The Probabilistic Relevance Framework: BM25 and Beyond

The Probabilistic Relevance Framework: BM25 and Beyond

Stephen Robertson

Microsoft Research
Cambridge CB3 0FB
UK
ser@microsoft.com

Hugo Zaragoza

Yahoo! Research
Barcelona 08028
Spain
hugoz@yahoo-inc.com

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 S. Robertson and H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundation and Trends® in Information Retrieval, vol 3, no 4, pp 333–389, 2009

ISBN: 978-1-60198-308-4

Foundations and Trends® in Information Retrieval

Volume 3 Issue 4, 2009

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, 2009, Volume 3, 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. 3, No. 4 (2009) 333–389
© 2009 S. Robertson and H. Zaragoza
DOI: 10.1561/1500000019

The Probabilistic Relevance Framework: BM25 and Beyond

Stephen Robertson[1] **and Hugo Zaragoza**[2]

[1] *Microsoft Research, 7 J J Thomson Avenue, Cambridge CB3 0FB, UK*
ser@microsoft.com

[2] *Yahoo! Research, Av. Diagonal 177, Barcelona 08028, Spain*
hugoz@yahoo-inc.com

Abstract

The Probabilistic Relevance Framework (PRF) is a formal framework for document retrieval, grounded in work done in the 1970–1980s, which led to the development of one of the most successful text-retrieval algorithms, BM25. In recent years, research in the PRF has yielded new retrieval models capable of taking into account document meta-data (especially structure and link-graph information). Again, this has led to one of the most successful Web-search and corporate-search algorithms, BM25F. This work presents the PRF from a conceptual point of view, describing the probabilistic modelling assumptions behind the framework and the different ranking algorithms that result from its application: the binary independence model, relevance feedback models, BM25 and BM25F. It also discusses the relation between the PRF and other statistical models for IR, and covers some related topics, such as the use of non-textual features, and parameter optimisation for models with free parameters.

Contents

1

Introduction

This monograph addresses the *classical* probabilistic model of information retrieval. The model is characterised by including a specific notion of relevance, an explicit variable associated with a query–document pair, normally hidden in the sense of *not observable*. The model revolves around the notion of estimating a probability of relevance for each pair, and ranking documents in relation to a given query in descending order of probability of relevance. The best-known instantiation of the model is the BM25 term-weighting and document-scoring function.

The model has been developed in stages over a period of about 30 years, with a precursor in 1960. A few of the main references are as follows: [30, 44, 46, 50, 52, 53, 58]; other surveys of a range of probabilistic approaches include [14, 17]. Some more detailed references are given below.

There are a number of later developments of IR models which are also probabilistic but which differ considerably from the models developed here — specifically and notably the *language model* (LM) approach [24, 26, 33] and the *divergence from randomness* (DFR) models [2]. For this reason we refer to the family of models developed here as the *Probabilistic Relevance Framework* (PRF), emphasising the importance of the relevance variable in the development of the models. We do not cover the development of other probabilistic models in the present survey, but some points of comparison are made.

This is not primarily an experimental survey; throughout, assertions will be made about techniques which are said to work well. In general such statements derive from experimental results, many experiments by many people over a long period, which will not in general be fully referenced. The emphasis is on the theoretical development of the methods, the logic and assumptions behind the models.

The survey is organised as follows. In Section 2 we develop the most generic retrieval model, which subsumes a number of specific instantiations developed in Section 3. In Section 4 we discuss the similarities and differences with other retrieval frameworks. Finally in Section 5 we give an overview of optimisation techniques we have used to tune the different parameters in the models and Section 6 concludes the survey.

References

[1] S. Agarwal, C. Cortes, and R. Herbrich, eds., *Proceedings of the NIPS 2005 Workshop on Learning to Rank*, 2005.

[2] G. Amati, C. J. van Rijsbergen, and C. Joost, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 357–389, 2002.

[3] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, and P. Williams, "Okapi at TREC-5," *The Fifth Text Retrieval Conference (TREC-5). NIST Special Publication* 500-238, pp. 143–165, 1997.

[4] F. V. Berghen, "Trust Region Algorithms," Webpage, http://www.lemurproject.org.

