Selective Web Information Retrieval

UNIVERSITY
of
GLASGOW

Vasileios Plachouras
Department of Computing Science
Faculty of Information and Mathematical Sciences
University of Glasgow

A thesis submitted for the degree of
Doctor of Philosophy

©Vasileios Plachouras, 2006

Abstract

One of the main challenges in Web information retrieval is the number of different retrieval approaches that can be used for ranking Web documents. In addition to the textual content of Web documents, evidence from the structure of Web documents, or the analysis of the hyperlink structure of the Web, can be used to enhance the retrieval effectiveness. However, not all the queries benefit equally from applying the same retrieval approach. An additional challenge is posed by the fact that the Web enables users to seek information by searching and browsing. Therefore, users do not only perform typical informational search tasks, but also navigational search tasks, where the aim is to locate a particular Web document, which has already been visited before, or which is expected to exist.

In order to alleviate these challenges, this thesis proposes selective Web information retrieval, a framework formulated in terms of statistical decision theory, with the aim to apply an appropriate retrieval approach on a per-query basis. The main component of the framework is a decision mechanism that selects an appropriate retrieval approach on a per-query basis. The selection of a particular retrieval approach is based on the outcome of an experiment, which is performed before the final ranking of the retrieved documents. The experiment is a process that extracts features from a sample of the set of retrieved documents. This thesis investigates three broad types of experiments. The first one counts the occurrences of query terms in the retrieved documents, indicating the extent to which the query topic is covered in the document collection. The second type of experiments considers information from the distribution of retrieved documents in larger aggregates of related Web documents, such as whole Web sites, or directories within Web sites. The third type of experiments estimates the

usefulness of the hyperlink structure among a sample of the set of retrieved Web documents. The proposed experiments are evaluated in the context of both informational and navigational search tasks with an optimal Bayesian decision mechanism, where it is assumed that relevance information exists.

This thesis further investigates the implications of applying selective Web information retrieval in an operational setting, where the tuning of a decision mechanism is based on limited existing relevance information and the information retrieval system's input is a stream of queries related to mixed informational and navigational search tasks. First, the experiments are evaluated using different training and testing query sets, as well as a mixture of different types of queries. Second, query sampling is introduced, in order to approximate the queries that a retrieval system receives, and to tune an ad-hoc decision mechanism with a broad set of automatically sampled queries.

The main contributions of this thesis are the introduction of the selective Web information retrieval framework and the definition of a range of experiments. In addition, this thesis presents a thorough evaluation of a set of retrieval approaches for Web information retrieval, and investigates the automatic sampling of queries in order to perform the training of a decision mechanism.

Overall, selective Web information retrieval is a promising approach, which can lead to improvements in retrieval effectiveness. The evaluation of the decision mechanism and the experiments shows that it can be successfully employed for a particular type of queries, as well as a mixture of different types of queries.

Acknowledgements

I would like to thank the following people:

My supervisors, Iadh Ounis and Keith van Rijsbergen. Iadh, thank you for being a great supervisor; your support and feedback through the journey of my Ph.D have been unsurpassed. Keith, I am grateful for your advice and the discussions we had; you have always pointed me to very interesting directions of thinking and research.

Gianni Amati for many helpful discussions and ideas related to research.

Mark Baillie and Tassos Tombros for reading parts of this thesis, and giving me very useful feedback, as well as discussing earlier papers related to this thesis. I would also like to thank Craig Macdonald, Ben He, and Christina Lioma for reading parts of this thesis, and for their support in the last stages of writing up.

All the members of the Information Retrieval Group, past and present, for making it a great place to work.

All those with whom I had a wonderful time in the past and at present. Those include Areti, Anna, Allan, Harold, Fr. Eirinaios and Alexandra.

My brother, Diamantis, for encouraging me to do a Ph.D.

Finally, I would like to thank my parents, Pavlos and Niki, for their unconditional support. Their care and love made it possible for me to complete this work.

Contents

1 Introduction **1**

1.1 Introduction 1

1.2 Motivation 2

1.3 Thesis statement 3

1.4 Thesis outline 4

2 Basic Concepts of Information Retrieval **6**

2.1 Introduction 6

2.2 Indexing 7

2.3 Matching 8

2.3.1 Best Match weighting models 10

2.3.2 Language modelling 12

2.3.3 Divergence From Randomness framework 13

2.4 Evaluation 19

2.5 About Web information retrieval 21

3 Web Information Retrieval **22**

3.1 Introduction 22

3.2 Differences between classical and Web information retrieval 22

3.2.1 Hypertext document model 23

3.2.2 Structure of the Web 24

3.2.3 Quality of information on the Web 26

3.2.4 Background of Web users 27

3.3 Web-specific sources of evidence 28

3.3.1 Document and Web site structure 28

3.3.2 Hyperlink structure analysis 29
3.3.3 User interaction evidence 35
3.4 Combination of evidence for Web information retrieval 35
3.4.1 Extending Hyperlink analysis algorithms 36
3.4.2 Implicit hyperlink analysis with anchor text 37
3.4.3 Network-based models 38
3.4.4 Combination of different retrieval techniques and representations 39
3.5 Evaluation 43
3.5.1 Experimental evaluation in Text REtrieval Conference 43
3.5.2 Search engine evaluation 45
3.6 Query classification and performance prediction 46
3.6.1 Identifying user goals and intentions 46
3.6.2 Predicting query performance and dynamic combination of evidence 48
3.7 Summary 50

4 Retrieval Approaches for Selective Web Information Retrieval 51
4.1 Introduction 51
4.2 Experimental setting 52
4.3 Document representations for Web information retrieval 54
4.3.1 Representing Web documents 55
4.3.2 Parameter setting 56
4.3.3 Evaluation results 61
4.3.4 Impact of query terms with high frequency on the Poisson-based models 64
4.3.5 Discussion and Conclusions 66
4.4 Combining document fields 67
4.4.1 Weighting models for field retrieval 68
4.4.2 Parameter setting for field-based weighting models 70
4.4.3 Evaluation of field-based weighting models 72
4.4.4 Discussion and conclusions 73
4.5 Query-independent evidence 74
4.5.1 URLs of Web documents 74
4.5.2 Hyperlink structure analysis 77

4.5.3 Evaluation of field retrieval with query-independent evidence . . 88
4.5.4 Summary and conclusions 91
4.6 Obtaining a realistic parameter setting 92
4.6.1 Using mixed tasks 92
4.6.2 Using mixed tasks and restricted optimisation 95
4.6.3 Conclusions 97
4.7 Potential improvements from selective Web information retrieval 97
4.8 Summary 99

5 A framework for Selective Web Information Retrieval 103
5.1 Introduction 103
5.2 Selective retrieval as a statistical decision problem 104
5.2.1 Selective Web information retrieval and related work 108
5.2.2 Decision mechanism with known states of nature 109
5.3 Retrieval score-independent experiments 110
5.3.1 Document-level experiments 110
5.3.2 Aggregate-level experiments 112
5.4 Retrieval score-dependent experiments 115
5.4.1 Divergence between probability distributions 116
5.4.2 Usefulness of hyperlink structure 117
5.5 Bayesian decision mechanism 122
5.5.1 Definition of the Bayesian decision mechanism 122
5.5.2 Application of the Bayesian decision mechanism 125
5.5.3 Density estimation 126
5.6 Summary 128

6 Evaluation of Selective Web Information Retrieval 130
6.1 Introduction 130
6.2 Evaluation methodology 131
6.2.1 Effectiveness of experiments $\mathcal{E}$ 131
6.2.2 Evaluation setting 132
6.2.3 Presentation and analysis of results 135
6.3 Evaluation of score-independent experiments 136
6.3.1 Document-level experiments 136

6.3.2 Aggregate-level experiments 141
6.3.3 Conclusions 151
6.4 Evaluation of score-dependent experiments 151
6.4.1 Setting the score distribution S_n 152
6.4.2 Evaluation results of experiments based on the usefulness of hyperlink structure $L(S_n, U_n)$ 154
6.4.3 Evaluation results of experiments based on the usefulness of hyperlink structure $L(S_n, U'_n)$ 157
6.4.4 Example of the usefulness of hyperlink structure experiments . . 159
6.4.5 Discussion 161
6.4.6 Conclusions 162
6.5 Document sampling 163
6.5.1 Revisiting the definition of experiments $\mathcal{E}$ 164
6.5.2 Description of experimental setting and presentation of results . 165
6.5.3 Document sampling for score-independent document-level experiments 166
6.5.4 Document sampling for score-independent aggregate-level experiments 167
6.5.5 Document sampling for score-dependent experiments 173
6.5.6 Discussion 175
6.5.7 Conclusions 178
6.6 Using retrieval approaches based on the same weighting model 179
6.7 Decision mechanism with more than two retrieval approaches 181
6.8 Discussion 185
6.9 Summary 188

7 Selective Web Information Retrieval with Limited Relevance Information 190
7.1 Introduction 190
7.2 Limited relevance information 191
7.2.1 Modelling limited relevance information 191
7.2.2 Experimental setting for limited relevance information 193
7.3 Evaluation of experiments $\mathcal{E}$ with limited relevance information 195

7.3.1 Score-independent experiments with limited relevance information 195
7.3.2 Score-dependent experiments with limited relevance information 196
7.3.3 Discussion and conclusions 198
7.4 Ad-hoc decision mechanism and query sampling 199
7.4.1 Ad-hoc decision mechanism 200
7.4.2 Query sampling 201
7.4.3 Evaluation of query sampling 204
7.4.4 Evaluation of ad-hoc decision mechanism 210
7.4.5 Conclusions 214
7.5 Summary 215

8 Conclusions and Future Work 217
8.1 Contributions and conclusions 217
8.1.1 Contributions 217
8.1.2 Conclusions 218
8.2 Future work 222

A Parameter settings and evaluation of retrieval approaches 225

B Evaluation of experiments $\mathcal{E}$ 236

References 280

List of Figures

2.1 The architecture of a basic information retrieval system. 7

3.1 Hubs and authorities as a bipartite graph. 32

4.1 The obtained mean average precision (MAP) for different c values tested during the two-step optimisation of full text retrieval with PL2 for the topic sets tr2001, td2004, hp2004 and np2004. 58

4.2 The obtained mean average precision (MAP) for different c values tested during the two-step optimisation of anchor text retrieval with PL2 for the topic sets tr2001, td2004, hp2004 and np2004. 60

4.3 The monotonically decreasing transformation for the URL path length, for $k_u = 1, 10$ and 100. 76

4.4 The Markov Chain representing the Web graph. 80

4.5 The extended Markov Chain including the clone states. 84

4.6 The monotonically increasing transformation of the hyperlink structure analysis scores, for $k_L = 1, 10$ and 100. 88

5.1 Selective application of retrieval approaches for three states of nature s_1, s_2, s_3 and three different retrieval approaches a_1, a_2, a_3. The loss associated with applying retrieval approach a_i when the true state of nature is s_j is denoted by $l(a_i, s_j)$. 106

5.2 The hyperlink graphs of the ranked documents, corresponding to the first three cases described in the Example 6. 121

5.3 Example of a Bayesian decision mechanism with 3 available actions, the corresponding posterior likelihoods and the loss 124

5.4 Box-and-whisker plots of the score-independent document-level experiment outcome values for the task td2003 127

6.1 Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.2. 139
6.2 Posterior likelihoods of the experiments $\mathcal{E}_{\exists(b)}$ and $\mathcal{E}_{\forall(b)}$ for the topic set hp2004. 140
6.3 Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.3. 144
6.4 Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.4. 147
6.5 Posterior likelihoods of the score-independent aggregate-level experiments $\mathcal{E}_{\exists b,avg(dom)}$ and $\mathcal{E}_{\exists b,avg(dir)}$, for the topic set hp2004, where one of the retrieval approaches PB2FU or DLHFA is selectively applied for each query. The posterior likelihoods for the domain and the directory based aggregates are presented on top, and the bottom diagram, respectively. . 149
6.6 Density estimates of the usefulness of the hyperlink structure experiments, according to whether an optimised or the default parameter setting is used. 153
6.7 Histogram summarising the relative differences between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.5. 157
6.8 Histogram summarising the relative differences between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.6. 159
6.9 Posterior likelihoods of the score-dependent experiments for the topic set td2003, where one of the retrieval approaches I(n_e)C2FU or DLHFP is selectively applied on a per-query basis. 160

6.10 Histogram summarising the relative differences between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.7. 168
6.11 Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.8. 171
6.12 Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.9. 173
6.13 Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.10. 175
6.14 Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.11. 177
6.15 Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.13. 185

List of Tables

2.1 The formulae of the weighting models PL2, PB2, I(n_e)C2, DLH, and BM25, respectively. 19

4.1 The search tasks and the corresponding topic sets from the TREC Web tracks. 53

4.2 The average length of documents for the different document representations in WT10g and .GOV test collections. The document length corresponds to the number of indexed tokens for each document, after removing stop words. 59

4.3 The average length of relevant documents for the different topic sets, and for the different document representations. The document length corresponds to the number of indexed tokens for each document, after removing stop words. 61

4.4 Evaluation of different document representations with the weighting models PL2, PB2, I(n_e)C2, DLH and BM25. 63

4.5 Mean Average Precision (MAP) for full text retrieval with the weighting models PL2 and PB2, when query terms with $\lambda > 1$ are employed for assigning weights to documents, or they are treated as stop words. . . . 66

4.6 Evaluation results of the best official submitted runs to the Web tracks from TREC-9 to TREC 2004. 67

4.7 Evaluation of the weighting models PL2F, PB2F, I(n_e)C2F, DLHF and BM25F. 72

4.8 Evaluation results of the combinations of field-based retrieval with the query-independent evidence from the URL path length, PageRank, and the Absorbing Model. 90

4.9 The evaluation of the field retrieval weighting models and their combination with the query-independent evidence for the mixed-type query sets, and for the query-type specific topic subsets. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task. 94

4.10 The evaluation of the field retrieval weighting models and their combination with the query-independent evidence for the mixed-type query sets and for the query-type specific topic subsets, with restricted optimisation. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task. 96

4.11 Potential for improvements in retrieval effectiveness from the selective application of two retrieval approaches on a per-query basis. The retrieval approaches are based on a restricted optimisation, as reported in Table 4.10. The table displays the pairs of retrieval approaches that result in the highest improvements in MAP for the tested topic sets. . . 100

5.1 Notation examples for the aggregate-level experiments. 114

6.1 The pairs of retrieval approaches employed by the Bayesian decision mechanism in the evaluation of the proposed experiments $\mathcal{E}$. The columns 'First approach' and 'Second approach' show the employed retrieval approaches and their MAP for the corresponding task within brackets. The column 'MAX' shows the maximum MAP that can be obtained by selectively applying one of the two retrieval approaches on a per-query basis. 134

6.2 Evaluation of score-independent document-level experiments $\mathcal{E}_{\exists(f)}$ and $\mathcal{E}_{\forall(f)}$ for combination of fields f, which result in at least one decision boundary for each tested topic set. 138

6.3 Evaluation of score-independent aggregate-level experiments with domains, which result in at least one decision boundary for each tested topic set. 143

6.4 Evaluation of score-independent aggregate-level experiments with directories, which result in at least one decision boundary for each tested topic set. 146

6.5 Evaluation of score-dependent experiments based on estimating the usefulness of the hyperlink structure $L(S_n, U_n)$, which result in at least one decision boundary for each tested topic set. 156

6.6 Evaluation of score-dependent experiments based on estimating the usefulness of the hyperlink structure $L(S_n, U'_n)$, which result in at least one decision boundary for each tested topic set. 159

6.7 The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs score-independent document level experiments with document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. 167

6.8 The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. The experiments compute the average domain or directory aggregate sizes. 170

6.9 The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. The experiments compute the standard deviation of the domain or directory aggregate sizes. 172

6.10 The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. The experiments compute the number of large domain or directory aggregates. 174
6.11 The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs the score-dependent experiments and document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. . 176
6.12 The number of times for which there is at least one decision boundary ('B>0'), or improvements in retrieval effectiveness ('+'), when the Bayesian decision mechanism selectively applies retrieval approaches, which use the same field-based weighting model. 180
6.13 Evaluation of the decision mechanism, which employs the retrieval approaches PL2FA, I(n_e)C2FU, and BM25FP, for the experiments that identify at least one decision boundary for all the tested tasks, and result in improvements in retrieval effectiveness for at least three tested tasks. 184

7.1 Evaluation of a decision mechanism with known states of nature for mixed-type queries. 194
7.2 Evaluation of the score-independent document-level and aggregate-level experiments with limited relevance information. 197
7.3 Evaluation of score-dependent experiments with limited relevance information. 198
7.4 Average and standard deviation for the length of the TREC 2003 and 2004 Web track queries. 206

7.5 Symmetric Jensen-Shannon (J-S) divergence between the distribution of experiment outcome values for the generated queries with STS, MTS, and ATS and the TREC 2003 and 2004 Web track queries (mq2003 and mq2004). The experiments are $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$. The mean and standard deviation of the query length distribution in MTS are denoted by μ and σ. 207
7.6 Evaluation of the ad-hoc decision mechanism with the experiments $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$. 212

A.1 Parameter values for retrieval from the full text, title, headings, and anchor text of documents, with the DFR weighting models PL2, PB2 and I(n_e)C2, and the weighting model BM25. 226
A.2 The values of the c parameters and the weights of the fields for the weighting models PL2F, PB2F and I(n_e)C2. 227
A.3 The weights of the anchor text and title fields for the weighting model DLHF. 228
A.4 The values of the parameters for the weighting model BM25F. 228
A.5 Precision at 10 retrieved documents (P10) for field retrieval and combination with query-independent evidence. 230
A.6 Mean reciprocal rank of the first retrieved relevant document (MRR1) for field retrieval and combination with query-independent evidence. . . 231
A.7 Number of retrieved relevant documents for field retrieval and combination with query-independent evidence. 232
A.8 The parameter values for the combination of the weighting models with the query-independent evidence. 233
A.9 The values of the parameters and the weights of the fields for the weighting models PL2F, PB2F, I(n_e)C2, DLHF and BM25F for training and evaluating with different mixed tasks. The parameter values used for the mixed tasks are the ones used for their corresponding subsets of tasks. . 234

A.10 The values of the parameters for the combination of each field retrieval weighting model and the query-independent evidence for training and evaluating with different mixed tasks. The parameter values used for the mixed tasks are the ones used for their corresponding subsets of tasks. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task. 234
A.11 The values of the parameters and the weights of the fields for the weighting models PL2F, PB2F, I(n_e)C2, DLHF and BM25F for training and evaluating with mixed tasks, and restricted optimisation. The parameter values used for the mixed tasks are the ones used for their corresponding subsets of tasks. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task. 235
A.12 The values of the parameters for the combination of each field retrieval weighting model and the query-independent evidence for training and evaluating with mixed tasks, and restricted optimisation. The parameter values used for the mixed tasks are the ones used for their corresponding subsets of tasks. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task. 235

B.1 Evaluation of experiments $\mathcal{E}_{\exists(f)}$ and $\mathcal{E}_{\forall(f)}$. 239
B.2 Evaluation of experiments $\mathcal{E}_{\exists(f),avg(dom)}$ and $\mathcal{E}_{\forall(f),avg(dom)}$. 240
B.3 Evaluation of experiments $\mathcal{E}_{\exists(f),std(dom)}$ and $\mathcal{E}_{\forall(f),std(dom)}$. 242
B.4 Evaluation of experiments $\mathcal{E}_{\exists(f),lrg(dom)}$ and $\mathcal{E}_{\forall(f),lrg(dom)}$. 244
B.5 Evaluation of experiments $\mathcal{E}_{\exists(f),avg(dir)}$ and $\mathcal{E}_{\forall(f),avg(dir)}$. 245
B.6 Evaluation of experiments $\mathcal{E}_{\exists(f),std(dir)}$ and $\mathcal{E}_{\forall(f),std(dir)}$. 247
B.7 Evaluation of experiments $\mathcal{E}_{\exists(f),lrg(dir)}$ and $\mathcal{E}_{\forall(f),lrg(dir)}$. 249
B.8 Evaluation of experiments $\mathcal{E}_{\exists(f),L(SU)_{pl}}$ and $\mathcal{E}_{\forall(f),L(SU)_{pl}}$. 250
B.9 Evaluation of experiments $\mathcal{E}_{\exists(f),L(SU)_{in}}$ and $\mathcal{E}_{\forall(f),L(SU)_{in}}$. 252
B.10 Evaluation of experiments $\mathcal{E}_{\exists(f),L(SU')_{pl}}$ and $\mathcal{E}_{\forall(f),L(SU')_{pl}}$. 254
B.11 Evaluation of experiments $\mathcal{E}_{\exists(f),L(SU')_{in}}$ and $\mathcal{E}_{\forall(f),L(SU')_{in}}$. 255

B.12 Evaluation of the score-independent document-level and aggregate-level experiments with limited relevance information. The table displays the evaluation results of a decision mechanism, which is trained and evaluated with different mixed tasks. 257
B.13 Evaluation of the score-dependent experiments with limited relevance information. The table displays the evaluation results of a decision mechanism, which is trained and evaluated with different mixed tasks. 258

Chapter 1

Introduction

1.1 Introduction

This thesis investigates the selective application of different approaches for information retrieval (IR) with documents from the World Wide Web (Web). The main argument of the thesis is that selective Web IR, a technique by means of which appropriate retrieval approaches are applied on a per-query basis, can lead to improvements in retrieval effectiveness. Two main issues are addressed. First, a range of retrieval approaches is evaluated for different test collections and search tasks, in order to establish the potential for improvements from selective Web IR. Second, a decision theoretical framework for selective Web IR is introduced and evaluated in both an optimal and a realistic setting, with limited relevance information.

The advent of the Web and the resulting wide use of Web search engines has resulted in a range of developments to combine and enhance classical IR techniques with Web-specific evidence. Most of the proposed approaches in the literature investigate a uniform combination of evidence, which is applied for all queries. Recent works have also focused on predicting the query difficulty, and proposing measures, which correlate statistical features of the retrieved documents for a particular query with the performance of a system. This thesis is focused on selectively applying the most effective retrieval approach on a per-query basis, in order to improve the retrieval effectiveness. The evaluation of selective Web IR is performed with different search tasks, as defined in the TREC 2003 and 2004 Web tracks (Craswell & Hawking, 2004; Craswell et al., 2003).

The remainder of the introduction describes the motivation for the work in this thesis, presents the statement of its aims and contributions, and closes with an overview of the structure for the remainder of the thesis.

1.2 Motivation

IR has been an active field of research for more than 30 years, starting as a need to search and locate information in the ever-growing body of scientific literature. While IR systems have always been useful in libraries, the advent of the Web made IR systems an essential tool for a wide range of people. Indeed, the Web was conceived as a virtual information space, which would facilitate sharing of information among scientists. At the beginning, finding information on the Web was a matter of keeping a set of pointers to interesting Web documents. However, as the number of Web documents grew rapidly, this became impractical. The first IR systems for searching the Web, also known as search engines, appeared as early as 1994 (McBryan, 1994). Today, there are several general purpose search engines as well as a large number of specialised search engines[1].

Classical IR systems have been primarily used in controlled settings where documents are rarely updated, information is considered to be reliable, and users are experts in the field of search. In contrast, the Web is a highly diverse and dynamic environment, where new information is published and existing information may be modified or become unavailable. In addition, the available information may be erroneous or intentionally misleading. The users that access the Web have a wide range of backgrounds and interests, making it impossible to assume that they have experience on the topic they search for, or on how to use a search engine effectively. They tend to formulate short queries and examine only the top ranked results (Jansen & Pooch, 2001; Silverstein et al., 1999). Furthermore, the queries are not always about finding out information related to a topic. Broder (2002) identified a taxonomy of three main types of Web search tasks. First, informational search tasks are about finding information and useful pointers about a topic. Second, navigational search tasks are about locating a particular Web document, that a user has visited in the past. Third, transactional search tasks are about accessing particular resources, or buying products.

[1] A extensive list can be found at http://www.searchenginewatch.com/links/ (visited on 17th October, 2005).

In addition, the Web offers a range of evidence that can be used to enhance the effectiveness of classical retrieval techniques, which are based on the analysis of the documents' textual content. A key element of the Web is the hypertext document model, which enables documents to directly reference other documents with hyperlinks. These hyperlinks can serve as navigational aids within a set of Web documents, or as pointers to other related Web documents. Similarly to work in the field of citation analysis for scientific journals (Garfield, 1972), a Web document pointed to by many other Web documents may be considered as popular, or authoritative. Evidence such as the popularity or authority of documents can be used to improve the effectiveness of a Web IR system, by first retrieving relevant documents of higher quality. Thus, relevance is not replaced, but only complemented.

