Overview of Special Issue

Donna Harman
National Institutes of Science & Technology
Gaithersburg MD, USA
dch34@cornell.edu

Diane Kelly
University of Tennessee
Knoxville TN, USA
dianek@utk.edu

(Editors)

Authors: James Allan, Nicholas J. Belkin, Paul Bennett, Jamie Callan, Charles Clarke, Fernando Diaz, Susan Dumais, Nicola Ferro, Donna Harman, Djoerd Hiemstra, Ian Ruthven, Tetsuya Sakai, Mark D. Smucker, Justin Zobel.

Introduction

This special issue of SIGIR Forum marks the 40th anniversary of the ACM SIGIR Conference by showcasing papers selected for the ACM SIGIR Test of Time Award from the years 1978-2001. These papers document the history and evolution of IR research and practice, and illustrate the intellectual impact the SIGIR Conference has had over time.

The ACM SIGIR Test of Time Award recognizes conference papers that have had a long-lasting influence on information retrieval research. When the award guidelines were created, eligible papers were identified as those that were published in a window of time 10 to 12 years prior to the year of the award. This meant that the first year this award was given, 2014, eligible papers came from the years 2002-2004. To identify papers published during the period 1978-2001 that might also be recognized with the Test of Time Award, a committee was created, which was led by Keith van Rijsbergen. Members of the committee were: Nicholas Belkin, Charlie Clarke, Susan Dumais, Norbert Fuhr, Donna Harman, Diane Kelly, Stephen Robertson, Stefan Rueger, Ian Ruthven, Tetsuya Sakai, Mark Sanderson, Ryen White, and Chengxiang Zhai.

The committee used citation counts and other techniques to build a nomination pool. Nominations were also solicited from the community. In addition, a sub-committee was formed of people active in the 1980s to identify papers from the period 1978-1989 that should be recognized with the award. As a result of these processes, a nomination pool of papers was created and each paper in the pool was reviewed by a team of three committee members and assigned a grade. The 30 papers with the highest grades were selected to be recognized with an award.

To commemorate the 1978-2001 ACM SIGIR Test of Time awardees, we invited a number of people from the SIGIR community to contribute write-ups of each paper. Each write-up consists of a summary of the paper, a description of the main contributions of the paper and commentary on why the paper is still useful. This special issue contains reprints of all the papers, with the

exception of a few whose copyrights are not held by ACM (members of ACM can access these papers at the ACM Digital Library as part of the original conference proceedings)[1].

As members of the selection committee, we really enjoyed reading the older papers. The style was very different from todays SIGIR paper: the writing was simple and unpretentious, with an equal mix of creativity, rigor and openness. We encourage everyone to read at least a handful of these papers and to consider how things have changed, and if, and how, we might bring some of the positive qualities of these older papers back to the SIGIR program.

Test of Time Awardees, 1978-2001

Stephen E. Robertson, C. J. (Keith) van Rijsbergen, and Martin. F. Porter. 1980. Probabilistic models of indexing and searching*. In Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (SIGIR '80). Butterworth & Co., Kent, UK, 35-56. ACM: http://dl.acm.org/citation.cfm?id=636673

Comments by: Djoerd Hiemstra

Brief summary: The paper discusses and evaluates approaches to estimating query term weights using probabilistic retrieval approaches. Approaches include simple *idf* weighting, variations of the binary independence model, and variations of the 2-Poisson model, for which the paper develops a new weighting scheme. The paper shows there is benefit gained from using term frequencies based on the 2-Poisson model, although its performance is "slightly disappointing" and does not beat simpler methods.

Contributions: The major theoretical contribution of the paper is the development of a practical weighting scheme for the 2-Poisson model, a model that was suggested some years earlier independently by Abraham Bookstein and Don Swanson [5], and by Stephen Harter [11]. But maybe the biggest contribution of the paper are almost 50 experiments on two test collection under various conditions: with vs. without relevance feedback information, and very interestingly, on train-test splits of test collections.

Why this is still useful to read: This paper shows the importance of properly publishing negative results. Its theoretical analysis of the 2-Poisson model, and its extensive experiments seem to put a final verdict on the usefulness of the 2-Poisson model: In the end, search quality did not improve. The authors point out that estimating the many parameters of the 2-Poisson model is problematic and they suggest to delve more deeply into the estimation problems as a further line of work. Don't think, however, that this further work is only a footnote in the history of IR. Getting the 2-Poisson model to work seems to have haunted Stephen Robertson for many years since 1980, and it ultimately inspired the development of the BM25 term weighting scheme, which

[1]These papers are denoted with * in the undernoted; For these papers, we have provided an additional ACM digital library URL.

he presented with Stephen Walker at SIGIR 1994: It is discussed later in this paper as 1994's SIGIR Test of Time Award.

Stephen E. Robertson, M. E. (Bill) Maron, and William S. Cooper. 1982. The unified probabilistic model for IR*. In Proceedings of the 5th annual ACM conference on Research and development in information retrieval (SIGIR '82). Springer-Verlag New York, Inc., New York, NY, USA, 108-117. DOI: https://doi.org/10.1007/BFb0036342 ACM: http://dl.acm.org/citation.cfm?id=636723

Comments by: Djoerd Hiemstra

Brief summary: The paper proposes models of information retrieval that unify Model 1, the probabilistic indexing model proposed by Maron and Kuhns in 1960 [17] and Model 2, the probabilistic retrieval model proposed by Robertson and Sparck-Jones in 1976 [22]. A new model, Model 3, is derived that is able to use interaction data from the individual user about other documents, as well as interaction data from other users about the individual document.

Contributions: Variants of Model 1 and Model 2 have been independently discovered several times and find applications that are relevant far beyond information retrieval. Today, they are probably best known as statistical language models and naive Bayes classifiers, respectively. Interestingly, gathering the above mentioned interaction data has only recently become feasible when web search engines started to log their user's queries and clicks. Still, unifying the models in a theoretically motivated way is largely unsolved, and the paper's road map is as relevant as it was 35 years ago.

