Statistics and retrieval: past and future

Karen Spärck Jones

Computer Laboratory, University of Cambridge
JJ Thomson Avenue, Cambridge CB3 0FD, UK

sparckjones@cl.cam.ac.uk

This paper in its final form will appear in *International Conference in Computing: Theory and Applications (Platinum Jubilee Conference of the Indian Statistical Institute)*, Kolkata, IEEE, 2007.

Abstract

Statistical approaches to document indexing and retrieval date back to the beginning of automation. This paper considers early ideas, how they developed, their status now, and the challenges to be tackled in the future.

1 Introduction

The Indian Statistical Institute was established well before the modern general-purpose digital computer. But computing has had a radical effect on statistics, not just because it makes it possible to analyse vast masses of data and model underlying processes far more effectively than before. Statistics and probability have invaded areas previously thought to be beyond their reach. In this paper I will look at the connection between computing and statistics as this has been developed in one particular area, namely document retrieval. This has been significant in its own right, but is also connected with progress on other tasks in natural language information processing (NLIP). Language is human beings' main way of communicating information, i.e. of meaningful expression. It may seem surprising that data about the distributional properties of language units like words has much to do with meaning. But work on document indexing and searching has shown that such statistical data can be put to work to meaningful ends, and has encouraged similar developments in other, related NLIP tasks like summarising and question answering, as well as in natural language processing generally.

The idea of automated document retrieval goes back to the beginning of modern computing, as illustrated by Bush's Memex (1945). Using statistics in retrieval goes back almost, if not quite so far, to the early 1950s. Some crucial ideas were first adumbrated then by H.P. Luhn (see Schultz 1968, p.84). In the half century since then these initial ideas have undergone very thorough development under the combined impact of clarification and refinement, underlying theory formulation, and extensive empirical testing. The key elements of these

ideas are now, moreover, entrenched in the most powerful and extensive modern information retrieval systems, namely the Web engines.

In this paper I will first review this historical development, showing how conceptual refinement, theoretical motivation, and practical experiment have interacted, and continue to spread in new directions. My account will be deliberately informal, and also notationally simplistic. I am not a mathematician, and owe everything when it comes to formal modelling to my colleagues, especially Stephen Robertson. The references are also indicative rather than comprehensive. I will then indicate the challenges that research on statistically-based retrieval has to address in the future.

2 Beginnings, to the 1970s

2.1 Key initial ideas

Luhn presented his essential ideas about document indexing and retrieval in Luhn (1957a, 1957b). Noting that

The problem of literature searching is to find those documents within a collection that have a bearing on a given topic. (1957a)

and claiming that

The ultimate benefits of mechanisation will be realised only if the characteristics of machines are better understood and systems are developed that exploit these characteristics to the fullest. (1957a)

he proceeded from arguing that as

Communication of ideas by way of words is carried out on the basis of statistical probability. (1957b)

and that

There is also the probability that the more frequently a notion and combination of notions occurs, the more importance the author attaches to them ... (1957a)

to propose the construction of an indexing vocabulary consisting of more frequent but not necessarily very common terms and, with index sets for extended text queries and document unlikely to coincide exactly, to envisage matching as

... carried out on a statistical basis by asking for a given degree of similarity. (1957a)

However, though these quotations have a modern feel about them they had, as Luhn treated them, important limitations. Assume, for the moment, that the ideas are applied to single words as index terms. The first limitation is that frequency is used only as a criterion for selecting an indexing vocabulary which is then applied in a presence/absence manner to documents. The second is that the notion of frequency is not sufficiently decomposed. Luhn

was primarily thinking about word frequency within individual documents, i.e. modern *tf*, and he seems to have assumed that this would be well correlated with occurrence in most documents, i.e. in modern terminology, with a high *df*.