Generally, the non-textual evidence have been used in a static manner, where they are applied for each query uniformly. However, their weaker nature in indicating the relevance of documents (Croft, 2000) suggests that alternative ways to incorporate them dynamically, according to the context of documents and queries, can lead to improvements in retrieval effectiveness. This thesis is also related to recent techniques for estimating the query difficulty, and consequently predicting the performance of an IR system. Estimators of the query difficulty have been based on the statistical properties of the query terms (He & Ounis, 2004), or on the co-occurrence of query terms in the retrieved documents (Yom-Tov et al., 2005).

1.3 Thesis statement

The statement of this thesis is that the retrieval effectiveness of an IR system can be enhanced by applying an appropriate retrieval approach on a per-query basis. This is investigated in the context of a framework for selective Web IR, where a decision mechanism selects appropriate retrieval approaches to apply on a per-query basis. The decision mechanism performs an experiment $\mathcal{E}$, which extracts features from a sample of the set of retrieved documents, and according to the outcome of $\mathcal{E}$, it applies an appropriate retrieval approach. If the experiment $\mathcal{E}$ is successful in identifying the most appropriate retrieval approaches, then the decision mechanism is expected to result in improved retrieval effectiveness, compared to the uniform application of a single retrieval approach.

The main contributions of this thesis are the following. A decision theoretical framework for selective Web IR is introduced. The framework is evaluated in a setting, where relevance information is assumed to exist, and it is shown that it is possible to obtain improvements in retrieval effectiveness from the selective application of different retrieval approaches. The evaluation of the proposed framework is also performed in a setting where limited relevance information exists. In this context, query sampling techniques are introduced and evaluated with respect to their effectiveness in setting up an ad-hoc decision mechanism. Moreover, a thorough evaluation of several retrieval approaches for Web IR is performed on different test collections and search tasks.

1.4 Thesis outline

The remainder of the thesis is organised in the following way.

- Chapter 2 provides a brief overview of the main concepts of IR. It describes a series of IR models, including those used in this thesis, as well as the experimental evaluation of IR systems.

- Chapter 3 presents in detail work related to Web IR. It provides an overview of the hypertext document model used for Web documents, and of the hyperlink structure of the Web. It discusses how particular features of the Web can be used to enhance the retrieval performance of Web IR systems. It also reviews issues related to the evaluation of Web IR systems, as well as the identification of the user goals, and the prediction of query performance.

- Chapter 4 investigates the potential for improvements in retrieval effectiveness from selective Web IR. First, it examines the effectiveness of performing retrieval with a range of weighting models from different representations of Web documents, such as the body, the title, the headings, and the anchor text of incoming hyperlinks. Next, it discusses the combination of different fields, which correspond to the text within particular tags of the HyperText Markup Language (HTML). The proposed approaches consider both the length normalisation and the weighting of the fields. The retrieval of documents with fields is further enhanced with query-independent evidence. The effectiveness of each retrieval approach is studied with respect to its optimal retrieval effectiveness for several types of search

tasks. This chapter also considers a realistic setting, where a restricted optimisation for mixed search tasks is performed for each retrieval approach. Finally, the chapter establishes the potential for improvements in retrieval effectiveness from applying selective Web IR.

- Chapter 5 introduces the framework for applying selective Web IR. First, it provides a description of the selection mechanism in terms of statistical decision theory. Then, the chapter defines a range of experiments, which aid the decision mechanism to select a retrieval approach to apply on a per-query basis. Finally, the chapter closes with the definition of an optimal Bayesian decision mechanism for the evaluation of the proposed experiments.

- Chapter 6 presents the evaluation of the proposed framework for selective Web IR. First, it employs the retrieval approaches described in Chapter 4, and evaluates the proposed experiments in the context of a Bayesian decision mechanism, as described in Chapter 5, with several types of search tasks. Second, the chapter investigates the use of small samples of documents in order to compute the experiments.

- Chapter 7 explores how selective Web IR can be applied when a retrieval system has only limited relevance information available. This corresponds to training and testing a decision mechanism with different sets of mixed tasks. The automatic generation of query samples is also investigated, in order to facilitate the training of an ad-hoc decision mechanism.

- Chapter 8 closes this thesis with the contributions and the conclusions drawn from this work, as well as possible directions of future work for extending the proposed framework for selective Web IR.

Chapter 2

Basic Concepts of Information Retrieval

2.1 Introduction

Information Retrieval (IR) deals with the efficient storage and access of information items (Baeza-Yates & Ribeiro-Neto, 1999). The information items can be text documents, images, video, *etc.* A common scenario of the use of an IR system is the following: while performing a task, a user needs to locate information in a repository of documents. The user expresses an *information need* in the form of a *query*, which usually corresponds to a bag of keywords. The user is only interested in the documents that are *relevant* to his information need. The ideal goal of an IR system is to return all the relevant documents, while not retrieving any non-relevant ones. Furthermore, the retrieved documents should be ranked from the most relevant to the least relevant. The above process is iterative in the sense that a user can refine the initial query, or provide feedback to the system, which leads to the retrieval process being performed again. This thesis is focused on retrieval from text and Web documents.

Automatically deciding whether a document is relevant to the information need of a user is not a straight-forward task, because of the inherent ambiguity in formulating a query for an information need, as well as the ambiguity of information in documents. This is a main difference between Information Retrieval and Data Retrieval, where the items to be retrieved must clearly satisfy a set of conditions, which can be easily verified (Van Rijsbergen, 1979). The current chapter provides an overview of basic concepts in IR regarding the indexing of documents (Section 2.2), the matching of

documents and queries (Section 2.3), and the evaluation of IR systems (Section 2.4).

2.2 Indexing

In order for an IR system to process queries from users, it is required to extract and store in an efficient way a representative for the documents to be searched. Creating the document representatives, or the document index, takes place in the indexing component of an IR system, as shown in Figure 2.1.

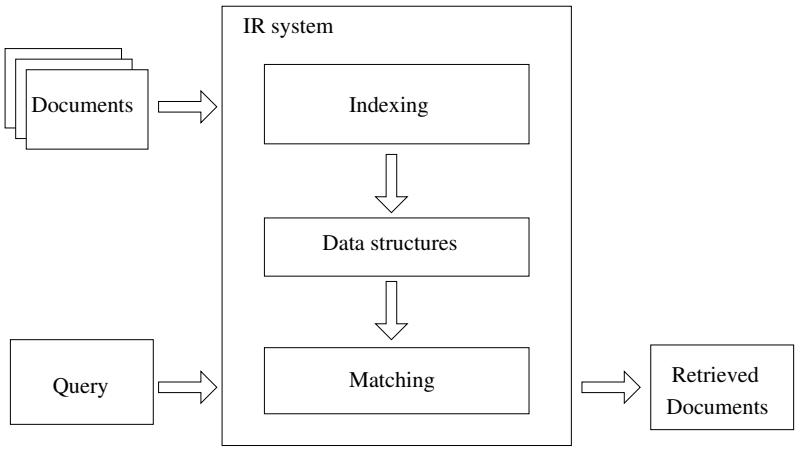

Figure 2.1: The architecture of a basic information retrieval system.

The simplest approach is to represent a document by its composing terms. However, not all the terms in a document carry the same amount of information about the topic of the document. Luhn (1958) proposed that the frequency of a term within a document can be used to indicate its significance in the document. In addition, there is a number of terms that appear very frequently in many documents, without being related to a particular topic. Such terms are called *stop words* and they can be discarded during the indexing process. A benefit from ignoring stop words during indexing is that the size of the generated document index is reduced.

Another common type of lexical processing of terms during indexing is stemming. The purpose of stemming is to replace a term by its stem, so that different grammatical forms of words are represented in the same way. For example, if the terms 'retriever', 'retrieval', and 'retrieving' appear in the text, they can be represented by the common stem 'retriev'. However, once these terms have been stemmed, any difference in their

meaning is lost. A widely used stemming algorithm for the English language was proposed by Porter (1980).

Instead of indexing single terms, more complicated strategies can be adopted, in which the indexing units are combinations of consecutive terms. For example, an IR system can index pairs of consecutive words, also known as *bigrams* (Manning & Schutze, 1999, ch. 6). The document index may also contain additional information, such as the positions of terms in a document, or whether the terms appear in particular fields of documents. For the purpose of this thesis, the documents are indexed using single terms, their frequencies and field information.

The output of the indexing process is a set of data structures that enables the efficient access of the document representatives. The most commonly used data structure is the inverted file (Frakes & Baeza-Yates, 1992), which stores the document identifiers that contain a particular term from the vocabulary of the indexed documents. Generally, the size of the inverted file is comparable to that of the document collection. However, it can be reduced by using appropriate compression techniques, based on encoding the integers that represent the document identifiers and the term frequencies with fewer bits. The commonly used encodings are the Elias gamma encoding for compressing the differences between a sequence of document identifiers, and the unary encoding for compressing term frequencies (Witten et al., 1994). These encodings achieve very good compression, but operate on a bit level, and require many operations for compressing and decompressing. Other compression techniques operate on bytes in order to exploit the optimised capacity of hardware to handle bytes (Williams & Zobel, 1999).

2.3 Matching

The second main component of an IR system, as shown in Figure 2.1, is the matching component that retrieves a set of documents for a given query. A user submits a query to an IR system, which aims to retrieve documents relevant to the query. Several models have been developed for matching documents to queries. The Boolean model (Belew, 2000), which is the oldest IR model, treats the query as a Boolean expression. An example of such a Boolean query is `information AND search AND (NOT storage)`. The Boolean model for this particular query would retrieve all the documents that contain

the terms *information*, *search*, but do not contain the term *storage*. The documents are presented to the user as a set, without any particular ranking. This lack of ranking of the results has been one of the main points of criticism for the Boolean model (Salton et al., 1983).

A different class of models is based on computing the similarity between the query and the documents. One such model is the vector space model (Salton & McGill, 1986), where both the queries and the documents are represented as vectors in the same space. The number of the dimensions of the vector space corresponds to the size of the vocabulary of the document index, or in other words, the number of distinct terms in the documents. The retrieved documents are ranked according to their similarity to the query, which corresponds to the distance between points in the vector space. Several distance functions can be defined and used to measure the similarity (Van Rijsbergen, 1979).

Another classical retrieval model is the probabilistic model (Robertson & Sparck Jones, 1976). This model is based on estimating the probability of relevance for a document, given a query. It assumes that there is some knowledge of the distribution of terms in the relevant documents and this distribution is refined through the iterative interaction with the user. Van Rijsbergen (1979) presents a decision theoretic interpretation of the probabilistic retrieval models, where a document is retrieved if the probability of being relevant to a given query is greater than the probability of the document being non-relevant. Through the definition of a loss function for the possible actions of retrieving or not retrieving a document, the number of retrieved documents can be adjusted in an appropriate way.

A series of simple and effective IR models have been based on the 2-Poisson indexing model (Harter, 1975), which aims to assign a set of *specialty* or useful index terms to documents. The set of *elite* documents, which are indexed with a particular specialty term, would be the answer to a query consisting of that specialty term. The specialty terms are identified by means of their different distributions in the elite documents, and in the documents that do not have the eliteness property. The two distributions are modelled as two different Poisson distributions. Robertson et al. (1981) combined the 2-Poisson model with the probabilistic model for retrieval, as described in Section 2.3.1.

The remainder of the current section describes particular families of IR models. Section 2.3.1 describes the family of *Best Match* (BM) models, which combines the

probabilistic model with the 2-Poisson model. A different family of IR models, based on language modelling, is briefly discussed in Section 2.3.2. Section 2.3.3 presents the Divergence From Randomness (DFR) framework of IR models, which is based on a generalisation of the 2-Poisson indexing model.

2.3.1 Best Match weighting models

Starting from a basic probabilistic model (Robertson & Sparck Jones, 1976), the weight of a term t in a document, assuming that the terms appear in documents independently from each other, is computed as follows:

$$w^{(1)} = \log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)} \tag{2.1}$$

where R is the number of relevant documents, r is the number of relevant documents that contain the query term t, N is the number of documents in the collection and n is the document frequency of the term t, or in other words the number of documents that contain the term t. When there is no relevance information, the above weight $w^{(1)}$ becomes (Croft & Harper, 1988):

$$w^{(1)} = \log \frac{N - n + 0.5}{n + 0.5} \tag{2.2}$$

which is similar to the inverse document frequency (idf): $\log \frac{N}{n}$

The above equations do not incorporate the within-document frequency of terms. Robertson et al. (1981) modelled the within-document term frequencies with two Poisson distributions: one distribution for modelling the occurrences of the term t in the relevant documents, and another distribution for modelling the occurrences of the term t in the non-relevant documents. This approach leads to the introduction of a substantial number of parameters that cannot be set in a straight-forward manner. For this reason, Robertson & Walker (1994) approximated the above model of term frequencies with a simple formula, that has a similar shape and properties. They identified four properties: (a) the weight should be zero when the term frequency is zero, (b) the weight should increase monotonically as the term frequency increases, (c) the weight should increase to an asymptotic maximum, and (d) this asymptotic maximum corresponds to the weight $w^{(1)}$. A formula that satisfies these properties is the following:

$$w = \frac{tf}{k_1 + tf} w^{(1)} \tag{2.3}$$

where k_1 is a parameter that controls the saturation of the term frequency *tf*.

By incorporating the frequency *qtf* of a term t in the query, and a correction for the length l of a document, Robertson and Walker derived the formula BM15 for computing the weight $w_{d,q}$ of a document for query:

$$w_{d,q} = \sum_{t \in q} w = \sum_{t \in q} \left(\frac{tf}{k_1 + tf} \cdot \frac{qtf}{k_3 + qtf} \cdot \log \frac{N - n + 0.5}{n + 0.5}\right) + k_2 \cdot nq \cdot \frac{\bar{l} - l}{\bar{l} + l} \tag{2.4}$$

where k_3 controls the saturation of the term frequency in the query, $\bar{l}$ is the average document length in the collection, k_2 is the weight of the document length correction, and nq is the number of terms in the query.

In addition, they introduced BM11, a different version of the formula that normalises the term frequency with respect to the document length:

$$w_{d,q} = \sum_{t \in q} \left(\frac{tf}{l/\bar{l} + tf} \cdot \frac{qtf}{k_3 + qtf} \cdot \log \frac{N - n + 0.5}{n + 0.5}\right) + k_2 \cdot nq \cdot \frac{\bar{l} - l}{\bar{l} + l} \tag{2.5}$$

Further research led to the introduction of the BM25 formula, which is a combination of BM11 and BM15, with the addition of the scaling factors $(k_1 + 1)$ and $(k_3 + 1)$ (Robertson et al., 1994):

$$w_{d,q} = \sum_{t \in q} \left(\frac{(k_1 + 1)tf}{(k_1(1 - b) + b\frac{l}{\bar{l}}) + tf} \cdot \frac{(k_3 + 1)qtf}{k_3 + qtf} \cdot \log \frac{N - n + 0.5}{n + 0.5}\right) + k_2 \cdot nq \cdot \frac{\bar{l} - l}{\bar{l} + l} \tag{2.6}$$

Indeed, if $b = 0$, then the formula BM15 is obtained, while if $b = 1$, then the formula BM11 is obtained. In most of the reported experiments, the document length adjustment $k_2 \cdot nq \cdot \frac{\bar{l}-l}{\bar{l}+l}$ has been ignored by setting $k_2 = 0$. In addition, when k_3 is very large, then the component $\frac{(k_3+1)qtf}{k_3+qtf} \approx qtf$.

In the Formulae (2.4), (2.5), and (2.6), when the document frequency $n > N/2$, the resulting weight of a particular query term in a document is negative. Fang et al. (2004) introduced a modified version of the BM25 formula, where $\log \frac{N-n+0.5}{n+0.5}$ is replaced with $\log \frac{N+1}{n}$, so that the computed weights are always positive. In the remainder of this thesis, when BM25 is employed for ranking documents and a term with a very high document frequency appears in a query, any resulting negative weight is ignored and it does not contribute to the weight of the document for the query.

2.3.2 Language modelling

The retrieval models based on the 2-Poisson indexing model make either explicit or implicit assumptions about the distribution of terms in documents. However, Ponte & Croft (1998) suggested that it is preferable to use the available data, instead of making any parametric assumptions about the distribution of terms. This view led to the application of language modelling for IR. In this approach, a data model is generated for each document. For a given query, the documents are ranked according to the probability that the corresponding document model generates the query. Ponte & Croft (1998) treated the queries as a set of words with binary weights. The probability of generating the query from a document language model corresponds to the product of the probabilities of generating each of the query terms times the product of the probability of not generating the terms that do not appear in the query.

Hiemstra (1998) modelled the queries as a sequence of terms and computed the probability of generating the query according to the product of the probabilities of generating the query terms from the document language model. In this approach, it is not necessary to consider the terms that do not appear in the query, and the resulting model is simpler than that of Ponte & Croft (1998).

In all language modelling approaches, there is the issue of assigning probabilities to the terms that do not appear in a document. Ponte & Croft (1998) suggested that it is harsh to assign a zero probability to a term that does not appear in a document. For this reason, smoothing techniques for the probability distribution of the language models have been employed. Ponte & Croft (1998) proposed to use the probability that a term occurs in the document collection, when it does not appear in a document. Hiemstra (1998) employed a smoothing approach based on the linear interpolation of the probabilities from the document model and the collection model. A study of the effectiveness of different smoothing techniques for language modelling in IR was conducted by Zhai & Lafferty (2001).

The ranking of documents according to the probability of generating the query has been criticised by Robertson (2002), because it implies that there is only one *ideal* document that is relevant to the query, and it could not be used to model relevance feedback. Further work with the language modelling has led to the introduction of approaches that are more similar to the probabilistic retrieval models. Lavrenko &

Croft (2001) introduced a language modelling approach, where relevance is explicitly modelled. The basic underlying assumption is that the information need of the user is described by a relevance language model. Then, the documents are ranked according to the probability that they generate the relevance language model. In addition, Lafferty & Zhai (2003) argued that the classical probabilistic model and the language models are equivalent from a probabilistic point of view, but differ in terms of statistical estimation: the probabilistic model estimates a model for relevant documents, based on a query, while language models estimate a model for relevant queries, based on a document. Moreover, Lafferty & Zhai (2001) employed Bayesian decision theory and introduced a language modelling framework, in which they estimated the information theoretic divergence between the document language models and the query language models.

2.3.3 Divergence From Randomness framework

Amati & Van Rijsbergen (2002) and Amati (2003) introduced the Divergence From Randomness (DFR) framework as a generalisation of the 2-Poisson model for generating IR weighting models. A central concept of the DFR framework is that a term is more informative when its distribution does not fit the probabilistic model that predicts a random occurrence of the term. The weight of a term t in a document is a function of two probabilities:

$$w = \big(1 - Prob_2(tf|E_t)\big) \cdot \big(- \log_2 Prob_1(tf|\text{Collection})\big) \tag{2.7}$$

In the above equation E_t stands for the *elite* set of documents, which is defined as the set of documents that contain the term t and tf is the observed within-document frequency of t.

2.3.3.1 Randomness models

The component $(-\log_2 Prob_1(tf|\text{Collection}))$ in Equation (2.7) corresponds to the informative content of the probability that a term appears with frequency tf in a document by chance, according to a given model of randomness. If the probability that a term occurs tf times in a document is low, then $-\log_2 Prob_1(tf|\text{Collection}))$ is high, and the term is considered to be informative. There are several randomness models that can be used to compute the probability $Prob_1$.

If the occurrences of a term are distributed according to a binomial model, then the probability of observing *tf* occurrences of a term in a document is given by the probability of *tf* successes in a sequence of F Bernoulli trials with N possible outcomes:

$$Prob_1(tf|\text{Collection}) = \binom{F}{tf} p^{tf} q^{F-tf} \tag{2.8}$$

where F is the frequency of a term in a collection of N documents, $p = \frac{1}{N}$ and $q = 1-p$. The informative content of this probability corresponds to $-\log_2 Prob_1(tf|\text{Collection})$.

If the maximum likelihood estimator $\lambda = \frac{F}{N}$ of the frequency of a term in the collection is low, or in other words $F \ll N$, then the Poisson distribution can be used to approximate the binomial model described above. In this case, the informative content of $Prob_1$ is given as follows:

$$-\log_2 Prob_1(tf|\text{Collection}) = tf \cdot \log_2 \frac{tf}{\lambda} + (\lambda - tf) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tf) \tag{2.9}$$

The Poisson model is denoted by P.

Another approximation of the binomial model is obtained by using the information theoretic divergence D and Stirling's formula of approximating factorials. In this case, the informative content of having *tf* occurrences of a term in a document is given as follows:

$$-\log_2(Prob_1(tf|\text{Collection}) = F \cdot \Big(D(\phi, p) + 0.5 \log_2 \big(2\pi \cdot \phi \cdot (1-\phi)\big)\Big) \tag{2.10}$$

where $\phi = \frac{tf}{F}$, $p = \frac{1}{N}$, and $D(\phi, p)$ is the Kullback-Leibler divergence of ϕ from p: $\phi \cdot \log_2 \frac{\phi}{p}$. This model is denoted by D.

Starting from the geometric distribution, a *tf-idf* model is generated, where the informative content of the probability that there are *tf* occurrences of a term in a document is given by:

$$-\log_2 Prob_1(tf|\text{Collection}) = tf \cdot \log_2 \frac{N+1}{n+0.5} \tag{2.11}$$

where n is the document frequency of the term in the document collection. This model is denoted by I(n). Alternatively, the document frequency n can be replaced with the expected document frequency n_e, which is given by the binomial law, as follows:

$$n_e = N \cdot \left(1 - \binom{F}{0} p^0 q^{F-0}\right) = N \cdot \left(1 - \left(\frac{N-1}{N}\right)^F\right) \tag{2.12}$$

In this case, the informative content of having *tf* occurrences of a term in a document is given by:

$$-\log_2 Prob_1(tf|\text{Collection}) = tf \cdot \log_2 \frac{N+1}{n_e + 0.5} \tag{2.13}$$

This model is denoted by I(n_e).

2.3.3.2 Aftereffect of sampling

In the basic Equation (2.7) of the DFR framework, the component $1 - Prob_2(tf|E_t)$ corresponds to the information gain obtained by considering a term to be informative for a document. If a term appears with a high frequency in a document, then it is almost certain that this term is informative for this document and the probability that this term occurs more times in the same document is high. At the same time, when a term appears frequently in a document, the associated information gain is lower. Therefore, the component $1 - Prob_2(tf|E_t)$ adjusts the importance of a term with respect to a document. One model for computing $Prob_2$ is the Laplace model (denoted by L), which corresponds to the conditional probability of having one more occurrence of a term in a document, where the term appears *tf* times already:

$$1 - Prob_2(tf|E_t) = 1 - \frac{tf}{1+tf} = \frac{1}{1+tf} \tag{2.14}$$

Another model for computing $Prob_2$ is the Bernoulli model (denoted by B), which is defined as the ratio of two binomial distributions:

$$1 - Prob_2(tf|E_t) = \frac{F}{n \cdot (tf+1)} \tag{2.15}$$

2.3.3.3 Document length normalisation

Before computing the weight of a term in a document with Equation (2.7), the term frequency *tf* can be normalised with respect to the length of the document. The length of the document simply corresponds to the number of indexed tokens. Amati (2003) assumed a decreasing density function of the normalised term frequency with respect to the document length and derived the following formula, which is called *normalisation 2*:

$$tfn = tf \cdot \log_2(1 + c \cdot \bar{l}/l) \tag{2.16}$$

where tfn is the normalised term frequency, l is the document length, $\bar{l}$ is the average document length in the document collection and c is a hyper-parameter. If $c = 1$, then Equation (2.16) becomes:

$$tfn = tf \cdot \log_2(1 + \bar{l}/l) \tag{2.17}$$

and it is called *normalisation 1*.

