Recommended Reading for IR Research Students

Editors:

Alistair Moffat
The University of Melbourne
Melbourne, Australia 3010
alistair@cs.mu.oz.au

Justin Zobel
RMIT University
Melbourne, Australia 3001
jz@cs.rmit.edu.au

David Hawking
CSIRO ICT Centre
Canberra, Australia 2601
david.hawking@acm.org

The Strategic Workshop on Information Retrieval at Lorne (SWIRL 2004) was held in Lorne, Australia, from 8–10 December 2004, see `http://www.cs.mu.oz/~alistair/swirl2004/` for further information. A total of 38 international and Australian researchers and Australian graduate students took part.

Prior to the workshop, participants at SWIRL were asked to provide two nominations, with supporting argument, for a corpus of "must read" papers that IR graduate students should be familiar with. Attendees chose papers that represented key breakthroughs, or represented work undertaken to a particularly high standard, or that presented established material in an innovative or accessible manner. The results are, we believe, illuminating and thought-provoking.

The IR community is fortunate in already having an excellent compendium of past research, the edited volume *Readings in Information Retrieval* (K. Sparck Jones and P. Willett, Morgan Kaufmann, 1997), which not only contains many interesting papers but has detailed commentaries on key sub-areas of IR. However, it is becoming dated; the most recent papers are now a decade old. Another benefit of *Readings* was that it made available many papers that, at the time, were otherwise difficult to access. Today, almost all of the papers nominated by SWIRL participants are available online – as indeed are most of the papers in *Readings* – and there is more value in compiling a list of recommended works than in providing the works themselves. We hope that our annotated reading list, admittedly a much less polished production than the earlier collection, provides a useful update to *Readings* and a valuable resource for graduate students.

Of the nominated papers, only five received more than one vote. The distribution of nominated papers by year is roughly skewed normal, with a median of 1998. The skew is in favor of recent papers. Only four venues provided more than one nominated paper: SIGIR (16), JASIST (4), IPM (3), and SIGIR Forum (2). However, TREC work featured prominently.

Each of the "commentaries" below is a contribution from a single individual; thus some papers have multiple commentaries. Some of the entries have been edited to reduce their length, while staying close to the author's original intention. Papers are listed in order of original publication date. The various commentaries were contributed by Vo Ngoc Anh, Peter Bruza, Jamie Callan, Charlie Clarke, Nick Craswell, Bruce Croft, Robert Dale, Sue Dumais, Luis Gravano, Dave Harper, Dave Hawking, Bill Hersh, Kal Järvelin, Gary Marchionini, Alistair Moffat, Doug Oard, Laurence Park, Edie Rasmussen, Steve Robertson, Mark Sanderson, Falk Scholer, Alan Smeaton, John Tait, Andrew Turpin, Phil Vines, Ellen Voorhees, Ross Wilkinson, Hugh Williams, and Justin Zobel.

Probabilistic models of indexing and searching
(*S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter*, SIGIR, 1981)

Commentary: This paper forms a link between on the one hand, the ideas on statistical indexing of Bookstein, Swanson, Kraft and Harter, and on the other hand, the probabilistic view of searching of Robertson, Sparck Jones, van Rijsbergen and others. The model of Harter et al. included a specific relation between a semantic notion (eliteness) and observable statistical data on term occurrence; this gave a handle on how to make use of within-document term frequency in the searching models. The paper starts with what I still think is a good way to develop the basic probabilistic searching model. The methods developed and tested in this paper were not in fact very successful. However, a considerably simplified version of the model was the basis for the Okapi BM25 scoring function, developed some years later (SIGIR 1994). The Harter 2-Poisson model can also be seen as a precursor to a simple language model. A further paper covering the development of the probabilistic searching models, from the binary independence model of Robertson and Sparck Jones through to Okapi BM25 is the two-part paper by Sparck Jones et al. [2000].

Term-weighting approaches in automatic text retrieval
(*G. Salton and C. Buckley*, IPM, 1988)

Commentary: While this paper is old – it uses only small collections, and good document length normalization techniques and more theoretically-motivated weighting techniques came after this – it is a classic paper for several reasons. First, the paper clearly demonstrates the importance of good weighting. Second, it explains the three components to term weights (document frequency, collection frequency, document length normalization) and defines the weighting nomenclature of "SMART triples" that still has some use today. Third, it is a good example of a retrieval experiment, especially demonstrating the need for testing on multiple collections.

Towards an information logic
(*C. J. van Rijsbergen*, SIGIR, 1989)

Commentary: This paper comes out of left field. It recasts the IR matching problem in terms of inference, instead of matching. An important point about this paper is that it tries to get a handle on the

issue of semantics of IR. The paper and others Keith wrote started a line of research into logic-based IR. Even though this research never led to major pragmatic developments, it allowed IR to be considered in a broader light. As IR blurs into areas such as text mining and knowledge discovery, it is possible that the philosophy behind this paper will be given a new lease of life.

Indexing by latent semantic indexing
(*S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman*, JASIS, 1990)

Commentary: By using singular value decomposition, the authors present a method, latent semantic indexing or LSI, to reduce the dimensionality of the original term-document matrix, creating a much smaller set of orthogonal factors. LSI offers an automated approach to indexing which is based on concepts rather than individual words, addressing the synonymy and (in part) the polysemy problems. Since the publication of this paper in 1990, LSI has been applied in a range of applications in information retrieval and related activities such as development of ontologies, text categorization, text mining and spam filtering.

Commentary: IR, as a field, hasn't directly considered the issue of semantic knowledge representation. The above paper is one of the few that does in the following way. LSI is latent semantic analysis (LSA) applied to document retrieval. LSA is actually a variant of a growing ensemble of cognitively-motivated models referred to by the term "semantic space". LSA has an encouraging track record of compatibility with human information processing across a variety of information processing tasks. LSA seems to capture the meaning of words in a way which accords with the representations we carry around in our heads. Finally, the above paper is often cited and interest in LSI seems to have increased markedly in recent years. The above paper has also made an impact outside our field. For example, recent work on latent semantic kernels (machine learning) draws heavily on LSI.

Basic local alignment search tool
(*S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman*, Journal of Molecular Biology, 1990)

Commentary: Genomic information retrieval is an emerging area of intersection between information retrieval and bioinformatics. Since the early 2000s, it has been a key SIGIR interest area. The seminal genomic IR paper is this original BLAST description, a paper cited in almost all bioinformatics papers since its publication. For those beginning Genomic IR, this is the key paper. BLAST is the Basic Local Alignment Search Tool, a heuristic approach to local alignment that is used to compare genomic DNA and protein sequences. With BLAST, users can search large genomic collections by providing a query sequence and view ranked results. Biologists use this tool to begin almost all explorations of unknown sequences, and most biologists are as familiar with its output as web users are with Google.

Retrieving records from a gigabyte of text on a minicomputer using statistical ranking
(*D. K. Harman and G. Candela*, JASIS, 1990)

Commentary: This paper documents a pragmatic encounter with – shock, horror – a gigabyte of text. But in 1990 a gigabyte was a lot, and while many of the techniques described in this paper have been refined, improved, or downright replaced, it still makes for interesting (and easy) reading, and sets the scene for what the field was like (and struggling with) just 15 years ago. Students new to the field might find the rapid pace of change within half a generation a sobering reminder of their real place in the grand scheme of things. So, while not really seminal, certainly worth a read, especially to people interested in efficiency.

