Full text available at: http://dx.doi.org/10.1561/1500000008

Statistical Language Models for Information Retrieval: A Critical Review

Statistical Language Models for Information Retrieval: A Critical Review

ChengXiang Zhai

University of Illinois at Urbana-Champaign
Urbana, IL 61801
USA
czhai@cs.uiuc.edu

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is C. Zhai, Statistical Language Models for Information Retrieval A Critical Review, Foundation and Trends® in Information Retrieval, vol 2, no 3, pp 137–213, 2008

ISBN: 978-1-60198-186-8

Foundations and Trends® in Information Retrieval
Volume 2 Issue 3, 2008

Editorial Board

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians
Foundations and Trends® in Information Retrieval, 2008, Volume 2, 4 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in
Information Retrieval
Vol. 2, No. 3 (2008) 137–213

DOI: 10.1561/1500000008

now
the essence of knowledge

Statistical Language Models for Information Retrieval A Critical Review

ChengXiang Zhai

University of Illinois at Urbana-Champaign, 201 N. Goodwin, Urbana, IL 61801, USA, czhai@cs.uiuc.edu

Abstract

Statistical language models have recently been successfully applied to many information retrieval problems. A great deal of recent work has shown that statistical language models not only lead to superior empirical performance, but also facilitate parameter tuning and open up possibilities for modeling nontraditional retrieval problems. In general, statistical language models provide a principled way of modeling various kinds of retrieval problems. The purpose of this survey is to systematically and critically review the existing work in applying statistical language models to information retrieval, summarize their contributions, and point out outstanding challenges.

Contents

1

Introduction

The goal of an information retrieval (IR) system is to rank documents optimally given a query so that relevant documents would be ranked above nonrelevant ones. In order to achieve this goal, the system must be able to score documents so that a relevant document would ideally have a higher score than a nonrelevant one.

Clearly the retrieval accuracy of an IR system is directly determined by the quality of the scoring function adopted. Thus, not surprisingly, seeking an optimal scoring function (retrieval function) has always been a major research challenge in information retrieval. A retrieval function is based on a retrieval model, which formalizes the notion of relevance and enables us to derive a retrieval function that can be computed to score and rank documents.

Over the decades, many different types of retrieval models have been proposed and tested. A great diversity of approaches and methodology has developed, but no single unified retrieval model has proven to be most effective. Indeed, finding the single optimal retrieval model has been and remains a long-standing challenge in information retrieval research.

The field has progressed in two different ways. On the one hand, theoretical models have been proposed often to model relevance through inferences; representative models include the logic models [27, 111, 115] and the inference network model [109]. However, these models, while theoretically interesting, have not been able to *directly* lead to empirically effective models, even though heuristic instantiations of them can be effective. On the other hand, there have been many empirical studies of models, including many variants of the vector space model [89, 90, 91, 96] and probabilistic models [26, 51, 80, 83, 110, 109]. The vector-space model with heuristic TF-IDF weighting and document length normalization has traditionally been one of the most effective retrieval models, and it remains quite competitive as a state of the art retrieval model. The popular BM25 (Okapi) retrieval function is very similar to a TF-IDF vector space retrieval function, but it is motivated and derived from the 2-Poisson probabilistic retrieval model [84, 86] with heuristic approximations. BM25 is one of the most robust and effective retrieval functions. Another effective retrieval model is divergence from randomness which is based on probabilistic justifications for several term weighting components [1].

While both vector space models and BM25 rely on heuristic design of retrieval functions, an interesting class of probabilistic models called language modeling approaches to retrieval have led to effective retrieval functions without much heuristic design. In particular, the query likelihood retrieval function [80] with Dirichlet prior smoothing [124] has comparable performance to the most effective TF-IDF weighting retrieval functions including BM25 [24]. Due to their good empirical performance and great potential of leveraging statistical estimation methods, the language modeling approaches have been attracting much attention since Ponte and Croft's pioneering paper published in ACM SIGIR 1998 [80]. Many variations of the basic language modeling approach have since been proposed and studied, and language models have now been applied to multiple retrieval tasks such as cross-lingual retrieval [54], distributed IR [95], expert finding [25], passage retrieval [59], web search [47, 76], genomics retrieval [129], topic tracking [41, 53, 99], and subtopic retrieval [122].