Why this is still useful to read: The paper is a beautiful sign of the times, when SIGIR was a more informal meeting. Clearly written on a typewriter, the paper refers to "the previous speaker" to provide context, after which the first author points out that he, being separated some 6000 miles from his co-authors, should be blamed for the paper's "worst infelicities". The paper invites the reader to critically discuss the implications of unified models. It presents lines of thought and hypotheses that would likely be dismissed as unsubstantiated by today's strict reviewing criteria, but that proved to be indispensable for bringing the field forward. A more extensive and less informal version of the paper was published by Information Technology [21].

Jeffrey Katzer, Judith Tessier, William Frakes, and Padmini Das-Gupta. 1983. A study of the overlap among document representations. In Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '83). ACM, New York, NY, USA, 106-114. DOI: https://doi.org/10.1145/511793.511809

Comments by: Nicholas J. Belkin

Brief summary: This paper reports on a project investigating the difference in documents retrieved (both relevant and not relevant) when different document representations are used. Four expert human search intermediates each conducted four different searches on 52 "real" information problems, elicited from 45 participants, in the PsychInfo database. Each of the four searches was conducted using a different document representation. The results demonstrate that, although precision and recall were similar, no matter the representation, there was little overlap in the (strictly relevant, less strictly relevant) documents retrieved when using the different representations.

Contributions: The study reported on in this paper was the first research, and this one of the first publications, to investigate whether different representational techniques for IR systems which led to similar effectiveness results, using the standard measures of recall and precision, actually did perform equally. The authors showed, for the first time, that different techniques led to different retrieved documents; this finding is the foundation for all subsequent research in IR on combination of evidence from different IR systems, and, indirectly, for theoretical contributions such as the principle of polyrepresentation. It is also an important document in evaluation research, pointing out that merely knowing some overall number(s) with respect to effectiveness, is insufficient for understanding and comparing system performance.

Why this is still useful to read: This paper is a very nice example of the need to question received wisdom in IR research, and, in particular, of the importance of understanding what actually lies beneath the surface of seemingly simple results. In this case, the authors were initially concerned with coming to a better understanding of what recall and precision were actually measuring, and how better to compare system performance. What they found, through detailed analysis, was something perhaps far more important in the long run; this paper is thus also a nice example of discovery beyond the original intentions of the research. But most of all, it's still useful to read as a foundational document.

Ellen M. Voorhees. 1985. The cluster hypothesis revisited. In Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '85). ACM, New York, NY, USA, 188-196.
DOI: https://doi.org/10.1145/253495.253524

Comments by: Donna Harman

Brief summary: This paper examines a new method for determining if the cluster hypothesis is true, looking at four of the old collections (MED, CACM, CISI and INSPEC), with results that show some differences based on the new test. It then reports on a series of experiments looking at retrieval using clustering for these collections. In particular, single-link clustering is used and two types of clustered searches are tried: one that retrieves entire low-level clusters until the specified number of documents are selected and one that retrieves only the top ranked documents from these clusters. These are compared with a straight sequential search and it is shown that the cluster searches are less effective, even when the cluster hypothesis applies.

Contributions: Because of the slowness of computers in 1980s, there was considerable work in clustering due to the need for efficiency, and there were also suggestions that it was more effective to search clusters. This paper looks at these issues across a series of very different collections and shows that it is important not only that similar relevant documents are clustered, but that similar non-relevant documents are excluded, and this is related both to the collection and to the individual queries. It is for this reason that the sequential search generally shows the best results (and therefore more research became channeled away from clustering!).

Why this is still useful to read: This is an in-depth look at the cluster hypothesis, both in how to measure it, but also in how it is related to search performance. There is considerable analysis of these effects across four small but very different collections. It is also an early look into how clustering affects search performance.

C. J. (Keith) van Rijsbergen. 1986. (invited paper) A new theoretical framework for information retrieval. In Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '86). ACM, New York, NY, USA, 194-200. DOI: https://doi.org/10.1145/253168.253208

Comments by: Nicola Ferro

Brief summary: The paper revises the foundations of IR to overcome the limits of purely keyword-based approaches and to introduce a formal notion of meaning and semantics in IR. A completely new formal framework is proposed in the context of logic by defining the Logical Uncertainty Principle which provides the basis for developing IR models rooted in logic reasoning.

Contributions: The paper (and its extension in The Computer Journal [24]) develops a very clean line of reasoning to show that the fundamental retrieval operation is the logical implication. Stemming from this observation, it shows how logical implication is fundamental in both boolean retrieval and coordination level matching. Then the paper explores which kind of logical implication is most suitable for retrieval and it shows that the uncertainty intrinsic in retrieval calls for a non-classical logic. The whole discussion is finally framed by the introduction of the Logical Uncertainty Principle, which grasps the uncertainty of relevance and introduces the notion of minimal revision as a way to actually evaluate a logical implication for IR. Besides its own contributions, the paper also sets the stage for an active area of research in IR devoted to develop a logic for information retrieval, lasting for more than 15 years [8, 9].

Why this is still useful to read: The paper is a masterpiece in its determination to revise the very fundamentals of the discipline and to introduce a new paradigm looking at IR from a different point of view. This is a very important lesson also today since we have keep a critical view on our own foundations and always ask ourselves whether we need to revise and go beyond them. This is even more true if you consider that IR is still lacking models able to fully explain the behaviour and performance of IR systems, as also pointed out by Norbert Fuhr [10]. Moreover, logic represents a

very important link between IR and databases, which make a much more prominent use of it, and a more unified vision to information access, both structured and unstructured, would represent an added value still today. Finally, nowadays, there is an ever growing interest in semantics also in IR - e.g. entity search, entity linking, and more - and some concepts in this paper might be helpful to further contribute to this current area of research.

Joel Fagan. 1987. Automatic phrase indexing for document retrieval. In Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '87). ACM, New York, NY, USA, 91-101. DOI: https://doi.org/10.1145/42005.42016

Comments by: Donna Harman

Brief summary: The experiments in this paper cover multiple ways of defining and using non-syntactic phrases as concepts in document retrieval, using 5 old test collections (CRAN, CACM, INSPEC, MED, and CISI). There is an extremely detailed analysis of why this (mostly) does not improve performance over a basic tf*idf run, and some suggestions of possible solutions.

Contributions: This paper is an extensive analysis of the issues in the use of nonsyntactic phrases across multiple test collections, including a very comprehensive comparison with related work and extensive failure analysis. The conclusions drawn from these experiments illustrate the difficulties even today in the use of phrases to index documents and led to a shift of research to more productive searching methods.