The picture is also complicated by the fact that Luhn recognised that words form notional classes, as in a thesaurus, which would be identified by human experts exploiting machine-generated concordances; and that he viewed the indexing vocabulary would thus be a 'descriptor' vocabulary, ideally with the equal discriminating power. Luhn also talked about indexing having more than one dimension (using "dimension" in an informal way), so words (or classes) would be associated with other words, perhaps within different discourse unit sizes. If the simple vocabulary gave one-dimensional indexing, then word pairs defined by tight proximitiy within a sentence would give (a particular form of) two-dimensional indexing.

Much of this was proposal rather than implementation for retrieval. However Luhn saw it as just one among a range of NLIP possibilities, including both support tool provision and task implementation, as listed in Luhn (1959). These included making KWIC indexes on the one hand, and automatic abstracting on the other, the latter based on extracting sentences with proximate associated *tf*-)significant words.

Finally, while Luhn not only sought to exploit statistical data but to motivate doing so, by referring to underlying probabilistic information in a manner associated with having a general, theoretical model of retrieval, he did not develop this idea in any way. It should also be noted that while Luhn's proposals for automatic abstracting naturally required machine-readable full text, and indeed were implemented for a set of conference papers in 1958 (see Schultz 1968, pp.145-163), his proposals for indexing, while well-suited to future full-text collections, would work differently in then practice. Thus he refers to having a sample of a document collection available as full text for the preliminary vocabulary analysis, so subsequent indexing applying the vocabulary would be done manually. However over time Luhn recognised, as apparent in Luhn (1961), that more work might be done automatically, albeit with simpler index terms, and also that frequency information might be explicitly used in matching.

Luhn was not the lone pioneer of automated retrieval. There was an active and innovatory community thinking about indexing, searching and matching in response to the growing technical literature, and Luhn's emphasis on the indexing vocabulary reflects this wider community's concerns. But his reference to statistical data, and position in a major computing company, were important stimuli to the growth of research on automated retrieval.

There were, in particular, three different lines of research, apparent in Luhn's work, to pursue for automated retrieval: on how specific techniques should be developed, on how effective they were in practice, and on how they embodied models of retrieval. These questions apply both to statistical approaches to retrieval in general, but also, more specifically, to ideas about individual terms and about term associations, whether by conjunction or substitution. Progress on all these fronts should, ideally, lead to full rather than only partial automation of indexing and searching.

2.2 Techniques

Term frequencies How were Luhn's initial ideas about term frequency followed up? The crucial work here was done by Salton and the SMART project (Salton 1968), even though his sources were only abstracts or very short full papers. Salton could gather *tf* information, and took the further step of exploiting it not to select a vocabulary, but to weight terms directly. The more a term occurs in a document, and also a text request, the more important it is,

and this can directly determine the document score using a similarity measure like cosine correlation.

However Luhn, as noted, did not explicitly use *df*, and nor did Salton. Responding to *df* was a separate development (Sparck Jones 1972, Robertson 2004), in part prompted by working with documents characterised by simple word lists without *tf* data. More importantly, the proposal was to use inverse *df*, *idf*, defined by a simple logarithmic function. The aim was to retain all query terms to ensure as much matching as possible, but with variable value. Salton's experiments showed that *tf* was helpful, Sparck Jones's that *idf* was, though both are extremely simple. The natural corollary was to combine them, as Salton did, leading to a matching function factoring in three basic retrieval data properties, namely *tf*, *df* and document length, *dl*. The generic function $tf * idf$ 'mod' dl has proved extremely robust, and has become entrenched in modern retrieval systems. There are many variants, so references to $tf * idf$ weighting should properly be to $tf * idf$-type weighting. The form given in Sparck Jones, Walker and Robertson (2000), also known as Okapi BM25, is sound and useful.

More broadly viewed, the very general notion of index term discrimination value, figuring in Luhn's work, could be given many particular interpretations. For example, if we rank terms in *df* order, then the most valuable terms are most likely to occur in the mid range of *df*s. Salton, e.g. in Salton (1975), examined many alternative definitions for discrimination value, some quite elaborate. But his experiments showed that the easily calculated $tf * idf$ with cosine matching worked well.