The setting of the hyper-parameter c has an impact on the retrieval effectiveness of the DFR weighting models that use *normalisation 2*, and it is collection-dependent. In order to tackle the problem of collection dependency, He & Ounis (2003) defined the normalisation effect as a function of the hyper-parameters related to the term frequency normalisation. The normalisation effect corresponding to the optimal setting of the hyper-parameters for a particular search task depends only on the type of the task and the type of the queries (i.e. short or long queries). Then, for a similar type of search task and for a similar type of queries, the hyper-parameters are set so that they result in the same normalisation effect. A refinement of the normalisation effect has been presented in (He & Ounis, 2005*b*). In addition, He & Ounis (2005*a*) proposed to set the hyper-parameters of term frequency normalisation by measuring the correlation of the document lengths and the normalised term frequencies.

In the remainder of this thesis, the hyper-parameter of *normalisation 2*, as well as any other parameters, are set so that the retrieval effectiveness is directly optimised. The details of this optimisation process are discussed in Section 4.3.2, page 56.

2.3.3.4 Divergence From Randomness weighting models

A DFR document weighting model is generated from a combination of a randomness model for computing $-\log_2 Prob_1(tf|Collection)$ in Equation (2.7), an aftereffect model for computing $1 - Prob_2(tf|E_t)$, and a term frequency normalisation model. For example, if the randomness model is the Poisson distribution (P), the information gain is computed with the Laplace model (L), and the term frequencies are adjusted according to *normalisation 2*, then the resulting DFR model is called PL2. The weight $w_{d,q}$ of document d for query q corresponds to the sum of the weights of each of the query terms. The formula of PL2 is given by combining Equations (2.7), (2.9), (2.14), and (2.16):

$$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{1}{tfn + 1}\big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn)\big) \tag{2.18}$$

where $qtfn = \frac{qtf}{qtf_{max}}$, qtf is the frequency of the term t in the query, and qtf_{max} is the maximum frequency of any term in the query.

If the Poisson randomness model (P) for computing $Prob_1(tf|\text{Collection})$ is combined with the Bernoulli model (B) for computing $Prob_2(tf|E_t)$ and *normalisation 2* for term frequency normalisation, then the resulting model is PB2 and its formula is the following:

$$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{F+1}{n \cdot (tfn+1)} \big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2 (2\pi \cdot tfn) \big) \tag{2.19}$$

Additional models can be generated from different combinations of basic models. The DFR model I(n_e)B2 is generated from the inverse expected document frequency model I(n_e) for computing $Prob_1(tf|\text{Collection})$, the Bernoulli model (B) for computing $Prob_2(tf|E_t)$, and *normalisation* 2. The formula of the model I(n_e)B2 is the following:

$$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{F+1}{n \cdot (tfn+1)} \big(tfn \cdot \log_2 \frac{N+1}{n_e + 0.5} \big) \tag{2.20}$$

A modification of I(n_e)B2 is generated if natural logarithms are used instead of logarithms base 2 in Equation (2.20). The resulting model is denoted by I(n_e)C2 and its formula is the following:

$$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{F+1}{n \cdot (tfn_e+1)} \big(tfn_e \cdot \ln \frac{N+1}{n_e + 0.5} \big) \tag{2.21}$$

where $tfn_e = tf \cdot \ln(1 + c \cdot (\bar{l}/l))$.

The four DFR models that are shown above, PL2, PB2, I(n_e)B2 and I(n_e)C2, employ *normalisation 2*, which introduces the only hyper-parameter required to be set by using relevance information. This hyper-parameter can be set either by measuring a collection independent quantity, such as the normalisation effect, or by directly optimising the retrieval effectiveness. Another interesting DFR weighting model can be generated with the hyper-geometric randomness model, which naturally incorporates a document length normalisation component. In this case, *normalisation 2* is not needed, and all the variables of the weighting model are computed from the collection statistics.

This model is denoted by DLH and its formula is the following:

$$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{1}{tf + 0.5} \Bigg(\log_2 (\frac{tf \cdot \bar{l}}{l} \cdot \frac{N}{F}) + \\ + (l - tf) \log_2 (1 - \frac{tf}{l}) + \\ + 0.5 \log_2 \left(2\pi tf (1 - \frac{tf}{l})\right) \Bigg) \tag{2.22}$$

Overall, the DFR framework provides an elegant and general way to generate IR models from basic probabilistic models. Similarly to the generation of the retrieval models, the DFR framework can be used to introduce weighting models for performing automatic query expansion, as discussed in Section 7.4.2.2, page 202.

Amati & Van Rijsbergen (2002) described a theoretically motivated derivation of BM25 within the DFR framework, where the resulting formula has an additional component compared to the original one. Regarding the relationship between the DFR framework and language modelling, Amati (2006) argued that the DFR weighting model DLH and language modelling are generated from the same probability space, but represent a frequentist and a Bayesian approach to the IR inference problem, respectively. Moreover, recent large-scale evaluations of several DFR weighting models and language modelling have shown that they result in similar retrieval effectiveness (Clarke et al., 2004). For these reasons, the employed weighting models in the remainder of the thesis are mostly based on the DFR framework, and not on language modelling.

The evaluation of selective Web IR in the subsequent chapters, can be performed with any retrieval model. For the purpose of this thesis, five weighting models are used. More specifically, the employed models are the DFR weighting models PL2, I(n_e)C2, PB2 and DLH, as well as the classical BM25. For ease of reference, the corresponding formulae are given in Table 2.1. These weighting models have been selected for several reasons. The weighting models PL2 and I(n_e)C2 are robust and perform well across a range of search tasks (Plachouras & Ounis, 2004; Plachouras, He & Ounis, 2004). The weighting model PB2 is selected in order to test the combination of the Poisson randomness model with the Bernoulli model for the after-effect. The weighting model DLH is particularly interesting, because it does not have any associated hyper-parameter. The weighting model BM25 is employed, because it has been frequently used in the literature. The employed weighting models are statistically independent, as it will be confirmed by the evaluation results in Chapter 4.

PL2	$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{1}{tfn+1} \left(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2 (2\pi \cdot tfn) \right)$
PB2	$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{F+1}{n \cdot (tfn+1)} \left(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2 (2\pi \cdot tfn) \right)$
I(n_e)C2	$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{F+1}{n \cdot (tfn_e+1)} \left(tfn_e \cdot \ln \frac{N+1}{n_e+0.5} \right)$
DLH	$w_{d,q} = \sum_{t \in q} qtfn \cdot \frac{1}{tf+0.5} \left(\log_2 (\frac{tf \cdot \bar{l}}{l} \cdot \frac{N}{F}) + (l - tf) \log_2 (1 - \frac{tf}{l}) + 0.5 \log_2 \left(2\pi tf (1 - \frac{tf}{l}) \right) \right)$
BM25	$w_{d,q} = \sum_{t \in q} \left(\frac{(k_1+1)tf}{(k_1((1-b)+b\frac{l}{\bar{l}}+tf} \cdot \frac{(k_3+1)qtf}{k_3+qtf} \cdot \log \frac{N-n+0.5}{n+0.5} \right) + k_2 \cdot nq \cdot \frac{l-\bar{l}}{l+\bar{l}}$

Table 2.1: The formulae of the weighting models PL2, PB2, I(n_e)C2, DLH, and BM25, respectively.

2.4 Evaluation

There are several IR models, based on different assumptions, or on combinations of theory and experimental data, as discussed in the previous section. A natural question that arises is how to evaluate and compare the different IR models. This has been an important issue that has attracted the interest of researchers from the early stages of IR.

The evaluation of the retrieval effectiveness of IR models has been based on measuring *precision* and *recall* on test collections, which consist of a set of documents, a set of topics and a set of relevance assessments. Precision is defined as the number of retrieved relevant documents over the total number of retrieved documents for a particular topic. Recall is defined as the number of retrieved relevant documents over the total number of relevant documents for a particular topic. The relevance assessments specify which documents are relevant to a particular topic. This approach was introduced in the Cranfield experiments, where the size of the test collections allowed the complete assessment of each document for all the topics (Cleverdon, 1997). However, as the number of documents that an IR system is expected to handle increased, the complete assessment of all documents became impractical and other approaches for the generation of relevance assessments were needed.

In the context of the Text REtrieval Conference (TREC), the relevance assessments are based on *pooling* (Harman, 1993), a technique developed from an idea of Spark-Jones & Van Rijsbergen (1976). The output of a set of IR systems is used to generate a pool of documents for each topic, by taking a number of top ranked documents from each system. In order to compare the retrieval effectiveness of the IR systems, it is sufficient to assess the documents in the generated pool, instead of assessing all the documents for relevance. However, Blair (2001) has pointed that as the size of the test

collections increases, the computed recall with pooling does not correspond to the real one, because only a small fraction of the documents are examined for relevance. In such a case, the recall maybe artificially boosted.

The evaluated search tasks in the initial TRECs were *ad-hoc* search and *routing* (Harman, 1993). The ad-hoc search involves matching an unknown set of topics against a known set of documents, while the routing task involves matching a known set of topics against a stream of documents. Subsequent TRECs introduced various *tracks* in order to evaluate different search tasks. For example, the Very Large Collection (VLC) and Web tracks, which ran from 1997 to 2004, were dedicated to the evaluation of IR systems with Web test collections for ad-hoc and Web search tasks. Hawking & Craswell (2005) provide a comprehensive presentation of both tracks until TREC 2003, and Craswell & Hawking (2004) give an overview of the Web track in TREC 2004. More details about the tracks are provided in Section 3.5.1, page 43.

There are several different measures that can be used to evaluate IR models with respect to precision and recall. Precision and recall are complementary concepts; the comparison of different IR models can only be made if both precision and recall are reported, or if the precision is reported at fixed recall points.

Average precision corresponds to the average of the precision after each relevant document is retrieved. For example, if an IR system retrieves three relevant documents for a topic, at ranks 2, 4 and 10, the average precision of the system for this particular topic is computed as $\frac{1}{3}(\frac{1}{2} + \frac{2}{4} + \frac{3}{10}) = 0.4333$. R-Precision is defined as the precision after R documents have been retrieved, where R corresponds to the number of relevant documents for a particular query.

For high precision search tasks, where there are few relevant documents, and it is important to retrieve a relevant document at the top ranks, an evaluation measure that is commonly used is the reciprocal rank of the first retrieved relevant document. If there is only one relevant document for a query, then this measure is equivalent to average precision. Another measure is precision at n retrieved documents (Pn), where n is a fixed number. This measure depends only on the number of relevant retrieved documents among the top n retrieved documents, and not on their ranking. The comparison of systems over a set of topics is performed by employing the mean of the above described evaluation measures, leading to mean average precision (MAP), mean reciprocal rank of the first retrieved relevant document (MRR1), and mean precision at

n retrieved documents. In addition, success at n retrieved documents (Sn) corresponds to the percentage of topics, for which a system retrieves at least one relevant document among the top n ranked ones.

Van Rijsbergen (1979) provided a comprehensive discussion on the evaluation of IR systems and related measures. More details about the evaluation measures that will be used in the subsequent chapters of this thesis will be given in Section 4.3.3, page 61.

2.5 About Web information retrieval

While classical information retrieval systems have dealt with reasonably sized test collections, and a variety of search tasks, involving *ad-hoc* and routing (Harman, 1993), as well as filtering (Lewis, 1996), the advent of the Web as a vast repository of information, has posed new challenges. One such challenge is the size of the Web, which is larger than any document test collection that has ever been used in IR experiments. Moreover, the hypertext document model used for Web documents offers several sources of evidence that can be exploited to enhance the retrieval effectiveness of IR systems. These challenges, as well as other related issues, are discussed in detail in the next chapter.

Chapter 3

Web Information Retrieval

3.1 Introduction

The Web can be considered as a large-scale document collection, for which classical text retrieval techniques can be applied. However, its unique features and structure offer new sources of evidence that can be used to enhance the effectiveness of IR systems. Generally, Web IR examines the combination of evidence from both the textual content of documents and the structure of the Web, as well as the search behaviour of users, and issues related to the evaluation of retrieval effectiveness.

This chapter presents an overview of Web IR. It discusses the differences between classical IR and Web IR (Section 3.2), a range of Web specific sources of evidence (Section 3.3), and the combination of evidence in the context of Web IR (Section 3.4). This chapter also provides a brief overview of work on the evaluation of Web IR systems (Section 3.5), as well as on query classification and performance prediction (Section 3.6).

3.2 Differences between classical and Web information retrieval

Classical IR systems have been often developed and used in the context of a controlled environment, such as a library, with a specific group of users and a document collection of moderate size. However, the Web represents a substantially different environment for IR systems. These differences are discussed with respect to three aspects: the hypertext document model (Section 3.2.1), the size and structure of the Web (Section 3.2.2), the quality of information on the Web (Section 3.2.3), and the background of Web users (Section 3.2.4).

3.2.1 Hypertext document model

The Web is based on a hypertext document model, where the documents are connected with directed hyperlinks. This results in a virtual network of documents. Hypertext was envisioned by Bush (1945) as a more natural way to organise, store and search for information, similar to the associative way in which the human mind works. A reader approaches a text by reading and understanding small sections of it, while discovering the connections between the exposed concepts in the text. The hypertext aids this process by making the connections between parts of the text explicit (Levy, 1995). In addition, it facilitates the reading of texts in non-linear ways, similarly to structures, such as the table of contents, or the indices in books (Belew, 2000).

The hyperlinks in hypertext systems can have explicit types. Trigg (1983) first noted the importance of making the type of links explicit. In his proposed hypertext model, the links are divided in two broad classes: internal substance links, and external commentary links. These two classes are further divided in subclasses, leading to an extensive taxonomy of link types. Similarly, Baron (1996) identified two main types of links, namely the organisational and the content-based links. The former type of links was used to organise and help navigation among hypertext documents, while the latter type was used for pointing to documents on similar topics. However, as with bibliographic references in scientific publications, some hypertext systems do not usually come with a set of typed links. The HyperText Markup Language (HTML) (Raggett et al., 1999), which is used to write Web documents, provides some functionality for defining typed links, but this mechanism is optional and it is not used consistently. The automatic inference of the link type is a difficult task, because it requires understanding the context of both the source and destination documents. Differently from identifying the type of hyperlinks, Allan (1996) investigated the automatic typed linking of related documents. After linking all pairs of documents, the similarity of which exceeds a threshold, the resulting graph is simplified by iteratively merging links. A type is assigned to the resulting links, according to a predefined taxonomy.

Hypertext alters the information search process by allowing a user to navigate through the document space by following hyperlinks. Navigating through hyperlinks may be sufficient for small collections of hypertext documents. However, as the number of documents increases, or when navigation is allowed across heterogeneous sets of

hypertext documents, users may not be able to locate information by merely following links, but instead, they may find themselves *lost in hyperspace* (Bruza, 1990; Guinan & Smeaton, 1992; Halasz, 1987). One aspect of this problem can be addressed by applying IR techniques to search for information, or locate starting points for browsing in hypertext documents.

3.2.2 Structure of the Web

The Web is a vast repository of information, the size of which is increasing continuously. Bharat & Broder (1998) estimated the size of the static Web in November 1997 to be approximately 200 million documents. Lawrence & Giles (1999) reported that the indexable part of the Web was about 800 million documents in February 1999. More recently, Gulli & Signorini (2005) reported that the indexable Web has more than 11.5 billion documents. All these estimates refer to the publicly available part of the Web, which is indexed by search engines. However, Raghavan & Garcia-Molina (2001) estimated that even more information is stored in databases or access-restricted Web sites, composing the *hidden Web*, which cannot be easily indexed by search engines.

In order to study its topology, the Web can be seen as a directed graph $\mathcal{G}(V, E)$, where the set of vertices V represents the Web documents, and the set of edges E represents the hypertext links[1] between Web documents. The number of links that point to a document d is the *indegree* of d, while the number of links that start from document d is the *outdegree* of d. The sum of the *indegree* and the *outdegree* of a document d is called the degree of d. Generally, complex interconnected systems can be modelled as random graphs. Initially, the research area of random graphs was explored by Erdös & Rényi (1959), who proposed the random graph model $\mathcal{G}_{m,t}$. This model describes graphs with m nodes, where a link exists between two randomly selected nodes with probability t. Random graphs have a short average distance between vertices. In addition, the indegrees and outdegrees of the vertices of random graphs follow a Poisson distribution.

Although the Web seems to be chaotic and to lack structure, because there is no single entity to organise the available information, its topology is similar to that of many other complex systems in nature, which display self-organising principles. Recent research has shown that the topology of many complex interconnected systems does not

[1] Hereafter, the terms hypertext link, hyperlink, and link will be used interchangeably.

fit the one predicted by the random graph model $\mathcal{G}_{m,t}$. Therefore, new graph models have been proposed to study these networks. First, Watts & Strogatz (1998) proposed a model with one free parameter. The Watts-Strogatz (WS) model is based on starting from an ordered finite-dimensional lattice and changing with a given probability one of the vertices connected by each edge. The WS model produces a range of graphs, between the extremes of an ordered finite-dimensional lattice and a random graph. It captures the properties of *small world* social networks, where people are more likely to know their neighbours than a random person that lives far away. The graphs generated by the WS model have a short average path length between vertices, similarly to random graphs. In addition, they have a high clustering coefficient, which corresponds to the fraction of transitive triplets of vertices, compared to random graphs generated by $\mathcal{G}_{m,t}$. Albert et al. (1999) estimated that the average path length between any two documents on the Web is 19 links (the estimated size of which was 800 million documents at the time). Adamic (2001) reported that the clustering coefficient of the graph generated by using hyperlinks across different Web sites has a significantly higher clustering coefficient than that of random graphs (0.081 vs. 0.00105).

The degree distribution of graphs generated by the WS model is similar to that of the random graphs $\mathcal{G}_{m,t}$. However, additional evidence obtained from the analysis of the Web showed that the degree distribution follows a power law (Barabási & Albert, 1999). In other words, the probability of a Web document having k incoming hyperlinks is proportional to $k^{-\gamma}$, where γ is a positive constant. Broder et al. (2000) reported that the distribution of indegrees and outdegrees of Web documents follow power laws with exponents $\gamma_{in} = 2.10$ and $\gamma_{out} = 2.72$, respectively. In addition, the number of pages in a site, as well as the number of visitors to a site follow similar power laws (Adamic, 2001). Faloutsos et al. (n.d.) also identified that the connections between Internet routers follow a power law. Pennock et al. (2002) observed that within some online communities of documents on the Web, the distribution of indegrees and outdegrees may deviate from a power-law and roughly follow a log-normal distribution.

Power law distributions have been observed in many highly complex networks arising in nature and human communities. Barabási & Albert (1999) attributed the origins of power laws in complex networks to two mechanisms: *growth* and *preferential attachment*. First, most real complex networks continuously grow with the introduction of

new nodes. Second, in most real networks, the likelihood of connecting to a node depends on the degree of the node. The nodes, which are linked to by many other nodes, are more likely to get a higher number of new links. Albert & Barabási (2002) provided an extensive survey on complex networks, and Barabási (2002) presented the historical background of studying complex networks, as well as related applications.

Broder et al. (2000) have studied the distribution of connected components of the Web, identifying that the Web graph consists mainly of four parts. The first is a large strongly connected component of Web sites, where it is possible to navigate between any two Web sites. The second part consists of documents that point to the large connected component, while the third part consists of documents that are pointed by other documents in the connected component. The last part consists of the rest of the documents on the Web.

Overall, the estimation of the size of the Web and the analysis of its structure, are very interesting issues for two main reasons. Search engines have to collect, or *crawl* the documents from the Web by following hyperlinks (Brin & Page, 1998; Heydon & Najork, 1999), differently from classical IR systems, where the documents are often readily provided. Therefore, studying the properties of the Web can enhance the effectiveness of crawling Web documents. Both the size and the structure of the Web may also be used to enhance retrieval effectiveness, as it will be described in Section 3.3.

3.2.3 Quality of information on the Web

Classical IR systems have been often used in controlled environments, where documents contain reliable information that rarely changes. However, the Web is a quite different environment, where no assumption can be made about the quality of Web documents.

The information available on the Web is very different from the information contained in either libraries or classical IR collections. A large amount of information on the Web is duplicated, and content is often mirrored across many different sites (Bharat & Broder, 1999; Shivakumar & Garcia-Molina, 1998). This redundancy ensures that the information is always available, even when some of the mirrors are out of service. However, search engines and IR systems need to take into account the duplication of Web documents, in order to reduce the required resources for crawling Web documents and to avoid returning duplicate Web documents in the results presented to users. Mirroring Web documents does not always result in exact duplicates of documents, as the

formatting may change, or dates on the pages may be updated. In such cases, it is necessary to detect duplicate or near duplicate pages. Scalable techniques (Bernstein & Zobel, 2004; Bharat & Broder, 1999) to do this are based on fingerprinting small parts of documents, and comparing the overlap between the generated fingerprints. The differences between the proposed techniques are mainly due to the selection of the parts of the documents to fingerprint, and the selection of the generated fingerprints to compare between documents.

In addition to duplication, the contents of Web pages are not guaranteed to be accurate. Indeed, Web pages may contain false or inaccurate information, due to unintentional errors by their authors, or due to intentional efforts to mislead users in visiting a particular website. Both the issues of duplication and quality of information are more significant in the case of the Web than in the case of classical hypertext systems (Spertus, 1997), or other types of document corpora, such as newswire articles and scientific publications.

3.2.4 Background of Web users

The Web is an open system accessible to anyone. Therefore, no assumption can be made about the users' expertise, experience or computer literacy. Hsieh-Yee (1993) reported differences in the search behaviour of novice and experienced searchers in a classical IR setting. Studies of query logs from Web search engines showed that the majority of the users provide short queries, browse only the top ranked documents and do not reformulate the original query (Silverstein et al., 1999). Jansen & Pooch (2001) provided a comprehensive review of user search behaviour studies.

Users perform search tasks of varying types. Broder (2002) has identified a taxonomy of three main types of search tasks on the Web: *informational*, *navigational*, and *transactional*. In informational tasks, users are looking for information about a particular topic. In navigational tasks, users are interested in viewing a Web document they have seen before, but do not remember its location, or they do not want to navigate back to that page. In transactional tasks, users are interested in making a transaction or obtaining an online resource. Rose & Levinson (2004) extended Broder's taxonomy by providing sub-types of informational and transactional, or resource, search tasks.

3.3 Web-specific sources of evidence

Web IR can exploit a range of sources of evidence, in addition to the textual content of documents. For example, evidence from the document structure, or the structure of the hyperlinks among documents, can be used to enhance retrieval effectiveness. This section presents an overview of the different sources of evidence that can be used for Web IR.

3.3.1 Document and Web site structure

Web documents are semi-structured in the sense that HTML offers basic structuring capabilities to the authors, even though the use of such capabilities is optional. Web documents may have titles and headings for improved readability. In addition, bold or italic typefaces can be used in order to emphasise specific parts of the document. Evidence about the formatting and the structure of the document has been used by commercial search engines as an indication of the importance of the text that appears with additional visual cues. For example, Brin & Page (1998) described that changes in the relative size, or the colour of the text, were stored in the index of an early version of the Google search engine.

The hypertext document model and the Web encourage authors to organise documents in several different ways. Documents on the Web are grouped in Web sites, where most of the documents cover either a specific topic, or a series of related topics. Within Web sites, documents are usually organised in a hierarchical directory structure. There have been several efforts towards the automatic identification of aggregates of hypertext, or Web documents. Botafogo & Shneiderman (1991) employed a graph theoretical approach in order to identify aggregates in hypertext documents. Eiron & McCurley (2003*a*), and Li et al. (2000), defined heuristics based on observations of the structure of Web sites. Grouping documents according to their domain has also been employed in order to limit the redundancy of retrieving many documents from a given site (Kwok et al., 2002). In classical IR systems, the retrieval unit most commonly corresponds to a whole document, such as a scientific publication or a news item. However, this is not necessarily the case for hypertext, where a document may correspond to several hypertext nodes. For example, Tajima et al. (1998) statically identified retrieval units as connected subgraphs, and proposed to retrieve those subgraphs that contain all the

query terms. Tajima et al. (1999) also identified the retrieval units from the set of retrieved documents for a query.

The URL of Web documents can be effectively used to detect documents that are likely to be home pages of Web sites. Westerveld et al. (2001) and Kraaij et al. (2002) identified four types of URLs for Web pages:

- root: a domain name. For example, such a URL would be http://ir.dcs.gla.ac.uk/.
- subroot: a domain name followed by a single directory. For example, such a URL would be http://ir.dcs.gla.ac.uk/terrier/.
- path: a domain name followed by a directory path of arbitrary depth. For example, such a URL would be http://ir.dcs.gla.ac.uk/terrier/doc/.
- file: a domain name followed by a path to a file other than the default file index.html. Such a URL would be http://ir.dcs.gla.ac.uk/terrier/people.html.