This survey is to systematically review this development of the language modeling approaches. We will survey a wide range of retrieval models based on language modeling and attempt to make connections between this new family of models and traditional retrieval models. We will summarize the progress we have made so far in these models and point out remaining challenges to be solved in order to further increase their impact.

The survey is written for readers who have already had some basic knowledge about information retrieval. Readers with no prior knowledge about information retrieval will find it more comfortable to read an IR textbook (e.g., [29, 63]) first before reading this survey. The readers are also assumed to have already had some basic knowledge about probability and statistics such as maximum likelihood estimator, but a reader should still be able to follow the high-level discussion in the survey even without such background.

The rest of the survey is organized as follows. In Section 2, we review the very first generation of language models which are computationally as efficient as any other existing retrieval model. The success of these early models has stimulated many follow-up studies and extensions of language models for retrieval. In Section 3, we review work that aims at understanding why these language models are effective and why they can be justified based on relevance. In Section 4, we review work on extending and improving the basic language modeling approach. Feedback is an important component in an IR system, but it turns out that there is some difficulty in supporting feedback with the first generation basic language modeling approach. In Section 5, we review several lines of work on developing and extending language models to support feedback (particularly pseudo feedback). They are among the most effective language models for retrieval. In Section 6, we further review a wide range of applications of language models to different special retrieval tasks where a standard language model is often extended or adapted to better fit a specific application. Finally, in Section 7, we briefly review some work on developing general theoretical frameworks to facilitate systematic applications of language models to IR. We summary the survey and discuss future research directions in Section 8.

References

[1] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information System*, vol. 20, pp. 357–389, 2002.

[2] J. Bai, J.-Y. Nie, G. Cao, and H. Bouchard, "Using query contexts in information retrieval," in *Proceedings of ACM SIGIR 2007*, pp. 15–22, 2007.

[3] K. Balog, L. Azzopardi, and M. de Rijke, "Formal models for expert finding in enterprise corpora," in *Proceedings of SIGIR-06*, 2006.

[4] A. Berger and J. Lafferty, "Information retrieval as statistical translation," in *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 222–229, 1999.

[5] A. L. Berger and J. D. Lafferty, "The Weaver system for document retrieval," in *Proceedings of TREC 1999*, 1999.

[6] J. Berger, *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlap, 1985.

[7] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," in *Neural Information Processing Systems (NIPS) 16*, 2003.

[8] D. Blei and J. Lafferty, "Correlated topic models," in *Proceedings of NIPS '05: Advances in Neural Information Processing Systems 18*, 2005.

[9] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[10] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

[11] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to

machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[12] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, USA, New York, NY: ACM, 2005.

[13] C. J. C. Burges, R. Ragno, and Q. V. Le, "Learning to rank with nonsmooth cost functions," in *Proceedings of NIPS 2006*, (B. Scholkopf, J. C. Platt, and T. Hoffman, eds.), pp. 193–200, 2006.

[14] G. Cao, J.-Y. Nie, and J. Bai, "Integrating word relationships into language models," in *Proceedings of the 2005 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 298–305, 2005.

[15] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proceedings of ICML 2007, Volume 227 of ACM International Conference Proceeding Sereies*, (Z. Ghahramani, ed.), pp. 129–136, 2007.

[16] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of SIGIR'98*, pp. 335–336, 1998.

[17] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Technical Report TR-10-98, Harvard University, 1998.

[18] K. Collins-Thompson and J. Callan, "Query expansion using random walk models," in *Proceedings of ACM CIKM 2005*, pp. 704–711, 2005.

[19] K. Collins-Thompson and J. Callan, "Estimation and use of uncertainty in pseudo-relevance feedback," in *Proceedings of ACM SIGIR 2007*, pp. 303–310, 2007.

[20] W. B. Croft and D. Harper, "Using probabilistic models of document retrieval without relevance information," *Journal of Documentation*, vol. 35, pp. 285–295, 1979.

