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

Full text available at: http://dx.doi.org/10.1561/1500000019
Foundations and Trends® in
Information Retrieval
Volume 3 Issue 4, 2009
Editorial Board

Editor-in-Chief:
Jamie Callan
Carnegie Mellon University
callan@cmu.edu

Fabrizio Sebastiani
Consiglio Nazionale delle Ricerche
fabrizio.sebastiani@isti.cnr.it

Editors
Alan Smeaton (Dublin City University)
Andrei Z. Broder (Yahoo! Research)
Bruce Croft (University of Massachusetts, Amherst)
Charles L.A. Clarke (University of Waterloo)
Ellen Voorhees (National Institute of Standards and Technology)
Ian Ruthven (University of Strathclyde, Glasgow)
James Allan (University of Massachusetts, Amherst)
Justin Zobel (RMIT University, Melbourne)
Maarten de Rijke (University of Amsterdam)
Marcello Federico (ITC-irst)
Norbert Fuhr (University of Duisburg-Essen)
Soumen Chakrabarti (Indian Institute of Technology)
Susan Dumais (Microsoft Research)
Wei-Ying Ma (Microsoft Research Asia)
William W. Cohen (CMU)
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.
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  
\hspace{1em} ser@microsoft.com
\(^2\) Yahoo! Research, Av. Diagonal 177, Barcelona 08028, Spain  
\hspace{1em} 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 1

2 Development of the Basic Model 3
   2.1 Information Needs and Queries 3
   2.2 Binary Relevance 4
   2.3 The Probability Ranking Principle 4
   2.4 Some Notation 5
   2.5 A Note on Probabilities and Rank Equivalence 12

3 Derived Models 15
   3.1 The Binary Independence Model 15
   3.2 Relevance Feedback and Query Expansion 17
   3.3 Blind Feedback 19
   3.4 The Eliteness Model and BM25 20
   3.5 Uses of BM25 28
   3.6 Multiple Streams and BM25F 29
   3.7 Non-Textual Relevance Features 33
   3.8 Positional Information 35
   3.9 Open Source Implementations of BM25 and BM25F 37

4 Comparison with Other Models 39
   4.1 Maron and Kuhns 39
   4.2 The Unified Model 40
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


References


References


References