Westerveld et al. (2001) found that the Web documents with root and subroot URLs are more likely to correspond to home pages of Web sites.

In addition to the type of URLs, another indication of whether a particular Web document is a home page of a Web site is given by the length of its URL (Savoy & Rasolofo, 2001). Because the URL of Web documents is likely to reflect the hierarchical directory structure of Web sites, documents that are higher in the hierarchy have shorter URLs. Savoy & Rasolofo (2001) defined the length as the number of "/" in the URL. Kamps et al. (2004*a*) considered the number of characters in a URL, and they also counted the number of "." in the domain name, and the number of "/" in the path of a URL. Plachouras et al. (2003) and Plachouras & Ounis (2004) used the length in characters of the URL path. Using evidence from the URLs of Web documents is further discussed in Section 4.5.1.

3.3.2 Hyperlink structure analysis

The analysis of the hyperlink structure of the Web has been based on citation analysis. For example, the impact factor of a journal, can be estimated by counting the number of times it is cited in other papers or journals (Garfield, 1972). Instead of just counting citations, Pinski & Narin (1976) suggested that the influence of a journal should depend on the influence of the journals that cite it. Therefore, the influence of a journal is

defined in a recursive manner. Geller (1977) provided additional insight by modelling the computation of the influence with Markov chains.

Similarly to citations, hyperlinks between Web documents can be exploited in order to estimate the importance of Web pages. From the perspective of a Web search engine, for each query there may be far more relevant documents than a user is willing to browse. Bar-Yossef et al. (2004) also suggested that the quality and the freshness of documents vary significantly. For this reason, when the number of retrieved documents for a query is large, a search engine should try to detect important documents that originate from the more authoritative, or trusted sources.

There have been various proposed ways to find the important documents within hypertext or Web documents. In an early study, Botafogo et al. (1992) investigated various structure-based measures to estimate the importance of documents in hypertext systems. They introduced measures to quantify the centrality of nodes in a hypertext, as well as measures related to the compactness and the linear ordering of a set of hypertext nodes. Moreover, Pirolli et al. (1996) analysed both the content and the link structure of Web documents within a single site in order to detect the most useful documents.

Carriere & Kazman (1997) were among the first to use evidence from the link structure of the Web, in order to rank Web documents. They proposed a method in which after a query is sent to a search engine, the obtained result set is extended with documents that are pointed to, or point to documents in the initial result set. Then, the extended result set is sorted according to the number of incoming hyperlinks of the documents. Two of the seminal works in the area of hyperlink structure analysis for ranking Web documents are the PageRank algorithm and the HITS algorithm, which are presented in the following sections.

3.3.2.1 PageRank

Brin & Page (1998) proposed PageRank, an algorithm for computing a global authority score for each document. While counting the number of links is expected to perform well in some cases (Amento et al., 2000), PageRank provides a more sophisticated way to rank Web documents, similar to the approaches proposed by Pinski & Narin (1976) and Geller (1977) in citation analysis. The PageRank score of a Web document depends on the PageRank scores of all the documents pointing to it. Documents with a high

PageRank are either pointed by many documents, or they are pointed by important documents. A simplified version of PageRank is defined as follows:

$$PR(i) = \sum_{d_j \to d_i} \frac{PR(j)}{outdegree_j} \tag{3.1}$$

where PR is the $N \times 1$ vector, which contains the PageRank values for each document, $outdegree_j$ is the outdegree of a document d_j, and N is the number of documents in the collection. The above equation can be expressed in terms of matrices. Let A to be a $N \times N$ matrix with the rows and columns corresponding to Web documents and $[a_{j,i}] = 1/outdegree_j$ if $d_j \to d_i$, otherwise $[a_{j,i}] = 0$. Then, Equation (3.1) can be written as $PR = cA^T \cdot PR$, so that PR is an eigenvector of A^T with eigenvalue c.

The simplified definition of PageRank in Equation (3.1) overestimates the PageRank values for documents without any outgoing hyperlinks, or for sets of documents that only link to each other. These problems are eliminated with the introduction of a vector E called *rank source*. Thus, Equation (3.1) can be expressed as follows:

$$PR(i) = (1 - prdf) \cdot E(j) + prdf \cdot \sum_{d_j \to d_i} \frac{PR(j)}{outdegree_j} \tag{3.2}$$

where $prdf \in [0, 1]$ is a constant called $dampingfactor$, and $E(j)$ is the score assigned to document j by the rank source E. The vector PR is the principal eigenvector of the matrix:

$$A = (1 - prdf) \cdot E + prdf \cdot M^T \tag{3.3}$$

where M is the matrix with elements:

$$m_{ij} = \begin{cases} \frac{1}{outdegree_i} & \text{if } d_i \to d_j \\ 0 & otherwise \end{cases} \tag{3.4}$$

PageRank scores correspond to the probability of visiting a particular node in a Markov chain for the whole Web graph, where the states represent Web documents, and the transitions between states represent hyperlinks. Alternatively, PageRank can be seen as a model of a random surfer, who browses Web documents and navigates by following hyperlinks. The random surfer chooses to browse a random Web document with some probability $1 - prdf$, instead of following a hyperlink. The introduction of this *jump* to a random Web document makes PageRank stable to small perturbations of the Web graph (Ng et al., 2001). Moreover, Diligenti et al. (2002) modified PageRank

in order to refine the random surfer model. They included more specific user actions, such as following a link forwards, backwards, or jumping to a random document.

The PageRank algorithm does not depend on the content of documents, nor on the queries of users. Instead, it computes a score for documents based only on the hyperlinks. The computation of PageRank at indexing time ensures that the computational overhead in applying it at query time is minimal. Section 3.4.1 discusses the extensions of PageRank with topical bias. Section 4.5.2 introduces a novel model for hyperlink analysis, the Absorbing Model, and evaluates the combination of content and hyperlink analysis.

3.3.2.2 Hubs and Authorities

Kleinberg (1998) proposed a more sophisticated algorithm for finding authoritative documents at query time. The algorithm, called Hyperlink-Induced Topic Search (HITS), is based on the spectral analysis of the adjacency matrix of the documents returned for a query. Documents on the Web may be *authorities*, or *hubs*. Authorities contain information about a specific topic, and hubs point to authorities on a specific topic. Between hubs and authorities, there is a *mutual reinforcing relation*, which is expressed as follows: good authorities are linked by many good hubs; and good hubs point to many good authorities. A graph representation of this structure corresponds to a bipartite graph, as shown in Figure 3.1, where hubs are presented on the left hand side, and authorities are shown on the right hand side.

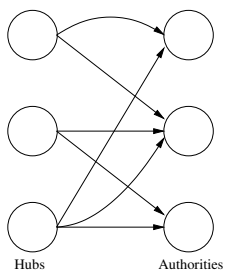

Figure 3.1: Hubs and authorities as a bipartite graph.

The HITS algorithm works as follows: a query is sent to a search engine and the top 200 retrieved documents form an initial *root set* of documents. This root set is expanded to a *base set* with documents that point to, or that are pointed to by the already retrieved documents. One imposed restriction is that a document in the root

set can only bring at most 50 documents in the base set. The generation of the base set of documents is performed in a similar way to the methodology proposed by Carriere & Kazman (1997). Each document d in the base set is associated with a hub value $h(d)$, and an authority value $a(d)$. If there is a hyperlink from document d_i to document d_j, then this is represented by $d_i \rightarrow d_j$. The values $h(d_i)$ and $a(d_i)$ for document d_i are iteratively updated, using the following formulae:

$$h(d_i) = \sum_{d_i \rightarrow d_j} a(d_j) \tag{3.5}$$

$$a(d_i) = \sum_{d_j \rightarrow d_i} h(d_j) \tag{3.6}$$

Kleinberg (1998) showed that the hub values $h(d)$ and the authority values $a(d)$ converge to the values $h^*(d)$ and $a^*(d)$, respectively, after iteratively performing the above computations.

If A is the adjacency matrix for the documents in the expanded set, then the Equations (3.5) and (3.6) can be written as $a \leftarrow A^T h$ and $h \leftarrow Aa$, where h, a are vectors of the hub and authority scores of documents. The vector h^* is the principal eigenvector of the matrix AA^T, and the vector a^* corresponds to the principal eigenvector of the matrix $A^T A$.

Bharat & Henzinger (1998) extended the original HITS algorithm in the following way. In order to diminish the effect of the mutual reinforcement relation between hubs and authorities, when there are many links from a site to another site, each link is given a weight inversely proportional to the number of links between the two sites.

Lempel & Moran (2000) introduced the Stochastic Approach for Link-Structure Analysis (SALSA), an algorithm that computes hub and authority scores for Web documents, differently from the HITS algorithm. In SALSA, the scores are computed from a two-step random walk, where alternately: (a) a randomly selected incoming hyperlink of document d is traversed backwards; (b) a randomly selected outgoing hyperlink of document d is traversed forwards. The authority scores correspond to the stationary distribution of the Markov chain resulting from performing first step (a), and then step (b). The hub scores correspond to the stationary distribution of the Markov chain resulting from performing first step (b), and then step (a). Borodin et al. (2001) suggested that SALSA is a one-step truncated version of HITS, where the authority of

a document depends only on its popularity in its immediate neighbourhood, while the authority of a document in HITS depends on the global link structure.

Cohn & Chang (2000) introduced Probabilistic HITS (PHITS) for calculating authority and hub scores for Web documents. PHITS is equivalent to the Probabilistic Latent Semantic Indexing (PLSI) proposed by Hofmann (1999) and it can be described as follows: the probability of the generation of a document d is $P(d)$. The probability of a factor, or a topic z to be associated with document d is $P(z|d)$. Given the associated topic z, the probability that there is a link to a document d is $P(d|z)$. The advantage of this model is that, apart from a measure of authority which is represented by the probability $P(d|z)$, other interesting measures can be extracted, such as the probability $P(z|d)$ that a document d is about topic z.

Borodin et al. (2001) provided heuristical refinements to the HITS and the SALSA algorithms. First, they introduced a hub-averaging version of HITS, where the authority scores of documents are computed in the same way as in HITS, but the hub scores of documents correspond to the average of the authority scores of all the documents pointed by the hub. In a second modification of the HITS algorithm, the authority scores depend only on the hubs with a score higher than the average hub score, and the hub scores are computed only from the scores of the top ranked authorities. Third, they extended the SALSA algorithm by allowing the authority scores to depend on documents in a broader neighbourhood of each document. They also introduced a Bayesian algorithm, where the prior probability of having a hyperlink from a hub to an authority is defined with respect to: (a) a parameter that represents the tendency of hubs to link to authorities; (b) a parameter that represents the level of authority. Then, the prior probability of having a hyperlink from a hub to an authority is conditioned on the observed data. The proposed algorithms are refinements of existing algorithms, but they have not been evaluated in a large-scale experiment.

The application of HITS and its extensions to an expanded set of retrieved documents for a query, means that the content of documents is implicitly considered by the algorithm. However, the associated computational cost of the algorithm at query time makes its application in an operational setting rather difficult. Another problem with HITS is related to topic drift, which occurs when the most prominent group of documents in the result set is not about the query topic, but dominates the results because it is more densely connected (Bharat & Henzinger, 1998). Section 3.4.1 describes

the extensions of the HITS algorithm, which employ the content of documents more explicitly.

3.3.3 User interaction evidence

In addition to evidence that can be obtained from the Web documents, another source of evidence is the information obtained from the visiting patterns of users in Web sites, or the click-through data obtained from the result pages of search engines. In a study of the usage patterns observed on Web sites, Huberman et al. (1998) found that the number of links followed by a user in a Web site is distributed according to an inverse Gaussian distribution, which means that most of the users follow few links, while there is a small number of users that will follow a high number of links. Pirolli & Pitkow (1999) suggested that characterising the visiting patterns of users can be used to enhance hyperlink structure analysis algorithms, which are based on a stochastic process of traversing hyperlinks.

Joachims (2002) employed clickthrough data from the logs of a metasearch engine, in order to adapt its retrieval function with a Support Vector Machine (SVM) classifier. The results from a controlled experiment with users suggested that the users viewed more retrieved documents after adapting the retrieval function with respect to the clickthrough data of the metasearch engine. Jiang et al. (2005) combined clickthrough data with associations between documents in order to alleviate the problem that a large amount of clickthrough data is required to improve the retrieval effectiveness. Joachims et al. (2005) performed a user study and suggested that the clickthrough data can be used as a relative indication of relevance for the retrieved documents.

3.4 Combination of evidence for Web information retrieval

The combination of different sources of evidence generally improves the retrieval effectiveness (Croft, 2000). The sources of evidence can be either different query representations, different document representations, or various retrieval techniques. Hyperlink analysis provides an estimation of the quality, or usefulness of documents, complementing the concept of relevance. The combination of evidence for IR, and specifically for Web IR, has been extensively investigated, with many different approaches proposed.

3.4.1 Extending hyperlink analysis algorithms

One approach to the combination of evidence from the content and hyperlink structure analysis is to refine already proposed hyperlink analysis algorithms with evidence from the content analysis, or the users' queries.

Extensions of PageRank focus more on biasing the PageRank scores towards a specific topic. Richardson & Domingos (2002) proposed a modified PageRank algorithm, where the random surfer is replaced by an intelligent surfer, who traverses links and jumps to documents according to the similarity of the latter to the query. A similar extension to PageRank has been proposed by Haveliwala (2002), where a set of PageRank scores biased towards specific topics is computed. Then at query time, the user profile determines the weight of each individual topic in the combination of the different PageRank scores. A drawback of these approaches is that a range of precomputed PageRank scores for various topics is required in order to be efficiently combined at query time.

There have been several proposed extensions to the HITS algorithm, aiming to incorporate more evidence from the textual content of documents in the algorithm. Chakrabarti et al. (1998) weighted the hyperlinks between documents according to the similarity between the query and a window of text surrounding the hyperlink. Bharat & Henzinger (1998) extended the original HITS algorithm by eliminating non-relevant documents from the expanded set of documents, and by regulating the influence of documents according to their relevance scores. The relevance scores correspond to the cosine similarity between the documents in the expanded set and a broad query, resulting from the concatenation of the first 1000 words from each document in the expanded set. They performed a user experiment, and reported considerable improvements in precision with respect to the original HITS algorithm. Li et al. (2002) investigated different similarity measures. Chakrabarti et al. (2001) used the Document Object Model (DOM) in order to detect *microhubs*, which correspond to focused hubs on a specific topic within a document. The proposed approach was successful at reducing the topic drift of the original HITS algorithm, by identifying more relevant hubs for a query.

Achlioptas et al. (2001) introduced a model for searching, where both the generation of links between pages and the distribution of terms are considered. They assumed a number of basic latent concepts, the combination of which results to every possible

topic. For each document, there is a vector for its authority on each topic and another vector for its quality as a hub on each topic. The inner product of the hub vector of one document with the authority vector of another document gives the expected number of links from the former to the latter document. There are two associated distributions of terms with each document. The first one determines the distribution of term frequencies for the authoritative terms, while the second one determines the distribution of term frequencies for the hub terms, e.g. the anchor text associated with hyperlinks. The use of latent topics is more evident in the continuation of PHITS (Section 3.3.2.2), where both content and link analysis are integrated by linearly combining PLSI and PHITS (Cohn & Hofmann, 2001). Both algorithms share the same space of latent topic factors, resulting in a principled integration of content and link analysis.

3.4.2 Implicit hyperlink analysis with anchor text

The algorithms HITS and PageRank, along with their extensions, explicitly employ the hyperlinks between Web documents, in order to find high quality, or authoritative Web documents. A form of implicit use of the hyperlinks in combination with content analysis is to use the anchor text associated with the incoming hyperlinks of documents. Web documents can be represented by an anchor text surrogate, which is formed from collecting the anchor text associated with the hyperlinks pointing to the document[1].

The anchor text of the incoming hyperlinks provides a concise description for a Web document. The used terms in the anchor text may be different from the ones that occur in the document itself, because the author of the anchor text is not necessarily the author of the document. Eiron & McCurley (2003*b*) found similarities in the distribution of terms between the anchor text of Web documents and the queries submitted to an intranet search engine by users. For these reasons, Web documents can be indexed with the anchor text of their incoming hyperlinks, in addition to their textual content. This approach has been used in Web search engines (Brin & Page, 1998; McBryan, 1994).

Craswell et al. (2001) showed that anchor text is very effective for navigational search tasks and more specifically for finding home pages of Web sites. Upstill et al. (2003) suggested that the anchor text of the incoming hyperlinks from documents outside the collection can enhance the retrieval effectiveness for home page finding

[1] In the remainder of the thesis, the anchor text surrogate of a document will be referred to as anchor text, unless otherwise stated.

in general Web collections. In the context of enterprise search, Hawking, Craswell, Crimmins & Upstill (2004) indicated that external link and anchor text evidence are less effective.

The distribution of terms in the anchor text has different characteristics from the distribution of terms in the body of Web documents. For example, the home page of a Web site may have several thousands of incoming hyperlinks with the same anchor text. As a consequence the terms of the anchor text would have a very high term frequency, that should not be penalised by the term frequency normalisation component of the used document weighting model (Hawking, Upstill & Craswell, 2004). Instead, the anchor text should be normalised differently from the text in the body of documents. This approach is further discussed in Section 3.4.4.2, and in Section 4.4, where the DFR framework is extended in order to allow the term frequency normalisation and weighting of different document fields.

3.4.3 Network-based models

Frei & Stieger (1995) used activation spreading of the retrieval scores along the semantic hyperlinks in a hypertext. They defined the semantic hyperlinks as hyperlinks that point to documents with similar, more detailed, or additional information. In the context of hypertext documents, Savoy (1996) suggested that constraints, such as avoiding to activate a document for which the number of links exceeds a given threshold, can also be used.

Savoy & Picard (2001) employed a spreading activation mechanism, based on the assumption that hypertext links between documents may contain some information about relevance. After retrieving a set of documents for a given query, the retrieval status value RSV_i of a retrieved document d_i is updated as shown in the following equation:

$$RSV_i := RSV_i + \kappa * \sum_{d_i \rightarrow d_j} RSV_j \tag{3.7}$$

where κ is a weighting parameter, and $d_i \rightarrow d_j$ denotes that there is a hyperlink from document d_i to document d_j. In their experiments, they considered links from or to the top 50 ranked documents only, based on the assumption that these documents are relevant.

Jin & Dumais (2001) employed a method similar to spreading activation. The combined score for document d_i depends on the score of document d_i and on a score based on the link structure. The latter score is computed by considering all documents $\{d_k\}$ that point to, or that are pointed by document d_i. For each such document d_k, its contribution to the combined score of document d_i depends on d_k's authority score, its similarity to the query, and also its similarity to document d_i.

Ribeiro-Neto & Muntz (1996) proposed a belief network model, which was extended to consider hyperlinks between Web documents (Silva et al., 2000). In the belief network model, the queries, the documents and the terms are treated as nodes in a network. For each document, the evidence associated with a document being either a hub, or an authority, are represented by two additional nodes in the network. From a theoretical point of view, the belief network model is more general than the Bayesian inference network proposed by Turtle & Croft (1991). However, both models are very similar for practical purposes.

3.4.4 Combination of different retrieval techniques and representations

This section describes the combination of evidence for Web IR from three different perspectives: the combination of the output of retrieval systems in metasearching; the combination of different document representations; and the combination of query-dependent and query-independent evidence.

3.4.4.1 Metasearching

Metasearching refers to the combination of the output of several IR systems. Saracevic & Kantor (1988) noted that the odds of a document being judged relevant increase monotonically with the number of retrieval systems in which the document appears to be relevant. Lee (1997) indicated that different systems may retrieve similar sets of relevant documents but different sets of non-relevant documents. As a consequence, the improvements in effectiveness may result from the detection of the non-relevant documents from the different systems.

Metasearching is performed by combining either the ranks or the scores of the retrieved documents. The retrieval scores of documents may be used without harming the retrieval effectiveness, when they are distributed similarly (Lee, 1997). When this

condition does not hold, the ranks of documents should be preferred, in order to remove the bias introduced by the different score distributions.

Aslam & Montague (2001) proposed a method for fusing ranked lists of documents obtained from search engines, by looking at the problem of fusion as a voting problem. They used Borda Count, where each voter ranks a fixed set of c candidates in order of preference and has at his disposal $\sum_{i=1}^{c} i$ votes. The top ranked candidate is given c votes, the second ranked candidate is given $c-1$ votes, etc. If the voter does not rank some candidates, then the remaining of the $\sum_{i=1}^{c} i$ votes are divided between the unranked candidates. Then, the candidates are ranked in order of total votes. Lebanon & Lafferty (2002) proposed a model for obtaining a probability distribution over the rankings of documents. Moreover, Fagin et al. (2003) performed a combination of several features using the ranks of documents. They employed various features, including the content of documents, the anchor text of incoming hyperlinks, PageRank, the length and depth of URLs, as well as the occurrence of query terms in the URL of Web documents.

The combination of ranked lists is particularly useful when combining the output of commercial Web search engines, which do not usually return the retrieval scores of documents (Meng et al., 2002). However, Craswell, Robertson, Zaragoza & Taylor (2005) suggested that using the scores is potentially more effective, because the scores contain more information than the ranks. Indeed, the ranks can be obtained from the scores, but it is not possible to obtain the original scores from the ranks.

Bartell et al. (1994) investigated the automatic combination of multiple retrieval techniques. They modelled the combination of evidence from different retrieval systems as the linear combination of the retrieval scores of each system. For example, for the query q and document d, the overall score $RSV_{\Theta,q}(d)$ for a combination of m systems is given by $\sum_{i=1}^{m} \Theta_i E_i(q,d)$, where $E_i(q,d)$ is the score assigned by the $i-$th system to the document d for query q, and Θ_i is the weight of the $i-$th system. A drawback of this approach is the need to calculate the m parameters Θ_i.

Shaw & Fox (1994) reported that the combination of scores performed better when the combined systems were related to different retrieval paradigms. They also suggested that the linear combination of scores was more effective than selecting one retrieval score from the available ones for each document.

Manmatha et al. (2001) introduced a methodology, which is based on the observation that the score distribution of relevant documents fits a Gaussian distribution, while the score distribution of non-relevant documents fits an exponential distribution. If the mean of the Gaussian distribution, and the point where the two distributions intersect are far from each other, then the retrieval system is expected to be more successful in separating the relevant documents from the non-relevant ones. When combining different retrieval techniques, the above described methodology could be applied to automatically set the weights of each search engine. However, a disadvantage of the approach is that the estimation of the parameters for the Gaussian and exponential distributions is computationally expensive. In addition, there is some variability in the results, due to the Expectation-Maximisation approach employed for the parameter estimation.

3.4.4.2 Combination of representations

Westerveld et al. (2001) and Kraaij et al. (2002) employed a mixture of language models for the content of Web documents and the corresponding anchor text of the incoming hyperlinks. Similarly, Ogilvie & Callan (2003) investigated the use of a mixture of one language model for each representation of documents. They found that combining low performing representations does not always improve performance. They also suggested that the mixture of language models is robust when low performing representations are incorporated among better performing ones.

Tsikrika & Lalmas (2004) employed a Bayesian inference network model (Turtle & Croft, 1991) in order to combine several representations of Web documents in a formal framework, and the obtained results showed improvements in early precision.

Robertson et al. (2004) suggested that the linear combination of the scores of different retrieval techniques may not be appropriate, because the linear combination of scores does not consider the non-linearities introduced by various weighting models, and it is difficult to interpret the resulting scores.

Zaragoza et al. (2004) proposed a modified version of the BM25 weighting model in order to handle terms from different fields, which correspond to the text within specific HTML tags. The extended version of BM25 assigns weights to terms using term frequency normalisation parameters and weights for each field separately. For each different field, the term frequencies are normalised and weighted, independently.

The term frequency that is used in the BM25 formula is the sum of the normalised and weighted frequencies. Section 4.4 provides more details about this extension of BM25, and introduces an extension to the DFR framework for performing per-field normalisation and weighting.

3.4.4.3 Combination of query-dependent and independent evidence

As indicated above, the combination of query-dependent evidence from the textual content of documents, and query-independent evidence, such as the hyperlink structure of Web documents, can be performed by aggregating ranked lists of documents. Alternatively, the combination of query-dependent and independent evidence can be achieved by first retrieving documents using query-dependent evidence, and then reranking the retrieved documents according to the query-independent evidence.