[21] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," in *Proceedings of the 25th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2002)*, pp. 299–306, August 2002.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[23] D. A. Evans and C. Zhai, "Noun-phrase analysis in unrestricted text for information retrieval," in *Proceedings of ACL 1996*, pp. 17–24, 1996.

[24] H. Fang, T. Tao, and C. Zhai, "A formal study of information retrieval heuristics," in *Proceedings of the 2004 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–56, 2004.

[25] H. Fang and C. Zhai, "Probabilistic models for expert finding," in *Proceedings of ECIR 2007*, pp. 418–430, 2007.

[26] N. Fuhr, "Probabilistic models in information retrieval," *The Computer Journal*, vol. 35, pp. 243–255, 1992.

[27] N. Fuhr, "Language models and uncertain inference in information retrieval," in *Proceedings of the Language Modeling and IR Workshop*, pp. 6–11, May 31–June 1 2001.

[28] J. Gao, J.-Y. Nie, G. Wu, and G. Cao, "Dependence language model for information retrieval," in *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 170–177, USA, New York, NY: ACM, 2004.

[29] D. Grossman and O. Frieder, *Information retrieval: Algorithms and heuristics.* Springer, 2004.

[30] D. Hiemstra, "A probabilistic justification for using tf x idf term weighting in information retrieval," *International Journal on Digital Libraries*, vol. 3, pp. 131–139, 2000.

[31] D. Hiemstra, "Using language models for information retrieval," PhD Thesis, University of Twente, 2001.

[32] D. Hiemstra, "Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term," in *Proceedings of ACM SIGIR 2002*, pp. 35–41, 2002.

[33] D. Hiemstra, "Statistical language models for intelligent XML retrieval," in *Intelligent Search on XML Data*, pp. 107–118, 2003.

[34] D. Hiemstra and W. Kraaij, "Twenty-One at TREC-7: Ad-hoc and cross-language track," in *Proceedings of Seventh Text REtrieval Conference (TREC-7)*, pp. 227–238, 1998.

[35] D. Hiemstra, S. Robertson, and H. Zaragoza, "Parsimonious language models for information retrieval," in *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 178–185, USA, New York, NY: ACM, 2004.

[36] T. Hofmann, "The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data," in *Proceedings of IJCAI' 99*, pp. 682–687, 1999.

[37] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of UAI 1999*, pp. 289–296, 1999.

[38] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of ACM SIGIR'99*, pp. 50–57, 1999.

[39] F. Jelinek, *Statistical Methods for Speech Recognition.* MIT Press, 1997.

[40] F. Jelinek and R. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *Pattern Recognition in Practice*, (E. S. Gelsema and L. N. Kanal, eds.), pp. 381–402, 1980.

[41] H. Jin, R. Schwartz, S. Sista, and F. Walls, "Topic tracking for radio, tv broadcast, and newswire," in *Proceedings of the DARPA Broadcast News Workshop*, pp. 199–204, 1999.

[42] R. Jin, A. G. Hauptmann, and C. Zhai, "Title language model for information retrieval," in *Proceedings of ACM SIGIR 2002*, pp. 42–48, 2002.

[43] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the ACM KDD 2002*, pp. 133–142, 2002.

[44] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-35, pp. 400–401, 1987.

[45] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 181–184, 1995.

[46] W. Kraaij, "Variations on language modeling for information retrieval," PhD Thesis, University of Twente, 2004.

[47] W. Kraaij, T. Westerveld, and D. Hiemstra, "The importance of prior probabilities for entry page search," in *Proceedings of ACM SIGIR 2002*, pp. 27–34, 2002.

[48] O. Kurland and L. Lee, "Corpus structure, language models, and *ad hoc* information retrieval," in *SIGIR '04: Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, pp. 194–201, ACM Press, 2004.

[49] O. Kurland and L. Lee, "PageRank without hyperlinks: Structural re-ranking using links induced by language models," in *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 306–313, USA, New York, NY: ACM, 2005.

[50] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," in *Proceedings of SIGIR'01*, pp. 111–119, September 2001.