A re-examination of relevance: Toward a dynamic, situational definition
(*L. Schamber, M. B. Eisenberg, and M. S. Nilan*, IPM, 1990)

Commentary: This landmark paper initiated the wave of relevance research to come during the next 13 years. It re-examined the literature made during 30 years, relying on the central works by Cuadra and Katter (1967), Rees and Schultz (1967), Cooper (1971), Wilson (1973), and Saracevic (1975). Essentially, the conclusions were as follows. (1) Relevance is a multidimensional cognitive concept. Its meaning is largely dependent on searchers' perceptions of information and their own information need situations. (2) Relevance assessments have multidimensional characteristics; Relevance is a dynamic concept. It can take many meanings, such as topically adequate, usefulness, or satisfaction. But relevance is also dynamic as assessments of objects may change over time. (3) Relevance is a complex but systematic and measurable phenomenon – if approached conceptually and operationally from the searchers' perspective. Schamber et al. [1990] stressed the importance of context and situation. They re-introduced the concept of "situational" relevance derived from Patrick Wilson's concept in 1973, originating from Cooper (1971). Context may come from the information objects or knowledge sources in systems, but may also be part of the actual information-seeking situation. Two lines of relevance research very fast followed the suggestions and conclusions in this paper. One track pursued the theoretical developments of relevance types, criteria and measurements, thereby bridging over to laboratory IR evaluations. The other line of research consists of empirical studies involving searchers in realistic settings.

Okapi at TREC-3
(*S. E. Robertson, S. Walker, M. M. Hancock-Beaulieu, and M. Gatford*, TREC-3, 1994)

Commentary: This paper (and associated TREC-3 runs) introduced the BM25 weighting formula and demonstrated that it worked very well. Comparing BM25's effectiveness against an arbitrarily chosen "tf.idf" formula provides newcomers to IR with a compelling argument for the value of well-thought-out mathematical models of retrieval. The performance of BM25 (and also Inquery) spurred an immediate advance in length normalization in the vector space model and has influenced many, many research papers and production retrieval systems. Ten years later BM25, with the original TREC-3 tuning parameters, still provides a credible baseline for many retrieval tasks, against which new models can be measured.

Collected papers about TREC-2
(*Various Authors*, IPM, 1995)

Commentary: TREC has since 1992 been the single most significant influence on research in information retrieval. It is not an exaggeration to say that, certainly in the early years, it was revolutionary. It provided a definitive mechanism for separating successful from unsuccessful approaches and provided the first clear

demonstration of many techniques that are now seen as foundational. Some of the papers in this special issue concern specific systems; written in the earliest years of TREC, they are now dated but they stand as a demonstration of the power of the TREC approach to identify strong IR techniques. Of particular interest are three papers. Harman's explanation of TREC is a summary of why TREC was necessary and of the rationale for its design. Sparck Jones's reflections on the TREC approach are valuable for the many insights into experimental methodology. And the Okapi work stands as a demonstration of clear explanation of and rationale for a specific, successful method.

Pivoted document length normalization
(*A. Singhal, C. Buckley, and M. Mitra*, SIGIR, 1996)

Commentary: This pair of papers (the other one being Sparck Jones, Walker, and Robertson [2000]) represent a substantial advance over tf.idf. When Amit unleashed a new approach to length normalization in TREC, we saw a significant improvement, and experimental evidence why this was the case. When Steve introduced the Okapi BM25 it provided a strongly grounded approach to tackle this problem – grounded in probabilistic retrieval theory. Together they show that by looking hard at the experimental data – not tweaking! – it is possible to come up with a significant advance on existing approaches, and that by looking hard at the theoretical underpinnings of information retrieval it is possible to elegantly and efficiently describe and compute such problems. Together they show the value of deeply experimentally based traditions with more model driven explorations can combine to provide us with collective insights into the uncertainties of associating information and need.

Commentary: Much of the work on improving the precision of the vector space method (VSM) has been performed as trial and error. A concept is thought of and tried out on a few document sets. If the precision increases, then it becomes the new vector space method standard for the time. The authors of this article have taken a different direction in improving the VSM. The probability of a document of a given length being retrieved is compared to the probability of a document of the same length being relevant. Their results show that the gradients of relevance and retrieval differ and they meet at a pivot point which is specific to each document set. By normalizing the document-term frequency weights by this pivoted slope (rather than the usual document vector norm) they achieve a significant improvement in retrieval. This work changed the way we think about document normalization and provides a reason as to why the change occurred. This normalization method is now used in the vector space and probabilistic document retrieval methods.

Filtered document retrieval with frequency-sorted indexes
(*M. Persin, J. Zobel, and R. Sacks-Davis*, JASIS, 1996)

Commentary: This paper (and the preliminary version of it by the first author in SIGIR'94) took up and ran with the idea of structuring the index to handle ranked queries as the number one goal, rather than Boolean ones. This simple change allowed a range of efficiency improvements, including dynamic pruning techniques. Other work then followed, suggesting other non-document based orderings for inverted lists. Anyone studying IR implementation needs to visit this paper, as the starting point for a whole thread of logical development.

Natural language processing for information retrieval
(*D. D. Lewis and K. Sparck Jones*, CACM, 1996)

Commentary: This nomination, like nomination of Hobbs et al. [1996], is made with an agenda in mind: we need to see more interaction between research in IR and research in NLP. This paper, written in 1996, was one of the first to argue for a research agenda that explored how NLP techniques could be used to improve information retrieval. The paper was written at a time when the world was in the process of moving to full text search, as opposed to surrogate search; the availability of indexes built from the full text of documents, rather than just abstracts and titles, opens up a range of new opportunities to apply ideas from natural language processing in order to improve indexing. The paper proposes three particular directions where NLP ideas might be explored: first, the use of parsing techniques to identify the appropriate terms to be used in indexing, rather than, for example, relying on simpler collocational criteria in determining compound terms; second, the use of NLP techniques to determine related terms (for example, semantically superordinate terms) to be used in indexing; and third, the use of NLP techniques to implement more sophisticated processing of user queries. The paper is a flag waving exercise in the sense that it suggests a number of directions that might be explored, but it leaves the research to be carried out by others; to my knowledge, the agenda proposed has not yet been fully explored.

FASTUS: A cascaded finite-state transducer for extracting information from natural-language text
(*J. R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. E. Stickel, and M. Tyson*, In *Finite-State Language Processing*, MIT Press, 1996)

Commentary: As with my nomination for Lewis and Sparck Jones [1996], this nomination is made with an agenda in mind; we need to see more interaction between research in IR and research in NLP. This paper remains one of the most cited foundational papers in information extraction (IE). On the face of it, information extraction, which is concerned with extracting from a document a set of predefined informational elements (typically, who did what to whom and when), does not have much to do with information retrieval as that task is commonly understood. Whereas IR is concerned with retrieving either documents or passages within documents, IE is concerned with extracting specific elements of information from a given document; IE is widely viewed as one of the more successful application areas to come out of NLP, and this paper just happens to be a good overview of the kinds of processing that are involved in building IE applications. I believe the paper (and the field of IE generally) is of interest to the IR community because it takes us beyond simple text retrieval to what we might think of as knowledge retrieval; combined with IR techniques to locate relevant documents, IE can deliver a summarization of essential content that meets the same set of needs as those addressed by IR more generally, that is, the management of and access to large document sets in a meaningful and useful manner.