Upstill et al. (2003) employed this approach to obtain an ideal evaluation of various query-independent sources of evidence for home page finding search topics. This was performed as follows. First retrieval from the content or the anchor text of documents was performed. Then the rank k of the correct answer for a topic was located. Finally, the top k ranked documents were reordered according to a query-independent source of evidence, such as PageRank, or the type of the URL. In a more realistic case, the number of top ranked documents to reorder is specified as a percentage of the total number of retrieved documents for a query. In addition, they suggested reordering the documents with a higher score than a percentage of the highest score assigned to a document for a query. The results indicated that the latter approach is more effective than the former one.

Westerveld et al. (2001) and Kraaij et al. (2002) employed the indegree of documents, as well as the type of URLs, in order to define the prior probability of a document being a home page, in the context of language modelling. Hauff & Azzopardi (2005) defined the document priors by considering both the age and popularity of Web documents, as estimated by a preferential attachment model for estimating the number of links to a Web document.

Amati et al. (2003) proposed the Dynamic Absorbing Model, which employs the retrieval score of a particular document as the prior probability of starting a random walk in the Absorbing Model. This results in a principled and theoretically motivated combination of content and hyperlink analysis.

Craswell, Robertson, Zaragoza & Taylor (2005) introduced a methodology, inspired by the work of Singhal et al. (1996) on pivoted document length normalisation, for finding appropriate functional forms to combine query-independent evidence with content retrieval. Then, they transformed the independent evidence into scores, which can be linearly combined with the analysis of query-dependent evidence. The combination of query-dependent and query-independent sources of evidence is further discussed in Section 4.5 of the next chapter.

3.5 Evaluation

The evaluation of Web IR systems has been primarily performed in the context of the TREC Very Large Collection and the Web tracks, which ran for eight consecutive years. Several studies have also been conducted in order to estimate the retrieval effectiveness of commercial search engines.

3.5.1 Experimental evaluation in Text REtrieval Conference

The Very Large Collection (VLC) track, followed by the Web track, have been dedicated to the evaluation of IR systems with Web test collections from TREC-6 until TREC 2004 (Craswell & Hawking, 2004; Hawking & Craswell, 2005). The definitions of the evaluated search tasks have evolved from standard ad-hoc search tasks with Web documents to Web-specific informational and navigational search tasks, similar to the search tasks specified by Broder (2002). The evaluation measures primarily used were mean average precision, the mean reciprocal rank of the first retrieved relevant document, precision at n retrieved documents, and success at n retrieved documents.

In the VLC track of TREC-6 (Hawking & Thistlewaite, 1997) the used test collection was the VLC collection, a set of 7.5 million Web and non-Web documents. In the VLC track of TREC-7 (Hawking et al., 1998*b*), the used test collection was the VLC2 collection, a set of 18.5 million Web documents crawled from the Internet Archive[1]. The evaluated search tasks were in the spirit of ad-hoc search tasks, and mainly focused on the scalability of the existing prototype IR systems.

In the Web tracks of TREC-8 (Hawking et al., 1999) and TREC-9 (Hawking, 2000), the used test collections were the VLC2 collection and two subsets of it. More specifi-

[1] http://www.archive.org

cally, in the Large Web task of TREC-8 the VLC2 collection was used to test whether the existing prototype IR systems would scale up to that amount of data. In addition, in the Small Web task of TREC-8, the WT2g collection, which corresponds to a subset of 250,000 Web documents and 2GB of data from the VLC2 collection, was used to perform ad-hoc retrieval. The results showed that hyperlink analysis-based approaches were not as effective as standard IR techniques for an ad-hoc search task (Hawking et al., 1999). In the Web track of TREC-9, a subset of 1.69 million Web documents and 10GB of data from the VLC2 collection, the WT10g collection, was used to perform ad-hoc retrieval. The results showed that standard IR techniques performed well for the ad-hoc search tasks (Hawking, 2000).

In the Web track of TREC 2001 (Hawking & Craswell, 2001), a homepage finding task was introduced, in addition to the ad-hoc search task with the WT10g collection. In this navigational task, the topics are about finding the homepage of a Web site, the name of which corresponds to the query. The results from this search task showed that both anchor text and the type of URL of Web documents improved the retrieval effectiveness.

For the Web tracks of TREC 2002 (Craswell & Hawking, 2002), TREC 2003 (Craswell et al., 2003), and TREC 2004 (Craswell et al., 2003), the used test collection was the .GOV collection, a partial crawl of the .gov domain from 2002. This collection consists of 1.24 million Web documents and 18GB of data. In the Web track of TREC 2002, the topic distillation task was introduced, where relevant documents are supposed to be useful resources about the query topic. However, due to the definition of what constituted a relevant document, the results from the evaluation of the task were similar to an ad-hoc retrieval task. In addition to the topic distillation task, there was a named page finding task, where the query topics were about finding a particular Web document, which is not necessarily a homepage. For this navigational task, the anchor text, and the document structure were effective sources of evidence (Craswell & Hawking, 2002).

In TREC 2003, the definition of the topic distillation task was refined to specify that relevant documents can only be homepages of relevant Web sites. The navigational task of the TREC 2003 Web track consisted of a mixture of named page finding and homepage finding topics. In both evaluated search tasks, the document structure and the anchor text of incoming hyperlinks resulted in important improvements in the retrieval effectiveness (Craswell et al., 2003).

The Web track in TREC 2004 consisted of a mixed query task, where topic distillation, named page finding, and homepage finding queries, are mixed in a single stream of queries. The IR systems are not aware of the query type during retrieval. This task is closer to the operational setting of a search engine, where users submit queries, without giving explicit evidence of the query type. The results showed that effective retrieval could be performed without classifying the mixed queries into the corresponding query types (Craswell & Hawking, 2004).

Summarising the findings from the VLC and the Web tracks, Hawking & Craswell (2005) suggested that the nature of the search task is very important in determining what sources of evidence will result in effective retrieval. Indeed, Web-specific evidence improved the retrieval effectiveness for informational and navigational search tasks, but not for typical ad-hoc search tasks.

The last Web track was run in TREC 2004. The evaluation of IR systems with Web data has also been performed in the Terabyte track of the TREC 2004 (Clarke et al., 2004) and TREC 2005 (Clarke et al., 2005), which employed the .GOV2 collection, a crawl of 25 million Web documents and 426GB of data from the .gov domain. The Enterprise Track in TREC 2005 (Craswell, de Vries & Soboroff, 2005) focused on email and expert search tasks. In the remainder of this thesis, the experimental setting is based on the Main Web task of TREC-9 and the Web tracks from TREC 2001 to TREC 2004. More details are presented in Section 4.2

3.5.2 Search engine evaluation

Gordon & Pathak (1999) proposed a list of features that should be considered in comparative evaluation studies of commercial search engines. They performed a comparison of seven commercial search engines and one subject directory, using genuine information needs and expert searchers. The relevance assessments were performed by the users who formulated the information needs. The evaluation was based on precision and recall, as well as the likelihood that documents that had been retrieved by one engine, were retrieved by others, as well. The authors found that, overall, the absolute precision of search engines is quite low. They also noted that different engines retrieve different relevant documents, suggesting that metasearching could potentially improve the retrieval effectiveness.

In another study, Hawking et al. (2001) extended the list of features for comparative studies of search engines, and performed an evaluation of 20 commercial search engines, including metasearch engines and subject directories. They used genuine queries obtained from search engine query logs. Therefore, the relevance assessments were not performed by the users with the original information need. The evaluation measures were mean average precision, the mean reciprocal rank of the first retrieved relevant document, precision at n retrieved documents, as well as the average precision from rank 1 to rank 5. The results showed that there are significant inter-correlations between the different evaluation measures, but that there are no statistically significant difference among the top performing search engines. Hawking et al. (2001) also suggested that the retrieval effectiveness of the search engines was lower than that of IR systems evaluated in the Large Web task of TREC-8.

Some of the proposed methodologies for the evaluation of IR systems, and Web search engines in general, have investigated the automatic evaluation of systems. Chowdhury & Soboroff (2002) compared IR systems by pooling randomly selected retrieved documents. Can et al. (2004) computed the similarity between an information need and the top ranked documents from a set of search engines. The most similar documents were considered to be relevant to the information need. For each search engine, precision and recall were calculated. Even though the automatic evaluation of IR systems, or search engines, is an interesting topic, it is doubtful whether it is possible to achieve results similar to the evaluation based on human assessors.

3.6 Query classification and performance prediction

As discussed in the previous sections, there are different types of search tasks performed by users of Web search engines, as well as different retrieval techniques that can be applied for Web IR. For these reasons, current research has focused on the identification of the users' goals and the performance prediction for IR systems.

3.6.1 Identifying user goals and intentions

Beitzel et al. (2003) identified navigational queries from a query log, by matching the queries to the titles of categories in edited taxonomies. Rose & Levinson (2004) refined the taxonomy of user goals, or Web search tasks, originally proposed by Broder

(2002). They added sub-types of user goals for the informational search tasks and the transactional search tasks. They manually identified the user goals from search engine query logs, by using: the submitted queries; the sets of retrieved documents; the documents that users clicked on; and other subsequent actions of the users. Additionally, they suggested that the successful identification of the user goal may result in applying different relevance ranking algorithms for different queries, depending on the user goal. Bomhoff et al. (2005) also proposed to identify the intentions of users by examining the query logs. In order to identify navigational, informational, and transactional queries, they looked at several features including: the terms and the length of queries; the fraction of the query terms that appear in the URLs of documents the users clicked on; information about the Web browser of users; part-of-speech information about the query terms; and a timestamp for the query.

In the context of TREC-style experiments, Kang & Kim (2003) proposed a query type classification method. The query types correspond to different search tasks, and a different combination of evidence is applied for each query type. They considered two search tasks, namely the ad-hoc search task and the homepage finding search task from the Web track of TREC 2001 (Hawking & Craswell, 2001). The two tasks differ considerably, since the first one is an ad-hoc search task, while the other is a navigational search task. For identifying the query type, they employed terms that are more likely to appear in homepages, part-of-speech information about the query terms, anchor text of the incoming hyperlinks of Web pages, and co-occurrence information for the query terms.

As described in Section 3.5.1, the mixed query task was introduced in the Web track of TREC 2004 in order to evaluate the performance of IR systems when a stream of different types of queries is available, without knowing explicitly the type of each query. The results showed that query type classification resulted in significantly better than random accuracy, but it did not help retrieval effectiveness (Craswell & Hawking, 2004). Section 5.2.1 discusses the differences between query type classification and selective Web IR, which is proposed in this thesis.

3.6.2 Predicting query performance and dynamic combination of evidence

In addition to identifying the user goals, some methodologies have been proposed to predict the performance of queries, and the dynamic combination of evidence. This has been partly motivated by the introduction of the TREC Robust track (Voorhees, 2003, 2004), where IR systems are required to provide a measure of confidence in the quality of the results for each query, thus predicting their performance on a per-query basis.

Cronen-Townsend et al. (2002) introduced the clarity score and measured the ambiguity of a query as the divergence of the language model of the top ranked documents, from the language model of the whole document collection. The clarity score was shown to be correlated with the query's average precision.

Amati et al. (2004) introduced query difficulty predictors, based on measuring the divergence between the query terms' distribution in the top ranked documents, and the whole collection. When the two distributions have a high divergence, then it is more likely that the query is easy and the system will perform well. Their experimental results suggest that the query difficulty predictors correlate with the mean average precision of the first-pass retrieval, but they cannot be used to predict the effectiveness of automatically applying query expansion.

He & Ounis (2004) defined and evaluated five pre-retrieval query performance predictors. Unlike the above two approaches, which depend on the assigned scores of the retrieved documents, the pre-retrieval predictors depend only on the collection statistics of the query terms, and they can be computed before performing retrieval. The proposed predictors include: the length of the query; the standard deviation of the query terms' *idf* values; the ratio of the maximum *idf* over the minimum *idf* values for the query terms; a predictor related to the number of retrieved documents; and a simplified version of the query clarity score that is based on maximum likelihood estimates instead of retrieval scores. It was found that the simplified query clarity score is more effective at predicting the query performance for short queries, while the standard deviation of the query terms' *idf* values correlates well with the performance of longer queries. Plachouras, He & Ounis (2004) introduced an additional pre-retrieval query performance predictor, which corresponds to the average inverse collection term frequency.

Instead of modelling the query ambiguity, Evans et al. (2002) distinguished three types of topics. The first type refers to monolithic topics, where the retrieved documents are similar. The second type refers to structured topics, that may contain several relatively monolithic subtopics. The third type refers to diffuse topics, which may retrieve highly dissimilar documents. The topic structure can be quantified by either considering the stability and the number of generated groups of documents using clustering, or by measuring the similarity between samples of retrieved documents and the rest of the retrieved documents. Similarly, Yom-Tov et al. (2005) estimated the query difficulty based on the number of documents that contain subsets of the query terms. In this way, they identified queries that are difficult, because some of their aspects dominate the results. The aim of predicting query difficulty in this approach, as well as in the previously described ones, is to disable the automatic query expansion for difficult queries, which usually leads to a degradation of precision.

In addition to performance prediction for retrieval, there has been work on predicting the quality of evidence from the hyperlink structure of the Web. In the context of Web document classification, Fisher & Everson (2003) focused on the relation between the effectiveness of hyperlink analysis and the density of hyperlinks in a test collection. Gurrin & Smeaton (2003) studied the effectiveness of hyperlink structure analysis as the size of a document collection increases.

Other proposed approaches include dynamically setting the weight of hyperlink analysis algorithms on a per-query basis. Amitay et al. (2002) and Amitay et al. (2003) set the weight of additional evidence from hyperlink analysis algorithms, such as HITS or SALSA, or other sources of evidence, such as the anchor text and the number of incoming hyperlinks, according to features of the set of retrieved documents. Plachouras & Ounis (2005) employed Dempster-Shafer theory to combine content and hyperlink analysis on a per-query basis, according to the specificity of each query. The employed hyperlink analysis algorithms were PageRank and the Absorbing Model. Section 5.2.1 discusses the differences between dynamically setting the weights of different combinations of evidence and selective Web IR. The latter is the proposed approach in this thesis.

3.7 Summary

This chapter has presented an overview of Web IR. It discussed the differences between classical IR and Web IR, with respect to the hypertext document model, the structure of the Web, the quality of information on the Web, and the Web users' background. Next, a range of Web-specific sources of evidence, as well as different methodologies to combine them for effective retrieval, were presented. The chapter continued with covering the evaluation of Web IR systems in an experimental setting, as well as the evaluation of Web search engines, and closed with a review of user goals prediction and query performance prediction. Overall, most of the discussed techniques either apply a particular retrieval approach for all queries, or they estimate the difficulty of a query, in order to apply automatic query expansion or not, or they identify the goal of the user. There has not been any extensive investigation and evaluation of the selective application of different retrieval techniques on a per-query basis, according to the appropriateness of each retrieval approach.

The remainder of this thesis presents selective Web IR, a novel framework, which aims to apply appropriate retrieval approaches on a per-query basis. Chapter 4 investigates the potential improvements obtained from selective Web IR. Chapter 5 presents a decision theoretical framework for selective Web IR. The evaluation of the proposed framework is presented in Chapter 6. Chapter 7 investigates the application of selective Web IR in a setting where limited relevance information is available.

Chapter 4

Retrieval Approaches for Selective Web Information Retrieval

4.1 Introduction

The aim of this thesis is to investigate the effectiveness of selective Web IR. However, before introducing a framework for selective Web IR, it is necessary to examine the potential of such an approach in improving the retrieval effectiveness. This chapter aims to establish this potential, by examining and evaluating a range of retrieval approaches.

This chapter starts with describing the experimental setting and the used search tasks from various TREC Web tracks in Section 4.2. Section 4.3 evaluates the effectiveness of retrieval from the full text of documents, and other document representations. These document representations correspond to the text of the title, and the heading HTML tags, as well as to the anchor text of the incoming hyperlinks. For each document representation, a range of statistically independent weighting models is evaluated. These models include the DFR weighting models PL2, PB2, I(n_e)C2, and DLH, as well as BM25 (Table 2.1 on page 19).

Next, Section 4.4 presents a new extension of the DFR framework, in order to allow the combination of different document fields, and to perform per-field normalisation of the term frequencies. Section 4.5 investigates the use of query-independent evidence for Web IR, including the URLs of Web documents, PageRank, and the Absorbing Model, a novel model for hyperlink structure analysis.

The introduced retrieval approaches are separately evaluated for the different types of ad-hoc, topic distillation, home page finding, and named page finding search tasks of the TREC Web tracks. The associated hyper-parameters are set in order to optimise precision for each evaluated task, and each weighting model. This allows for the comparison of the retrieval approaches on the basis of their optimal performance.

Section 4.6 evaluates the proposed retrieval approaches in a different setting, in order to reduce any overfitting effects from the optimisation process. First, the hyper-parameters of the retrieval approaches are set in order to optimise precision for different sets of mixed tasks. Second, the optimisation process is stopped early, before converging to the optimal setting of the hyper-parameters.

This chapter closes with establishing the potential for improvements in retrieval effectiveness from selective Web IR in Section 4.7. The results from this chapter provide a further motivation for the introduction of the decision theoretical framework for selective Web IR in the next chapter.

4.2 Experimental setting

The retrieval approaches presented in this chapter are evaluated using the standard TREC Web test collections, and the associated search tasks from the TREC Web tracks. Table 4.1 presents an overview of the tasks used in the TREC Web tracks with the WT10g and the .GOV test collections. The tasks from the earlier Very Large Collection tracks are not used, because their aim was primarily to test whether IR systems would scale to process a large amount of data (Hawking & Craswell, 2005). Moreover, the WT2g test collection is not employed, because it corresponds to a small subset of the WT10g collection, and it has been used only once in TREC-8 for an ad-hoc search task (Hawking et al., 1999).

The WT10g test collection consists of 1,692,096 Web documents, and 10GB of data (Bailey et al., 2003). The topic relevance tasks tr2000 and tr2001, used in TREC-9 (Hawking, 2000) and TREC 2001 (Hawking & Craswell, 2001), respectively, correspond to ad-hoc search tasks. The task hp2001, which is associated with the WT10g test collection, corresponds to the home page finding task of TREC 2001 (Hawking & Craswell, 2001), where the topics are about finding the home page of a Web site, the name of which corresponds to the query.

Name	Used in	Task	Collection	Topics
tr2000	TREC-9	Topic relevance	WT10g	451-500 (50 topics)
tr2001	TREC 2001	Topic relevance	WT10g	501-550 (50 topics)
hp2001	TREC 2001	Home page finding	WT10g	EP1-EP145 (145 topics)
td2002	TREC 2002	Topic distillation	.GOV	551-600 (50 topics)
np2002	TREC 2002	Named page finding	.GOV	NP1-NP150 (150 topics)
td2003	TREC 2003	Topic distillation	.GOV	TD1-TD50 (50 topics)
ki2003	TREC 2003	Known-item finding	.GOV	NP151-NP450 (300 topics)
mq2004	TREC 2004	Mixed query	.GOV	WT04-1-WT04-225 (225 topics)
hp2003	TREC 2003	Home page finding	.GOV	subset of ki2003 (150 topics)
np2003	TREC 2003	Named page finding	.GOV	subset of ki2003 (150 topics)
mq2003	TREC 2003	Mixed query	.GOV	td2003 and ki2003 (350 topics)
td2004	TREC 2004	Topic distillation	.GOV	subset of mq2004 (75 topics)
hp2004	TREC 2004	Home page finding	.GOV	subset of mq2004 (75 topics)
np2004	TREC 2004	Named page finding	.GOV	subset of mq2004 (75 topics)

Table 4.1: The search tasks and the corresponding topic sets from the TREC Web tracks.

The .GOV test collection consists of 1,247,753 Web documents, and 18GB of data. The associated topics with .GOV have been used for the topic distillation (td2002) and named page finding (np2002) tasks of TREC 2002 (Craswell & Hawking, 2002), the topic distillation (td2003) and known-item finding (ki2003) tasks of TREC 2003 (Craswell & Hawking, 2002), and the mixed query task (mq2004) of TREC 2004 (Craswell & Hawking, 2002). The tasks hp2003 and np2003 correspond to the home page and named page finding topics of the TREC 2003 known item finding task ki2003, respectively. The task mq2003 corresponds to the set of topics from td2003 and ki2003. The tasks td2004, hp2004, and np2004 correspond to the topic distillation, home page finding and named page finding topics of the mixed query task mq2004 of TREC 2004 Web track, respectively.

The proposed retrieval approaches in Sections 4.3, 4.4, and 4.5 will be evaluated separately for the different types of tasks: tr2000, tr2001, hp2001, td2002, np2002, td2003, hp2003, np2003, td2004, hp2004, and np2004. Sections 4.6 and 4.7 will also consider the mixed search tasks mq2003 and mq2004, focusing on Web-specific search tasks.

Savoy & Picard (2001) highlighted that removing stop words and applying stemming has a positive effect on the precision of a retrieval system for the TREC-7 ad-hoc retrieval task with the WT2g collection. On the other hand, Hawking et al. (1998*a*) suggested that stop words can be indexed and stemming can be applied during retrieval,

if necessary. This is more similar to the indexing approach taken by commercial Web search engines, where stop words are usually indexed, and weak stemming may be applied. For Web specific tasks, such as topic distillation, named page and home page finding tasks, there has not been any clear indication that removing stop words and applying stemming harm the retrieval effectiveness (Craswell & Hawking, 2004).

The WT10g and .GOV test collections are indexed by processing the text, which is visible on a Web browser application. Stop words are removed, and the stemming algorithm of Porter (1980) is applied during indexing. In addition, certain restrictions are applied in order to reduce the number of non-informative terms in the generated document index. First, tokens which are longer than 20 characters are discarded. Next, tokens that contain more than three same consecutive characters, or more than four numerical digits, are discarded. This restriction is not applied when indexing the anchor text of incoming hyperlinks of documents. Indeed, the anchor text is intentionally used to concisely describe other Web documents. Therefore, it is considered to be more informative. The applied indexing restrictions aim to reduce the number of non-informative terms in the document index, and also result in reducing the size of the generated data structures, similarly to the approach of static pruning of the inverted index (Carmel et al., 2001). In this thesis, indexing and retrieval have been performed with the Terrier IR platform (Ounis et al., 2005).

4.3 Document representations for Web information retrieval

This section investigates the retrieval effectiveness of different document representations for Web documents. The employed document representations include the full text of Web documents, the title, the headings, and the anchor text of incoming hyperlinks.

For each document representation, a range of five weighting models is tested. The first four models are derived from the Divergence From Randomness (DFR) framework (Section 2.3.3, page 13): PL2, PB2, I(n_e)C2, and DLH. The fifth weighting model is BM25 (Section 2.3.1, page 10). The formulae of the five weighting models are given in Table 2.1 on page 19. As discussed in Chapter 2, these weighting models have been selected for several reasons. The weighting models PL2 and I(n_e)C2 are robust and perform well across a range of search tasks (Plachouras & Ounis, 2004; Plachouras, He

& Ounis, 2004). The weighting model PB2 is selected in order to test the combination of the Poisson randomness model with the Bernoulli model for the after-effect. The weighting model DLH is particularly interesting, because it does not have any associated hyper-parameter. The weighting model BM25 is employed, because it has been frequently used by many participants of TREC.

In order to compare the effectiveness of each weighting model, a two-step optimisation process is employed to set the hyper-parameters of the weighting models PL2, PB2, I(n_e)C2, and BM25. For each tested task, the hyper-parameters are set in order to optimise the precision of the weighting models. Note that the weighting model DLH does not have any associated hyper-parameter, and therefore, no optimisation is required.

4.3.1 Representing Web documents

The analysis of the textual content of documents is necessary for matching documents to the users' queries. There are several different representations for Web documents. The first representation corresponds to the full text of Web documents. In addition, particular features of HTML can be employed to define other document representations.

HTML is a markup language that is used for authoring Web documents (Raggett et al., 1999). It provides a set of tags for specifying the structure of Web documents, as well as the way they should be rendered by a Web browser application. The HTML tags convey information about the textual content of documents, which can be used to improve the retrieval effectiveness in navigational, and informational search tasks (Craswell & Hawking, 2002, 2004; Craswell et al., 2003; Hawking & Craswell, 2001). For example, the text within the tags ⟨TITLE⟩ and ⟨/TITLE⟩ corresponds to the title of the Web document. Jin et al. (2002) observed that the user queries are more similar to the titles of documents than to the actual documents, and they suggested that both the queries and the titles provide concise descriptions of information. The text within the different heading tags (for example ⟨H1⟩ and ⟨/H1⟩) usually corresponds to the titles of a Web document's sections.