Why this is still useful to read: In addition to the deep discussion of this type of phrase indexing, the paper is a classic example of how to design and run complex and complete experiments, along with how to analyze the results and present them in a comprehensive manner. Because these collections were so small, it was possible to deconstruct the searches to understand the real issues with retrieval. Note that a slightly extended version of this paper was published by the Journal of Association for Information Science, March 1989, pages 115-132.

Karen Sparck Jones. 1988. A look back and a look forward. In Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '88). ACM, New York, NY, USA, 13-29. DOI: https://doi.org/10.1145/62437.62438

Comments by: Justin Zobel

Context: This article was written when information retrieval had already been developing for 30 years, but just before a collapse in the cost of (and rise in the size of) disk storage, the advent of the

web, and the arrival of high-performance desktop computing. These factors would dramatically change information retrieval over the following years.

Summary & Contribution: This paper is a retrospective on a thread of research, at that time already 25 years deep, coupled to forward-looking speculation on promising avenues and reflections on why some directions might not be worth pursuing. But this paper is not easy to summarise! It is a conversation ranging across many aspects of information retrieval, and how research in the field might be pursued, and needs to be read carefully to appreciate its breadth of insight. It highlights the extent to which foundational and insightful ideas were explored in the first decades of research in the field.

Why this is still useful to read: It offers a perspective on the (early) history of information retrieval, and documents the evolution of principles that continue to be of interest. It offers a range of insights into how research should be done some of them based on an open assessment of why some certain work was unsuccessful. Indeed, there are many valuable lessons here, including the value of questioning the assumptions on which work is based, and the fact that ideas often emerge before we have the data, systems, technology, or principles to fully evaluate them. It is an argument for open, reflective approach to research focused on creating understanding, not on short-term achievement which leads to the lack of sustained contribution made by researchers pursuing a succession of quick papers; Sparck Jones called these 'one-off feasibility sketches'.

Most significant of all is the explanation of the significance of experimentation, and that it is critical in grounding speculation and abstraction. As Sparck Jones notes, 'plausible arguments are not enough in IR'. This and other reflections remain pertinent today.

Donna Harman. 1988. Towards interactive query expansion. In Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '88). ACM, New York, NY, USA, 321-331. DOI: https://doi.org/10.1145/62437.62469

Comments by: Ian Ruthven

Brief summary: This experimental paper looks at useful ways to select new query terms to improve an initial query, either when some relevant documents have been found or when no relevant documents have been found. A series of experiments clearly demonstrates the significant potential retrieval improvements that can be gained by viewing relevance feedback as an interactive process.

Contributions: This paper has many contributions based on the core aim of understanding how to rank expansion terms to a manageable list from which a searcher can select a subset to add to an initial query. The clear outcomes are that small lists of expansion terms can be more effective than long lists of expansion terms, that searcher selection of expansion terms can be more effective than automatic expansion and that searcher interaction can be encouraged by such system guidance in query modification.

Why this is still useful to read: This is a very elegant paper in a classical experimental IR tradition. It takes a clear research question and systematically unravels the various sub-questions to provide a comprehensive set of results. The experimentation is based on a solid and objective view of the state of the art at the time and, as well as answering many of its questions, it also opens up a whole new area of interactive IR research, namely how to offer interactive query expansion terms to a searcher.

George W. Furnas, Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter, and Karen E. Lochbaum. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '88). ACM, New York, NY, USA, 465-480. DOI: https://doi.org/10.1145/62437.62487

Comments by: Fernando Diaz

Brief summary: The authors use the singular value decomposition of the term-document collection matrix in order to address problems such as vocabulary mismatch. The approach allows a shared lower-dimensional representation of terms and documents. Experiments include both standard ad hoc text retrieval as well as expert finding.

Contributions: Latent semantic analysis contributed a well-grounded approach to joint term and document representation in a lower dimensional space. Although LSA arrived after a long history of work in document and term clustering, the majority of this prior research independently studied either document or term clustering. In addition to initiating research in similar approaches based on linear algebra, this paper has direct connection to modern techniques like probabilistic latent semantic analysis [13], latent Dirichlet allocation [4], and distributed term representations [18].

Why this is still useful to read: As if contributing an entirely new formalism were not enough, the authors also offer a model for a solid SIGIR paper, including a grounding in lab studies, theoretical analysis, and extensive experimentation. Concepts in this paper continue to resonate within the information retrieval, natural language processing, and machine learning communities. That said, the extensive subsequent work from the authors is also worth revisiting since it provides a better understanding of when LSA works and when it does not.

Richard K. Belew. 1989. Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. In Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '89). ACM, New York, NY, USA, 11-20. DOI: https://doi.org/10.1145/75334.75337

Comments by: Mark D. Smucker

Brief summary: This paper introduces a connectionist (neural network) representation for information retrieval that adapts to usage with the goal of improving its retrieval quality. The connectionist approach is fully taken with words, documents, and authors all interconnected and part of the retrieval. After a user enters a query, a spreading activation retrieval is performed and shown visually to the user. The user then browses for information and provides relevance feedback on the various network nodes that deserve more and less weight. The feedback not only is used to refine the results, but it is also used to modify the neural network's connection weights.

Contributions: The chief contribution of this paper is its demonstration of how an information retrieval system could learn about the relationships between words, documents, authors, etc. given the interactions that users have with the system. Not only could the system learn which documents were considered relevant to which queries, the system also could learn stems of words and other associations merely from repeated usage of many users. An important distinction from other relevance feedback work is that Belew's system foresaw how today's web search engines would utilize the interactions of millions of users to adapt their various functions. A simple search of a patent database in 2017 shows the numerous adaptive IR techniques employed by today's web search engines.

Why this is still useful to read: The paper is rich with ideas and insightful discussion. The user interface is radically different from today's ten blue links and is supportive of a more relaxed search experience were feedback is integral to browsing and search. The user interface also provides a direct view for the user into the mechanism of retrieval and why certain documents are retrieved. Likewise, while we typically think of relevance feedback at the document level, here we see the idea of a user providing feedback at the feature level.

