

Donald A. Metzler Jr.

Yahoo! Research
2821 Mission College Blvd.
Santa Clara, CA 95054

metzler@yahoo-inc.com
<http://ciir.cs.umass.edu/~metzler/>
Phone: (408) 336-0073

RESEARCH INTERESTS I am interested in theoretical and practical information retrieval problems, as well as machine learning applied to large scale textual applications. My research has focused on retrieval models, query/document representations, term weighting, term proximity models, and learning to rank (machine learned ranking functions). I am always interested in expanding the breadth and the depth of my research into other related areas or fields of study. I am also passionate about seeing my research applied to real world problems, especially those dealing with large, complex data sets. Along these lines, I have experience evaluating and designing novel search algorithms for news search, web search, summarization, and online advertising systems.

EMPLOYMENT Research Scientist, September 2007 – Present
Search and Computational Advertising Group
Yahoo! Research, Santa Clara, CA

Research Intern, June 2006 – August 2006
Adaptive Systems and Interaction Group
Microsoft Research, Seattle, WA

Research Assistant, September 2002 – August 2007
Center for Intelligent Information Retrieval
University of Massachusetts, Amherst, MA

EDUCATION Ph.D., Computer Science, September 2007
University of Massachusetts, Amherst, MA
Thesis: *Beyond Bags of Words: Effectively Modeling Dependence and Features in Information Retrieval*
Advisor: W. Bruce Croft
Committee Members: James Allan, John Buonaccorsi, Andrew McCallum

M.S., Computer Science, May 2005
University of Massachusetts, Amherst, MA

B.S., Computer Science and Mathematics, Summa Cum Laude, May 2002
Rose-Hulman Institute of Technology, Terre Haute, IN

NOTABLE RESEARCH PROJECTS **Web Search Relevance**, 2007-Present.
Yahoo! Research, Santa Clara, CA

Researched, developed, and implemented a variety of techniques for improving the textual matching component of Yahoo!'s web search engine. One line of research focused on improving term proximity scoring by extending elements of my thesis research to a real web environment. The other line of research aimed at building enriched document representations of web documents. I developed a method for overcoming the anchor text sparsity problem that involves aggregating anchor text along the web graph. Both lines of research, which targeted improving relevance for long and/or difficult queries, yielded significant improvements in search relevance.

Computational Advertising, 2007-Present.

Yahoo! Research, Santa Clara, CA

Developed state-of-the-art retrieval methods for sponsored search and contextual advertising. For sponsored search, our group developed a highly effective method for expanding tail queries in real time by leveraging significant offline processing. For contextual advertising, we proposed using site-level information to improve page-level term weighting. Finally, our group proposed a supervised approach for automatically determining the quality of a set of ads in order to reduce the number of non-relevant ads shown to users. Each of these projects yielded significant improvements in ad relevance.

Markov Random Fields for Information Retrieval, 2004-2007.

Advisor: W. Bruce Croft

University of Massachusetts, Amherst, MA

Developed an information retrieval model based on Markov Random Fields that generalizes and naturally extends the language modeling framework for information retrieval. The model provides an effective means of modeling general dependencies between query terms as well as the facility to include a wide range of other features, such as named entities, part of speech information, and term proximity. The model has been shown to consistently and significantly improve effectiveness over state-of-the-art retrieval models.

Matching Short Segments of Text, Summer 2006.

Advisors: Susan Dumais, Chris Meek

Microsoft Research, Redmond, WA

Investigated how to effectively match short segments of text, such as queries to queries, queries to ad keywords, or queries to image captions, by producing richer representations. Such representations were generated using resources such as query search logs and web search results.

RECAP: A System for Identifying Text Reuse, Fall 2004.

Advisors: W. Bruce Croft, Alistair Moffat, Justin Zobel

Royal Melbourne Institute of Technology, Melbourne, Australia

University of Melbourne, Melbourne, Australia

Investigated similarity measures for finding various types of reuse within a collection of documents. Examples of reuse include exact copies, minor edits, and paraphrases. We found that existing similarity measures are good at finding exact copies and minor edits, but perform poorly for more complex forms of reuse.

The Indri Search System, 2003-2004.