The anchor text, which appears within the tags ⟨A⟩ and ⟨/A⟩ in the source documents of incoming hyperlinks, functions as a brief description of a document. Eiron & McCurley (2003*b*) suggested that the anchor text exhibits similarities to the user queries on a statistical and grammatical level. In order to provide a concise description

of Web documents, anchor text tends to be short and may contain abbreviated terms and acronyms. Compared to the titles of Web documents, Eiron & McCurley pointed that there are as many anchor texts as the incoming hyperlinks of a document, while there can be only one title for a document. The anchor text has been shown to be effective for navigational tasks, such as named page finding (Craswell & Hawking, 2002, 2004; Craswell et al., 2003) and home page finding (Craswell et al., 2001), as well as for informational tasks, such as topic distillation, when there is a bias towards the home pages of Web sites (Craswell & Hawking, 2004; Craswell et al., 2003).

In order to establish the retrieval effectiveness of the different HTML tags, the documents are represented only by the text within the corresponding tags. In addition to the full text representation, the other three document representations correspond to: the text within the title tags; the text within the heading tags (⟨H1⟩ and ⟨/H1⟩ to ⟨H6⟩ and ⟨/H6⟩); and the anchor text of the incoming hyperlinks.

4.3.2 Parameter setting

The evaluation of the different document representations is performed with a range of weighting models. However, the retrieval effectiveness of the weighting models depends on the setting of any associated hyper-parameters. In order to compare the document representations, the hyper-parameters are set in order to optimise the retrieval effectiveness of the weighting models. This allows for the comparison of the weighting models on the basis of their optimal performance.

The employed weighting models include four DFR weighting models: PL2, PB2, I(n_e)C2, and DLH. The weighting model BM25 is used as well. Their formulae are given in Table 2.1, page 19. All the employed weighting models, with the exception of DLH, have associated hyper-parameters that need to be estimated. The DFR weighting models PL2, PB2, and I(n_e)C2 have one associated hyper-parameter, c, which is related to the *normalisation 2* from Equation (2.16) on page 15. This parameter takes real values greater than zero. The considered parameters for the weighting model BM25 are b, which is related to the term frequency normalisation, and k_1, which is a saturation factor for the term frequency. The parameters k_2, which is related to a correction of the weights due to the different lengths of documents, and k_3, which is related to the importance of the term frequency in a query, are set equal to 0 and 1000, respectively (Robertson et al., 1994).

The values of the parameter c for the models PL2, PB2, and I(n_e)C2, and the parameters b and k_1 for the model BM25, are independently set for every tested task, after performing a one-dimensional optimisation for the DFR models, and a two-dimensional optimisation for BM25. Each optimisation maximises the mean average precision (MAP).

The direct optimisation of MAP is preferred over more classical optimisation techniques, such as maximum likelihood estimation, for two main reasons suggested by Metzler & Croft (2005). First, the training data, which corresponds to the available relevance information, is a very small sample of the event space of documents and queries. Therefore, the maximum likelihood estimation is less likely to result in a good estimate of the parameters. Second, the maximisation of the likelihood for generating the training data does not necessarily mean that a metric, such as MAP, is optimised. Therefore, it may be more useful to optimise a particular retrieval effectiveness metric, such as MAP.

The direct optimisation of MAP for each tested task results in optimal retrieval effectiveness. However, a potential problem is the overfitting of the weighting models to each task. For this reason, the optimisation in Section 4.6 is performed with different training and testing tasks of mixed query types. The optimisation process is also stopped after a given number of iterations.

The optimisation involves two steps. In the first step, a simulated annealing algorithm (Press et al., 1992) is applied. Its output is used as a starting point for the second step, where the applied optimisation algorithm is based on a combination of heuristics to avoid local maxima (Yuret, 1994). The optimisation is performed at least twice for each of the tested topic sets, in order to increase the chances of finding a global maximum for MAP, and the most effective parameter values are selected. The optimal c values for the DFR models PL2F, PB2F and I(n_e)C2F, as well as the parameters b and k for BM25, are shown in Table A.1 of Appendix A. The weighting model DLH does not have any associated hyper-parameter, because the hypergeometric distribution naturally incorporates term frequency normalisation in the model, as discussed in Section 2.3.3.

Figure 4.1 shows the tested values of c during the optimisation of full text retrieval with PL2 for the tasks tr2001, td2004, hp2004, and np2004. The c parameter is set to higher values for the topic-relevance topics, than for the topic distillation, or any of the

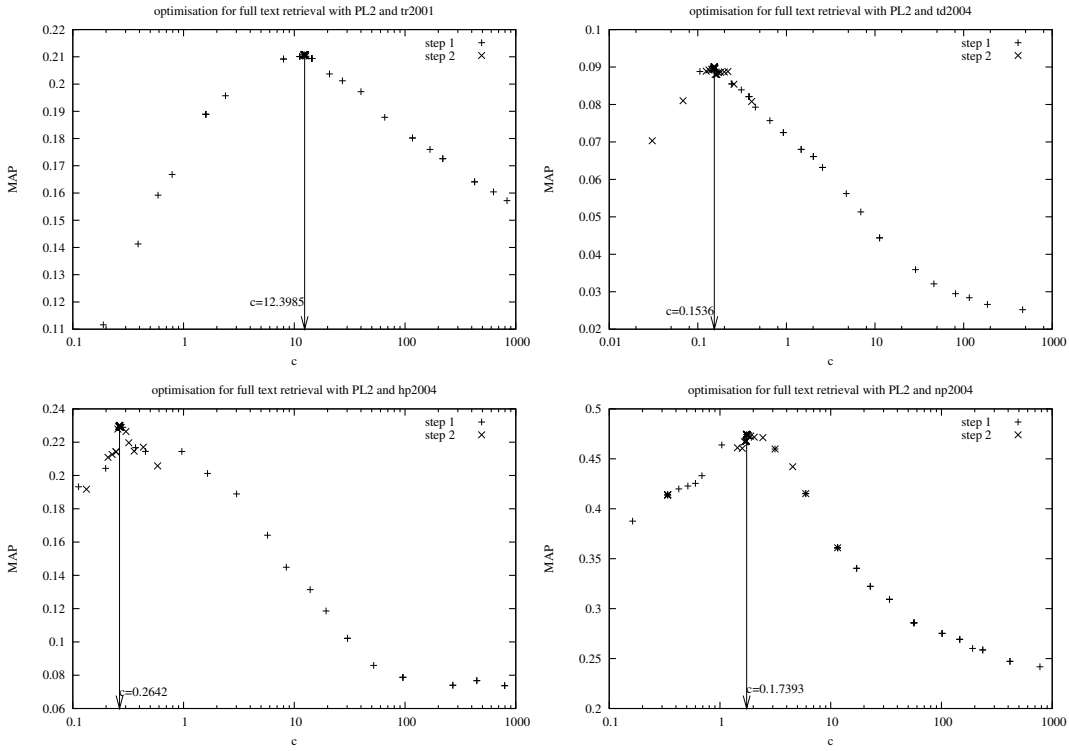

Figure 4.1: The obtained mean average precision (MAP) for different c values tested during the two-step optimisation of full text retrieval with PL2 for the topic sets tr2001, td2004, hp2004 and np2004.

navigational search topics. This is related to both the average length of documents in the collection (Table 4.2), and the average length of the relevant documents (Table 4.3). The document length corresponds to the number of indexed tokens for a particular document representation. Regarding the WT10g collection and the ad-hoc task tr2001, the average length of documents (394.87 from Table 4.2) is lower than the average length of relevant documents (1689.98 from Table 4.3), and the optimal c value is relatively high ($c = 12.3985$ from the top left diagram in Figure 4.1). For the .GOV collection and the topic distillation task td2004, where there is a bias towards the home pages of Web sites, the average length of documents (726.71 from Table 4.2) is higher than the average length of relevant documents (494.28 from Table 4.3). In this case, the optimal value c is relatively low ($c = 0.1536$ from the top right diagram in Figure 4.1).

The above dependence between the length of the relevant documents, the average document length, and the parameter c is explained with respect to the formula of *normalisation 2* from Equation (2.16):

$$tfn = tf \cdot log_2\big(1 + c \cdot (\bar{l}/l)\big)$$

where l is the document length, $\bar{l}$ is the average document length, c is a hyper-parameter, tf is the term frequency, and tfn is the normalised term frequency. When a low c value is used and $\bar{l}/l > 1$, then $tfn/tf = \log_2\big(1 + c \cdot (\bar{l}/l)\big) \approx 1$. When a low c value is used and $\bar{l}/l < 1$, then $tfn/tf = \log_2(1 + c \cdot (\bar{l}/l)) < 1$. Thus, low c values favour short documents, and penalise longer ones. When a high c value is used, for either $\bar{l}/l < 1$, or $\bar{l}/l > 1$, $tfn/tf = \log_2(1 + c \cdot (\bar{l}/l)) > 1$. Therefore, high c values correspond to a weaker normalisation of the term frequencies, because $tfn/tf > 1$ regardless of the ratio $\bar{l}/l$.

Average document length		
Document representation	WT10g	.GOV
Full-text	394.87	726.71
Title	4.42	4.12
Heading	25.41	10.14
Anchor text	13.50	21.41

Table 4.2: The average length of documents for the different document representations in WT10g and .GOV test collections. The document length corresponds to the number of indexed tokens for each document, after removing stop words.

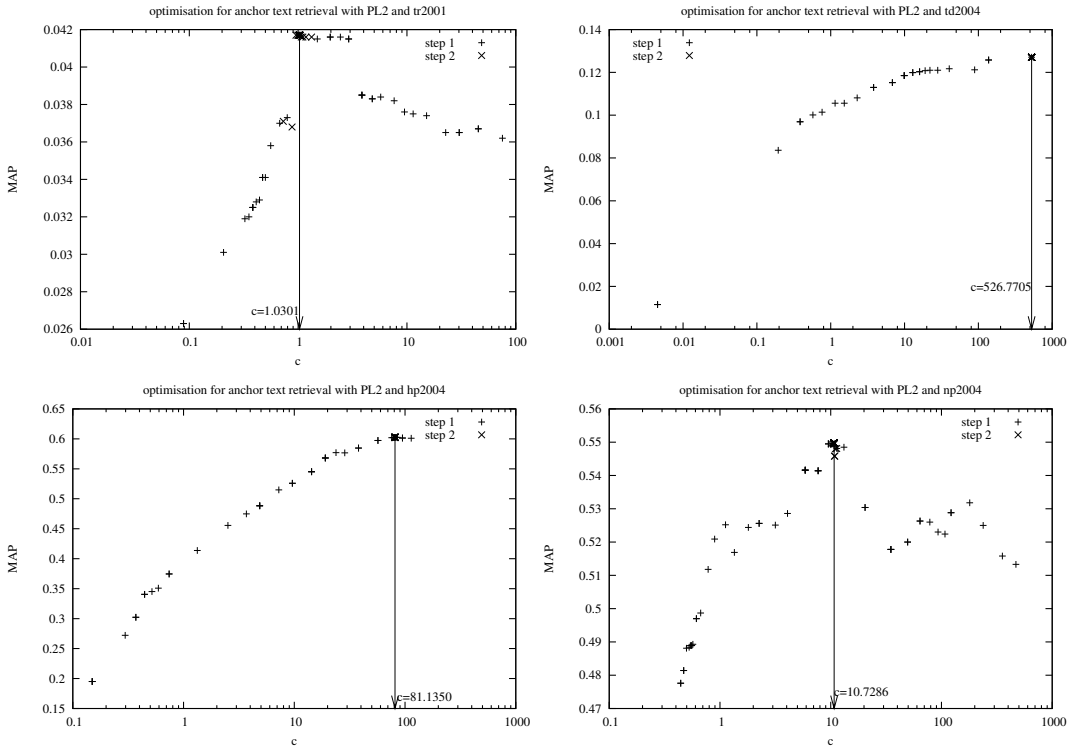

Figure 4.2: The obtained mean average precision (MAP) for different c values tested during the two-step optimisation of anchor text retrieval with PL2 for the topic sets tr2001, td2004, hp2004 and np2004.

Topic relevance		Topic distillation		Home page finding		Named page finding	
Task	Avg. Length	Task	Avg. Length	Task	Avg. Length	Task	Avg. Length
Full-text							
tr2000	2016.76	td2002	1315.33	hp2001	204.70	np2002	782.37
tr2001	1689.98	td2003	539.66	hp2003	266.12	np2003	834.71
		td2004	494.28	hp2004	357.10	np2004	923.39
Title							
tr2000	5.72	td2002	4.88	hp2001	4.05	np2002	5.49
tr2001	4.18	td2003	4.48	hp2003	4.88	np2003	4.60
		td2004	4.73	hp2004	4.69	np2004	4.79
Headings							
tr2000	78.78	td2002	14.69	hp2001	13.63	np2002	40.12
tr2001	56.25	td2003	8.99	hp2003	3.31	np2003	11.67
		td2004	18.42	hp2004	2.82	np2004	14.71
Anchor text							
tr2000	14.47	td2002	84.82	hp2001	1264.65	np2002	63.41
tr2001	11.42	td2003	300.62	hp2003	3258.32	np2003	79.62
		td2004	346.49	hp2004	902.20	np2004	218.30

Table 4.3: The average length of relevant documents for the different topic sets, and for the different document representations. The document length corresponds to the number of indexed tokens for each document, after removing stop words.

Figure 4.2 displays the range of tested c values for the optimisation of PL2 with the anchor text document representation, for the topic sets tr2001, td2004, hp2004, and np2004. In particular for the task td2004, where there is a bias towards home pages, the optimal c value is 526.7705 (top right diagram in Figure 4.2). For the tasks td2004, the average anchor text length of the relevant documents is 346.49 (Table 4.3), while the average anchor text length in .GOV is 21.41 (Table 4.2).

The optimal c values obtained for the full text and the anchor text representations of documents, regarding the task td2004, indicate that different representations of documents require different normalisation settings. This is in agreement with Hawking, Upstill & Craswell (2004), who suggested applying different length normalisation for the full text representation and the anchor text representation of Web documents. The benefit from such an approach is that the documents with a high number of hyperlinks, and consequently, a significant amount of anchor text, are not penalised by document length normalisation.

4.3.3 Evaluation results

This section presents the evaluation results obtained with the different representations of documents and the various used weighting models.

Several measures have been employed for the evaluation of IR systems in the TREC Web tracks. Mean average precision has been used to evaluate ad-hoc tasks. The mean reciprocal rank of the first retrieved relevant document (MRR1) has been used to evaluate navigational tasks. For the evaluation of topic distillation tasks, precision at 10 (P10), mean average precision (MAP) and R-precision (R-Prec) have been employed. In order to have a consistent setting, the evaluation in this thesis is performed using mean average precision. The mean reciprocal rank of the first retrieved relevant document is equivalent to mean average precision when there is only one relevant document for each query. In addition, precision at 10 is expected to correlate with average precision, even though the optimal setting for average precision does not necessarily correspond to the optimal setting for precision at 10.

Table 4.4 shows the mean average precision (MAP) for all the tested topics and the different representations of documents. Each row shows the achieved MAP by the five tested weighting models for a task and a document representation. Each column shows the achieved MAP by a particular weighting model for all the tested tasks and document representations. The entries in bold show the weighting model that results in the highest MAP for each task and representation of documents.

Row	Task	Mean Average Precision (MAP)				
		PL2	PB2	I(n_e)C2	DLH	BM25
		Full text				
1	tr2000	0.2038	0.1923	0.2073	0.1606	**0.2102**
2	tr2001	0.2107	0.2032	**0.2132**	0.1746	**0.2132**
3	td2002	**0.1997**	0.1909	0.1983	0.1738	0.1989
4	td2003	**0.1245**	0.1108	0.1167	0.1091	0.1234
5	td2004	0.0901	0.0844	0.0927	0.0856	**0.0956**
6	hp2001	0.3355	0.3280	0.3524	0.3331	**0.3552**
7	hp2003	0.2528	0.2190	0.2624	0.2608	**0.2893**
8	hp2004	0.2300	0.2074	**0.2335**	0.1956	0.2276
9	np2002	0.5651	0.5432	**0.5785**	0.5034	0.5771
10	np2003	0.5185	0.4850	0.5237	0.5095	**0.5309**
11	np2004	0.4744	0.4508	0.4614	0.4029	**0.4853**
		Title				
12	tr2000	0.0281	0.0264	0.0282	0.0284	**0.0297**
13	tr2001	0.0214	0.0208	0.0208	0.0175	**0.0224**
14	td2002	0.0512	**0.0537**	0.0528	0.0501	0.0514
15	td2003	0.0759	0.0759	0.0758	0.0661	**0.0789**
16	td2004	0.0641	0.0640	0.0640	0.0571	**0.0650**

continued on next page

continued from previous page						
Row	Task	Mean Average Precision (MAP)				
		PL2	PB2	I(n_e)C2	DLH	BM25
		Title				
17	hp2001	**0.3288**	0.3194	0.3230	0.3066	0.3287
18	hp2003	0.2796	0.2765	0.2860	0.2726	**0.2974**
19	hp2004	0.3026	0.3095	0.3020	0.3009	**0.3130**
20	np2002	**0.4014**	0.4000	0.3958	0.3974	0.3996
21	np2003	0.4147	0.4136	**0.4148**	0.3975	0.4115
22	np2004	0.4282	0.4287	**0.4288**	0.4267	0.4276
		Headings				
23	tr2000	0.0501	0.0463	0.0480	0.0474	**0.0511**
24	tr2001	0.0527	0.0554	0.0578	0.0527	**0.0578**
25	td2002	0.0422	0.0420	**0.0432**	0.0401	0.0425
26	td2003	**0.0684**	0.0680	0.0682	0.0415	0.0676
27	td2004	**0.0397**	0.0383	0.0393	0.0336	0.0379
28	hp2001	0.1555	0.1506	0.1607	0.1549	**0.1633**
29	hp2003	**0.1174**	0.1116	0.1113	0.1084	0.1173
30	hp2004	0.1027	0.0994	0.0995	0.1037	**0.1060**
31	np2002	0.1928	0.1855	0.1946	0.1882	**0.1952**
32	np2003	0.2432	0.2330	0.2341	0.2362	**0.2510**
33	np2004	**0.3419**	0.3194	0.3209	0.3204	0.3389
		Anchor text				
34	tr2000	0.0328	0.0222	0.0244	0.0274	**0.0402**
35	tr2001	0.0417	0.0352	0.0378	0.0267	**0.0436**
36	td2002	0.0663	0.0563	0.0652	0.0581	**0.0669**
37	td2003	0.1433	0.1216	0.1239	0.1216	**0.1437**
38	td2004	**0.1271**	0.1149	0.1126	0.1013	0.1261
39	hp2001	0.5219	0.4828	0.5265	0.4337	**0.5383**
40	hp2003	**0.6675**	0.6317	0.6423	0.4365	0.6655
41	hp2004	0.6025	0.5159	0.5711	0.4329	**0.6043**
42	np2002	0.4476	0.3297	0.4008	0.4287	**0.4630**
43	np2003	0.4939	0.4187	0.4797	0.4885	**0.5060**
44	np2004	**0.5498**	0.4298	0.4544	0.5176	0.5225

Table 4.4: Evaluation of different document representations with the DFR models PL2, PB2, I(n_e)C2, DLH and the weighting model BM25. The bold entries correspond to the weighting model that results in the highest MAP. The parameter values of the models are given in Table A.1 on page 226.

For the topic relevance tasks, tr2000 and tr2001, it can be seen that the full text representation is more appropriate than any of the title, headings, or anchor text representations of documents (rows 1-2 vs. rows 12-13, 23-24, and 34-35 in Table 4.4). In particular, the achieved MAP with the anchor text and title representations is less than 0.0500 (rows 12-13, and 34-35 in Table 4.4). The topic distillation task td2002 is more similar to an ad-hoc retrieval task, such as the topic relevance tasks tr2000 and tr2001. Therefore, the full text representation of documents is the most effective one

(row 3 vs. rows 14, 25, and 36 in Table 4.4).

For the topic distillation tasks td2003 and td2004, the most effective document representation is the anchor text (rows 37-38 vs. rows 4-5, 15-16, and 26-27 in Table 4.4), due to the fact that for those particular tasks, the relevant documents are restricted to be the home pages of Web sites about the query topic. For the same reason, the anchor text representation of documents is the most effective for the home page finding topic sets (rows 39-41, vs. rows 6-8, 17-19, and 28-30 in Table 4.4). For the home page finding tasks, the title representation results in similar levels of MAP as the retrieval from the full text representation (rows 17-19 vs. rows 6-8 in Table 4.4). This indicates that the title is an adequate description for the name of a Web site's home page, even though its size is limited, and the frequencies of its terms are distributed almost uniformly.

For the named page finding tasks np2002, np2003 and np2004, the most effective representations of documents are the full text and the anchor text representations (rows 9-11, and 42-44 from Table 4.4, respectively). However, there is no document representation that outperforms the other ones consistently.

With respect to the different weighting models that are evaluated, Table 4.4 shows that in most of the cases the weighting models PL2, I(n_e)C2 and BM25 outperform the weighting models PB2 and DLH. For the title and headings representations, where the distribution of term frequencies are more uniform than the content representation, the different weighting models have very similar performance. It should be noted that the small differences in retrieval effectiveness may be attributed to the parameter estimation process.

Before closing with a discussion and some conclusions from the evaluation of the different document representations, the next section investigates an implication of using the Poisson randomness model in the DFR weighting models PL2 and PB2, when the query terms have extremely high frequency in the test collection.

4.3.4 Impact of query terms with high frequency on the Poisson-based models

Considering the weighting models PL2 and PB2 (Table 2.1 on page 19), the Poisson distribution is an approximation of the Bernoulli distribution and $\lambda = \frac{F}{N}$ is the maximum likelihood estimator of the distribution's mean and variance, where F is the frequency of a term in the document collection, and N is the number of documents in

the collection. When $\lambda \ll 1$, or equivalently $F \ll N$, then the Poisson distribution provides a good approximation of the Bernoulli distribution. This is the case for terms with a low frequency in a large document collection.

When the term frequency F is comparable to the size N of the document collection, or equivalently when λ is close to 1, then the Poisson does not provide a good approximation of the Bernoulli distribution. This situation is more likely to appear in the context of the .GOV collection, which is a domain specific collection of documents from governmental organisations. Therefore, terms such as `national` or `federal` are very likely to occur many times, because their distribution reflects the topics of the documents. The application of stemming transforms these terms to `nation` and `feder`, respectively, and results in a further increase of their frequency.

Example 1 The query NP167 from the known item finding task ki2003 of the TREC 2003 Web track is: `Federal Deposit Insurance Corporation`, and it corresponds to a home page finding query for the .GOV test collection. The term `Federal` is stemmed to `feder`, which appears 1,465,491 times in the .GOV collection. The number of documents in .GOV is 1,247,753. Therefore, $\lambda = \frac{F}{N} = \frac{1{,}465{,}491}{1{,}247{,}753} > 1$. □

The terms for which $\lambda > 1$ can be considered as stop words during retrieval. Therefore, these terms can be ignored when assigning weights to documents. Table 4.5 shows the retrieval effectiveness of the weighting models PL2 and PB2 for full text retrieval, when scores are assigned for all the query terms, irrespectively of the value of λ, and when scores are not assigned for the terms which result in $\lambda > 1$. In each case, the weighting models have been optimised with respect to MAP, as described in Section 4.3.2. The c values shown in Table 4.5 correspond to cases where scores are not assigned for the terms with $\lambda > 1$.

For all the tested topic sets, there are only small differences in MAP resulting from either assigning weights for the query terms with $\lambda > 1$, or ignoring these terms. This suggests that the weights assigned to documents for a term t when $\lambda > 1$ do not have an important effect on the resulting MAP. For the remainder of this thesis, when the weighting models PL2 or PB2 are employed, all the query terms will be used to assign weights to documents, irrespectively of the associated λ value.

