Given the underlying shifts in the context it only needed the two earthquake triggers supplied by the Text REtrieval Conferences (TRECs) and the Web to bring the retrieval strategies previously confined to the research laboratory out onto the operational stage.

The design, scale and range of the TREC effort on retrieval evaluation have made what can be done with the text-derived and statistics-driven work of previous research quite clear; and the Web has provided new applications to exploit this. Early Web search engines were not tied down by the prior commitments and presuppositions of the bibliographic services. Their builders were open to ideas from computing research, so statistical techniques were applied in system design, for example in AltaVista, and they have a key role, albeit in a different form, in Google.

The TREC evaluations themselves, over more than a decade, tested indexing and searching devices far more thoroughly than ever before (TREC-2, 1995, TREC-6, 2000, Voorhees and Harman, in press). With world-wide participation, and very significant contributions from Europe, they have also brought multilingual operations into the hitherto English-centred evaluation world. As importantly, their data and findings have stimulated extensive further work, both along existing lines (e.g. Spärck Jones et al. 2000), and in newer ones (e.g. Dumais et al. 2002).

In 'mainstream' retrieval, TREC has confirmed earlier beliefs and findings on the value of the 'basic' statistical strategies, albeit with some development in scaling up to very large files. TREC has continued to cast doubt on the added value, for ad hoc topic searching, of structured classifications and thesauri or (to use the currently fashionable term) ontologies, whether manually or automatically constructed, and on the value of sophisticated natural language processing for retrieval. More importantly, in the TREC Web track experiments, where the older research ideas have been applied to far more challenging and timely data than before, these methods have maintained their standing. These experiments have shown that hyperlink information, the Web's real indexing novelty, does not imply better performance, for topic searching, than ordinary content terms, though it is helpful for the more specific task of finding homepages (Hawking and Craswell 2002). The TREC tests as a whole also show that statistical methods, especially when enhanced by

simple feedback, can bootstrap respectable performance from a poor initial request. This matters because the user's contribution and effort are important for effective retrieval, but cannot be guaranteed present. It is also worth noting that TREC confirmed early on that statistical retrieval strategies developed for English applied elsewhere, for instance to Chinese (Wilkinson 1998)

The TREC experiments have served to endorse not only the computational technologies applied, but also the IR theories underlying them. In general, the established statistically-based approaches - the Vector Space Model, the Probabilistic Model, the Inference Model - have performed well, and much alike, not surprisingly since they tend to use the same facts about about terms and documents in similar ways. The older Boolean Model has barely figured in TREC, and tests with a Non-Classical Logic Model have so far been limited. The most interesting recent development has been the introduction, or rather import, of a new model, the so-called Language Model. Language Modelling has performed well in TREC, and has stimulated new debate on appropriate models for retrieval (Croft and Lafferty, in press).

This model, like others, is a probabilistic one. Initially developed and established as highly effective for speech recognition (Young and Chase 1998), it has been applied in appropriate forms to a range of language and information processing tasks including translation and summarising, as well as retrieval (Brown et al. 1992, Banko et al. 2000, Knight and Marcu 2000, Berger and Lafferty 1999, Miller et al. 1999, Ponte and Croft 1998, Hiemstra 2002).

If we take all retrieval models as characterising the relation between a query and a (relevant) document, we can relate the Language Model to the other previous statistically-based models as follows. The Vector Space Model treats this relation as an object proximity relationship (Salton et al. 1975, Salton 1975). The Inference Model views the query document relation as a connectivity one (Turtle and Croft 1990). The Non-Classical Logic Model takes the query document relationship as a proof one, with the document proving the query (van Rijsbergen 1986). The Probabilistic Model has a generative relation from a query (along with relevance) to a document (Robertson et al. 1981, see also Fuhr 1989, Kwok 1995).

In the Language Model there is also a generative relationship, but the other way round, from the document to the query, i.e. the query is thought of as derived from the document (and relevance), in the same sort of way that in speech the heard sounds are generated from a word string. The other tasks to which the LM has been applied are given an analogous generative or derivational characterisation, though is not always very intuitive, so in translation the *source* text is seen as generated by the desired target text, and in summarising the full document is seen as generated by the desired summary text. In all of these applications of Language Modelling the task process is one of recovering the unknown original, given a more or less corrupted or defective received version; the system learns to do this by extensive training on prior instances of pairs, e.g. in the summarising case of full texts and their human abstracts, in the retrieval case of relevant documents and queries.

Language Modelling, like Vector Space Modelling for instance, is a quite general, abstract approach to information characterisation and processing, with many potential task applications. Whether it is superior to others as a theoretical foundation for retrieval, and if so in what way, is still a matter for argument (Croft and Lafferty, in press). But as technology it has shown its power, in both speech and other cases, to exploit large masses of training material very effectively, and to allow good probability estimation. It has also, in TREC, performed very well so far, as well as though not consistently and significantly better than, the best of the other models. Thus what its overall contribution to retrieval systems will be is not yet clear.