Self-indexing inverted files for fast text retrieval
(*A. Moffat and J. Zobel*, ACM TOIS, 1996)

Commentary: This work describes in detail how to build a document index for fast text retrieval. It begins by covering the lower level of compressing sequences of positive integers using gamma,

delta and Golomb coding of d-gaps. It explains how queries are resolved using simple Boolean and ranked document retrieval methods. It then goes on to discuss fast retrieval methods using skipping for Boolean queries and reduced-memory ranking (the Quit and Continue methods) for ranked queries. Experimental results are given for query times, storage required and precision of the query results, showing that the methods provided are useful in building a large scale document retrieval system. This article covers the whole automatic document indexing and querying process and is very useful for those who wish to implement their own system. This work also forms the basis of large scale information retrieval and should be read by those who wish to enter the field.

Commentary: This paper provides a good example of research (and research communication) into the efficiency of information retrieval systems, presenting an alternative structure for compressed inverted indexes that allows fast query processing with small sacrifices in index size. In achieving the goal of improving the retrieval efficiency, the paper provides a useful introduction to the main components of information retrieval systems: indexing, index compression, and the query evaluation process. It motivates the new compressed index structure by analyzing the operations needed during Boolean and ranked query evaluation, covering also some pruning techniques that can be efficiently applied for the latter without degradation on retrieval effectiveness. The paper is very well structured and written. I especially value the presentation method for introducing new index structures – a combination of motivation, description, formal analysis, and practical experiments.

Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory
(*P. Ingwersen*, Journal of Documentation, 1996)

Commentary: This is a key paper in the development of a holistic cognitive theory for information retrieval interaction. It draws from a number of theories of IR and Information Science and synthesizes them towards a holistic cognitive theory of IR. The theories (or approaches) covered include IR matching models, user-oriented IR research, prior cognitive IR research, informetrics and information seeking. The paper discusses several types of information needs (requests) and several levels of their representation. Mainstream IR focuses on well-defined and stable topical requests while Ingwersen also discusses the importance (and realism) of covering also ill-defined and variable requests, which may be represented at request level, problem/goal level or work task level. Further, the paper discusses the concept of poly-representation of documents and suggests that the cognitive overlaps of different representations (for example, from cognitively different origins) be employed in retrieval. Similarly, various request versions and representation levels provide possibilities for poly-representation. All this is significant if IR research seeks to be more aware of the variety of situations and contexts where IR is applied. And it should.

A survey of multilingual text retrieval
(*D. Oard and B. Dorr*, University of Maryland Technical Report, 1996)

Commentary: Cross-lingual information retrieval has attracted tremendous interest over the past decade, spurred on first by TREC and later the CLEF and NTCIR conferences aimed at European and Asian language CLIR respectively. Yet there is very little in the way of survey papers in Cross lingual, or Multi-lingual information retrieval. The most commonly cited paper, even in relatively recent literature is this unrefereed 1996 report by Doug Oard and Bonnie Dorr. Although Oard has published a number of more recent papers, this one appears to be the most commonly cited because of the length and amount of detail it contains. Regrettably some of the information is becoming dated, and underlines the dire need for a good survey paper in this area.

Simple, proven approaches to text retrieval
(*S. E. Robertson and K. Sparck Jones*, Cambridge Technical Report, 1997)

Commentary: My second nomination is something that has never been published (!) but I always point students at. Its a really simple DIY guide to building an IR system though it is a pity nobody has updated it to 2004 weighting formulae. I like it because it is simple and straightforward and something any undergraduate or graduate student can pick up, read, understand, and then do it. Its something a student should see early in life and always have to hand. There are very many papers in IR which are tough to read and deservedly so because their material is difficult, but there are very few simple, classical, starter papers, and this is the best one I have found.

Commentary: This paper provides a brief but well informed and technically accurate overview of the state of the art in text retrieval, at least up to 1997. It introduces the ideas of terms and matching, term weighting strategies, relevance weighting, a little on data structures and the evidence for their effectiveness. In my view it does an exemplary job of introducing the terminology of IR and the main issues in text retrieval for a numerate and technically well informed audience. It also has a very well chosen list of references. Many of my graduate students come from more conventional computer science backgrounds and are unfamiliar with the ideas it contains. I think it would provide a very useful early chapter for a readings book.

An informal information-seeking environment
(*D. G. Hendry and D. J. Harper*, JASIS, 1997)

Commentary: This paper describes an information-seeking environment, in which its design is informed both by carefully framed user needs and by an analysis of design alternatives using Green's Cognitive Dimensions Framework (CDF). Information-seeking is presented as a problem-solving activity, and the authors argue for under-determined, flexible interfaces to support informal problem-solving practices when searching. This environment emphasizes a particular cognitive dimension, namely secondary notation, in that the user can employ space, layout and locality to represent and organize his/her search activities. Effectively, the authors view the environment as a "spreadsheet for information organization and retrieval" (Marchionini), in which the display "talks back" (Schön) to people about their work. This paper is recommended to anyone interested in designing information-seeking environments (or information retrieval interfaces) where they are concerned with what users want to do with such systems, and where they wish to understand the consequences of high-level design decisions on user behavior and performance. The paper challenges the designer to be free of the strait-jacket of the classical information retrieval interface (query input, result list, document display), which arguably limits information seeking activities, and to widen the space of design possibilities for information-seeking environments.

Improved algorithms for topic distillation in a hyperlinked environment
(*K. Bharat and M. Henzinger*, SIGIR, 1998)

Commentary: There was other early work on hyperlink-based ranking, notably by Brin and Page [1998] and Kleinberg [1999]. Such papers tended to include no effectiveness evaluation, or very non-standard evaluation. To this day, there is a genre of paper that introduces, for example, a new PageRank variant, and completely fails to test its impact on retrieval effectiveness. Bharat and Henzinger use standard IR evaluation methodology, but with web-specific judging instructions to identify "relevant hubs and authorities". It introduces new ranking methods and demonstrated their superiority. It offers insights which remain true today, into the problem domain and specific problems of link analysis. Although it fails to compare link-based ranking to pure content-based ranking (this wasn't done until SIGIR'01), it is an early and important crossover between hyperlink-based ranking and IR methodology.

How reliable are the results of large-scale information retrieval experiments?
(*J. Zobel*, SIGIR, 1998)

Commentary: Together with Voorhees [1998], this paper heralds the start of a series of papers appearing in SIGIR that investigated an aspect of IR that had been little examined up to this point, namely the evaluation of retrieval effectiveness. Researchers had been using the TREC collections for nearly a decade assuming that the collections were OK even though relevant documents were being sampled form the collection using pooling. Apart from the original British Library reports from Sparck Jones and van Rijsbergen, and some checks conducted by Donna Harman in the early years of TREC, no one had really looked carefully at the reliability of the QREL sets being produced by pooling. Zobel [1998] provided such re-assurance demonstrating through the results of carefully conducted experiments why it was OK to rely on pooling. The paper is a tour de force of experiments exploring a range of topics relating to test collections, such as proposing a more efficient method of locating relevant documents from within pools, but the core result, pooling is a good way of finding QRELS is the finding the paper should be remembered for.