	Assign scores for $\lambda > 1$		Do not assign scores for $\lambda > 1$			
	MAP		MAP		c	
Task	PL2	PB2	PL2	PB2	PL2	PB2
tr2000	0.2038	0.1923	0.2029	0.1950	12.0603	53.8243
tr2001	0.2107	0.2032	0.2103	0.2054	11.9829	10.7955
td2002	0.1997	0.1909	0.2030	0.1938	1.2796	1.0132
td2003	0.1245	0.1108	0.1245	0.1108	0.4133	0.2614
td2004	0.0901	0.0844	0.0909	0.0868	0.2086	0.1424
hp2001	0.3355	0.3280	0.3328	0.3278	0.3663	0.3400
hp2003	0.2528	0.2190	0.2446	0.2045	0.3128	0.2200
hp2004	0.2300	0.2074	0.2295	0.2136	0.7904	0.5988
np2002	0.5651	0.5432	0.5636	0.5403	2.0209	1.4632
np2003	0.5185	0.4850	0.5193	0.4800	1.4065	1.1433
np2004	0.4744	0.4508	0.4644	0.4342	2.8387	1.9617

Table 4.5: Mean Average Precision (MAP) for full text retrieval with the weighting models PL2 and PB2, when query terms with $\lambda > 1$ are employed for assigning weights to documents, or they are treated as stop words.

4.3.5 Discussion and Conclusions

In order to put the obtained results in the context of the various TREC Web tracks, Table 4.6 presents the official measure of evaluation of the best official submitted runs to the corresponding TREC Web track for each of the tested topic sets. Wherever the official evaluation measure is not mean average precision (MAP), and if it is available in the TREC proceedings, then it is reported in addition to the official evaluation measure.

The evaluation results for tr2000, tr2001, and td2002 have shown that the full text representation of documents is very effective for ad-hoc tasks. For example, full text retrieval with the weighting model PL2 outperforms the best performing run submitted to the topic distillation task td2002 in the TREC 2002 Web track (0.1997 from row 3 in Table 4.4 vs. 0.1571 from row 3 in Table 4.6). For the topic distillation task td2004, the anchor text representation of documents with PL2 results in lower MAP than that of the best performing runs (0.1271 from row 38 in Table 4.4 vs. 0.1791 from row 5 in Table 4.6). For the named page finding task np2004, all the four different document representations result in lower MAP than the best performing runs in TREC (rows 11, 22, 33, and 44 from Table 4.4 vs. row 11 from Table 4.6).

Overall, this section has presented the first step towards the introduction of effective retrieval approaches for Web IR. It has examined four document representations including: the full text; the title; the headings; and the anchor text of Web documents. The hyper-parameters of the weighting models PL2, PB2, I(n_e)C2, and BM25 have

Row	Tasks	Run name	Official Evaluation Measure	MAP (if available and not official evaluation measure)
1	tr2000	j2cbt9wcs1	MAP=0.2011	-
2	tr2001	fub01be2	MAP=0.2226	-
3	td2002	thutd5	P10=0.2510	MAP=0.1571
4	td2003	csiro03td03	R-Prec=0.1636	MAP=0.1543
5	td2004	uogWebCAU150	MAP=0.1791	-
6	hp2001	tnout10epCAU	MRR1=0.774	-
7	hp2003	csiro03ki01	MRR1=0.815	-
8	hp2004	MSRC04C12	MRR1=0.749	MAP=0.7351
9	np2002	thunp3	MRR1=0.719	-
10	np2003	LmrEq	MRR1=0.688	-
11	np2004	MSRC04B2S	MRR1=0.731	MAP=0.7232
12	mq2004	MSRC04B2S	Avg=0.546	MAP=0.5389

Table 4.6: Evaluation results of the best official submitted runs to the Web tracks from TREC-9 to TREC 2004. For the mixed query task mq2004, Avg stands for the average of MAP for the topic distillation and MRR1 for the home page and the named page finding tasks.

been set in order to directly optimise mean average precision. Note that the weighting model DLH is parameter-free, as discussed in Section 2.3.3.4, and it does not require any optimisation. The evaluation results have shown that full text retrieval is very effective for ad-hoc search tasks. However, there is room for improvements in retrieval effectiveness for Web specific tasks. In addition, this section has shown that the weighting models PL2 and PB2 are robust when assigning weights for query terms with very high collection frequencies.

In order to improve the effectiveness of the employed retrieval approaches, the combination of document representations, or fields, is introduced in the next section.

4.4 Combining document fields

The effectiveness of each document representation has been evaluated separately so far. This section investigates the improvements in retrieval effectiveness from the combination of different fields, in order to obtain a better representation for documents. The combination of fields is achieved by extending the evaluated weighting models in Section 4.3. The employed fields are: the body of Web documents; the anchor text of incoming links; and the title of Web documents. The body field is defined as the text between the HTML tags `<BODY>` and `</BODY>`. Compared to the full text document representation, the body field includes the headings, but not the title of Web

documents.

4.4.1 Weighting models for field retrieval

This section extends the DFR framework with a new normalisation method, which takes into account the fields of Web documents, that is the terms that appear within particular HTML tags. This new normalisation method applies term frequency normalisation and weighting for a number of different fields. The per-field normalisation has been similarly applied in (Zaragoza et al., 2004) using the BM25 formula. In this thesis, a different document length normalisation formula is used.

Per-field normalisation is useful in a Web context, where different document fields need to be combined. There are several ways to combine the information from different fields of documents. One approach involves performing retrieval independently from each field and then, merging the ranked lists of results (Fagin et al., 2003). The combination of the different fields can be achieved as the linear combination of relevance scores for each of the document representations (Gao et al., 2001; Kamps et al., 2003; Savoy et al., 2003; Tomiyama et al., 2003). In the context of language modelling, the combination of fields, or different document representations, is achieved with a linear combination of language models computed for each of the fields or document representations (Ogilvie & Callan, 2003).

Plachouras et al. (2003, 2002) extended documents with the anchor text of their incoming hyperlinks and treated the anchor text as a field of the document, effectively adding the frequencies of terms from the body and the anchor text. In addition, Plachouras, He & Ounis (2004) re-weighted the documents according to the importance of fields and increased the documents' scores by a certain percentage when a query term appeared in a particular field.

Robertson et al. (2004) suggested that it is more appropriate to weight and combine the frequencies of terms from different fields in a *pseudo-frequency*, before applying a term weighting model. Hawking, Upstill & Craswell (2004) suggested that terms in the body and the anchor text of Web documents are distributed very differently. For example, a term may occur many times in a document, because of the document's verbosity. On the other hand, a term appearing many times in the anchor text of a document's incoming hyperlinks represents votes for this document. Thus, performing normalisation and weighting independently for the various fields allows to take into

account the different characteristics of the fields, and to achieve their most effective combination.

The per-field *normalisation 2F* extends *normalisation 2* from Equation (2.16) (Amati & Van Rijsbergen, 2002), so that the normalised term frequency tfn corresponds to the weighted sum of the normalised term frequencies tf_f for each used field f:

$$tfn = \sum_f \Big(w_f \cdot tf_f \cdot \log_2(1 + c_f \cdot \frac{\bar{l_f}}{l_f}) \Big), \quad (c_f > 0) \tag{4.1}$$

where w_f is the weight of field f, $\bar{l_f}$ is the average length of field f in the collection, l_f is the length of field f in a particular document, and c_f is a hyper-parameter for each field f. Note that *normalisation 2* is a special case of *normalisation 2F*, when the entire document is considered as one field, with weight 1.

After defining *normalisation 2F*, the DFR weighting model PL2 (Table 2.1 on page 19) can be extended to PL2F by replacing tfn from Equation (4.1) in the following formula:

$$w_{d,q} = \sum_{t \in q} \frac{qtfn}{tfn + 1} \Big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn) \Big)$$

where $w_{d,q}$ corresponds to the relevance score of document d for query q, $\lambda = \frac{F}{N}$ is the mean and variance of a Poisson distribution, F is the total term frequency in the collection, and N is the number of documents in the collection. In addition, $qtfn$ is the normalised query term frequency, given by $qtfn = \frac{qtf}{qtf_{max}}$, where qtf is the query term frequency, and qtf_{max} is the maximum query term frequency among the query terms.

The weighting models PB2 and I(n_e)C2 (Table 2.1 on page 19) are extended in a similar way by replacing tfn from Equation (4.1). The extended models are denoted by PB2F, and I(n_e)C2F.

The weighting model DLH is extended by replacing the frequency tf of a term t with the weighted sum of the frequencies tf_f of t in each field f:

$$tf = \sum_f w_f \cdot tf_f \tag{4.2}$$

and it is denoted by DLHF.

Zaragoza et al. (2004) proposed BM25F, an extension of BM25 with per-field normalisation. The formula of BM25F is given below:

$$w_{d,q} = \sum_{t \in q} \frac{tfn}{k_1 + tfn} \cdot \log \frac{N - n + 0.5}{n + 0.5} \quad \text{where } tfn = \sum_f w_f \cdot \frac{tf_f}{(1 + b_f(l_f/\bar{l_f})} \tag{4.3}$$

where: b_f is a field-dependent normalisation parameter, similar to the parameter b of BM25 (Equation (2.6) on page 11); k_1 is a parameter that controls the saturation of tfn, similar to the parameter k_1 of BM25; $\bar{l_f}$ is the average length of the field f in the document collection; and l_f is the length of f in a particular document. The parameter w_f is the weight of the field f.

In Equation (4.3), the frequency of a term in the query qtf is ignored. In order to make the comparison of the employed weighting models more fair, the following formula is used for the conducted evaluation of BM25F, where the original query term frequency component from BM25 (Equation (2.6) on page 11) is added:

$$w_{d,q} = \sum_{t \in q} \frac{tfn}{k_1 + tfn} \cdot \frac{(k_3+1)\,qtf}{k_3 + qtf} \cdot \log \frac{N - n + 0.5}{n + 0.5} \quad \text{where } tfn = \sum_f w_f \cdot \frac{tf_f}{(1 + b_f(l_f/\bar{l_f})} \tag{4.4}$$

The value of k_3 is set to 1000, in the same way as described in Section 4.3. This value of k_3 essentially means that $\frac{(k_3+1)qtf}{k_3+qtf} \approx qtf$.

There are three different fields considered in this thesis: the body of Web documents; the anchor text of incoming hyperlinks; and the title of Web documents. The body field corresponds to the text within the HTML tags `<BODY>` and `</BODY>`. It includes the headings, but not the title of Web documents. The next sections discuss the setting of the hyper-parameters for the field-based models (Section 4.4.2), present the evaluation results (Section 4.4.3), and provide a discussion and conclusions (Section 4.4.4).

4.4.2 Parameter setting for field-based weighting models

This section focuses on the parameter setting of the models for combining different Web document fields, which were presented in Section 4.4.1. The per-field normalisation of term frequencies, and the weighting of the fields result in the introduction of an additional number of hyper-parameters in the weighting models. For example, in the case of BM25, there are two hyper-parameters, if only b and k_1 are considered, while there is only one hyper-parameter in the DFR models PL2, PB2 and I(n_e)C2. The weighting model DLH does not have any hyper-parameter. If the body, anchor text and title fields are considered, then the weighting model BM25F has seven parameters: the parameter k_1; the parameters b_b, b_a, b_t; and the weights w_b, w_a, w_t for each of the body, anchor text, and title fields, respectively. The weighting models PL2F, PB2F and I(n_e)C2F have six hyper-parameters: the parameters c_b, c_a, c_t and the weights w_b, w_a, w_t

for each of the body, anchor text and title fields, respectively. The weighting model DLHF has only 3 hyper-parameters w_b, w_a, w_t related to the weights of the fields.

The values for these parameters are set experimentally, as suggested by Zaragoza et al. (2004). For the case of BM25F, a two-dimensional optimisation for the parameters b_f and k_1 is performed for each field f. The weight of the field f is set equal to 1, and the weights of the other fields are set equal to zero. With this first step, the optimised value for b_f is set. Next, the weights of the fields are set equal to 1, and a one-dimensional optimisation for the parameter k_1 is performed, using the already optimised values for b_f. The third and last step of the optimisation process involves setting the weights w_f of the fields. The weight of the body w_b is set equal to 1, and a two-dimensional optimisation is performed in order to set the weights w_a and w_t. During the optimisation of the weights w_a and w_t, the value of k_1 is adjusted by taking into account the difference in the average term frequencies due to the field weights (Robertson et al., 2004):

$$k_1 := k_1 \cdot \frac{\text{weighted average term frequency}}{\text{unweighted average term frequency}} \tag{4.5}$$

The hyper-parameters of the weighting models PL2F, PB2F and I(n_e)C2F are set in a similar way. First, the parameter c_f is set for each field separately, by setting the weight of f equal to 1 and the weights of the other fields equal to 0. Next, the weight of the body w_b is set equal 1, and a two-dimensional optimisation is performed in order to set the weights w_a and w_t for the anchor text and the title fields, respectively. Regarding the weighting model DLHF, the weight of the body w_b is set equal to 1 and there is only a two-dimensional optimisation to set w_a and w_t.

Overall, setting the hyper-parameters for BM25F involves 4 two-dimensional optimisations for the parameters b_f and the field weights w_f, and 1 one-dimensional optimisation for the parameter k_1. In the case of the DFR models that employ *normalisation 2F*, it is necessary to perform 3 one-dimensional optimisations for the term frequency normalisation parameters c_f and 1 two-dimensional optimisation for the weights w_f. Furthermore, the weighting model DLHF requires only one two-dimensional optimisation. Therefore, optimising the DFR weighting models is less computationally demanding than optimising BM25F, because of the lower number of two-dimensional optimisations. All the optimisations have been performed with the same two-step process described in Section 4.3.2. Table A.2 on page 227 displays the parameter values

for the weighting models PL2F, PB2F and I(n_e)C2F. Tables A.3 on page 228 and A.4 on page 228 display the parameter values for the weighting models DLHF and BM25F, respectively.

4.4.3 Evaluation of field-based weighting models

This section presents the evaluation results for the field-based document models. Table 4.7 presents the evaluation of the weighting models PL2F, PB2F, I(n_e)C2F, DLHF and BM25F, where documents with the body, anchor text and title fields are considered. The bold entries correspond to the weighting model that achieved the highest MAP for a particular task. The Tables A.5, A.6, and A.7 in the Appendix A contain the precision at 10, the mean reciprocal rank of the first retrieved relevant document, and the number of retrieved relevant documents for the evaluated retrieval approaches.

		Mean Average Precision				
Row	Task	PL2F	PB2F	I(n_e)C2F	DLHF	BM25F
1	tr2000	0.2047	0.1927	0.2066	0.1699	**0.2097**
2	tr2001	0.2144	0.2083	0.2199	0.1833	**0.2231**
3	td2002	**0.2155**	0.2115	0.2020	0.1764	0.2133
4	td2003	0.1745	0.1650	0.1577	0.1571	**0.1876**
5	td2004	0.1483	0.1316	0.1400	0.1343	**0.1497**
6	hp2001	0.6450	0.6252	0.6787	0.5534	**0.6874**
7	hp2003	0.7281	0.6743	0.7201	0.6244	**0.7446**
8	hp2004	0.6559	0.5889	0.6519	0.5770	**0.6731**
9	np2002	0.7174	0.6888	**0.7302**	0.5829	0.7277
10	np2003	**0.7657**	0.7199	0.7068	0.5963	0.7138
11	np2004	**0.7437**	0.7189	0.7048	0.5354	0.7163

Table 4.7: Evaluation of the weighting models PL2F, PB2F, I(n_e)C2F, DLHF and BM25F.

The weighting models PL2F, I(n_e)C2F and BM25F perform well for the tasks tr2000, tr2001, and td2002 (rows 1-3 in Table 4.7). With respect to the topic distillation tasks td2003 and td2004, the weighting model BM25F outperforms the other four weighting models (rows 4-5 in Table 4.7). For the same tasks, PL2F is the most effective weighting model among the DFR weighting models. For the home page finding topic sets, the most effective weighting model is BM25F (rows 6-8 in Table 4.7). Regarding the named page finding tasks, the best performing models are I(n_e)C2F for np2002 (row 9 in Table 4.7), and PL2F for the np2003 and np2004 tasks (rows 10-11 in Table 4.7). The weighting model DLHF is outperformed by the other four weighting models for both the home page finding and the named page finding tasks (rows 6-11

from Table 4.7). Overall, the results confirm that the evaluated weighting models are statistically independent, since they are based on different probabilistic models, and they result in different performance.

4.4.4 Discussion and conclusions

The combination of content retrieval with information from different fields of documents results in very good performance and improvements in retrieval effectiveness for Web specific search tasks, compared to the results obtained with retrieval from each document representation separately (Table 4.4). For the ad-hoc retrieval tasks, employing field-specific term frequency normalisation and weighting of the different fields result in small improvements of retrieval effectiveness. For example, full text retrieval with the weighting model PL2 for the task tr2001 resulted in MAP 0.2107 (row 2 in Table 4.4), while the field-based weighting model PL2F for the same task resulted in MAP 0.2144. The improvements from using field-based weighting models are greater for the home page finding and named page finding tasks. For example, retrieval from the anchor text document representation with PL2 for the task np2004 resulted in MAP 0.5498 (row 44 row in Table 4.4). The field-based weighting model PL2F for the same task np2004 resulted in MAP 0.7437 (row 11 in Table 4.7), which represents an improvement of 35% in MAP.

With respect to the performance of Web track runs in TREC2004 (Craswell & Hawking, 2004), PL2F achieves higher MAP for the task np2004 (0.7437 from row 11 in Table 4.7), than the most effective submitted run (0.7232 from row 11 in Table 4.6). However, there is still room for improvements regarding the home page finding tasks. The MAP of the field-based weighting model I(n_e)C2F for the task hp2004 is 0.6519 (row 8 in Table 4.7), while the most effective submitted run in the same task of TREC 2004 achieved 0.7351 (row 8 in Table 4.6).

Overall, per-field normalisation has been shown to be particularly effective. The evaluation results have shown that the most effective field-based weighting models are PL2F, I(n_e)C2F, and BM25F. A comparison of the weighting model BM25F and the DFR weighting models, which employ *normalisation 2F*, shows that none of the models outperforms the other ones for all the tested tasks consistently. A drawback of the per-field normalisation is the introduction of additional hyper-parameters in the weighting models. With respect to the introduced parameters for the term frequency

normalisation and the weighting of fields, the DFR weighting models with *normalisation 2F* have an advantage of fewer hyper-parameters, compared to BM25F.

4.5 Content retrieval with query-independent evidence

The previous sections have focused on employing query-dependent evidence in order to retrieve and rank documents. The assigned weight to the retrieved Web documents depends on the distribution of the query terms in the body, as well as in the title and the anchor text of incoming hyperlinks. In addition to the query-dependent evidence, and as discussed in Sections 3.3 and 3.4, the ranking of Web documents can be further enhanced by using other query-independent sources of evidence, such as the URL of Web documents (Section 4.5.1), or the analysis of the hyperlink structure of the Web (Section 4.5.2). Section 4.5.3 presents the evaluation results from combining field-based weighting models and the employed query-independent sources of evidence. Finally, Section 4.5.4 closes with a summary and some conclusions.

4.5.1 URLs of Web documents

In order to be able to locate and browse a certain Web document, it is necessary to have a way to uniquely identify it. This is achieved with the Uniform Resource Locators (URL) (Berners-Lee et al., 1994). The general syntax of a URL for an available resource on a network is `<scheme>:<scheme specific part>`, where `<scheme>` specifies the scheme, or the protocol to use for accessing the resource, and `<scheme specific part>` is specified by the particular protocol. For example, the URL for a Web document, which is called `news.html` and it is stored in the root directory / of the host `www.dcs.gla.ac.uk`, is `http://www.dcs.gla.ac.uk:80/news.html`. In this URL, `news.html` corresponds to the path of the URL, `www.dcs.gla.ac.uk` corresponds to the fully qualified name of the network server that hosts the Web page, and `http` corresponds to the HyperText Transfer Protocol (HTTP) that is used for requesting and transferring Web documents. In addition, the number 80 corresponds to the standard port that the HTTP server is listening to and it is usually not included in the URL.

In the context of Web Information Retrieval, the URL of a Web document can be used as a query-independent indication of the functionality of a Web document

within a group of related Web documents, which form a Web site. This is based on two observations. First, two common conventional filenames for home pages are `index.html` or `default.html`. Second, due to common practice in the organisation of Web documents in Web sites, the entry point or home page of a Web site is more likely to be in the root directory of a Web site.

Westerveld et al. (2001) and Kraaij et al. (2002) considered both observations, in order to compute the prior probability that a Web document with a certain type of URL is the home page of a Web site. They identified four types of URLs and found that the Web pages with a root URL, such as `http://domain/`, are highly likely to be home pages. Then, they used these prior probabilities in a language modelling approach for the home page finding task of the TREC 2001 Web track.

Tomlinson (2005) assigned a distinct term for each type of URLs. During indexing, the term that corresponded to the type of the document's URL was added to the index. Then, during retrieval, the *idf* of the terms, which corresponded to the types of URLs, were used for weighting the documents.

The second observation has been used in order to employ evidence from the length of a URL, or the length of parts of the URL. For example, Savoy & Rasolofo (2001) counted the number of '`/`' in a URL. Kamps et al. (2004*b*) also defined the URL length in terms of the number of characters, and the number of '.' in the domain name of the URL.

In this thesis, the query-independent evidence from the URL of Web documents is based on the length of the URL path (Plachouras & Ounis, 2004; Plachouras et al., 2003; Plachouras, He & Ounis, 2004). For example, the path length of the URL `http://www.dcs.gla.ac.uk/news.html` corresponds to the length in characters of the string `news.html`, which is 9 characters. This choice is justified by the fact that employing the fully qualified domain name may bias the resulting scores towards the Web sites that have been present for a longer period of time, and had an advantage in registering shorter domain names. However, the length of the domain name does not provide any indication about which Web document of the Web site corresponds to the home page.

The combination of the URL path length with query-dependent evidence requires the URL path length to be transformed into an appropriate score. More specifically, the URL length score should be lower for the Web documents with a longer URL path, and it should be higher for the Web documents with short URLs, that are more likely

to correspond to home pages. An appropriate transformation is given by the following formula (Zaragoza et al., 2004):

$$\mathrm{URL}(d) = \frac{k_u}{k_u + URLpathlen(d)} \tag{4.6}$$

where URL(d) is the URL-related score assigned to document d, $URLpathlen(d)$ corresponds to the length in characters of the URL path of document d and k_u is a parameter, which controls the saturation of URL(d) with respect to the URL path length. When the parameter k_u takes small values with respect to the length of the URL path of documents, the documents with short URL paths are more favoured. For the higher values of k_u, the effect of the URL path length is smoothed and the resulting score is less biased towards the documents with shorter URL paths. This is shown in Figure 4.3 for three different values of k_u.

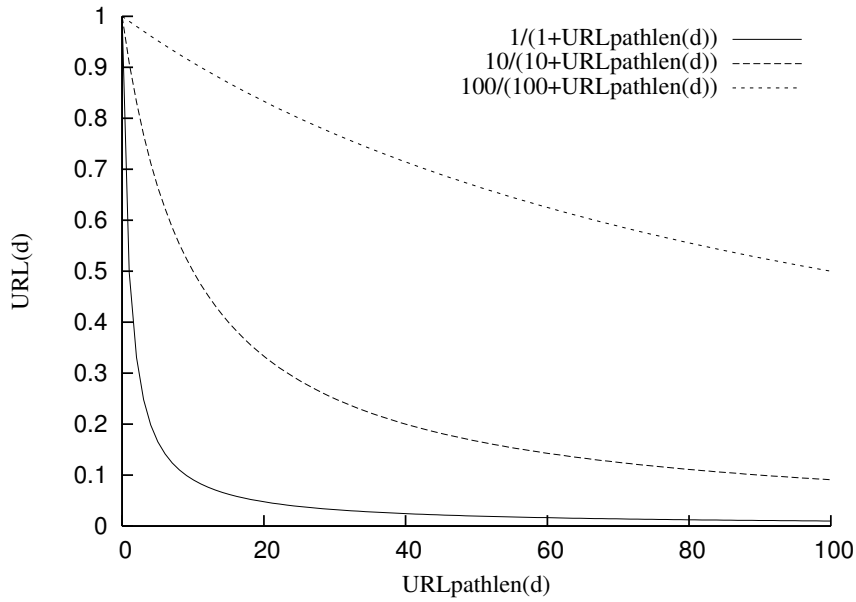

Figure 4.3: The monotonically decreasing transformation for the URL path length, for $k_u = 1, 10$ and 100.