Annelise Mark Pejtersen. 1989. A library system for information retrieval based on a cognitive task analysis and supported by an icon-based interface. In Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '89). ACM, New York, NY, USA, 40-47. DOI: https://doi.org/10.1145/75334.75340

Comments by: Ian Ruthven

Brief summary: This paper shows that studying how and why people use specific strategies when interacting with information can inform the design of interfaces that support users natural strategies for finding information. Leading from a series of task analyses, this paper proposes a novel metaphorical interface for interacting with fiction and describes its evaluation within public library settings.

Contributions: This paper is (still) rare in starting from a principled series of cognitive studies in operational information environments that led to a set of design principles that were manifested in a novel interactive system. This system, the BOOKHOUSE system, offers an information system that was radically different from what went before in terms of design and interaction and was then evaluated over a six-month period.

Why this is still useful to read: This paper has three valuable contributions. Firstly, it shows the progression of an intellectual journey from studying cognitive needs in real-life search environments to interaction designs based on ways users would find natural to interact with information systems. Secondly, it shows that search systems can be radically different in design by using metaphors that people would find comfortable, familiar and help them make use of their existing search skills. Thirdly, it was one of the few systems to be solidly evaluated within a real-life library environment, over a sustained period, and in the environment in which the original ideas originated.

Howard Turtle and W. Bruce Croft. 1990. Inference networks for document retrieval. In Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '90). ACM, New York, NY, USA, 1-24. DOI: https://doi.org/10.1145/96749.98006

Comments by: Jamie Callan

Brief summary: This paper defines the inference network approach to document retrieval. It addresses the use of multiple representations of the document content; document attributes; dependencies among documents; multiple representations of the information need by query variants; and structured queries with probabilistic query operators. It also shows how several earlier probabilistic models are subsumed by this new approach to probabilistic retrieval.

Contributions: This paper integrated into a single, well-organized probabilistic framework a set of ideas that the IR community had been exploring for quite some time. It was the theoretical foundation for the widely-used Inquery and Indri search engines, which were unique in the research community for their ability to easily mix multiple representations of document content, complex document structure, and complex query structure; and for West Publishings WIN search engine, which was one of the first large commercial natural language search engines. Even today, there are few competing retrieval models that offer such a comprehensive set of capabilities.

Why this is still useful to read: Few retrieval model papers address such a broad range of topics; even fewer do so as concisely. Most of the issues and forms of evidence addressed by this paper are still relevant today.

Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. Scatter/Gather: a cluster-based approach to browsing large document collections. In

Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '92). ACM, New York, NY, USA, 318-329. DOI: https://doi.org/10.1145/133160.133214

Comments by: Susan Dumais

Brief summary: The paper describes Scatter/Gather, a technique that uses document clustering to support browsing over a large document collections. Initially, a collection of documents is "scattered" into clusters, then related clusters are gathered, this subset is scattered, etc. To support interactive browsing, the authors developed a fast clustering algorithms and techniques to summarize the content of clusters.

Contributions: In information retrieval, clustering had traditionally been used to improve the efficiency or the effectiveness of document retrieval. Scatter/Gather showed how document clustering could be used to support interactive browsing of large collections. To enable real-time interaction, they developed a linear-time algorithm, Buckshot, which clusters a random subset of documents and the results to form the initial cluster centers. They used a slower but more accurate technique for identifying cluster centers, Fractionization, for the initial clustering of the collection.

Why this is still useful to read: An important motivation for Scatter/Gather was to support a continuum from browsing to searching, and this is still an interesting but unsolved area. The 1992 paper focused on the browsing scenario, and a subsequent paper (SIGIR 1996) showed how Scatter/Gather could be used to organize subsets of documents returned by more focused searches.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers*. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94). Springer-Verlag New York, Inc., New York, NY, USA, 3-12. DOI: https://doi.org/10.1007/978-1-4471-2099-5_1 ACM: http://dl.acm.org/citation.cfm?id=188495

Comments by: Fernando Diaz

Brief summary: The authors introduce uncertainty sampling for active learning of text classifiers. The technique, which can apply to any model that outputs class probabilities, results in improved performance for sequential relevance feedback tasks, where a model incorporates feedback incrementally. Experiments demonstrate that uncertainty sampling outperforms baselines (random- and relevance-based sampling) with far less data. This paper should be read with its corrigendum, published in SIGIR Forum the following year, which describes a bug in the experiments and corrects results.

Contributions: Prior to this work, gathering training data for text classification domains with skewed class distributions followed sampling either randomly, by relevance, or according to heuristics. Lewis and Gale present an elegant, intuitive, simple idea for active learning. Although other approaches have followed since-primarily in the machine learning community-uncertainty sampling is a concept present in many and a baseline to compare against. This work also is one of the few I have read where the authors discovered and published a corrigendum to their original paper.

Why this is still useful to read: Uncertainty sampling remains an important and basic concept in active learning. While perhaps not the strongest active learning algorithm, it provides a gentle introduction. More importantly, the scientific honesty of the authors in publishing a corrigendum should be encouraged within the community. There are likely many published results that deserve similar companion articles.

Stephen E. Robertson and Stephen Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval*. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94). Springer-Verlag New York, Inc., New York, NY, USA, 232-241. DOI: https://doi.org/10.1007/978-1-4471-2099-5_24 ACM: http://dl.acm.org/citation.cfm?id=188561

Comments by: Charles Clarke

Brief summary: A classic of probabilistic information retrieval, the paper starts with theoretical insights into weighting functions for retrieval and proceeds to develop these insights into the BM (Best Match) family of ranking functions. The first insight relates to term relevance saturation, such that the influence of term frequency should tend to an asymptotic maximum. The second insight leads to a document length normalization method that reflects a trade-off between scope and verbosity. The proposed ranking functions are tested on TREC-2 collections, showing substantial improvements over previous probabilistic methods.

Contributions: Soon after the publication of this paper, the family would grow to include BM25 (introduced at TREC-3 in November 1994) and later BM25F (at CIKM 2004). These ranking functions have proven themselves time and again, on many test collections and publications. Lucene and most other open source search engine support BM25 as a key ranking function. In most learned rankers, both academic and commercial, BM25 typically emerges as feature of high importance.