Variations in relevance judgments and the measurement of retrieval effectiveness
(*E. M. Voorhees*, SIGIR, 1998)

Commentary: Until 1998, test collections were often criticized for their relevance judgments, people would say that the judgments were unreliable because almost all test collections were formed with relevance judgments (QRELS) made by a single person. "People's judgment of relevance vary, therefore the QRELS of test collections like TREC are unreliable", people would say. There was some past work on this matter conducted by Salton using the early very small test collections, but nothing had been tried on the larger collections like TREC. Voorhees produced a paper at SIGIR 1998 that answered the concerns of test collections' critics. Showing that when different relevance assessors judge documents for relevance, there is a large level of disagreement between assessors, but the variation rarely changes the relative ranking of systems. In other words, if System A is found to be better than System B on a test collection using one set of relevance judgments and then the judgments are replaced with those of another assessor, on the new version of the test collection, System A will remain measured better than System B. What sets this paper apart from others is not just the significance of the result, but the experiments used to produce the result. Voorhees was the first to use the corpus of past TREC results to test her ideas out and produce a wealth of data showing that test collections were OK. The Voorhees paper also heralds the start of series of excellent papers from Voorhees and Buckley across most of the subsequent SIGIRs, each of which used past TREC result data to tackle important topics such as stability of evaluation measures, reliability of significance measures, all of which deserve recognition.

Advantages of query biased summaries in information retrieval
(*A. Tombros and M. Sanderson*, SIGIR, 1998)

Commentary: Google's sweep to dominance in web retrieval is most often attributed to its use of the PageRank(tm) algorithm. This is simplistic: Google brought several new features into the marketplace and each was a key driver in its adoption as the search engine of choice. One of the most significant advantages of Google – and a technique adopted by all search engines since – is its query-biased summaries. These allow users to more effectively judge whether documents are relevant or not, and also to identify key details such as email addresses from home pages without retrieving the page itself. Tombros and Sanderson's proposal and experimental evaluation of query-biased summaries is seminal IR work. In the paper, they describe how to compute, rank, and display query-biased summaries, and show they are an effective mechanism for users. The paper is required reading for anyone building a retrieval engine.

A language modeling approach to information retrieval
(*J. Ponte and W. B. Croft*, SIGIR, 1998)

Commentary: This paper introduced the language modeling approach to IR and has become one of the primary sources for a range of research in recent years. Besides this historical value, the paper has an interesting way of presenting the material – the way that might benefit beginning theoretical IR researchers. Step by step, the authors walk the reader through a discussion of existing retrieval models, arguing for a move to language modeling. The new approach itself is described in detail, with motivation provided for each of the decisions taken. In short, the paper is necessary for the beginning researchers to understand the language modeling approach as well as understanding many theoretical aspects of information retrieval research.

Commentary: My first nomination is Ponte and Croft's Language Modeling paper from SIGIR'98 and I include this because it was seminal (cliché!) in that it was the first real exposition of LM in information retrieval. There was a smattering of subsequent papers at other conferences in the period following, but this was the first to a real IR audience and I remember coming out of the auditorium afterwards and thinking "gosh, that was interesting". There are probably better LM papers, and more advanced developments in LM and other applications, but this one sticks in my mind.

The anatomy of a large-scale hypertextual web search engine
(*S. Brin and L. Page*, WWW7, 1998)

Commentary: This paper (and the work it reports) has had more impact on everyday life than any other in the IR area. A major contribution of the paper is the recognition that some relevant search

results are greatly more valued by searchers than others. By reflecting this in their evaluation procedures, Brin and Page were able to see the true value of web-specific methods like anchor text. The paper presents a highly efficient, scalable implementation of a ranking method which now delivers very high quality results to a billion people over billions of pages at about 6,000 queries per second. It also hints at the technology which Google users now take for granted: spam rejection, high speed query-based summaries, source clustering, and context(location)-sensitive search. IR and bibliometrics researchers had done it all (relevance, proximity, link analysis, efficiency, scalability, summarization, evaluation) before 1998 but this paper showed how to make it work on the web. For any non-IR engineer attempting to build a web-based retrieval system from scratch, this must be the first port of call.

Commentary: The original Google paper is such an obvious candidate that I hesitated to nominate it, since I'm certain that others will as well. The web site for the WWW7 conference no longer appears to be functioning, but typing "Brin pagerank" into Google produces the paper as the top hit. The fact that the paper can be referenced in this fashion demonstrates its importance. Unfortunately, it is not very "polished". Various aspects of algorithms, architecture and low-level data structures are mixed up together and covered in different amounts of detail. Nonetheless, there are few papers that have had the same level of impact on both research and practice. It appears as the 63rd most cited paper on CiteSeer, and "Google" is now a verb. Ideally, a volume of IR background reading would contain an entire section on web-related methods, including papers by Kleinberg and Henzinger, along with this one.

Commentary: This paper has been enormously influential for obvious reasons. While this paper does not follow many traditional information retrieval evaluation and presentation conventions, its impact has turned it into a must-read paper for anyone interested in web search.

Commentary: Web search has illustrated several things including the importance of non-content factors in ranking (beyond "aboutness" to use Hutchins terminology), and of course issues of scale in crawling, index building, and querying. I considered three papers (Brin and Page, as listed; Page, Brin, Motwani and Winograd, "The PageRank citation ranking: Bringing order to the web", and Kleinberg, "Authoritative sources in a hyperlinked environment"). I picked the Brin and Page paper because of the breadth of topics it covers. Published in WWW7 in Brisbane, this paper provides a high-level overview of an early implementation of the Google web search engine. It also highlights the importance of non-content factors (for example, PageRank) along with a variety of content matches (anchor text, plain text in large font, plain text, etc) and proximity to arrive at an overall ranking. To me, this is the single biggest contribution of web search systems to the IR community and has implications beyond the web. Finally, there is some discussion of results presentation issues such as grouping by site and summarization. There is no systematic evaluation of the ranking algorithm in this paper, but I still think that the breadth of topics covered in this paper make it a must read. It's also interesting to re-read Appendix A (Advertising and Mixed Motives), five years and one IPO after that fact. Currently, the predominant business model for commercial search engines is advertising. "The goals of the advertising business model do not always correspond to providing quality search to users ... we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers ... But we believe the issue of advertising causes enough mixed incentives that it is crucial to have a competitive search engine that is transparent and in the academic realm." (Kleinberg's paper is much more scholarly, citing work in social networks, bibliometrics, hypertext, and so on, but it is focused almost entirely on properties of the web graph, and in practice this is only a small part of what goes into building a good web search engine.)

Exploring the similarity space
(*J. Zobel and A. Moffat*, SIGIR Forum, 1998)

Commentary: This paper exposes the myriad of weighting schemes for measuring similarity used in the IR literature in a clear, systematic way. It provides a good framework for rigorously testing similarity schemes, from basic cosine measures through to Okapi and beyond. It also provides a convenient, succinct notation for describing weighting schemes, removing the "black magic" from IR engines (if the notation is employed!). To my mind, it puts a stop to any serious research in the area of fiddling weighting schemes in order to improve recall-precision. Any new changes should be slotted into the framework, and compared with the results obtained using the existing schemes as described.