[5] F. V. Berghen, "CONDOR: A constrained, non-linear, derivative-free parallel optimizer for continuous, high computing load, noisy objective functions," PhD thesis, Université Libre de Bruxelles, 2004.

[6] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.

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

[8] D. Bodoff and S. E. Robertson, "A new unified probabilistic model," *Journal of the American Society for Information Science and Technology*, vol. 55, pp. 471–487, 2004.

[9] P. Boldi and S. Vigna, "MG4J at TREC 2005," in *The Fourteenth Text Retrieval Conference (TREC 2005) Proceedings, NIST Special Publication* 500-266, 2005. http://mg4j.dsi.unimi.it/.

[10] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 22, p. 89, 2005.

[11] W. Cooper, "Some inconsistencies and misidentified modelling assumptions in probabilistic information retrieval," *ACM Transactions on Information Systems*, vol. 13, pp. 110–111, 1995.

[12] N. Craswell, S. E. Robertson, H. Zaragoza, and M. Taylor, "Relevance weighting for query independent evidence," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 472–479, ACM, 2005.

[13] N. Craswell, H. Zaragoza, and S. E. Robertson, "Microsoft Cambridge at TREC-14: Enterprise track," in *The Fourteenth Text Retrieval Conference (TREC 2005)*, 2005.

[14] F. Crestani, M. Lalmas, C. J. van Rijsbergen, and I. Campbell, ""Is this document relevant? ... probably": A survey of probabilistic models in information retrieval," *ACM Computing Surveys*, vol. 30, no. 4, 1998.

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

[16] W. Feller, *An Introduction to Probability Theory and Its Applications,* vol. 1. Wiley, 1968.

[17] N. Fuhr, "Probabilistic Models in Information Retrieval," *The Computer Journal*, vol. 35, no. 3, 1992.

[18] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, "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*, pp. 465–480, ACM, 1988.

[19] S. P. Harter, "A probabilistic approach to automatic keyword indexing (parts 1 and 2)," *Journal of the American Society for Information Science*, vol. 26, pp. 197–206 and 280–289, 1975.

[20] D. Hiemstra, S. E. Robertson, and H. Zaragoza, "Parsimonious language models for information retrieval," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 178–185, ACM, 2004.

[21] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, ACM, 1999.

[22] Indri. Homepage. http://www. lemurproject.org/indri.

[23] T. Joachims, H. Li, T. Y. Liu, and C. Zhai, "Learning to rank for information retrieval (LR4IR 2007)," *SIGIR Forum*, vol. 41, no. 2, pp. 58–62, 2007.

[24] J. Lafferty and C. Zhai, "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*, ACM, 2001.

[25] J. Lafferty and C. Zhai, "Probabilistic relevance models based on document and query generation," in *Language Modelling for Information Retrieval*, (W. B. Croft and J. Lafferty, eds.), pp. 1–10, Kluwer, 2003.

[26] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120–127, ACM, 2001.

[27] Lemur Toolkit. Homepage. http://www.lemurproject.org.

[28] H. Li, T. Y. Liu, and C. Zhai, "Learning to rank for information retrieval (LR4IR 2008)," *SIGIR Forum*, vol. 42, no. 2, pp. 76–79, 2008.

[29] Lucene. Homepage. http://lucene.apache.org/.

[30] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM*, vol. 7, no. 3, pp. 216–244, 1960.

[31] D. Metzler, "Automatic feature selection in the Markov random field model for information retrieval," in *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pp. 253–262, ACM New York, NY, USA, 2007.

[32] D. Metzler and W. B. Croft, "A Markov random field model for term dependencies," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 472–479, ACM, 2005.

[33] D. Metzler, T. Strohman, and B. Croft, *Information Retrieval in Practice*. Pearson Education (US), 2009.

[34] MG4J: Managing gigabytes for java. Homepage. http://mg4j.dsi.unimi.it/.

[35] Okapi-Pack. Homepage. http://www.soi.city.ac.uk/ andym/OKAPI-PACK.

[36] J. R. Pérez-Agüera and H. Zaragoza, "UCM-Y!R at CLEF 2008 Robust and WSD tasks," *CLEF 2008 Workshop*, 2008.