Term associations The situation for associations was, not suprisingly, much more complicated, with far more possibilities to explore, but depressingly sparse data. Luhn, like others at the time, believed human judgement was needed to form word classes, albeit exploiting machine-generated data. But there was the obvious challenge of whether, given raw distributional data about term cooccurrences, class formation could also be automated. This interest in automatic thesaurus construction also figured in other areas of NLIP, while NLIP was one generic area for the application of general-purpose automatic classification techniques in which there was great interest at the time, with trials in fields as diverse as medicine and archaeology (see e.g. Sneath and Sokal 1973).

It is important to note that terminology has changed since then, for example in the way "clustering" is used, and has been further muddied by that associated with machine learning. I am using "classification" here, as it was then, as a quite general term. It thus subsumes distinctions between classification types like hierarchical and non-hierarchical, between constructing a set of classes (whether by supervised or unsupervised methods) and assigning to a set of classes, and between specific definitions of what constitutes a class and specific methods for finding classes according to a given definition. In application areas like constructing a retrieval thesaurus, classification tended to refer to unsupervised methods generating overlapping classes, but there were many variations. More importantly the stimulus supplied by automation led to investigations both with existing statistical approaches like Latent Class Analysis and Factor Analysis and to work on new approaches like those represented by work on Clumps. Developing computationally viable classification algorithms was particularly important. This work is well illustrated in Stevens, Giuliano and Heilprin (1965).

The problem was that different classifications look equally plausible, and their real merits can only be demonstrated in actual retrieval. The belief that word classifications are required for effective indexing and retrieval was deeply entrenched, and is reflected in the many ex-

periments carried out throughout the 1960s. But it was impossible to demonstrate any really worthwhile performance gains with automatic thesauri, especially those relying on relatively sophisticated views of classification, though it was acknowledged that this was possibly due to the small, and hence not (statistically) very informative, source data used. The results were certainly felt to be counterintuitive. They did, however, have the important result that analysing and trying to explain test outcomes led to a more careful study of the goals and conditions of retrieval and thus to a concern with models of retrieval. For example, it appeared that term associations (whether conjunctive or substitutive) were only of value when extremely strong. What did this imply about the retrieval situation?

Theories and models It seems to be the case that earlier theories in the retrieval (or rather library) area were not only informal, they were also primarily theories of *indexing*. The presumption was that indexing, rightly conceived, subsumed other aspects of the situation as a whole like matching procedures to obtain appropriate documents. It was obvious what the aim of the whole was, and hence what these aspects were, though with, for example, hierarchical classifications schemes it was recognised that generalisation could imply matches less close to the user's concerns. This emphasis appears not only in Vickery's essentially pre-automation 'On retrieval system theory' (Vickery 1961); it continued long after, for example in Salton's emphasis in Salton (1968) on associative indexing as a central notion and in Salton (1975) on a theory of *indexing*. However the central retrieval notion of relevance appears for example in the 'scale of relevance' procedure used in Joyce and Needham (1958).

Thus in general, even when the concern was with automation, and on the precise definitions and processes this implied, the work focused mainly on specific formulae which were not embedded in a larger encompassing theory where other important notions were explicitly represented: these were implicitly assumed.

There were nevertheless important early moves to formulate a theoretical approach to retrieval that both incorprated key retrieval concepts and was well fitted to the statistical and probabilistic view which Luhn invoked. Maron and Kuhns (1960) make this quite explicit in their title: 'On relevance, probabilistic indexing and information retrieval'. Thus Maron and Kuhns refer to the fundamental notion of 'relevance number' which provides a means of ranking documents in search output by their probability of relevance, where the selection of documents to be ranked, which is designed to capture closeness of meaning in query and document, is based on statistics. Their (Bayesian) probabilistic model is characterised in terms of three types of event: approaching the document collection with a request, expressing an information request with a term, and judging a proffered document relevant. The joint probability over these three types is estimated using distributional data about terms, covering both frequencies of occurrence and frequencies of co-occurrence, to determine query-document closeness. These data can also be exploited to elaborate initial requests or matching document sets.