Document expansion for speech retrieval
(*A. Singhal and F. Pereira*, SIGIR, 1999)

Commentary: The key innovation in ASR was that it became possible to create ASR systems that could automatically transcribe broadcast news, and that in turn made it possible for the first time to build systems that could effectively search large collections of useful spoken content. There were three research threads that deserve mention. The most fundamental was the effective integration of ASR with IR, which was pursued vigorously by a small number of teams (mostly ASR teams, since IR was the easier of the two problems) in the TREC "Spoken Document Retrieval" (SDR) track. The best ASR systems during this period were built at the University of Cambridge in the UK, so one of their later TREC papers might be a good choice. An alternative is the paper selected here, covering document expansion for ASR-based search, which offers some nice insights into the structure of the problem. The second thread was the integrated use of ASR with other sources of evidence (for example, face recognition and video OCR) to search video. This is presently the focus of the TRECVID evaluation, but the seminal work in that area is unquestionably the CMU Informedia project because that is the first time that a team had the resources to attack that challenge at large scale. The third thread was the Topic Detection and Tracking (TDT) evaluations, which introduced a focus on clustering similar news stories and detecting stories on novel topics. Score normalization turned out to be the key to this process, and the BBN team was the first to find a truly effective solution to that problem. The BBN chapter of the TDT book edited by Allan might therefore be a good choice.

Information retrieval as statistical translation
(*A. Berger and J. D. Lafferty*, SIGIR, 1999)

Commentary: The paper provides one of the earliest formal treatments of the language model approach to IR. Viewed by many in the other language technology communities as a seminal IR paper because it "spoke their language". It also introduced translation probabilities into the retrieval model, and this was subsequently used heavily in cross-lingual retrieval work.

User interfaces and visualization
(*M. Hearst*, In *Modern Information Retrieval*, Addison-Wesley Longman, 1999)

Commentary: I wanted something about user interfaces for search. There are lots of individual papers on specific aspects of the problem, but I nominate Hearst's chapter in Baeza-Yates and Ribeiro-Neto's book since it presents an overview of research and innovation in interfaces for search. It covers a wide range of techniques for query specification and refinement, and for browsing and searching collections. Current web search interfaces are about as impoverished as one can get — people are provided with a small rectangle in which to enter the query, shown a list of results, and if the search doesn't return what they want they just have to try again. Researchers have explored a variety of techniques for improved query specification, results presentation and interaction, many of which are reviewed in this chapter.

Grouper: A dynamic clustering interface to web search results
(*O. Zamir and O. Etzioni*, WWW8, 1999)

Commentary: This paper is not perfect. I was tempted to recommend the Cutting, Karger, Pedersen and Tukey paper, "Scatter/Gather: a cluster-based approach to browsing large document collections", in SIGIR'92 (pages 318–329) instead, but went with this one as it offers three interesting lessons for readers. First, it implements a clustering technique that is practical for the WWW environment. The notion of implementing something that works rather than aiming for an optimal solution goes against IR goals of discovering scalable theories, however, it provides a stepping stone to developing a larger IR environment that actually affects the lives of people. In this sense, the paper demonstrates good engineering rather than theory. Second, the paper offers an actual user interface that leverages clusters for search results. The mix of phrases for labeling the clusters, sample titles, cluster size, and query refinement on the interface makes this is very low-tech but high-information design. I also like the fact that the authors have exposed their numerous design decisions along the way. Note that if the suffix tree clustering itself is the breakthrough, then the authors' SIGIR 1998 paper would be a better choice. However, because it is the whole system that is the lesson here, I strongly prefer this paper. Third, there is an evaluation. It is a somewhat novel evaluation in that it uses a comparison of server logs for two systems and tries to get at some search path criteria rather than reducing everything to recall or precision surrogates for performance. They tried some simple but interesting metrics like click distance within results. One disappointment I have in this work is that, to my knowledge, they have not done the user studies and follow up work that they say they will do in the paper.

Authoritative sources in a hyperlinked environment
(*J. M. Kleinberg*, JACM, 1999)

Commentary: Kleinberg's work on hubs and authorities was a seminal paper in showing how the information inherent in the underlying network structure of the web could be exploited. Kleinberg bases his model on the authorities for a topic, and on hubs – pages that link to a large number of thematically related authorities. He observes that hubs are in equilibrium with, and confer authority on, the sites to which they link, that is, they have a mutually reinforcing relationship. This work was significant in providing an algorithmic approach to quantifying the quality of web pages, a key issue in the web environment where the massive size of the database, information redundancy and the uncertain quality and source of information make retrieval difficult. Related work (Bharat and Henzinger [1998]; Chakrabarti's "Clever" system; Brin and Page [1998] and PageRank) has applied similar methods to resource discovery on the web. (This is actually a second paper on Kleinberg's work, the original was a conference presentation 1998. This version has greater detail.)

Variations in relevance judgments and the measurement of retrieval effectiveness
(*E. M. Voorhees*, IPM, 2000)

Commentary: Evaluation is an important component of the IR field, and most evaluation is done using the Cranfield methodology. This paper addresses one of the major concerns about the appropriateness of the Cranfield methodology by confirming that while relevance judgments *do* depend on the assessor, the relative quality of retrieval runs is stable despite these changes. This result holds for different collections, different evaluation measures, different types of judgments, and different types of assessors. The paper also shows that the upper bound on the effectiveness of retrieval systems as measured by recall/precision is limited by this disagreement among humans, and therefore systems cannot hope to reach the theoretical limits of "perfect" precision or recall.

A probabilistic model of information retrieval: development and comparative experiments. Parts I and II
(*K. Sparck Jones, S. Walker, and S. E. Robertson*, IPM, 2000)

Commentary: This two-part paper presents a probabilistic retrieval model. It begins from first principles, and derives formulations that culminate in the Okapi BM25 ranking function. As such, it draws together developments and experiences from over a decade of IR research. The paper is important because: it explains the successful Okapi BM25 ranking function; a probabilistic model of retrieval is derived from first principles; important assumptions underlying the model are explained; the paper systematically shows how additional sources of information (for example, relevance information and term frequencies) can be incorporated into the model; comprehensive experiments, based on the TREC framework, are included to illustrate the impact that different parameters have on overall performance.

Commentary: See under Singhal, Buckley, and Mitra [1996].

Evaluating evaluation measure stability
(*C. Buckley and E. Voorhees*, SIGIR, 2000)

Commentary: This paper investigates the stability of widely-used IR evaluation measures such as mean average precision (MAP), precision at 10 documents retrieved, and R-precision. By calculating error rates based on runs submitted to the TREC Query track, the authors demonstrate that the stability of different measures can vary significantly. For example, MAP is shown to have a low error rate when 50 topics are used. Precision at 10 documents retrieved, on the other hand, has a substantially higher error rate. This paper is important because: it gives an overview of the experimental methodology used to evaluate the performance of information retrieval systems; the assumptions underlying the commonly-used evaluation measures are investigated; important limitations are demonstrated, assisting IR researchers to conduct meaningful experiments for the evaluation of new ideas; it promotes thinking

about the meaning of the numbers, rather than just looking at the numbers themselves; a sizeable bibliography of important related papers that consider experimentation in IR is included.

Do batch and user evaluations give the same results?
(*W. Hersh, A. Turpin, S. Price, D. Kraemer, B. Chan, L. Sacherek, and D. Olson*, SIGIR, 2000)

Commentary: It is not often that one would point to a "failed" experiment as a key piece of work in a field. However, the failure of the users in the experiments reported in this work to gain the benefit predicted in batch experiments is a key piece of research. Our colleagues in Library Sciences keep on pointing out the importance of the people who use information retrieval systems, and too little of our work takes into account how people use information retrieval systems. Even enormously successful information retrieval experiments – Google – still spends comparatively little effort on understanding its user's behaviors, and this paper points out the risks of such effort.