Why this is still useful to read: Apart from the tremendous practical impact of BM25, the paper remains a joy to read. The careful progression from step to step, and the use of the 2-Poisson model as inspiration for the term saturation model shows the clear connections between the theory and the highly practical outcome. Anyone involved in search, either from an academic standpoint or an industrial standpoint, will regularly encounter BM25. Without an understanding of the

theory, the details of BM25 may seem arbitrary. After reading the paper, the elegant simplicity of the method becomes clear.

James P. Callan, Zhihong Lu, and W. Bruce Croft. 1995. Searching distributed collections with inference networks. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '95). ACM, New York, NY, USA, 21-28. DOI: https://doi.org/10.1145/215206.215328

Comments by: James Allan

Brief summary: This paper presents a model for addressing the distributed information retrieval problem and carries out an extensive set of experiments to demonstrate the models effectiveness. The model is built upon the formal inference network approach to information retrieval embodied in the INQUERY retrieval system. The experiments are based on a 3-gigabyte collection of just over a million documents, divided into 17 sub-collections by document source.

Contributions: The paper develops and evaluates methods for collection representation, collection ranking, collection selection, and ranked list merging, yielding solid results. It showed that a document relevance belief function could be adapted to estimate a belief that a collection contained relevant documents. It demonstrated that it was not necessary to normalize scores across all collections, but that weighting document scores by the belief in their containing collection was equally good and more efficient. The paper also showed that precision was only slightly impacted when a limited number of collections were selected. A final set of experiments explored limiting the number of terms used to represent a collection (small impact except in extreme settings) and reducing the number of documents retrieved from each collection (minimal impact on recall and precision with a 50% reduction in candidate documents retrieved).

Why this is still useful to read: Although web-scale search engines have shown that with sufficient resources it is possible and effective to build a monolithic and massive search index, it is still common to have situations where collection information is not easily centralized. This paper lays out a way to think about this broad problem and shows how to evaluate its various aspects. Moreover, this paper is a solid example of how to take a formal approach to a problem (document retrieval) and adapt it to another problem (collection retrieval). Finally, it shows the importance of extensive experiments to validate a model and understand the impact of parameter values.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96). ACM, New York, NY, USA, 4-11. DOI: https://doi.org/10.1145/243199.243202

Comments by: Donna Harman

Brief summary: The paper examines in depth three different methods of query expansion using different test collections. The first is global document analysis which uses the entire corpus to generate concepts using noun groups and the fixed length windows surrounding them to create a concept database (a type of automatic thesaurus). The original queries are then matched against this database to select appropriate concepts for addition to (and/or reweighing of) the query terms. The second method is local feedback (also known as pseudo relevance feedback), in which concepts are generated based on the top retrieved documents. The third method (local context analysis) is a new combination of the two in which the top documents are used to mine top passages, which are then examined for new concepts. Both TREC-3 and TREC-4 test collections were used, along with the WEST collection.

Contributions: In addition to the new method (local context analysis), the paper details many of the issues surrounding these three methods. For the TREC-3 and TREC-4 collections, the local context analysis works very well, (23% improvement), but much less well for the WEST collection. The global analysis works poorly, whereas local feedback works well also for TREC but not for WEST. Some excellent analysis is done as to why the difference in the collections, and it is shown that local context analysis is a more robust method when there are fewer relevant documents in the top set.

Why this is still useful to read: This paper looks into the details of what types of concepts are best for addition to (or reweighting of) query terms. This was done at a time before language modeling came into vogue, and it is interesting to see what concepts are useful and how the different methods find these concepts. Although the two local methods both work well, the concepts that are found by these methods are not the same! The paper also is an excellent example of a thorough investigation of an area, with the use of different collections, and how to analyze the results in order for readers to understand the various issues.

Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96). ACM, New York, NY, USA, 21-29. DOI: https://doi.org/10.1145/243199.243206

Comments by: Mark D. Smucker

Brief summary: Dealing with differing document lengths is an important part of retrieval methods. Pivoted document normalization allows cosine similarity to prefer certain documents using the known probability of relevance given a document's length, which can be computed from a test collection.

Contributions: A significant contribution of this paper is that it makes clear that characteristics of documents can influence their probability of relevance independent of the query. Here, the document feature of interest is document length, but the same idea can be applied to many different attributes, e.g. the spaminess of a web page should decrease its probability of relevance.

Why this is still useful to read: This paper makes clear that retrieval methods can have inherent biases built into them. For example, relative to their prior probability of relevance, cosine similarity favors short documents, and this bias was corrected using pivoted length normalization. Once the authors discovered this phenomena and corrected for it, several other methods were either found to only be effective because they had indirectly corrected for it or the methods were found to improve since they no longer had to compensate for the issue [6, pg. 308].

James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 37-45. DOI: https://doi.org/10.1145/290941.290954

Comments by: Fernando Diaz

Brief summary: The authors present methods for monitoring a stream of text news articles. The authors develop tasks beyond classic filtering by focusing on discovery (i.e. event detection) and organization (i.e. event tracking). The work summarizes the results of the pilot topic detection and tracking (TDT) program.

Contributions: Before this paper, most of the work with text streams followed Luhn's model of information filtering [15]. This allowed the community to use very similar techniques to those used in ad hoc retrieval [3]. The TDT work first introduced in this paper presented unique aspects of the streaming task. To date, the evaluation methodologies used for streaming text analysis (e.g. temporal summarization, realtime summarization) appeal to insights from TDT. Algorithms for these tasks use concepts first presented in this work.

Why this is still useful to read: Increasingly streamed data (e.g. social media) has resulted in renewed interest in TDT, even though the core tasks remain much less well understood than ad hoc retrieval. This paper provides a primer for how to evaluate and build these systems.

Krishna Bharat and Monika R. Henzinger. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 104-111. DOI: https://doi.org/10.1145/290941.290972

Comments by: Charles Clarke

Brief summary: Written at a time when Web link analysis was still emerging as a ranking method, the paper extends one of these early methods (Kleinberg, 1998) in three important ways: 1) reducing the influence of a single node, 2) weighting nodes by document-query similarity,

and 3) graph pruning. The work is evaluated on a realistic web collection, showing substantial improvements in precision.

Contributions: While the specific ideas in the paper have been extended and superseded by other methods, the paper informed and inspired much of this work, with over 1000 citations on Google Scholar.