[51] J. Lafferty and C. Zhai, "Probabilistic relevance models based on document and query generation," in *Language Modeling and Information Retrieval*, (W. B. Croft and J. Lafferty, eds.), pp. 1–6, Kluwer Academic Publishers, 2003.

[52] V. Lavrenko, "A generative theory of relevance," PhD Thesis, University of Massachusetts, Amherst, 2004.

[53] V. Lavrenko, J. Allan, E. DeGuzman, D. LaFlamme, V. Pollard, and S. Thomas, "Relevance models for topic detection and tracking," in *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 115–121, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002.

[54] V. Lavrenko, M. Choquette, and W. B. Croft, "Cross-lingual relevance models," in *Proceedings of ACM SIGIR 2002*, pp. 175–182, 2002.

[55] V. Lavrenko and W. B. Croft, "Relevance-based Language Models," in *Proceedings of SIGIR'01*, pp. 120–127, September 2001.

[56] D. D. Lewis, "Representation and learning in information retrieval," Technical Report 91-93, University of Massachusetts, 1992.

[57] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, pp. 577–584, 2006.

[58] X. Li and W. B. Croft, "Time-based language models," in *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 469–475, USA, New York, NY: ACM, 2003.

[59] X. Liu and W. B. Croft, "Passage retrieval based on language models," in *CIKM '02: Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 375–382, USA, New York, NY: ACM, 2002.

[60] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," in *SIGIR '04: Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, pp. 186–193, ACM Press, 2004.

[61] D. MacKay and L. Peto, "A hierarchical Dirichlet language model," *Natural Language Engineering*, vol. 1, pp. 289–307, 1995.

[62] R. E. Madsen, D. Kauchak, and C. Elkan, "Modeling word burstiness using the Dirichlet distribution," in *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, pp. 545–552, USA, New York, NY: ACM, 2005.

[63] C. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[64] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *AAAI-1998 Learning for Text Categorization Workshop*, pp. 41–48, 1998.

[65] Q. Mei, H. Fang, , and C. Zhai, "A study of Poisson query generation model for information retrieval," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 319–326, 2007.

[66] Q. Mei and C. Zhai, "A mixture model for contextual text mining," in *Proceedings of KDD '06*, pp. 649–655, 2006.

[67] Q. Mei, D. Zhang, and C. Zhai, "A general optimization framework for smoothing language models on graph structures," in *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 611–618, USA, New York, NY: ACM, 2008.

[68] D. Metzler and W. B. Croft, "A Markov random field model for term dependencies," in *Proceedings of the 2005 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 472–479, 2005.

[69] D. Metzler, V. Lavrenko, and W. B. Croft, "Formal multiple-Bernoulli models for language modeling," in *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 540–541, USA, New York, NY: ACM, 2004.

[70] D. H. Miller, T. Leek, and R. Schwartz, "A hidden Markov model information retrieval system," in *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 214–221, 1999.

[71] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the UAI 2002*, pp. 352–359, 2002.

[72] R. Nallapati and J. Allan, "Capturing term dependencies using a language model based on sentence trees," in *CIKM '02: Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 383–390, USA, New York, NY: ACM, 2002.

[73] R. Nallapati, B. Croft, and J. Allan, "Relevant query feedback in statistical language modeling," in *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 560–563, USA, New York, NY: ACM, 2003.

[74] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modeling," *Computer Speech and Language*, vol. 8, pp. 1–38, 1994.

[75] K. Ng, "A maximum likelihood ratio information retrieval model," in *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, (E. Voorhees and D. Harman, eds.), pp. 483–492, 2000.

[76] P. Ogilvie and J. Callan, "Experiments using the Lemur toolkit," in *Proceedings of the 2001 TREC conference*, 2002.

[77] P. Ogilvie and J. Callan, "Using language models for flat text queries in XML retrieval," in *Proceedings of the Initiative for the Evaluation of XML Retrieval Workshop (INEX 2003)*, 2003.

[78] P. Ogilvie and J. P. Callan, "Combining document representations for known-item search," in *Proceedings of ACM SIGIR 2003*, pp. 143–150, 2003.