Commentary: This paper, and its companion in SIGIR 2001, are important because they provide solid evidence that optimizing IR engines using batch experiments in the ad-hoc style of TREC does not necessarily translate into an improved IR engine for users. Unfortunately they do not offer any real reasons why the "improved" systems do not translate into real improvements, either perceived or actual. Hopefully these papers will cause some researchers to stop the relentless pursuit of ever higher precision on known query sets, and concentrate on human factors in the retrieval process.

Content-based image retrieval at the end of the early years
(*A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain*, IEEE PAMI, 2000)

Commentary: This paper is a comprehensive and authoritative survey paper of CBIR up to pretty much its date of publication in 2000. It is an exemplary survey paper, and also appeared at a turning point in its field. In particular it appeared just as really effective systems appeared which focused on allowing the user to explore images in feature space, rather than categorical or analytic keyword searching. Subsequent advances within feature space searching have tended to be incremental refinements, not presenting fundamental advances on the work reviewed in the paper. For this reason I believe this paper will become a classic, representing not just a state of the art, but also the possibilities and limitations in terms of retrieval effectiveness within a given set of technical limitations. It really defines what can be done in (especially) still image retrieval without the introduction of deep semantics or a surrogate for them, that is automatic keyword image indexing. It contains definitions of such specialist terms as the "semantic gap" and makes some perceptive comparisons between image and text retrieval which also broadly apply to other forms of multimedia retrieval (for example music).

Distributed information retrieval
(*J. Callan*, In *Advances in Information Retrieval*, Kluwer Academic Publishers, 2000)

Commentary: This paper is on distributed information retrieval. In distributed information retrieval, the documents of interest appear scattered over multiple text databases, which can be heterogeneous along a number of dimensions, such as topical focus, intended audience, supported query evaluation models, and degree of "cooperation" in the distributed retrieval process. Unfortunately, text databases on the web are often not "crawlable" by traditional methods, so search engines largely ignore the database contents. Furthermore, these "hidden-web" databases are often of high quality, which highlights the importance of distributed information retrieval research to – in the not-so-distant future – fully integrate search over the crawlable web with search over the hidden-web database contents. This very nicely written paper presents an overview of key research addressing the main challenges in distributed information retrieval.

A study of smoothing methods for language models applied to ad hoc information retrieval
(*C. Zhai and J. Lafferty*, SIGIR, 2001)

Commentary: This paper is one of the better descriptions of the generative approach to using statistical language models for probabilistic information retrieval. However, what I find most interesting about this paper is that it shows that (i) even this "more principled" approach to probabilistic IR requires careful tuning for success, and (ii) the basic theory offers little guidance about how to do the tuning. The result is a clean theory, with knobs for tuning, and ad-hoc methods for doing the tuning. I think this paper reveals both the strengths and weaknesses of the generative approach to using statistical language modeling for IR.

Relevance-based language models
(*V. Lavrenko and W. B. Croft*, SIGIR, 2001)

Commentary: One criticism of most statistical language modeling approaches to information retrieval is that they are essentially word-matching models, like the much maligned (although very successful) vector-space retrieval model. There is no place in the model for the user, the user's (unspecified) information need, or the concept of relevance. This paper is the beginning of a line of influential research from Lavrenko and Croft that bridges the gap between classical models of probabilistic information retrieval and the newer statistical language modeling approaches to probabilistic information retrieval.

Cross-lingual relevance models
(*V. Laverenko, M. Choquette, and W. B. Croft*, SIGIR, 2002)

Commentary: There have been many papers written about Cross-lingual information retrieval (CLIR) in recent years. Most of the papers discuss some improved technique, together with a specific collection on which the experiments are conducted. Inevitably the technique leads to an improvement compared to the baseline. While such results do provide information about successful techniques, they often lack an overarching theoretical framework (a not uncommon problem in IR!). There are a number of approaches to CLIR, generally using one or more of dictionary, corpus, or parallel texts as resources to facilitate CLIR. The above paper represents one of the few attempts to construct a formal model for the cross-lingual retrieval process, and analyze different techniques within that framework. It is also interesting because it does not rely on any translation mechanism, but models the probability of a document in language A being relevant to a query in language B. The supporting experiments use the TREC 9 Chinese CLIR task.

Factors associated with success for searching MEDLINE and applying evidence to answer clinical questions
(*W. R. Hersh, M. K. Crabtree, D. H. Hickam, L. Sacherek, C. P. Friedman, P. Tidmarsh, C. Moesback, and D. Kraemer*, Journal of the American Medical Informatics Association, 2002)

Commentary: This is one of several investigations looking at how real users do with IR systems. I choose this paper because the general IR community is less likely to be familiar with it (as compared to other papers coming out of the TREC Interactive Track that had similar findings). This paper assessed the factors associated with the correct answering of clinical questions by advanced medical and nurse practitioner students. They used a state of the art MEDLINE system. The research found that a substantial number of questions were answered incorrectly even aided by the use of an IR system and that one of groups, nurse practitioner students, obtained no benefit from the system at all. The study also looked at the relationship between recall/precision and successful use of the system, finding there was no relationship whatsoever, giving further credence to the notion that these measures are not that important in the larger searching environment.

A taxonomy of web search
(*A. Broder*, SIGIR Forum, 2002)

Commentary: I believe this paper is important because it looks at the environment most people use to do IR, the web, and analyzes what they do with it. It is important to realize that IR is an important part of using the web, but not all of it.

Stuff I've seen: A system for personal information retrieval and re-use
(*S. Dumais, E. Cutell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins*, SIGIR, 2003)

Commentary: I selected this paper because it addresses an increasingly important problem – how to search and manage personal collections of electronic information. So the primary reason to choose it is that it addresses an important user-centered problem. Secondly, as in the first paper, this paper presents a practical user interface to make the system useful. Third, the paper includes large scale, user-oriented testing that demonstrates the efficacy of the system. Fourth, the evaluation uses both quantitative and qualitative data to make its case. I think this paper is destined to be a classic because it may eventually define how people manage their files for a decade. Moreover, it is well-written and can serve as a good model for developers doing system design and evaluation, and for students learning about IR systems and evaluation.

On collection size and retrieval effectiveness
(*D. Hawking and S. E. Robertson*, Information Retrieval, 2003)

Commentary: This paper is an exemplar of good research method in information retrieval. The authors take hypotheses put forward by participants in a TREC track (VLC at TREC 6, 1997) and devise experiments to test each of the hypotheses. The experiments are diverse, ranging from obvious tests on effectiveness to approaches based on deep insight into how retrieval processes work. The writing is lucid, the conclusions are clear and thoroughly justified, and the presentation is refreshingly free of prior bias towards one point of view or another. The paper is also an exemplar of the fact that an important result does not have to be a demonstration of a novel technique or of an innovation of some kind. It shows that experimental confirmation (or rebuttal) of previous theories can be a valuable contribution.

A noisy-channel approach to question answering
(*A. Echihabi and D. Marcu*, ACL, 2003)

Commentary: This paper is not well-known in the IR community, but it describes a statistical, language-model approach to question answering can be as effective as more knowledge-based approaches. Given the increasing importance of QA and the overlap with IR, it is critical to show that the statistical approaches of IR are not superseded by more language and knowledge-based approaches. This paper shows that there is a common basis for these two tasks and that a statistical framework can be used to capture and use a lot of linguistic knowledge.