Advisor: W. Bruce Croft

University of Massachusetts, Amherst, MA

Created the retrieval model used in the Indri search system. The model combines the language modeling and inference network approaches to information retrieval. The model uses formal language modeling estimates and provides a powerful structured query language. The system, which was evaluated at the 2004-2006 TREC Terabyte Tracks and the 2005 TREC Robust Track, was shown to be highly effective.

An Inference Network Approach to Image Retrieval, 2003-2004.

Advisor: R. Manmatha

University of Massachusetts, Amherst, MA

Developed a multi-modal image retrieval system based on the inference network model for information retrieval. The system provides a robust query language for querying a set of images using both textual keywords and images. Users may request images of tigers by providing the system with an image of a tiger or by entering the keyword 'tiger'.

Question Classification, 2002-2003.

Advisors: W. Bruce Croft, Andrew McCallum

University of Massachusetts, Amherst, MA

Developed a question classification system based on support vector machines. The system made use of a novel set of domain-specific, linguistically inspired features, such as question word type, bigrams, and the WordNet hypernyms of the question headword. Results showed significant improvements over using standard unigram features, especially for "What..?" questions.

TEACHING
EXPERIENCE

Teaching Assistant, Graduate Information Retrieval, Fall 2006.

Instructor: James Allan

University of Massachusetts, Amherst, MA

Helped create homework problems. Responsible for grading homeworks and helping students with homework-related questions. Gave several guest lectures related to probabilistic information retrieval.

BOOKS

Croft, W.B., Metzler, D., and Strohman, T. "Search Engines: Information Retrieval in Practice," Addison Wesley, 2009.

THESES

Metzler, D., "Beyond Bags of Words: Effectively Modeling Dependence and Features in Information Retrieval," Ph.D. Dissertation, University of Massachusetts, Amherst, MA, 2007.

JOURNAL
ARTICLES

Metzler, D. and Croft, W.B., "Linear Feature-Based Models for Information Retrieval," in *Information Retrieval*, 10(3), 257-274, 2007.

Metzler, D. and Croft, W.B., "Analysis of Statistical Question Classification for Fact-based Questions," *Information Retrieval*, 8(3), 481-504, 2005.

Metzler, D. and Croft, W.B., "Combining the Language Model and Inference Network Approaches to Retrieval," *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval*, 40(5), 735-750, 2004.

CONFERENCE
PAPERS

Broder, A. Ciccolo, P., Gabrilovich, Josifovski, V., Metzler, D., Riedel, L., and Yuan, J. "Online Expansion of Rare Queries for Sponsored Search," to appear in the *Proceedings of the ACM Conference on the World Wide Web*, 2009.

Broder, A. Ciaramita, M., Fontoura, M., Gabrilovich, E., Josifovski, V., Metzler, D., Murdock, V., and Plachouras, V. "To Swing or not to Swing: Learning when (not) to Advertise," in the *Proceedings of the ACM Conference on Information and Knowledge Management*, 2008.

Metzler, D. "Generalized Inverse Document Frequency," in the *Proceedings of the ACM Conference on Information and Knowledge Management*, 2008.

Metzler, D., Strohman, T., and Croft, W.B., "A Statistical View of Binned Retrieval Models," in the *Proceedings of the 30th European Conference on Information Retrieval*, 2008.

Metzler, D. "Automatic Feature Selection in the Markov Random Field Model for Information Retrieval," in the *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 253-262, 2007.

Metzler, D and Croft, W.B. "Latent Concept Expansion Using Markov Random Fields," in the *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 311-318, 2007.

Metzler, D., Dumais, S., and Meek, C. "Similarity Measures for Short Segments of Text," in the *Proceedings of the 29th European Conference on Information Retrieval*, 16-27, 2007.

Diaz, F., and Metzler, D. "Pseudo-Aligned Multilingual Corpora," in the *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2727-2732, 2007.

Metzler, D., and Croft, W.B. "Beyond Bags of Words: Modeling Implicit User Preferences in Information Retrieval," in the *Proceedings of the AAAI'06 Nectar Track*, 2006.

Diaz, F. and Metzler, D. "Improving the Estimation of Relevance Models Using Large External Corpora," in the *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 154-161, 2006.

Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., and Zobel, J. "Similarity Measures for Tracking Information Flow," in the *Proceedings of the 14th ACM Conference on Information and Knowledge Management*, 517-524, 2005.

Metzler, D. and Croft, W.B., "A Markov Random Field Model for Term Dependencies," in the *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 472-479, 2005.

Metzler, D., and Manmatha, R., "An Inference Network Approach to Image Retrieval," in the *Proceedings of the International Conference on Image and Video Retrieval*, 42-50, 2004

UNREFEREED
CONFERENCE
PAPERS

Turtle, H. and Metzler, D., "CIIR Experiments for TREC Legal 2007," in the *Proceedings of the 2007 Text REtrieval Conference*, 2007.

Metzler, D., Strohman, T., and Croft, W.B., "Indri at TREC 2006: Lessons Learned From Three Terabyte Tracks," in the *Proceedings of the 2006 Text REtrieval Conference*, 2006.

Metzler, D., Strohman T., Zhou, Y., and Croft, W.B., "UMass Robust 2005: Using Mixtures of Relevance Models for Query Expansion," in the *Proceedings of the 2005 Text REtrieval Conference*.

Metzler, D., Strohman T., Zhou, Y., and Croft, W.B., "Indri at TREC 2005: Terabyte Track," in the *Proceedings of the 2005 Text REtrieval Conference*.

Metzler, D., Strohman T., Turtle H., and Croft, W.B., "Indri at TREC 2004: Terabyte Track," in the *Proceedings of the 2004 Text REtrieval Conference*.

- WORKSHOP PAPERS Metzler, D. and Kanungo, T. "Machine Learned Sentence Selection Strategies for Query-Biased Summarization," in the *Proceedings of the ACM SIGIR Learning to Rank for Information Retrieval Workshop*, 2008.
- DEMOS AND SHORT PAPERS Metzler, D. "Using Gradient Descent to Optimize Language Modeling Smoothing Parameters" in the *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 687-688, 2007.
- Metzler, D., "Estimation, Sensitivity, and Generalization in Parameterized Retrieval Models," in the *Proceedings of the 15th ACM Conference on Information and Knowledge Management*, 2006.
- Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., and Zobel, J., "The Recap System for Identifying Information Flow," *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 678-678, 2005.
- Strohman, T., Metzler, D., Turtle, H., and Croft, W.B., "Indri: A Language-Model Based Search Engine for Complex Queries," in the online *Proceedings of the International Conference on Intelligence Analysis*, 2005.
- Metzler, D., Lavrenko, V., and Croft, W. B., "Formal Multiple-Bernoulli Models for Language Modeling," *Proceedings of ACM SIGIR 2004*, 540-541, 2004.
- TECHNICAL REPORTS Metzler, D., "Estimation, Sensitivity, and Generalization in Parameterized Retrieval Models (Extended Version)," CIIR Technical Report, 2006.
- Donald Metzler, "Direct Maximization of Rank-based Metrics," CIIR Technical report.
- Strohman, T., Metzler, D., Turtle, H., and Croft, W.B., "Indri: A Language-Model Based Search Engine for Complex Queries (Extended Version)". CIIR Technical Report.
- PATENTS FILED *Identifying and Expanding Implicitly Temporally Qualified Queries*
Inventors: Rosie Jones, Donald Metzler, and Fuchun Peng
- Phrase Identification using Break Points*
Inventors: Hadar Shemtov, Tapas Kanungo, Rajhans Samdani, and Donald Metzler
- Prediction of a Degree of Relevance Between Query Rewrites and a Search Query*
Inventors: Evgeniy Gabrilovich, Donald Metzler, Vanja Josifovski, Andrei Broder, Vassilis Plachouras, Vanessa Murdock, and Massimiliano Ciaramita
- Rare Query Expansion by Web Feature Matching*
Inventors: Donald Metzler, Lance Riedel, Evgeniy Gabrilovich, and Vanja Josifovski
- System and Method for Automatically Ranking Lines of Text*
Inventors: Tapas Kanungo and Donald Metzler
- System and Method for Improved Search Relevance Using Proximity Boosting*
Inventors: Fuchun Peng, Xing Wei, Yumao Lu, Xin Li, Donald Metzler, Hang Cui, Beniot Dumoulin
- Systems And Methods For Building A Prediction Model To Predict A Degree Of Relevance Between Digital Ads And A Search Query Or Webpage Content*