Maron and Kuhns did not, however, take the further step of exploiting actual relevance judgements as opposed to predicted relevance values. This was first done in Rocchio (1966)'s development of the relevance feedback techniques that because a feature of the SMART system. The development of a general model providing a grounding for the determining the probability that a document is relevant and incorporating relevance feedback was done by Robertson and appeared in Robertson and Sparck Jones (1976). This builds on the *idf* aspect of term weighting but exploits relevant-document rather than all-document term distributions.

As is evident from Maron and Kuhns, these general models could incorporate the notion of term association, however specifically defined, and thus in principle allow for term associations expressing conjunctive or substitutive relationships. These approaches already allow, via (best) multiple term matching, for associations, and for especially valuable ones through relevance feedback. However while term classes could be plugged in as terms, there was never any serious attempt to motivate chosen forms of term classification as natural specific interpretations of the generic probabilistic formulae. Equally, as is evident in van Rijsbergen (1st ed 1975), incorporating term dependencies directly into probabilistic retrieval rapidly becomes extremely hairy.

The other major model dating from the 1960s was that subsequently referred to as Salton's vector space model (see e.g. Salton 1978, also Dubin 2004). This viewed the entities available for retrieval operations - terms. documents, requests, etc., as objects in an information space characterised using the very general and familiar notions of vector space. As Dubin points out, this was primarily, and certainly initially, a model for computational processes, not a theory of retrieval per se. The retrieval context is assumed pertinent to such operations as modifing document or request vector representations to move them closer to one another. The approach is a natural corollary of Salton's starting point in associative indexing. Salton (1979), however, maintains that the vector space model is a retrieval model. But as such it is essentially based on a notion of closeness of meaning, and there is no direct grounding in such a key notion as relevance.

Moreover, whatever grounding a theory might have, there was still the problem of how the theory related, in detail, to system performance (Robertson 1977). The experimental work of the period addressed this, explicitly or implicitly, and the progress that was made was in large part owing to the concurrent development of evaluation methodology, test data resources, etc.

3 Transition phase

The main features of the period from the later 1970s to the early 1990s were strongly connected with one another.

First, there was a gradual extension of retrieval system testing to larger data data sets and to different data sets, including very fine grained comparisons between different strategies, for example for term weighting (Salton and Buckley 1988). These experiments confirmed earlier findings that, for the data scales and types involved, simple term weighting was more valuable than using term classes or complex (multi-word) term units.

Second, with the spread of automation, the attitudes and concerns that had developed in the retrieval research community were increasingly recognised as legitimate and interesting in their own right, rather than as mere hangers on to traditional library activities. Some of the outcomes of decades of research were also gradually filtering through into operational practice, especially where full text was becoming available (see Tenopir and Cahn 1994).

Third, the period saw further development of retrieval system theories, which reflected, and also encouraged, these changes in perspective and status. They included both work on existing theories, like Robertson's probablistic one and Salton's vector space model, and work on new ones, notably on probabilistic inference nets (Turtle and Croft 1992), and on non-classical logics (van Rijsbergen 1986). The first two treated the retrieval situation as predicting document relevance from document properties, and about descriptive resemblances between

document and queries, respectively. The last two treated retrieval as a matter of connecting documents with queries, and as proving requests from documents. These new models, like the older ones, however, still exploit the basic statistical data with which automated retrieval system research began.

4 The recent past

The period from the early 1990s to the present has seen a step change in retrieval research, and one which has been particularly important from the statistical angle. This step change has had several stimuli.