Relevance models in information retrieval
(*V. Lavrenko and W. B. Croft*, In *Language Modeling for Information Retrieval*, Kluwer Academic Publishers, 2003)

Commentary: In early approaches to applying language modeling in information retrieval, the notion of relevance had not been explicitly modeled. In particular, it has been difficult to capture processes such as relevance feedback in the language modeling framework. In this important and ground-breaking paper, the authors develop a formal model which effectively integrates the classical probabilistic model of retrieval with recent developments in estimation techniques, arising from work on language modeling. There are two main theoretical foundations for the new relevance (language) model (actually, models). One, the classical probabilistic approach, as expressed in the Probability Ranking Principle, which proposes that documents are ranked according to $P(R \mid \mathit{Document}, \mathit{Query})$, where R is the class of relevant documents. And, two, the various generative language models, which attempt to estimate $P(\mathit{Query} \mid \mathit{Document})$. They propose a basic relevance model, and then two distinct approaches based on this model: the probability ratio approach and the cross-entropy approach. Much of the theoretical part of the paper is devoted to estimating relevance models, both with and without examples of the set of relevant documents, and to exploring the role of smoothing in addressing the problem of high variance in maximum likelihood estimators. Further, two approaches to estimating probabilities of words in the unknown set of document relevant to a query are presented. In the experiments, they compare the new relevance model approach against the best performing baseline approaches, and demonstrate that the new approach significantly outperforms the already excellent performance of the baselines. The paper provides an excellent and clear description of the new generative relevance models, supported by a comprehensive set of experiments. Importantly, the authors provide insightful analysis and argument as to why particular approaches do in fact outperform others. This paper appears in a collection of papers that grew out of a workshop held in May/June 2001 at Carnegie Mellon University. This collection of papers is recommended to those researchers that intend developing or applying language modeling in IR.

Simple BM25 extension to multiple weighted fields
(*S. Robertson, H. Zaragoza, and M. Taylor*, CIKM, 2004)

Commentary: Due to its simplicity, effectiveness and theoretical underpinning, the BM25 measure is now widely used in IR re-

search, and a paper describing the measure is an absolute requirement for a volume of IR background reading. However, the SIGIR '94 paper by Robertson and Walker (reproduced in Sparck Jones and Willett) is slightly out-of-date and should be replaced by a more current paper. Just before the deadline for SWIRL homework, I received a preprint of the nominated paper, and I realized that it was an ideal candidate. In addition to presenting an up-to-date version of BM25, it provides valuable insight into how the measure is used in practice, including examples of parameter tuning. The paper discusses the problem of extending BM25 to structured documents, where terms appearing in certain fields (for example, titles and anchors) must be given a greater weight. The simplicity of the solution should be an inspiration to any new researcher.

Interactive cross-language document selection
(*D. W. Oard, J. Gonzalo, M. Sanderson, F. López-Ostenero, and J. Wang*, Information Retrieval, 2004)

Commentary: It is hard to point to a single seminal work in CLIR because the innovations were introduced sequentially. The first paper in the modern evolution of the community was a 1990 conference paper by the Bellcore group on cross-language LSI. Introduction of CLIR in TREC in 1996 led to a stream of innovations that depended on the availability of a large test collection, including (1) cross-language blind relevance feedback, introduced by Ballesteros and Croft in 1997, (2) structured queries, introduced by Pirkola in 1998 (building on earlier work by Hull), (3) bidirectional translation, introduced by McCarley in 1999, and the use of translation probabilities trained on parallel corpora, introduced separately by three TREC teams (BBN, TNO, and UMass) in 2000. Of these, the effective use of translation probabilities was the ultimate key to success – effective use of translation probabilities has a greater beneficial effect than any other single issue in CLIR. So if I were to recommend two "must-read" papers in CLIR, I would choose the Ballesteros and Croft SIGIR 1998 paper (one year after their first one, and thus better developed and also incorporating structured queries) and one of the three initial parallel corpus papers from TREC-9 (or, in every case, later published journal articles). Of the three, the BBN paper was the most accessible, but it had the unusual feature that it adopted an HMM rather than a language model as a point of departure (mathematically, this choice led to the same result, though). The TNO and UMass papers were cast in the language modeling framework that has come to dominate resent research in IR, so one of them might be a better choice for a "must read" volume where there are sure to be other language modeling papers that will set the reader up to understand that framework well. Finally, it is important to note that the basic structure of the IR problem breaks down in cases when the searcher cannot read the document's language. This has been the focus of the CLEF interactive track, the most interesting result of which is that current MT technology is good enough to support interactive selection of documents, and led to the paper recommended here. Another possible choice with far less detail and more recent results using a QA task (which proved to be quite interesting) would be the CLEF-2004 interactive track overview paper.

Other comments

Commentary: Other topics that I considered included: machine learning especially for text categorization, clustering, and information extraction; retrieval from structured data (the web is a special case of this); models for IR including language models and dimension reduction (LSI/PLSI/topics); personal information management; use of redundancy for QA or information extraction (KnowItAll); and analysis of novelty (a la MMR). And, for fun, I also looked at the 100 most cited papers in CiteSeer and scanned for IR-related ones, `http://citeseer.ist.psu.edu/source.html`. I realize that this is biased by the nature of the papers that they index (very few HCI/NLP papers, for example), by the age of papers, etc. Yet, three IR-related papers appeared in the Top100. This list only includes documents in the CiteSeer.IST database. Citations where one or more authors of the citing and cited articles match are not included. The data is automatically generated and may contain errors. The list is generated in batch mode and citation counts may differ from those currently in the CiteSeer.IST database, because the database is continuously updated. At rank 53 was "Indexing by Latent Semantic Analysis", Deerwester, Dumais, Furnas et al. (1990); at rank 63, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", [Brin and Page, 1998]; and at rank 77, "Authoritative Sources in a Hyperlinked Environment", Kleinberg (1997).

Commentary: I gave serious consideration to three other papers: "Matching Words and Pictures", Barnard et al., J. Machine Learning Research, 2003, a seminal paper on automatic keyword indexing of images, but rather rambling, strays beyond IR and has some methodological problems; "The Automatic Derivation of IR encodements for machine-readable text", by H. P. Luhn, from *Readings in IR*, always worth another read, unique and stimulating brief early work in some ways more relevant in the world of the semantic web than years ago; and "On relevance. . . ", Maron and Kuhns, again in *Readings*, another astonishing piece of work for its day showing real insight into problems, which in some cases only really impinged on the practical search world with search engines.

Commentary: There are, of course, several other seminal papers that without question should be included in our consideration that address other topics. Most notable among those are the original Brin and Page paper on PageRank, and something from the pioneering work on blind relevance feedback (the UMass LCA paper comes to mind, but I suspect that there is something that predates it), and something on statistical significance testing (perhaps Hull's well received SIGIR paper). One of Hersh's two SIGIR papers that was motivated by the results of the TREC interactive track would also be an excellent choice, and would something on QA. Thinking even more broadly, the impact of TREC on research in IR has been so fundamental that a "must read" volume without a TREC overview would clearly be incomplete. Voorhees [2000] might be the right one to pick for that.

Acknowledgments

SWIRL 2004 was funded by the Australian Academy of Technological Sciences and Engineering (`http://www.atse.org.au/`). We gratefully acknowledge the support of Frontiers of Science and Technology Mission and Workshop component of the Innovation Access Program. part of the Australian Government's Innovation Statement, *Backing Australia's Ability*. We also thank the CSIRO ICT Centre, RMIT University, and the University of Melbourne for additional financial support.