Why this is still useful to read: In introductory courses, link analysis is often presented through the lens of the original PageRank algorithm. Given the great success of PageRank as the foundation for Google, students often overlook the large body of related work emerging around the same time. A deeper understanding of link analysis requires a return to classic papers, such as this one, particularly since they may point towards research directions that were not fully explored.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 335-336. DOI: https://doi.org/10.1145/290941.291025

Comments by: Jamie Callan

Brief summary: This paper defines Maximum Marginal Relevance (MMR), a technique for list re-ranking that reduces redundancy without reducing relevance. It discusses how to use MMR to improve search engine rankings, and to improve single- and multi-document extractive summarization. This two-page poster abstract shows that short papers can have a major impact on the field.

Contributions: Legend has it that MMR was inspired by frustration with the long lists of nearly-identical results returned by the commercial search engines of that time. MMR was a simple technique that improved the quality of search results and extractive text summarization, as well as inspiring a large body of work on diversification. For many years, it was the baseline to beat.

Why this is still useful to read: The paper is short, clear, and straight to the point. Two problems that seem very different are shown to be ranking problems that benefit from diversity-based re-ranking.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 275-281. DOI: https://doi.org/10.1145/290941.291008

Comments by: Djoerd Hiemstra

Brief summary: Inspired by models used in speech recognition, Ponte and Croft infer a model of language generation for each document and estimate the probability of generating the query according to each of the models. Documents are ranked by these probabilities. The paper shows that the language modeling approach significantly outperforms standard tf.idf weighting on two TREC collections and query sets.

Contributions: The contribution of language models to information retrieval cannot be overstated. The Ponte and Croft 1998 paper set the stage for language models in the field by introducing generative models for query generation, the use of simple tf-estimators, and the need for probability smoothing. In that same year, two research groups (BBN and Twente/TNO who both independently developed similar models) tested language models successfully at TREC. By 2001, SIGIR organized two sessions on language models and in 2003 the main research roadmap for the field was called "Challenges in Information Retrieval and Language Modeling" [2]. The language modeling approach contributed to several practical implementations, including Lemur, Lucene and Terrier.

Why this is still useful to read: The paper does an excellent job in motivating a new way of thinking about models of information retrieval by putting the language modeling approach into context of the literature at the time. Language models are simple models for probabilistic indexing that need very few assumptions, for instance, they do not need parametric distributions, they do not assume documents are members of predefined classes, and they do not even explicitly model relevance. Language model's lack of these assumptions mark an important departure of indexing models like the 2-Poisson model which inspired for instance the BM25 term weighting formula.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). ACM, New York, NY, USA, 50-57. DOI: https://doi.org/10.1145/312624.312649

Comments by: Susan Dumais

Brief summary: The paper describes Probabilistic Latent Semantic indexing (PLSI), a variant of Latent Semantic Indexing (LSI), that is based on a statistical latent class or aspect model for co-occurrence data. When applied to term-document matrices in information retrieval, PLSI led to improvements in retrieval accuracy compared to vector space term-matching and LSI models on four test collections (MED, CRAN, CACM, CISI).

Contributions: In contrast to LSI, PLSI provides a probabilistic formalism for reduced-dimensional representations as well as a generative model of the documents in the collection. The paper also

develops a generalization of maximum likelihood estimation, called tempered EM, for mixture models. Experimental results show consistently better than LSI or term-matching,

Why this is still useful to read: The paper provides a nice overview of the advantages of using probabilistic models for dimension reduction. PLSI has been applied broadly beyond document retrieval, including to language modeling and collaborative filtering. In addition, PLSI is closely related to nonnegative matrix factorization (NMF) and one of the inspirations for Latent Dirichlet Allocation (LDA).

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). ACM, New York, NY, USA, 222-229. DOI: https://doi.org/10.1145/312624.312681

Comments by: James Allan

Brief summary: This paper develops an approach to information retrieval based on statistical machine translation, where the query is viewed as a "translation" of an imagined relevant document. The paper explores a few approaches based on IBMs Model 1 approach (Brown et al, Computational Linguistics 1990) that leverages unigram translation probabilities. The authors create synthetic query-document pairs to determine translation probabilities. The resulting approaches consistently out-perform a tf*idf baseline.

Contributions: This paper successfully tackles the task of treating monolingual information retrieval as a statistical translation process. It demonstrates that this approach is effective. It develops an approach to constructing training data that is "synthetic" in that it does not use annotation, but is nonetheless effective for this task. This approach is similar in spirit and is motivated by the language modeling approach of Ponte and Croft (SIGIR 1998), but the translation model allows it to directly incorporate synonymy and polysemy and thus better bridge the so-called vocabulary gap. In that sense, it is an instance of query expansion in the language modeling framework.

Why this is still useful to read: This paper is an excellent example of bridging two computational areas, deriving a formal model from statistical machine translation to be used for information retrieval, implementing the model, constructing required training data, and evaluating the results. Even though the idea of using translation within a single language feels counterintuitive initially, the models are clean and easy to understand, making this an excellent introduction to the area. Readers of this paper should be motivated to consider the relationship between this approach and various query expansion techniques.

Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In Proceedings of

the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). ACM, New York, NY, USA, 230-237. DOI: https://doi.org/10.1145/312624.312682

Comments by: Paul Bennett

Brief summary: This paper presents an empirical comparison between several neighborhood-based recommendation methods and investigates the impact of several design choices before making practical recommendations for use in recommender systems. The paper described a framework for neighborhood-based recommendation given a set of user-item ratings as: (1) weighting users based on similarity to current user; (2) selecting a subset of users (neighbors) to use to predict (for current user and possibly current item); (3) predicting using a weighted combination of similar user's ratings (possibly with normalization). The authors then investigate the choice of similarity measures, use of variance weighting, significance weighting, and prediction combination method. Several popular choices of similarity measures were investigated and both Pearson's and Spearman's correlation were shown to be better than other investigated choices and comparable to each other. More interestingly, the authors noted that many similarity measures do not account for the amount of data that are used in computing the similarity and proposed significance weighting which gives weight dependent on the amount of data. While the particular significance weighting here (giving low weight to any pair of users whose number of common ratings were below 50) are simple, the notion of significance weighting remains important in neighborhood-based methods. In computing similarity, the authors also investigated variance weighting where the score increases the influence of items with high variance over the entire population when computing the similarity with the intuition such items are segmenting the audience. For example, nearly everyone likes Titanic but Sleepless in Seattle has high variance since it separates those who prefer action movies to those who prefer romantic movies. The methods the authors investigated for variance weighting showed no effect but the effort to identify key items that segment the population or could be used for rapidly profiling a user in a cold-start scenario remain relevant [23]. Finally, when producing the prediction, users scores may be centered around different points. So, when combining predictions, two approaches are deviation from the mean prediction of a user versus a variance normalized z-score. If some users effectively have a "smaller scale" the z-score variant upweights deviations from their mean to be comparable to users who have a "larger scale".