The first, and most obvious for researchers, has been the institution of major retrieval system evaluation programmes. The DARPA/NIST Text Retrieval Conferences (TRECs), running annually since 1992 (Voorhees and Harman 2005), have had an enormous impact on retrieval research, both through the task evaluations themselves and through the participation of many teams in the same evaluations. This has accelerated progress by establishing sound methods (and endorsing their underlying models) and by encouraging the rapid exchange of experience, whether of ideas to explore or techniques to adopt. TREC was initially designed to test existing research techniques on a far larger scale than before, with large document and request sets and with full text material. But it has become a research driver rather than follower by setting challenging new tasks. Both in confirming the value of earlier statistical methods, albeit sometimes with adjustment of detail, and in leading naturally to statistical responses to increasing scale, TREC has further emphasised the contribution that statistical approaches to retrieval make.

Second, the arrival of the Web, and Web search engines, which have both exploited existing research ideas, notably on term weighting, and implemented other novel ideas of their own, has served not merely to make the whole business of retrieval more visible but to provoke new research, for example by embracing a wide range of data types and retrieval activities. This was symbolised by the inclusion within TREC of tasks specifically focusing on Web material and user needs. For the same reasons as TREC, but much more so, the Web has encouraged statistical information processing.

Third, there has concurrent progress in automating other NLIP activities, from speech transcription to information extraction and question answering. This has been partly attributable to the development of sound methods and robust tools for symbolic natural language processing, for instance sentence parsing, and partly to an appreciation of the contribution statistical approaches can make, whether as very effective in themselves, as in speech transcription, or in combination with symbolic processing, as in probabilistic parsing.

Fourth, there has been a surge of activity in statistically-based machine learning, made more useful by more powerful machines as well as very large data sets: the Web is exploited as a data resource for NLIP machine learning, for example. Thus modern forms of data interpretation like Latent Semantic Analysis or Support Vector Machines are widely used.

These four developments have interacted with one another, with valuable outcomes. In the NLIP area, they have encouraged the development of systems that embrace multiple tasks, like retrieval and summarising; and they have spread basic notions like statistically-based term weighting from one area to another. For example, recent work on extractive summarising makes use of term weighting ideas developed for text retrieval, in an updated version of the simple relationship recognised in Luhn (1958). Even within the retrieval area,

broadly speaking, there has been increasing research, exploiting shared techniques, on variant subtasks like filtering as opposed to ad hoc searching.

This interaction has, moreover, been not only at the techniques level but at the model level. One of the most significant developments in retrieval in the last decade has been the application of so-called Language Modelling. Language Modelling was originally proposed for speech transcripion, and proved extremely effective for this. It uses learnt ngram patterns to guide interpretation at multiple levels in speech processing, i.e. for establishing likely sound (type) sequences and likely word sequences. It has spread from there into, for example, translation and summarising, and into retrieval. At the abstract Bayesian level it is as acceptable as the abstract vector approach. It does, however, present problems when given a substantive interpretation as a retrieval model. Thus it is a 'reconstructive' model. For speech it is reasonable to say that the received specific sounds should be reconstructed as the 'real' word sequence. In the retrieval case the reconstructive model is predicting the query given the document: thus document A is a better candidate (i.e. implicitly more relevant) that document B as an account of what request R really is, which presumes that the user 'really' knows what he is looking for.

The retrieval interpretation for Language Modelling, and its relationship to Robertson's older probabilistic model, have been intensively discussed (Croft and Lafferty 2003). It is possible to argue that while relevance is overt in the latter and not the former, this is more appearance than reality and that both models are actually taking relevance as a hidden variable. It is certainly the case that all of the retrieval models currently deployed, including van Rijsbergen's non-classical logic, tend to perform in a similar way because they make similar use of the same distributional data. However Language Modelling has the advantage of offering a consistent way of treating multiple types of information entity and sets of these, and of long practice with estimation techniques.

5 Challenges

It is evident that statistically-based approaches to retrieval have made very significant progress, in both formulation and implementation, since Luhn. However they are faced with major new challenges.

These are all functions of increasing data scale and increasing data heterogeneity.