Nominated papers

S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *J. of Molecular Biology*, 215:403–410, 1990.

A. Berger and J. D. Lafferty. Information retrieval as statistical translation. In *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 222–229, Berkeley, CA, August 1999. ACM Press, NY. URL `http://citeseer.ist.psu.edu/berger99information.html`.

K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 104–111, Melbourne, Australia, August 1998. ACM Press, NY. URL `http://gatekeeper.dec.com/pub/DEC/SRC/publications/monika/sigir98.pdf`.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In P. Thistlewaite and H. Ashman, editors, *Proc. 7th World Wide Web Conf. (WWW7)*, pages 107–117, Brisbane, Australia, April 1998. URL `http://decweb.ethz.ch/WWW7/1921/com1921.htm`.

A. Broder. A taxonomy of Web search. *SIGIR Forum*, 36(2), Fall 2002. URL `http://sigir.org/forum/F2002/broder.pdf`.

C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In Emmanuel Yannakoudakis, Nicholas J. Belkin, Mun Kew Leong, and Peter Ingwersen, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 33–40, Athens, Greece, September 2000. ACM Press, NY. URL `http://doi.acm.org/10.1145/345508.345543`.

J. Callan. Distributed information retrieval. In W. Bruce Croft, editor, *Advances in Information Retrieval*, chapter 5, pages 127–150. Kluwer Academic Publishers, 2000. URL `http://www-2.cs.cmu.edu/~callan/Papers/ciir00.ps.gz`.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic indexing. *J. of the American Society for Information Science*, 41(6):391–407, 1990.

S. Dumais, E. Cutell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins. Stuff I've seen: A system for personal information retrieval and re-use. In *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 72–79, 2003.

A. Echihabi and D. Marcu. A noisy-channel approach to question answering. In *Proc. 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003.

D. K. Harman and G. Candela. Retrieving records from a gigabyte of text on a minicomputer using statistical ranking. *J. of the American Society for Information Science*, 41(8):581–589, August 1990.

D. Hawking and S. E. Robertson. On collection size and retrieval effectiveness. *Information Retrieval*, 6(1):99–150, January 2003. URL `http://www.kluweronline.com/issn/1386-4564`.

M. Hearst. User interfaces and visualization. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 257–323. Addison-Wesley Longman, 1999. URL `http://www.sims.berkeley.edu/~hearst/irbook/chapters/chap10.html`.

D. G. Hendry and D. J. Harper. An informal information-seeking environment. *J. of the American Society for Information Science*, 48(11):1036–1048, 1997.

W. Hersh, A. Turpin, S. Price, D. Kraemer, B. Chan, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In Emmanuel Yannakoudakis, Nicholas J. Belkin, Mun Kew Leong, and Peter Ingwersen, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 17–24, Athens, Greece, September 2000. ACM Press, NY. URL `http://medir.ohsu.edu/~hersh/sigir-00-batcheval.pdf`.

W. R. Hersh, M. K. Crabtree, D. H. Hickam, L. Sacherek, C. P. Friedman, P. Tidmarsh, C. Moesback, and D. Kraemer. Factors associated with success for searching MEDLINE and applying evidence to answer clinical questions. *J. of the American Medical Informatics Association*, 9(3):283–293, May/June 2002. URL `http://medir.ohsu.edu/~hersh/jamia-02-irfactors.pdf`.

J. R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. E. Stickel, and M. Tyson. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*, pages 383–406. MIT Press, 1996. URL `http://citeseer.nj.nec.com/hobbs96fastus.html`.

P. Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *J. of Documentation*, 52(1):3–50, 1996.

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. of the ACM*, 46(5):604–632, 1999.

V. Laverenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 175–182, 2002. URL `http://ciir.cs.umass.edu/pubfiles/ir-251.pdf`.

V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New Orleans, LA, September 2001. ACM Press, NY.

V. Lavrenko and W. B. Croft. Relevance models in information retrieval. In W. Bruce Croft and John Lafferty, editors, *Language Modelling for Information Retrieval*, pages 11–56. Kluwer Academic Publishers, 2003.

D. D. Lewis and K. Sparck Jones. Natural language processing for information retrieval. *Communications of the ACM*, 39(1):92–101, 1996. URL `http://citeseer.ist.psu.edu/lewis96natural.html`.

A. Moffat and J. Zobel. Self-indexing inverted files for fast text retrieval. *ACM Transactions on Information Retrieval*, 14(4):349–379, October 1996. URL `http://doi.acm.org/10.1145/237496.237497`.

D. Oard and B. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, 1996. URL http://www.glue.umd.edu/~dlrg/filter/papers/mlir.ps.

D. W. Oard, J. Gonzalo, M. Sanderson, F. López-Ostenero, and J. Wang. Interactive cross-language document selection. *Information Retrieval*, 7(1-2):205–228, 2004.

M. Persin, J. Zobel, and R. Sacks-Davis. Filtered document retrieval with frequency-sorted indexes. *J. of the American Society for Information Science*, 47(10):749–764, October 1996.

J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 275–281, Melbourne, Australia, August 1998. ACM Press, NY. URL http://doi.acm.org/10.1145/290941.291008.

S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proc. 13th Conf. on Information and Knowledge Management (CIKM)*, Washington D.C., November 2004.

S. E. Robertson and K. Sparck Jones. Simple, proven approaches to text retrieval. Technical Report UCAM-CL-TR-356, Cambridge Computer Laboratory, May 1997. URL http://www.cl.cam.ac.uk/TechReports/UCAM-CL-TR-356.pdf.

S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. Probabilistic models of indexing and searching. In *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1981. URL http://portal.acm.org/citation.cfm?id=636673.

S. E. Robertson, S. Walker, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. TREC-3*, November 1994. URL http://trec.nist.gov/pubs/trec3/papers/city.ps.gz. NIST Special Publication 500-225.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5): 513–523, 1988.

L. Schamber, M. B. Eisenberg, and M. S. Nilan. A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management*, 26(6):755–776, 1990.

A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland, August 1996. ACM Press, NY. URL http://doi.acm.org/10.1145/243199.243206.

A. Singhal and F. Pereira. Document expansion for speech retrieval. In *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 34–41, Berkeley, CA, August 1999. ACM Press, NY.

A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (12):1349–1380, December 2000. URL http://ieeexplore.ieee.org/xpl/abs_free.jsp?arNumber=895972.

K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. Parts I and II. *Information Processing and Management*, 36(6):779–840, 2000.

A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 2–10, Melbourne, Australia, August 1998. ACM Press, NY.

C. J. van Rijsbergen. Towards an information logic. In N. J. Belkin and C. J. van Rijsbergen, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 77–86, Cambridge, MA, June 1989. ACM Press, NY.

Various Authors. Collected papers about TREC-2. *Information Processing and Management*, 31(3):269–453, May 1995. URL http://www.sciencedirect.com/science/journal/03064573.

E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 315–323, Melbourne, Australia, August 1998. ACM Press, NY.

E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

O. Zamir and O. Etzioni. Grouper: A dynamic clustering interface to web search results. In *Proc. 8th World Wide Web Conf. (WWW8)*, 1999. URL http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html.

C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New Orleans, LA, September 2001. ACM Press, NY.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, NY.

J. Zobel and A. Moffat. Exploring the similarity space. *SIGIR Forum*, 32(1):18–34, Spring 1998.