Contributions: In addition to being one of the first empirical analyses of the performance of recommender systems over (at the time) a large amount of historical data from users of a movie prediction site (1K users, 100K ratings), this paper also demonstrated that significance weight could be more impactful on neighborhood-based recommendation than the choice of similarity measure. While both methods (mean-difference and z-score) of dealing with different users effective scales were comparable here, calling out the need to deal with these differing scales was in important contribution.

Why this is still useful to read: This is a seminal paper on neighborhood-based collaborative filtering. Ning et al [19] has a recent survey of the state-of-the-art and current applications in neighborhood-based recommendation. In that paper, they argue that while model-based methods

have outperformed neighborhood methods for prediction accuracy neighborhood methods remain relevant for their ability to better provide serendipitous recommendations and leverage local similarity as well as their simplicity, justifiability, efficiency and stability. The efficiency may be particular useful in interactive or session-based recommender systems. In modern neighborhood-based methods, the notions of similarity can be implicit, e.g. as in graph methods which use random walks over the bipartite user-item graph. Significance weighting, the importance of which was demonstrated in this paper, continues to be important for both neighborhood-based recommendation [19] and the IR community [16]. Furthermore, this was the first work to acknowledge and use the spread in individual rating scales through z-score normalization. While z-score showed no improvement over mean-difference here it was later showed to have an impact [12]. Interestingly in the future work section, the authors also mention that applications of SVD to recommendation were being investigated hinting at the beginning of matrix factorization approaches to recommendation. Finally, the paper is one of the earliest uses of coverage of all items in recommender systems which would later be recast as a notion of aggregate (population-level) diversity which measures how much the entire inventory has been seen by users [1].

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '00). ACM, New York, NY, USA, 33-40. DOI: https://doi.org/10.1145/345508.345543

Comments by: Justin Zobel

Context: When TREC commenced in 1992, with collections two orders of magnitude larger than had been routinely available to researchers at the time, the impact on IR research was immediate. But in some ways the community was not ready for this leap. Not only were there technical obstacles for many laboratories, but much of the knowledge of experimental design was based on very small collections. The team that developed the TREC experimental methodology made a range of decisions based on best information available it is remarkable how many of these remain valid. But in those initial years of TREC, some of the decisions simply had to be taken on faith; there was insufficient evidence to determine if they were the best practice, or the best choices available within the programs fixed resources. Even in this paper from 2000, eight years later, these practices could still be accurately described as rules-of-thumb.

Summary & Contributions: In this paper, Buckley and Voorhees showed that the TREC data (runs of queries by systems, combined with relevance judgments) could be used to make sensitive assessments of the reliability of the evaluation measures. Some of the specific conclusions have (of course!) been superseded by newer work, but this paper helped establish a methodology for considering how system performance should be evaluated, and also helped consolidate evaluation as a field of study in its own right.

Why this is still useful to read: Like many papers that have endured, it is solidly founded on reflection and analysis. The authors have not simply attacked a current problem, but have sought

to understand ways in which current practice might be flawed, and have developed methodologies to examine whether those flaws are indeed present.

Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '00). ACM, New York, NY, USA, 41-48. DOI: https://doi.org/10.1145/345508.345545

Comments by: Tetsuya Sakai

Brief summary: The authors propose IR evaluation methods that consider graded relevance. Their first proposal is to draw precision-recall curves for each relevance level separately. Their second proposal is to compute cumulative gain (CG) and discounted cumulative gain (DCG) and to visualise the gap between the (D)CG curve for the system being evaluated and that for an ideal ranked list. A case study concerning structural queries is reported.

Contributions: This SIGIR 2000 paper proposed CG and DCG; their follow-up ACM TOIS paper proposed normalised CG (nCG) and normalised DCG (nDCG) in 2002. While nCG is basically the same as normalised Sliding Ratio proposed by Pollock in 1968 (where an ideal ranked list, which Pollock calls the master list, is defined with "all documents in the library" (!) "ordered in decreasing master value"), the novel idea of nDCG brought about a paradigm shift: IR researchers started to take up evaluation and optimisation based on graded-relevance measures instead of adhering to classical binary-relevance measures such as recall and precision. Indeed, the proposal of DCG in the SIGIR 2000 paper was extremely timely: according to Hawking and Craswell (The "TREC book", Chapter 5. p.204, 2005), at the Infonortics Search Engines meeting held in April 2000, "TRECers" and a panel of search engine representatives got together, where the latter stressed the importance of relevance assessments with multiple levels. Several years later, a version of nDCG (See [7]) became the de facto standard in web search evaluation as well as many other IR evaluation tasks.

Why this is still useful to read: While IR system evaluators tend to "forget about the user" (or at least, try to eliminate user factors) and focus on single numbers such as nDCG, the paper stresses the importance of visualisation from the user's point of view. Figure 3 in this paper is always worth going back to—the DCG graph is much more intuitive than a precision-recall graph in which precision is interpolated for each recall value and do not necessarily represent anything practical for the user. Finally, there is this perfect statement in the paper: "It is a consistent and statistically significant difference but are the users able to notice it?" Thus, the authors clearly distinguish between statistical significance and practical significance, and stress that the latter is what really matters.

John Lafferty and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In Proceedings of the 24th annual

international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). ACM, New York, NY, USA, 111-119. DOI: https://doi.org/10.1145/383952.383970

Comments by: Nicola Ferro

Brief summary: The paper improves language models for IR by formulating them into an elegant risk minimization framework which allows us to derive both traditional probabilistic models and different types of language models. Moreover, the paper introduces a clean way to perform query expansion by leveraging the proposed framework and to address the training problems which previous approaches suffered from.