Scale In principle, having more data allows more accurate capture and characterisation of the structure underlying the surface appearance. However, when the real number of things involved gets large, small percentage errors mean many unwanted things. Retrieval, in general, is not just a very selective task in that there are typically relatively few relevant documents in a large file; it is a highly constrained task in that users are typically unwilling even to inspect many documents. Thus one in five non-relevant in the top ranked search output is acceptable, but three is not, whether or not there are twenty relevant in the file and all are retrieved above rank 30.

So while larger dataset make it possible to get more reliable relevance probability estimates, it is natural to consider using primary features, i.e. terms, that are more refined than the single words familiar from research. The implication in particular is that there is mileage to be got from phrases (even ones defined by pure proximity without any text parsing); and there is some evidence from TREC, for example, that using phrasal terms is helpful, partic-

ularly at high precision. Similarly, indexing and matching at the passage rather than whole text level may offer more search discrimination.

Heterogeneity There are far greater challenges, however, in coping with heterogeneity. Retrieval research was classically directed towards files of technical papers, and assumed relatively homogeneous user populations. The Web has changed all that. Further, heterogeneity takes several shapes.

First, there is much greater heterogeneity in document files, in all their aspects: subject, individual size, etc. Web material also varies greatly in format, reflecting many and complex notions of discourse. Earlier retrieval research concentrated primarily on the basic running text of a document. This was partly because significant document constituents like tables or formulae were not available, but much more because running text in itself, regardless of particular format features, was regarded as the main vehicle for document content. It was indeed recognised that, e.g., titles could be especially important, but format implications were not systematically explored. Modern document representation schemes exploiting markup languages offer a very rich notion of format, encompassing both physical structure manifest as 'fields', for example title, abstract, etc., but also logical structure manifest in such feature types as typographic emphasis, or URLs, to which the notion of field may be extended.

This feature variety presents problems for the statistical approaches hitherto confined to simpler document forms. Specifically, how is information about the different feature types, or fields, to be combined, and especially to be combined in a way which encompasses degrees of importance, perhaps imposed by the user? The probabilistic networks used in the INQUERY system (Croft 2000) were specifically intended to allow for multiple forms of document (or request) description, and at the mechanical level can do this very well. Manipulating multiple fields in a way which is properly grounded in retrieval model foundations is much more difficult. Some work on this has been done (e.g. Robertson, Zaragoza and Taylor 2004, see also Sparck Jones 2005), but there is much more to do. The scope for enalysis is well illustrated by the statement that Google uses more than a hundred feature types for scoring.

The second major form of heterogeneity is in user needs (which vary more than their explicit verbal requests). Is a two-word Web request wanting pages about a topic, or a directory-type inquiry, for example? Web engines apply heuristics, e.g. to establish that requests interpretable as names are for home pages. But there are far more need types than these two. The presumption here is that where individual user requests do not in themselves provide many clues as to underlying needs, it may be possible to mine the behavioural data over time that online searching provides, notably what pages are inspected, to acquire the information to estimate probable underlying needs. User search behaviour and interaction with systems has long been a research concern, and has already been incorporated into statistical approaches via relevance feedback. But there is much more to investigate here, and especially for researchers seeking not merely to establish, as commercial system operators may, *that* something works, but *why* it works.

6 Conclusion

What research on automated retrieval as achieved in the past half century can be illustrated in a very striking way. The International Conference on Scientific Information in 1958 was a response to the perceived need for new responses to the growing volume of literature. At

this meeting Bar-Hillel, a prominent pundit, stated categorically that indexing could never be automated because it required understanding of meaning, and only humans could supply that. Yet by that same meeting Luhn had already published his early ideas, and he presented auto-abstracts for some of the conference papers. Within ten years Salton's first book (Salton 1968) examined automatic indexing and matching ideas, and reported experiments with them, in detail. Since then, statistical approaches to retrieval, that treat meaning implicitly rather than explicitly, but are none the less effective for that, are wholly established. The 1958 conference was, moreover, despite Bar-Hillel, positive about the potential for computing, even if it did not include any papers on the potential for statistical methods. By now, in the major international ACM-SIGIR conferences, statistical papers dominate: thus in 2005 nearly all of the technique and system papers used statistical approaches of one sort or another. For example there are papers using Latent Semantic Indexing, Language Modelling, Logistic Regression, K-means Clustering, Support Vector Machines and Maximum Entropy Modelling. Further, the motivation for continuing to develop statistical approaches to retrieval, and in NLIP generally, is stronger than ever.