Inventors: Evgeniy Gabrilovich, Vassilis Plachouras, Andrei Broder, Vanessa Murdock, Donald Metzler, Vanja Josifovski, Massimiliano Ciaramita, and Marcus Fontoura

Systems and Methods for Predicting a Degree of Relevance Between Digital Ads and Web-page Content

Inventors: Evgeniy Gabrilovich, Vassilis Plachouras, Andrei Broder, Vanessa Murdock, Donald Metzler, Vanja Josifovski, Massimiliano Ciaramita, and Marcus Fontoura

Systems and Methods for Predicting a Degree of Relevance Between Digital Ads and a Search Query

Inventors: Evgeniy Gabrilovich, Vassilis Plachouras, Andrei Broder, Vanessa Murdock, Donald Metzler, Vanja Josifovski, Massimiliano Ciaramita, and Marcus Fontoura

Systems And Methods For Query Expansion In Sponsored Search

Inventors: Andrei Z. Broder, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Lance Riedel and Peter Ciccolo

INVITED
TALKS

“To Swing or Not to Swing: Learning when (not) to Advertise”
National University of Singapore
Singapore, July 2008.

“Overview of Computational Advertising”
University of Massachusetts
Amherst, MA, March 2008

“An Overview of the Indri Search Engine”
University of Illinois
Champaign, IL, March 2005.

PROFESSIONAL
ACTIVITIES

Professional Membership

- Association for Computing Machinery (ACM)
- ACM Special Interest Group on Information Retrieval (SIGIR)
- ACM Special Interest Group on the World Wide Web (SIGWEB)

Program Committee Member (Chairs/co-Chairs)

- ACM SIGIR Poster Track (2009)

Program Committee Member

- ACM Conference on Web Search and Data Mining (WSDM) (2008, 2009)
- ACM International Conference on Research and Development in Information Retrieval (SIGIR) (2007, 2008, 2009)
- Empirical Methods for Natural Language Processing Conference (EMNLP) (2008)
- European Association for Computational Linguistics Conference (EACL) (2008)
- European Conference on Information Retrieval (ECIR) (2008)
- International Conference on Machine Learning (ICML) (2008)
- International World Wide Web Conference (WWW) (2008)
- North American Chapter of the Association for Computational Linguistics - Human Language Technologies Conference (NAACL-HLT) (2007)

Program Committee Member (Workshops, Posters, Demos)

- Beyond Binary Relevance Workshop (SIGIR 2008, 2009)
- Human Language Technologies Conference, Demo Track (2008)
- Information Retrieval for Advertising Workshop (SIGIR 2008, 2009)
- Learning to Rank Workshop for Information Retrieval (SIGIR 2007, 2008)
- Workshop on Web Search Result Summarization and Presentation (WWW 2009)

Journal Reviewer

- ACM Transactions on Information Systems (TOIS)
- Data Mining and Knowledge Discovery (DMKD)
- Electronics and Telecommunications Research Institute Journal (ETRI)
- Foundations and Trends in Information Retrieval (FnTIR)
- IEEE Transactions on Knowledge and Data Engineering (TKDE)
- Information Processing and Management (IP&M)
- Information Processing Letters (IPL)
- Journal of Artificial Intelligence Research (JAIR)
- Journal of Machine Learning Research (JMLR)
- Pattern Recognition Letters (PATREC)

Conference Reviewer

- ACM International Conference on Knowledge Discovery and Data Mining (KDD) (2003)
- ACM International Conference on Research and Development in Information Retrieval (SIGIR) (2005)
- European Conference on Information Retrieval (2005)

Other Service

- National Science Foundation Panelist

AWARDS AND
HONORS

- Microsoft Research / Live Labs Graduate Fellow, 2006-2007
- Best Student Paper, ACM SIGIR 2005
- Passed Ph.D. Qualifier with Distinction, 2005