Contents

1 Introduction 1
1.1 Contributions 2
1.1.1 A new model of relevance 2
1.1.2 A new generative model 2
1.1.3 Minor contributions 3
1.2 Overview 3

2 Relevance 7
2.1 The many faces of relevance 7
2.1.1 A simple definition of relevance 7
2.1.2 User-oriented views of relevance 8
2.1.3 Logical views of relevance 9
2.1.4 The binary nature of relevance 10
2.1.5 Dependent and independent relevance 10
2.2 Attempts to Construct a Unified Definition of Relevance 12
2.2.1 Relevance in this book 15
2.3 Existing Models of Relevance 15
2.3.1 The Probability Ranking Principle 15
2.3.2 The Classical Probabilistic Model 17
2.3.3 The Language Modeling Framework 26
2.3.4 Contrasting the Classical Model and Language Models . 32

3 A Generative View of Relevance 37
3.1 An Informal Introduction to the Model 37
3.1.1 Representation of documents and requests 38
3.1.2 Advantages of a common representation 39
3.1.3 Information retrieval under the generative hypothesis . . 42
3.2 Formal Specification of the Model 44
3.3 Representation of Documents and Queries 45
3.3.1 Document and query generators 45
3.3.2 Relevant documents 45

3.3.3 Relevance in the information space . . . 46
3.3.4 Relevant queries . . . 46
3.3.5 Summary of representations . . . 47
3.4 Probability Measures . . . 47
3.4.1 Distribution over the representation space . . . 48
3.4.2 Distribution over documents and queries . . . 49
3.4.3 Significance of our derivations . . . 52
3.4.4 Summary of probability measures . . . 52
3.5 Relevance Models . . . 53
3.5.1 Frequentist interpretation: a sampling game . . . 54
3.5.2 Bayesian interpretation: uncertainty about relevance . . . 55
3.5.3 Multi-modal domains . . . 55
3.5.4 Summary of relevance models . . . 57
3.6 Ranked Retrieval . . . 57
3.6.1 Probability ranking principle . . . 57
3.6.2 Retrieval as hypothesis testing . . . 59
3.6.3 Probability ratio or KL-divergence? . . . 65
3.6.4 Summary of ranking methods . . . 67
3.7 Discussion of the Model . . . 68

4 Generative Density Allocation . . . 71
4.1 Problem Statement . . . 71
4.1.1 Objective . . . 72
4.2 Existing Generative Models . . . 72
4.2.1 The Unigram model . . . 73
4.2.2 The Mixture model . . . 74
4.2.3 The Dirichlet model . . . 76
4.2.4 Probabilistic Latent Semantic Indexing (pLSI) . . . 78
4.2.5 Latent Dirichlet Allocation . . . 79
4.2.6 A brief summary . . . 80
4.2.7 Motivation for a new model . . . 81
4.3 A Common Framework for Generative Models . . . 82
4.3.1 Unigram . . . 83
4.3.2 Dirichlet . . . 84
4.3.3 Mixture . . . 85
4.3.4 pLSI . . . 86
4.3.5 LDA . . . 88
4.3.6 A note on graphical models . . . 90
4.4 Kernel-based Allocation of Generative Density . . . 91
4.4.1 Delta kernel . . . 92
4.4.2 Dirichlet kernel . . . 94
4.4.3 Advantages of kernel-based allocation . . . 96
4.5 Predictive Effectiveness of Kernel-based Allocation . . . 99
4.6 Summary . . . 101

5 Retrieval Scenarios 103
5.1 Ad-hoc Retrieval 104
  5.1.1 Representation 104
  5.1.2 Examples of Relevance Models 107
  5.1.3 Experiments 108
5.2 Relevance Feedback 116
  5.2.1 Representation 116
  5.2.2 Experiments 118
5.3 Cross-Language Retrieval 119
  5.3.1 Representation 119
  5.3.2 Example of a cross-lingual relevance model 124
  5.3.3 Experiments 125
  5.3.4 Significance of the cross-language scenario 131
5.4 Handwriting Retrieval 131
  5.4.1 Definition 132
  5.4.2 Representation 133
  5.4.3 Experiments 136
5.5 Image Retrieval 138
  5.5.1 Representation 140
  5.5.2 Experiments 143
5.6 Video Retrieval 147
  5.6.1 Representation 148
  5.6.2 Experiments 149
5.7 Structured search with missing data 151
  5.7.1 Representation of documents and queries 153
  5.7.2 Probability distribution over documents and queries . . . 153
  5.7.3 Structured Relevance Model 154
  5.7.4 Retrieving Relevant Records 154
  5.7.5 Experiments 155
5.8 Topic Detection and Tracking 159
  5.8.1 Definition 159
  5.8.2 Representation 162
  5.8.3 Link detection algorithm 163
  5.8.4 Experiments 165

6 Conclusion 175
6.1 Limitations of our Work 179
  6.1.1 Closed-universe approach 179
  6.1.2 Exchangeable data 179
  6.1.3 Computational complexity 180
6.2 Directions for Future Research 181
  6.2.1 Relevance-based indexing 181
  6.2.2 Hyper-linked and relational data 181
  6.2.3 Order-dependent data 183
  6.2.4 Dirichlet kernels 183

References 185

Index 195