References

Bush, V. 'As we may think', *Atlantic Monthly*, 176, 1945, 101-108.

Croft, W.B. (Ed.) *Advances in information retrieval*, Dordrecht: Kluwer, 1000.

Croft, W.B. and Lafferty,J. (Eds.), *Language modelling for information retrieval*, Dordrecht: Kluwer, 2003.

Dubin, D. 'The most influential paper Gerard Salton never wrote' *Library Trends*, 52, 2004, 748-764.

Joyce, T. and Needham, R.M. 'The thesaurus approach to information retrieval', *American Documentation*, 9, 1958, 192-197.

Luhn, H.P. 'A statistical approach to mechanised literature searching', Report RC-3, IBM Corporation, Yorktown Heights, NY. (1957a) (Reprinted in Schultz, 1968)

Luhn, H.P. 'A statistical approach to mechanised encoding and searching of literary information', *IBM Journal of Research and Development*, 1, 1957, 309-317. (1957b) (Reprinted in Schultz, 1968)

Luhn, H.P. 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2, 1958, 159-162. (Reprinted in Schultz, 1968)

Luhn. H.P. 'Potentialities of auto-encoding of scientific literature', Report RC-101, IBM Corporation, Yorktown Heights, NY, 1959. (Reprinted in Schultz, 1968)

Luhn, H.P. 'The automatic derivation of information retrieval encodements from machine-readable texts', in *Information retrieval and machine translation*, (Ed. A. Kent), Vol. 3, Pt 2, 1961, 1021-1028.

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

Rijsbergen, C.J. van, *Information retrieval*, London: Butterworths, 1st Ed. 1975.

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

Robertson, S.E. 'Theories and models in information retrieval', *Journal of Documentation*, 33, 1977, 126-148.

Robertson, S.E. 'On theoretical arguments for IDF weighting', *Journal of Documentation*, 60, 2004, 503-520.

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

Robertson, S.E., Zaragoza, H. and Taylor, M. 'Simple BM25 extension to multiple weighted fields', *Proceedings of the 13th ACM Conference on Information and Knowledge Management*, 2004.

Rocchio, J.J. *Document retrieval systems - optimisation and evaluation*, (thesis), Report ISR-10, Harvard Computation Laboratory, 1966.

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

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

Salton, G. 'Mathematics and information retrieval', *Journal of Documentation*, 35, 1979, 1-29.

Salton, G. and Buckley, C. 'Term-weighting approaches in automatic text retrieval', *Information Processing and Management*, 24, 1988, 513-523.

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

Sneath, P.H.A. and Sokal, R.R. *Numerical taxonomy*, San Francisco: Freeman, 1973

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

Sparck 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, 779-808 and 809-840.

Sparck Jones, K. *Wearing proper combinations*, Report 655, Computer Laboratory, University of Cambridge, 2005.

Stevens, M.E., Giuliano, V.E. and Heilprin, L.B. (Eds.) *Statistical association methods for mechanised documentation*, Publication 269, National Bureau of Standards, Washington DC, 1965.

Tenopir, C. and Cahn, P. 'TARGET and FREESTYLE: DIALOG and Mead join the relevance ranks', *Online*, 18, 1994, 31-47.

Turtle, H. and Croft, W.B. 'Inference networks for document retrieval', *Proceedings of the 13th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1992, 279-290.

Vickery, B.C. *On retrieval system theory*, London: Butterworths, 1961.

Voorhees, E.M. and Harman, D.K. (Eds.) *TREC: Experiment and evaluation in information retrieval*, Cambridge MA: MIT Press, 2005.