Contributions: The paper significantly contributed to the shaping of language models. In particular, it introduces a risk minimization framework based on the Bayesian decision theory which provides a unified vision on several IR models, ranging from the vector space model and classical probabilistic models to other language models, such as the seminal one by Ponte & Croft [20]. The benefits of this new framework go beyond providing a unified vision of different IR models. Indeed, it introduces for the first time the notion of a query language model, which complements the document language model, and proposes the Kullback-Leibler divergence between these two models as a means to rank documents. This new framework opened up a further possibility for a better and more naturally modelling of query expansion. Finally, it exploits Markov chains to estimate model parameters, reducing the need of huge amounts of training data.

Why this is still useful to read: The paper represents a cornerstone in the evolution of language models and provides an extremely valuable example on how to extend an existing framework in a clean and sound way. Moreover, the proposed extensions have been leveraged and adapted to many other contexts, such as entity search, multilingual information retrieval, or ranking on RDF graphs. Finally, the idea of using Markov chains to estimate model parameters might be further explored also today, for example, to incorporate other kinds of user dynamics into language models, such as signals coming from user interaction with an IR system.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). ACM, New York, NY, USA, 120-127. DOI: https://doi.org/10.1145/383952.383972

Comments by: Fernando Diaz

Brief summary: The authors present two methods for pseudo-relevance feedback in the language modeling approach to retrieval. Experiments demonstrate efficacy for retrieval and topic tracking.

Contributions: Beyond the theoretical contribution of incorporating relevance into language modeling, the relevance model method provides strong empirical performance. The first method, RM1, reduces to a weighted combination of the top retrieved document vectors, making implementation trivial for most retrieval systems. If one interpolates the original query model with the relevance model, performance is frustratingly hard to beat. Moreover, contrary to the intuition that query expansion should be conservative, adding a few words from a few documents, Lavrenko and Croft advocate massive query expansion with tens to hundreds of words and as many documents.

Why this is still useful to read: Relevance models remain, especially in their RM3 form, state of the art performance for most retrieval tasks. This paper and its extended treatment in "A Generative Theory of Relevance" [14] provide insight into how to perform effective automatic query expansion. I have run into many papers that include relevance model runs with poorly chosen hyperparameters, demonstrating clear misunderstanding of the algorithm.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). ACM, New York, NY, USA, 334-342. DOI: https://doi.org/10.1145/383952.384019

Comments by: Jamie Callan

Brief summary: This paper is a careful study of the role of smoothing in the query-likelihood retrieval model. It studies three smoothing methods (Jelinek-Mercer, Dirichlet, absolute discounting), queries of different lengths, and collections with different average document lengths to tease apart the effectiveness and contribution of each smoothing method.

Contributions: The role of smoothing in retrieval models was not well-understood when this paper was published. Zhai and Lafferty showed that smoothing plays two important roles. Query modeling explains the importance of each term to the query; this role is similar to idf. Estimation improves the maximum likelihood probability estimates in document language models. The paper shows that query modeling is essential for long queries, but less important for short queries. Jelinek-Mercer smoothing is shown to be more effective for query modeling; and Dirichlet smoothing for improving document language models. The paper discusses the sensitivity of each smoothing method to parameter values, and suggests effective ranges for each parameter.

Why this is still useful to read: This paper encourages one to think about why smoothing in the query likelihood model works, and how it can be improved. It justifies the default smoothing parameter settings that are used by many researchers. It also reminds us that default values, which many of us use without much thought, aren't good choices for all queries and collections.

References

[1] Gediminas Adomavicius and YoungOk Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Trans. on Knowl. and Data Eng.*, 24(5):896–911, 2012.

[2] James Allan et al. Challenges in information retrieval and language modeling. In *ACM SIGIR Forum*, volume 37, pages 31–47. ACM, 2003.

[3] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38, 1992.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[5] Abraham Bookstein and Don R. Swanson. Probabilistic models for automatic indexing. *Journal of the Association for Information Science*, 25(5):312–316, 1974.

[6] Chris Buckley. *The SMART project at TREC*, pages 301–320. MIT Press, 2005.

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 89–96, New York, NY, USA, 2005. ACM.

[8] F. Crestani, M. Lalmas, and C.J. van Rijsbergen. *Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information*. The Information Retrieval Series. Springer US, 2012.

[9] Fabio Crestani, Sandor Dominich, Mounia Lalmas, and Cornelis Joost van Rijsbergen. Mathematical, logical, and formal methods in information retrieval: An introduction to the special issue. *J. Am. Soc. Inf. Sci. Technol.*, 54(4):281–284, 2003.

[10] Norbert Fuhr. Salton award lecture information retrieval as engineering science. *SIGIR Forum*, 46(2):19–28, 2012.

[11] Stephen P. Harter. A probabilistic approach to automatic keyword indexing. part i and ii. *Journal of the Association for Information Science*, 26(5):197–206, 280–289, 1975.

[12] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[13] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[14] Victor Lavrenko. *A Generative Theory of Relevance*. PhD thesis, University of Massachusetts Amherst, 2004.

[15] H. P. Luhn. A business intelligence system. *IBM J. Res. Dev.*, 2(4):314–319, 1958.

[16] Hao Ma, Irwin King, and Michael R. Lyu. Effective missing data prediction for collaborative filtering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 39–46, New York, NY, USA, 2007. ACM.

[17] Melvin E. Maron and J. Lary Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244, 1960.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[19] Xia Ning, Christian Desrosiers, and George Karypis. *A Comprehensive Survey of Neighborhood-Based Recommendation Methods*, pages 37–76. Springer US, Boston, MA, 2015.

[20] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

[21] Stephen E. Robertson, Melvin E. Maron, and William S. Cooper. Probability of relevance: a unification of two competing models for document retrieval. *Information technology: research and development*, 1(1):1–21, 1982.

[22] Stephen E. Robertson and Karen Sparck-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[23] Mingxuan Sun, Fuxin Li, Joonseok Lee, Ke Zhou, Guy Lebanon, and Hongyuan Zha. Learning multiple-question decision trees for cold-start recommendation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 445–454, New York, NY, USA, 2013. ACM.

[24] C. J. Van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481, 1986.