What the recent work on Language Modelling has emphasised, however, is first, the value of the very large training data sets that are now available for system development

and customisation. The work on machine learning, text mining, and the like which has been done in the last decade has shown how valuable such large data resources are in building description (and hence discrimination) systems, whether as data extraction feature sets, categorisation rules, or grammars. Further, it appears that very heterogenous data, hitherto thought of as a source of confusion rather than clarification, can be readily digested, to very nutritious effect, by such learning systems. Though document collections of familiar kinds are more varied than is often recognised, the sheer variety of the Web has been seen as a challenge rather than opportunity for indexing and search techniques. In fact, as operational systems such as Autonomy's suggest, the range of material in a large file can promote, rather than undermine, the characterisation of topics and concerns for effective information management.

The recent work on Language Modelling has also, by being applied to multiple tasks, reinforced the other important development that TREC has helped to foster, for example through its filtering and question-answering tracks, namely making progress with multi-task systems. While those engaged with document retrieval in earlier decades recognised that documents might be sought for a variety of purposes, and information services might accommodate multiple tasks so that, for example, it might be possible to request a document be translated, advances in information and language processing are now stimulating genuinely integrated, and hence truly flexible, multi-task systems. These can take advantage on the one hand of of common formal models and computational techniques, as illustrated by Language Modelling, applying them to different component tasks. But they can also take advantage, on the other hand, of techniques and resources developed for particular purposes and simply incorporate them in larger systems. This is manifest in the wide application of stemmers, part of speech taggers, named entity recognisers and the like.

The Mitre MiTAP illustrates this most recent progress towards multi-task systems very well (Damianos et al. 2002). It exploits a range of devices, and components, developed across the whole information and language processing field, and supports a range of tasks including retrieval, translation and summarising, within the framework of a single convenient interface, in a substantial, fully operational system. It is not, of course, just a statistically-based system: it incorporates parsing, for example. But it makes use of statistical as well as symbolic methods, most obviously in its retrieval sub-component, but also, and more interestingly, elsewhere. Thus as one small example of the way that text-statistic notions dating back to the 1960s have found a modern home, the MiTAP summariser uses *tf*idf*-type weighting.

References

Atherton, P. Ed.) *Classification research*, Copenhagen: Munksgaard, 1965.

Banko, M., Mittal, V. and Witbrock, M. 'Headline generation based on statistical translation', *ACL 2000: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, 318-325.

Belkin, N.J., Seeger, T. and Wersig, G. 'Distributed expert problem treatment as a model for information system analysis and design', *Journal of Information Science*, 5, 1983, 153-167.

Bely, N. et al. *Procédures d'analyse sémantique appliquées a la documentation scientifique*, Paris: Gauthier-Villars, 1970.

Berger, A. and Lafferty, J. 'Information retrieval as statistical translation', *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, 222-229.

Biebricher, B. et al. 'The automatic indexing system AIR/PHYS - from research to application', *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1988, 333-342.

Brown, P.F. et al. 'Class-based n-gram models of natural language', *Computational Linguistics*, 18, 1992, 467-680.

Classification Research Group, *Classification and information control: papers representing the work of the CRG from 1960 to 1968*, London: The Library Asssociation, 1969.

Cleverdon, C.W. 'The Cranfield tests on index language devices', *Aslib Proceedings*, 12, 1967, 173-193.

Croft, W.B. 'A model of cluster searching based on classification', *Information systems*, 5, 1980, 189-195.

Croft, W.B. and Lafferty, J. Eds.) *Language modelling for information retrieval*, Dordrecht: Kluwer, in press.

Croft, W.B., Turtle, H.R and Lewis, D.D. 'The use of phrases and structured queries in information retrieval', *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1991, 32-45.

Damianos, L. et al. 'MiTAP for biosecurity: a case study', *The AI Magazine*, 23 (4), 2002, 13-29.

Doszkocs, T.E. 'CITENLM: natural language searching in an online catalogue', *Information Technology and Libraries*, 2, 1983, 364-380.

Dumais, S., et al. 'Web question answering: is more always better?', *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, 291-298.

Fuhr, N. 'Models for retrieval with probabilistic indexing', *Information Processing and Management*, 25, 1989, 55-72.

Hawking, D. and Craswell, N. 'Overview of the TREC-2001 Web track', *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, Ed. E.M. Voorhees and D.K. Harman, Special Publication 500-250, Gaithersburg, MD: National Institute for Standards and Technology, 2002.

Hiemstra, D. 'Term-specific smoothing for the language modelling approach to information retrieval: the importance of a query term', *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, 35-41.

Knight, K. and Marcu, D. 'Summarisation beyond sentence extraction: a probabilistic approach to sentence compression', *Artificial Intelligence*, 139, 2002, 91-107.