[79] J. Ponte, "A language modeling approach to information retrieval," PhD Thesis, University of Massachusetts at Amherst, 1998.

[80] J. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the ACM SIGIR'98*, pp. 275–281, 1998.

[81] T. Qin, T.-Y. Liu, X.-D. Zhang, Z. Chen, and W.-Y. Ma, "A study of relevance propagation for web search," in *Proceedings of SIGIR 2005*, pp. 408–415, 2005.

[82] L. R. Rabiner, "A tutorial on hidden Markov models," *Proceedings of the IEEE*, vol. 77, pp. 257–285, 1989.

[83] S. Robertson and K. Sparck Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, vol. 27, pp. 129–146, 1976.

[84] S. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," in *Proceedings of SIGIR'94*, pp. 232–241, 1994.

[85] S. E. Robertson, "The probability ranking principle in IR," *Journal of Documentation*, vol. 33, pp. 294–304, December 1977.

[86] S. E. Robertson, S. Walker, K. Sparck Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *The Third Text Retrieval Conference (TREC-3)*, (D. K. Harman, ed.), pp. 109–126, 1995.

[87] J. Rocchio, "Relevance feedback in information retrieval," in *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323, Prentice-Hall Inc., 1971.

[88] T. Roelleke and J. Wang, "A parallel derivation of probabilistic information retrieval models," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 107–114, New York, NY, USA: ACM, 2006.

[89] G. Salton, *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[90] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[91] G. Salton, C. S. Yang, and C. T. Yu, "A theory of term importance in automatic text analysis," *Journal of the American Society for Information Science*, vol. 26, pp. 33–44, Jan–Feb 1975.

[92] A. Shakery and C. Zhai, "A probabilistic relevance propagation model for hypertext retrieval," in *Proceedings of CIKM 2006*, pp. 550–558, 2006.

[93] A. Shakery and C. Zhai, "Smoothing document language models with probabilistic term count propagation," *Information Retrieval*, vol. 11, pp. 139–164, 2008.

[94] X. Shen, B. Tan, and C. Zhai, "Context-sensitive information retrieval using implicit feedback," in *Proceedings of SIGIR 2005*, pp. 43–50, 2005.

[95] L. Si, R. Jin, J. P. Callan, and P. Ogilvie, "A language modeling framework for resource selection and results merging," in *Proceedings of CIKM 2002*, pp. 391–397, 2002.

[96] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21–29, 1996.

[97] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 279–280, 1999.

[98] K. Sparck Jones, S. Robertson, D. Hiemstra, and H. Zaragoza, "Language modeling and relevance," in *Language Modeling for Information Retrieval*, (W. B. Croft and J. Lafferty, eds.), pp. 57–72, 2003.

[99] M. Spitters and W. Kraaij, "Language models for topic tracking," in *Language Modeling for Information Retrieval*, pp. 95–124, 2003.

[100] M. Srikanth and R. Srihari, "Biterm language models for document retrieval," in *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 425–426, USA, New York, NY: ACM, 2002.

[101] M. Srikanth and R. Srihari, "Exploiting syntactic structure of queries in a language modeling approach to IR," in *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 476–483, USA, New York, NY: ACM, 2003.

[102] M. Srikanth and R. Srihari, "Incorporating query term dependencies in language models for document retrieval," in *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 405–406, USA, New York, NY: ACM, 2003.

[103] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of KDD'04*, pp. 306–315, 2004.

[104] T. Strzalkowski and B. Vauthey, "Information retrieval using robust natural language processing," in *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pp. 104–111, Morristown, NJ, USA: Association for Computational Linguistics, 1992.

[105] B. Tan, X. Shen, and C. Zhai, "Mining long-term search history to improve search accuracy," in *KDD*, pp. 718–723, 2006.

[106] T. Tao, X. Wang, Q. Mei, and C. Zhai, "Language model information retrieval with document expansion," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 407–414, Morristown, NJ, USA: Association for Computational Linguistics, 2006.

[107] T. Tao and C. Zhai, "Mixture clustering model for pseudo feedback in information retrieval," in *Proceedings of the 2004 Meeting of the International Federation of Classification Societies*, Spriner, 2004.