[37] J. R. Pérez-Agüera, H. Zaragoza, and L. Araujo, "Exploiting morphological query structure using genetic optimization," in *NLDB 2008 13th International Conference on Applications of Natural Language to Information Systems*, Lecture Notes in Computer Science (LNCS), Springer Verlag, 2008.

[38] J. Pérez-Iglesias, "BM25 and BM25F Implementation for Lucene," Webpage, http://nlp.uned.es/∼jperezi/Lucene-BM25.

[39] PF-Tijah. Homepage. http://dbappl.cs.utwente.nl/pftijah.

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

[41] S. E. Robertson, "On term selection for query expansion," *Journal of Documentation*, vol. 46, pp. 359–364, 1990.

[42] S. E. Robertson, "Threshold setting and performance optimization in adaptive filtering," *Information Retrieval*, vol. 5, pp. 239–256, 2002.

[43] S. E. Robertson, M. E. Maron, and W. S. Cooper, "The unified probabilistic model for IR," in *Proceedings of Research and Development in Information Retrieval*, (G. Salton and H.-J. Schneider, eds.), pp. 108–117, Berlin: Springer-Verlag, 1983.

[44] S. E. Robertson and K. Sparck Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, 1977.

[45] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter, "Probabilistic models of indexing and searching," in *Information Retrieval Research (Proceedings of Research and Development in Information Retrieval, Cambridge, 1980)*, (R. N. Oddy, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, eds.), pp. 35–56, London: Butterworths, 1981.

[46] S. E. Robertson and S. Walker, "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*, pp. 232–241, ACM/Springer, 1994.

[47] S. E. Robertson and S. Walker, "On relevance weights with little relevance information," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 16–24, ACM, 2007.

[48] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau, "Okapi at TREC," in *The First Text Retrieval Conference (TREC-1), NIST Special Publication* 500-207, pp. 21–30, 1992.

[49] S. E. Robertson and H. Zaragoza, "On rank-based effectiveness measures and optimization," *Information Retrieval*, vol. 10, no. 3, pp. 321–339, 2007.

[50] S. E. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management*, pp. 42–49, ACM, 2004.

[51] R. Song, M. J. Taylor, J. R. Wen, H. W. Hon, and Y. Yu, "Viewing term proximity from a different perspective," *Advances in Information Retrieval (ECIR 2008),* Springer LNCS 4956, pp. 346–357, 2008.

[52] K. Sparck Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments. Part 1," in *Information Processing and Management*, pp. 779–808, 2000.

[53] K. Sparck Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments. Part 2," in *Information Processing and Management*, pp. 809–840, 2000.

[54] T. Tao and C. Zhai, "An exploration of proximity measures in information retrieval," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 295–302, ACM, 2007.

[55] M. Taylor, H. Zaragoza, N. Craswell, S. E. Robertson, and C. Burges, "Optimisation methods for ranking functions with multiple parameters," in *Fifteenth Conference on Information and Knowledge Management (ACM CIKM)*, 2006.

[56] Terrier. Homepage. http://ir.dcs.gla.ac.uk/terrier.

[57] R. van Os D. Hiemstra, H. Rode, and J. Flokstra, "PF/Tijah: Text search in an XML database system," *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)*, pp. 12–17, http://dbappl.cs.utwente.nl/pftijah, 2006.

[58] C. J. van Rijsbergen, *Information Retrieval.* Butterworth, 1979.

[59] E. M. Voorhees and D. K. Harman, "Overview of the eighth text retrieval conference (TREC-8)," *The Eighth Text Retrieval Conference (TREC-8), NIST Special Publication* 500-246, pp. 1–24, 2000.

[60] Wumpus. Homepage. http://www.wumpus-search.org/.

[61] Xapian. http://xapian.org.

[62] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. E. Robertson, "Microsoft Cambridge at TREC 2004: Web and HARD track," in *The Thirteenth Text Retrieval Conference (TREC 2004), NIST Special Publication,* 500-261, 2005.

[63] Zettair. Homepage. http://www.seg.rmit.edu.au/zettair.