Kwok, K.L. 'A network approach to probabilistic information retrieval', *ACM Transactions on Information Systems*, 13, 1995, 324-353.

Luhn, H.P. 'A statistical approach to mechanised encoding and searching of literary information', *IBM Journal of Research and Development*, 1, 1957, 309-317.

Maron, M.E. and Kuhns, J.L. 'On relevance, probabilistic indexing and information retrieval', *Journal of the ACM*, 7, 1960, 216-244.

Miller, D.R.H. Leek, T. and Schwartz, R.M. 'A hidden Markov model retrieval system', *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, 214-221.

Mooers, C.N. 'Zatocoding applied to mechanical organisation of knowledge', *American Documentation*, 2, 1951, 20-32.

Needham, R.M. and Spärck Jones. K. 'Keywords and clumps', *Journal of Documentation*, 20, 1964, 5-15.

J.M. Ponte and W.B. Croft, 'A language modelling approach to information retrieval', *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, 275-281.

Porter, M.F. 'An algorithm for suffix stripping', *Program*, 14, 1980, 130-137.

Rijsbergen, C.J. van *Information retrieval*, 2nd ed., London: Butterworths, 1979.

Rijsbergen, C.J. van 'A non-classical logic for information retrieval', *The Computer Journal*, 29, 1986, 481-485.

Robertson, S.E., van Rijsbergen, C.J. and Porter, M.F. 'Probabilistic models of indexing and searching', in *Information retrieval research*, (Ed. R.N. Oddy et al.), London:Butterworths, 1981.

Robertson, S.E. and Spärck Jones, K. 'Relevance weighting of search terms', *Journal of the American Society for Information Science*, 27, 1976, 129-146.

Salton, G. *Automatic information organisation and retrieval*, New York: McGraw-Hill, 1968.

Salton, G. Ed.) *The SMART retrieval system: experiments in automatic document processing*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

Salton, G. *A theory of indexing*, Philadelphia: Society for Industrial and Applied Mathematics, 1975.

Salton, G. 'Another look at automatic text-retrieval systems', *Communications of the ACM*, 29, 1986, 648-656.

Salton, G. remarks at the meeting on 30 Years of Information Retrieval at Cornell: A SMART Celebration, Department of Computer Science, Cornell University, April 1995 (video record).

Salton, G. and Buckley, C. 'Term weighting approaches to automatic information retrieval', *Information Processing and management*, 24, 1988, 269-280.

Salton, G., Buckley, C. and Smith, M. 'On the application of syntactic methodologies in automatic text analysis', *Information Processing and Management*, 26, 1990, 73-92.

Salton, G., Wong, A. and Yang, C.S. 'A vector space model for automatic indexing', *Communications of the ACM*, 18, 1975, 613-620.

Schultz, C.K. (Ed.), *H.P. Luhn : Pioneer of information science*, New York: Spartan, 1968.

Spärck Jones, K. 'The Cranfield tests', in *Information retrieval experiment*, (Ed. K. Spärck Jones), 1981.

Spärck Jones, K. 'Search term relevance weighting - some recent results', *Journal of Information Science*, 1, 1980, 325-332.

Spärck Jones, K. and Barber, E.O. 'What makes an automatic keyword classification effective?', *Journal of the American Society for Information Science*, 22, 1971, 66-75.

Spärck Jones, K., Walker, S. and Robertson, S.E. 'A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2', *Information Processing and Management*, 36, 2000, 779-840.

Stevens, M.E. *Automatic indexing: a state-of-the-art report*, Monograph 91, Washington, DC: National Bureau of Standards, 1965.

Stevens, M.E., Guiliano, V.E. and Heilprin, L.B. *Statistical association methods for mechanised documentation*, Symposium Proceedings 1964, Miscellaneous Publication 269, National Bureau of Standards, Washington DC., 1965.

Taube, M., Gull, C.D. and Wachtel, I.S. 'Unit terms in coordinate indexing' *American Documentation*, 3, 1952, 213-.

TREC-2, Special Issue on the Second Text REtrieval Conference (TREC-2), *Information Processing and Management*, 31, 1995, 269-448.

TREC-6, Special Issue on the Sixth Text REtrieval Conference (TREC-6), *Information Processing and Management*, 36, 2000, 1-204.

Turtle, H.R. and Croft, W.B. 'Inference networks for document retrieval', *Proceedings of the13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1990, 1-24.

Voorhees, E.M. and Harman, D.K. Eds.) *TREC: Experiment and evaluation in information retrieval*, Cambridge, MA: MIT Press, in press.

Wilkinson, R. 'Chinese document retrieval at TREC-6', *The Sixth Text REtrieval Conference*, (Ed. E.M. Voorhees and D.K. Harman), Special Publication 200-240, National Institute of Standards and Technologoy, Gaithersburg, MD, 1998, 25-30.

Young, S.J. and Chase, L.L. 'Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes', *Computer Speech and Language*, 12, 1998, 263-279.