[108] T. Tao and C. Zhai, "Regularized estimation of mixture models for robust pseudo-relevance feedback," in *Proceedings of ACM SIGIR 2006*, pp. 162–169, 2006.

[109] H. Turtle and W. B. Croft, "Evaluation of an inference network-based retrieval model," *ACM Transactions on Information Systems*, vol. 9, pp. 187–222, July 1991.

[110] C. J. van Rijbergen, "A theoretical basis for the use of co-occurrence data in information retrieval," *Journal of Documentation*, pp. 106–119, 1977.

[111] C. J. van Rijsbergen, "A non-classical logic for information retrieval," *The Computer Journal*, vol. 29, no. 6, pp. 481–485, 1986.

[112] X. Wang, H. Fang, and C. Zhai, "Improve retrieval accuracy for difficult queries using negative feedback," in *CIKM '07: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 991–994, USA, New York, NY: ACM, 2007.

[113] X. Wei and W. Bruce Croft, "LDA-based document models for ad-hoc retrieval," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 178–185, USA, New York, NY: ACM, 2006.

[114] S. K. M. Wong and Y. Y. Yao, "A probability distribution model for information retrieval," *Information Processing and Management*, vol. 25, pp. 39–53, 1989.

[115] S. K. M. Wong and Y. Y. Yao, "On modeling information retrieval with probabilistic inference," *ACM Transactions on Information Systems*, vol. 13, pp. 69–99, 1995.

[116] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proceedings of the SIGIR'96*, pp. 4–11, 1996.

[117] J. Xu and W. B. Croft, "Cluster-based language models for distributed retrieval," in *Proceedings of the SIGIR'99*, pp. 254–261, 1999.

[118] J. Xu, R. Weischedel, and C. Nguyen, "Evaluating a probabilistic model for cross-lingual information retrieval," in *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 105–110, USA, New York, NY: ACM, 2001.

[119] H. Zaragoza, D. Hiemstra, and M. E. Tipping, "Bayesian extension to the language model for ad hoc information retrieval," in *Proceedings of ACM SIGIR 2003*, pp. 4–9, 2003.

[120] C. Zhai, "Fast statistical parsing of noun phrases for document indexing," in *5th Conference on Applied Natural Language Processing (ANLP-97)*, pp. 312–319, March 31–April 3 1997.

[121] C. Zhai, "Risk minimization and language modeling in text retrieval," PhD Thesis, Carnegie Mellon University, 2002.

[122] C. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," in *Proceedings of ACM SIGIR'03*, pp. 10–17, August 2003.

[123] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, pp. 403–410, 2001.

[124] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *Proceedings of ACM SIGIR'01*, pp. 334–342, September 2001.

[125] C. Zhai and J. Lafferty, "Two-stage language models for information retrieval," in *Proceedings of ACM SIGIR'02*, pp. 49–56, August 2002.

[126] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions on Information Systems*, vol. 2, pp. 179–214, 2004.

[127] C. Zhai and J. Lafferty, "A risk minimization framework for information retrieval," *Information Processing Management*, vol. 42, pp. 31–55, 2006.

[128] C. Zhai, X. Lu, X. Ling, A. Velivelli, X. Wang, H. Fang, and A. Shakery, "UIUC/MUSC at TREC 2005 Genomics Track," in *Proceedings of TREC 2005*, 2005.

[129] C. Zhai, T. Tao, H. Fang, and Z. Shang, "Improving the robustness of language models — UIUC TREC 2003 robust and genomics experiments," in *Proceedings of TREC 2003*, pp. 667–672, 2003.

[130] C. Zhai, A. Velivelli, and B. Yu, "A cross-collection mixture model for comparative text minning," in *Proceeding of the 10th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 743–748, 2004.

[131] Y. Zhang, J. Callan, and T. Minka, "Redundancy detection in adaptive filtering," in *Proceedings of SIGIR'02*, pp. 81–88, August 2002.

[132] Y. Zhou and W. B. Croft, "Document quality models for web ad hoc retrieval," in *CIKM '05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 331–332, USA, New York, NY: ACM, 2005.