The URL-related score URL(d) for the document d is linearly combined with the corresponding content analysis score as follows:

$$w_{d,q,URL} = w_{d,q} + \omega_u \mathrm{URL}(d) \tag{4.7}$$

where $w_{d,q}$ is the content analysis score assigned to the document d, using any of the retrieval approaches described in Sections 4.3 and 4.4, $w_{d,q,URL}$ is the combined score for document d, and ω_u is the weight of the URL-related score in the linear combination.

Plachouras & Ounis (2004) have also experimented with multiplying the content analysis scores $w_{d,q}$ with the URL-related score as follows:

$$w_{d,q,URL} = w_{d,q} \cdot \frac{1}{\log_2(1 + URLpathlen(d))} \tag{4.8}$$

This approach has been particularly effective (Craswell & Hawking, 2004), but it has to be applied for the top ranked documents only, because it alters the score distribution significantly. On the other hand, the linear combination used in Equation (4.7) is more robust, because it does not alter significantly the original distribution of scores. Hence, it can be applied for all retrieved documents.

4.5.2 Hyperlink structure analysis

This section focuses on the effectiveness of combining content analysis with query-independent evidence from the analysis of the hyperlink structure. The hyperlinks that exist between Web documents can be considered as an indication that the author of the source Web document believes the destination Web document is related to the source one, or it is worth viewing. When a particular Web document has a significant number of incoming hyperlinks from other Web documents, this suggests that it is either a popular Web document, or it is an authoritative document.

In order to compute a popularity, or authority score for Web documents, in a query-independent way, the Web graph can be modelled as a Markov chain. The probability of entering a particular state in the Markov chain stands for the popularity or the authority-based score of Web documents. For example, the PageRank scores of Web documents correspond to the probability of visiting the state that represents the Web document, in a Markov chain for the whole Web (Page et al., 1998). However, the hyperlink structure of the Web is not necessarily appropriate in order to define a Markov chain. For this reason, PageRank introduces a transformation with which any Web document, even the ones without any incoming hyperlinks, can be visited with a finite probability. This transformation corresponds to the rank source E, which was described in Section 3.3.2.1.

This section also introduces the Absorbing Model, a novel hyperlink structure analysis model, which employs a different transformation of the Web graph in order to define a Markov chain. Instead of adding a small but finite probability to the probability of visiting any state in the Markov chain, the Absorbing Model introduces the *clones*, a set of virtual states that have a one-to-one correspondence with the states of the original Web documents.

The remainder of the section is organised as follows. Sections 4.5.2.1 and 4.5.2.2 present the basic definitions for Markov chains. Section 4.5.2.3 discusses the transformations of the Web graph that are required to define a Markov chain. The Absorbing Model and its instantiation with static priors are introduced in Sections 4.5.2.4 and 4.5.2.5, respectively. The introduced notation and terminology for Markov chains are similar to that used by Feller (1957). Finally, the combination of PageRank and the Absorbing Model with the field-based weighting models is discussed in Section 4.5.2.6.

4.5.2.1 Markov chains

Each document is considered as a possible outcome of the retrieval process. Therefore, the documents are orthogonal, or alternative states d_k, which have a prior probability p_k to be retrieved. The prior probability p_k is defined by the system. Each pair of documents (d_i, d_j) has an associated transition probability $p_{ij} = p(d_j|d_i)$ of reaching the document d_j from the document d_i. This conditional probability $p(d_j|d_i)$ can be also interpreted as the probability of having the document d_j as outcome, when the document d_i is the evidence.

Both prior and transition probabilities must satisfy the conditions of a probability space, which are:

$$\sum_k p_k = 1 \tag{4.9}$$

$$\sum_j p_{ij} = 1 \tag{4.10}$$

Condition (4.10) imposes that each state d_i must have access to at least one state d_j for some j, where it is possible that $i = j$.

In order to obtain a more compact representation of probabilities for arbitrary sequences of states, it is useful to express the prior probabilities as a row vector P and

the transition probabilities as a row-by-column matrix M, as follows:

$$P = [\ p_k\ ] \tag{4.11}$$

$$M = [\ p_{ij}\ ] \tag{4.12}$$

Then, let M^n be the matrix product rows-into-columns of M with itself n-times:

$$M^n = [\ p_{ij}^n\ ] \tag{4.13}$$

In a first order Markov chain, the probability of any walk from a state d_i to a state d_j depends only on the probability of the last visited state. In other words, the probability of any sequence of states $(d_1, \ldots, d_n)$ is given by the relation:

$$p(d_1, \ldots, d_n) = p_1 \prod_{i=1}^{n-1} p(d_{i+1}|d_i) \tag{4.14}$$

where p_1 is the prior probability of document d_1. It is possible to define Markov chains of higher order, where the probability of a walk depends on more of the visited states than just the last one. In this thesis, only first-order chains are considered for the purpose of hyperlink structure analysis.

In terms of matrices, the element p_{ij}^n of the product M^n corresponds to the probability $p(d_i, \ldots, d_j)$ of reaching the state d_j from d_i by any random walk, or sequence of states $(d_i, \ldots, d_j)$ made up of exactly n states.

If $p_{ij}^n > 0$ for some n, then the state d_j is *reachable* from the state d_i. A set of states $C = \{d_i\}$ is said to be *closed* if any state inside C can reach all and only all other states inside C. The states in a closed set are called *persistent* or *recurrent* states, since a random walk, starting from the state d_i and terminating at state d_j, can be ever extended to pass through d_i again. Indeed, from the definition of the closed set, the probability $p_{ji}^m > 0$ for some m. If a single state forms a closed set, then it is called *absorbing*, since a random walk that reaches this state cannot visit any other states. A state, which is not in any closed set, is called *transient* and it must reach at least one state in a closed set. Thus, there is a random walk, starting from the transient state d_i, that cannot be ever extended to pass through d_i again.

A useful property of Markov chains is the decomposition characterisation. It can be shown that all Markov chains can be decomposed in a unique manner into non-overlapping closed sets $C_1, C_2, \ldots, C_n$ and a set T that contains all and only all the transient states of the Markov chain (Feller, 1957). If this decomposition results in a single closed set C, then the Markov chain is called *irreducible*.

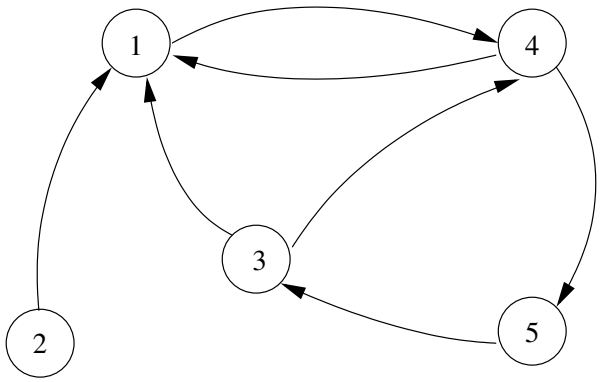

Figure 4.4: The Markov Chain representing the Web graph.

Example 2 Figure 4.4 provides an illustration of the different types of states in a Markov chain. The directed graph may be seen as a Markov Chain consisting of the states $1, 2, 3, 4$ and 5. The arcs represent the possible transitions between the states in the Markov chain. According to the terminology given above for Markov chains, states $1, 3, 4, 5$ form a closed set and they are persistent states. State 2 is a transient state. Therefore, this Markov chain is irreducible, as it can be decomposed in a non-empty set of transient states and a single set of persistent states. If the arc from state 5 to state 3 is replaced by an arc from 5 to itself, then state 5 becomes an absorbing state. □

4.5.2.2 Classification of states

According to Equation (4.14), the probability of reaching the state d_j from any initial state by any random walk $w = (d_i, \ldots, d_j)$ is given below:

$$\sum_i \sum_w p(d_i, \ldots, d_j) = \sum_i \sum_{n=1}^{\infty} p_i p_{ij}^n = \sum_i p_i \left(\sum_{n=1}^{\infty} p_{ij}^n \right) \tag{4.15}$$

However, in a Markov chain, the limit $\lim_{n\to\infty}\sum_n p_{ij}^n$ does not always exist, or it can be infinite. The limit does not exist when there is a state d_i such that $p_{ii}^n = 0$ unless n is a multiple of some fixed integer $t > 1$. In this case, the state d_i is called *periodic*. Periodic states are easily handled: if t is the largest integer which makes the state d_i periodic, then it is sufficient to use the probabilities p_{kj}^t as new transition probabilities p'_{kj}. With the new transition probabilities, p'^n_{ii} will be greater than 0 and the periodic states d_j will become aperiodic. Hence, it may be assumed that all states in a Markov chain are aperiodic (Feller, 1957).

Recurrent states in a finite Markov chain have the limit of p_{ij}^n greater than 0 if the state d_j is reachable from d_i, while for all transient states this limit is 0:

$$\lim_{n\to\infty} p_{ij}^n = 0 \text{ if } d_j \text{ is transient} \tag{4.16}$$

$$\lim_{n\to\infty} p_{ij}^n > 0 \text{ if } d_j \text{ is persistent and } d_j \text{ is reachable from } d_i \tag{4.17}$$

In an irreducible finite Markov chain, all nodes are persistent and the probability of reaching them from an arbitrary node of the graph is positive. In other words, $\lim_{n\to\infty} p_{ij}^n > 0$ and $\lim_{n\to\infty} p_{ij}^n = \lim_{n\to\infty} p_{kj}^n = u_j$ for all i and k. Due to this property, an irreducible Markov chain possesses an invariant distribution, that is a distribution u_k such that:

$$u_j = \sum_i u_i p_{ij} \qquad \text{and} \qquad u_j = \lim_{n\to\infty} p_{ij}^n \tag{4.18}$$

In the case of irreducible Markov chains, the vector P of prior probabilities does not affect the unconditional probability of entering an arbitrary state, since all rows are identical in the limit matrix of M^n. Indeed:

$$\lim_{n\to\infty}\sum_i p_i p_{ij}^n = \lim_{n\to\infty}\sum_i p_i p_{kj}^n = \lim_{n\to\infty} {p_{kj}}^n \sum_i p_i = u_j\left(\sum_i p_i\right) = u_j \tag{4.19}$$

Because of this property, the probability distribution u_j in a irreducible Markov chain is called *invariant* or *stationary* distribution.

If the distribution $\lim_{n\to\infty}\sum_i p_i p_{ij}^n$ is taken to assign weights to the nodes, then it is equivalent to the invariant distribution u_j in the case of an irreducible Markov chain. More generally, if the Markov chain is not irreducible or does not possess an invariant distribution, then $\lim_{n\to\infty}\sum_i p_i p_{ij}^n$ can be still used to define the distribution of the node weights. However, it will depend on the prior distribution p_i.

4.5.2.3 Modelling the hyperlinks of the Web

Markov chains can be applied to model the hyperlinks between documents on the Web. Let R be the binary accessibility relation between the set of documents. More specifically, it is $R(d_i, d_j) = 1$ if there is a hyperlink from document d_i to document d_j, and 0 otherwise.

Let $o(i)$ be the number of documents d_j which are accessible from d_i:

$$o(i) = |\{d_j : R(i,j) = 1\}| \tag{4.20}$$

This is equal to the outdegree of a Web page. The probability p_{ij} of a transition from document d_i to document d_j is defined as follows:

$$p_{ij} = \frac{R(i,j)}{o(i)} \tag{4.21}$$

The above definition of p_{ij} assumes that there is an equal probability to make a transition from document d_i to any of the documents pointed to by d_i, irrespectively of their content, or the type of the hyperlink.

There are two main implications from using the transition probabilities defined in Equation (4.21) in order to model the Web graph as a Markov chain. First, there are Web documents that do not contain any hyperlinks to other documents. Such documents can be plain text files that do not contain any HTML markup. In this case, the Equation (4.10) is not satisfied and the transition probabilities defined in Equation (4.21) cannot be used in order to define a Markov chain from the Web graph. Even if all the Web documents have hyperlinks to other Web documents and the Equation (4.10) is satisfied, all the transient states in the resulting Markov chain would have $\lim_{n\to\infty} p_{ij}^n = 0$, independently from the number of their incoming hyperlinks. Therefore this limit cannot be used as a score, since only persistent states would have a significant probability of being visited during a random walk.

There are two possible ways to overcome the above two implications. First, all the states can be linked by assigning a new probability $p_{ij}^* \neq 0$ in a suitable way, such that $|p_{ij}^* - p_{ij}| < \epsilon$. In this way all states become persistent. In other words the Web graph is transformed into a single irreducible closed set, namely the set of all states. Therefore, all states receive a positive probability that they will be visited in a random walk in the Markov chain. This approach is used in PageRank, where the assumed random

surfer may randomly jump with a finite probability to any Web document. Second, the original graph G can be extended to a new graph G^*. The new states of the extended graph G^* are all and only all the persistent states of the graph G^*. The scores of all the states in the original graph, irrespectively of whether they are transient or persistent, will be uniquely associated to the scores of these persistent states in the new graph. The latter is the approach that is used to define the Absorbing Model.

4.5.2.4 The Absorbing Model

The Absorbing Model is based on a simple transformation of the Web graph. The original graph G is projected onto a new graph G^*, the decomposition of which is made up of a set of transient states $T = G$ and a set $\{C_i, \ldots, C_n\}$ of absorbing states, in other words a set of singular closed sets. The state C_i is called the *clone* of state d_i of the original graph G. Any state in G has direct access only to its corresponding clone, but not to other clones. Since the clones are absorbing states, they do not have direct access to any state except to themselves. The Absorbing Model is formally introduced as follows:

Definition 1 Let $G = (V, R)$ be the graph consisting of the set $V = \{d_i\}$ of N documents and the binary accessibility relation $R(d_i, d_j) = 1$ if there is a hyperlink from d_i to d_j and 0 otherwise. The graph G is extended by introducing N additional states $d_{N+i}, i = 1, \ldots, N$, called the *clone nodes*. These additional nodes are denoted as: $d_{N+i} = d_i^*$ and the accessibility relation R is extended in the following way:

$$\begin{aligned}
&R(d_i^*, d) = R(d, d_i^*) = 0, d \neq d_i^*, i = 1, \ldots, N \text{ except for:}\\
&R(d_i, d_i^*) = 1\\
&R(d_i^*, d_i^*) = 1
\end{aligned}$$

The transition probability p_{ij} from state d_i to state d_j is:

$$p_{ij} = \frac{R(d_i, d_j)}{|\{d_j : R(d_i, d_j) = 1\}|}$$

where the denominator stands for the number of the possible transitions from state d_i.

The following example illustrates the transformation of the graph according to the definition of the Absorbing Model.

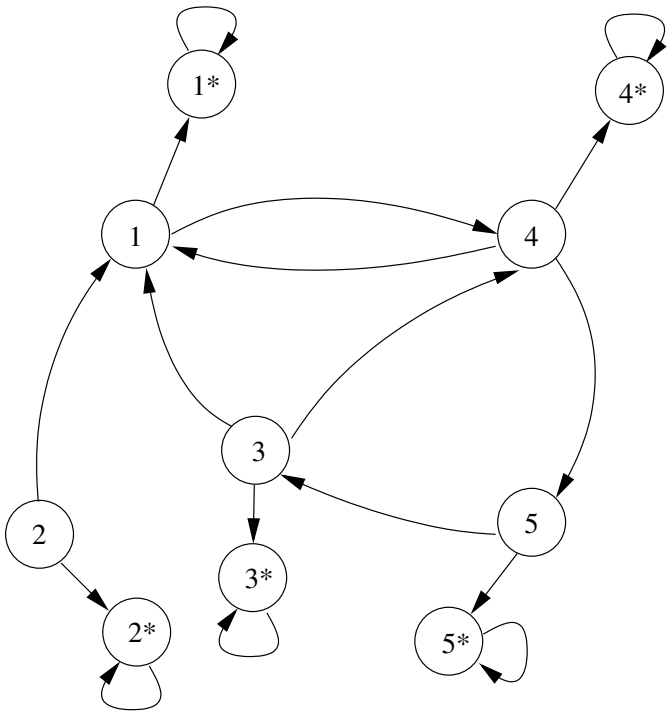

Figure 4.5: The extended Markov Chain including the clone states.

Example 3 Figure 4.4 shows a graph that represents a part of the Web. Figure 4.5 shows the same graph, transformed according to the definition of the Absorbing Model. In this case, the states 1 to 5 become transient and the only persistent states are the newly introduced states 1^* to 5^*. The introduced transformation results in removing any absorbing states from the original Web graph, as there are no closed sets consisting of any of the original states. □

With the introduction of the clone nodes, all the original states $d_j, j = 1, \ldots, N$ become transient, while all the clone states $d_j^*, j = 1, \ldots, N$ are the only persistent states. In other words, the probabilities of visiting a state in the original Markov chain become:

$$p_{jk}^n \to 0, \;\; k = 1, \ldots, N \tag{4.22}$$

while for the clone states it is:

$$p_{jk}^n \to u_{jk}, \;\; k = N+1, \ldots, 2N \tag{4.23}$$

where u_{jk} stands for the probability that a random walk starting from state d_j will pass through state d_k. The Absorbing Model score $s(d_k)$ of a state d_k is given by the unconditional probability of reaching its clone state d_k^*:

$$s(d_k) = \sum_j p_j u_{jk^*} \tag{4.24}$$

where $k^* = k + N$ and $k = 1, \ldots, N$.

Intuitively, the Absorbing Model score measures the probability of a user being "absorbed" by a Web document, while he is browsing other documents in its vicinity. This probability depends on both the incoming and the outgoing hyperlinks. If a document has many outgoing links, then its Absorbing Model score is low, while if it has few outgoing links, it is more probable that its Absorbing Model score will be higher. Additionally, documents with a significant number of incoming links, have a high Absorbing Model score, while documents without incoming links have a lower score. Therefore, the higher values of the Absorbing Model score can be considered as evidence of authority for documents.

The Absorbing Model has two main qualitative differences from PageRank. First, while in PageRank the scores depend mainly on the quality of the incoming links of a document, in the Absorbing Model the document's score is affected by its outgoing links. The second difference is that PageRank scores correspond to the stationary probability distribution of the Markov chain resulting from the Web graph after adding a link between every pair of documents. On the other hand, the Absorbing Model does not possess a stationary distribution, and therefore, the Absorbing Model scores depend on the prior probabilities of the documents. Depending on the way the prior probabilities are defined, different extensions to the model maybe introduced. For example, the use of the content retrieval scores as the prior probabilities results in a simple and principled way to dynamically combine content and link analysis (Amati et al., 2003), called the Dynamic Absorbing Model.

On the other hand, if the prior probabilities are independent of the content retrieval, the Static Absorbing Model can be defined, as it will be seen in the next section, the Absorbing Model scores can be computed offline, similarly to the case of PageRank. This flexibility of the Absorbing Model enables its application in either a query-dependent, or a query-independent way.

4.5.2.5 Definition of the Static Absorbing Model

From the possible ways to define the prior probabilities independently of the queries, such as the document's length, or its URL type, one option is to assume that they are uniformly distributed. This approach reflects the concept that all the documents have an equal chance of being retrieved, without taking into account any of their specific characteristics. As a consequence, the prior probabilities are defined as follows:

Definition 2 (Static mode priors) The prior probability that a document d_k is retrieved is uniformly distributed over all the documents:

$$p_k = \frac{1}{N} \; k = 1, \ldots, N \tag{4.25}$$

where the number N refers to the total number of states in the original graph, that is the total number of documents. The prior probability for the clone nodes is set equal to zero.

When the static mode priors are employed, the Absorbing Model score $s(d_j)$ of a document d_j is given from Equations (4.24) and (4.25) as follows:

$$s(d_j) = \sum_i p_i u_{ij^*} = \sum_i \frac{1}{N} u_{ij^*} \propto \sum_i u_{ij^*} \tag{4.26}$$

In other words, the Absorbing Model score $s(d_j)$ for a document d_j is the probability of accessing its clone node d_j^* by performing a random walk, starting from any state with equal probability. The interpretation of this score is derived in a straight-forward manner from the intuitive description of the Absorbing Model in Section 4.5.2.4: a document has a high Absorbing Model score if there are many paths leading to it. As a result, a random user would be *absorbed* by the document, while browsing the documents in its vicinity.

4.5.2.6 Combination of field retrieval with PageRank or the Absorbing Model

It is necessary to combine the hyperlink analysis with the content analysis of Web documents, similarly to the case of using evidence from the URLs of Web documents in Section 4.5.1. In the case of combining the scores, a transformation of the hyperlink structure analysis scores is required, because the content and hyperlink structure

analysis scores follow different distributions. Indeed, Manmatha et al. (2001) modelled the content analysis score distribution of the retrieved documents as a mixture of two distributions: a Gaussian distribution for the scores of the relevant documents, and a exponential distribution for the scores of the non-relevant documents. On the other hand, Pandurangan et al. (2002) suggested that the values of PageRank follow a power law. Therefore, there are only few Web documents with a high PageRank score, while most of the Web documents have a low score.

Plachouras et al. (2005) experimented with a Cobb-Douglas utility function, where the content and hyperlink analysis scores are multiplied:

$$w_{d,q,L} = w_{d,q} \cdot LS(d) \tag{4.27}$$

In order to address the difference in the score distributions, they transformed the hyperlink analysis scores in the following way:

$$w_{d,q,L} = w_{d,q} \cdot \log_2(shift \cdot LS(d)) \tag{4.28}$$

where $shift$ is a parameter and LS corresponds to the score computed by a hyperlink structure analysis method, such as PageRank (PR), or the Static Absorbing Model (SAM). The transformation resulted in better retrieval effectiveness compared to Equation (4.27). However, a limitation of this approach is that multiplying the content analysis scores and the transformed hyperlink analysis scores greatly changes the document ranking and boosts non-relevant documents to the top ranks of the results.

Craswell, Robertson, Zaragoza & Taylor (2005) proposed that the scores computed by the hyperlink structure analysis methods, are transformed with a saturating function of the following form:

$$\mathrm{L}(d) = \frac{LS(d)}{k_L + LS(d)} \tag{4.29}$$

where k_L is the saturating parameter. The effect of the saturating parameter k_L in the transformation is shown in Figure 4.6. For the low values of k_L, the score $\mathrm{L}(d)$ is effectively inversely proportional to the hyperlink analysis score $LS(d)$. For the higher values of k_L, the relation between the score $\mathrm{L}(d)$ and the hyperlink analysis score $LS(d)$ is almost linear. Differently from the URL-based scores, where the URL-based score is a monotonically decreasing function of the URL path length, the applied transformation to the hyperlink structure scores is a monotonically increasing function.

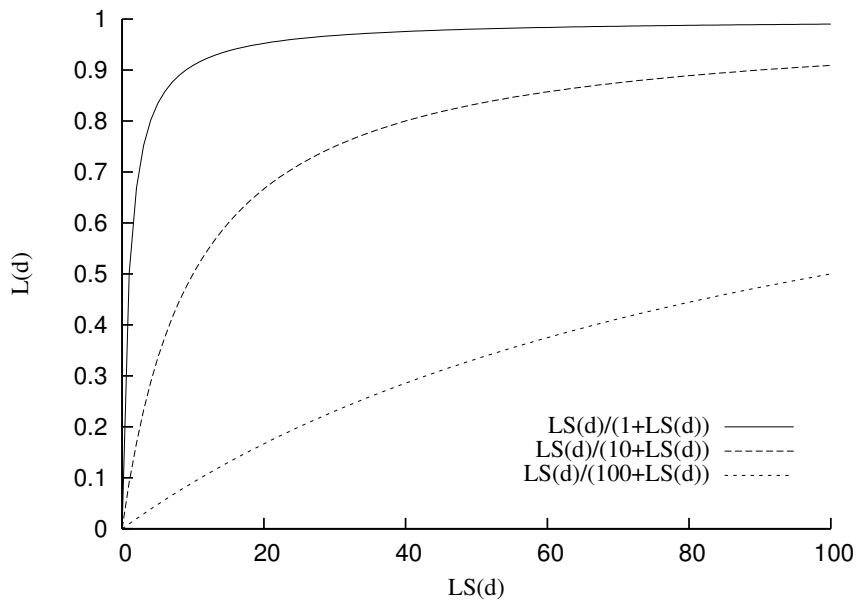

Figure 4.6: The monotonically increasing transformation of the hyperlink structure analysis scores, for $k_L = 1, 10$ and 100.

Similarly to Section 4.5.1, the hyperlink analysis score is linearly combined with the content analysis score, as follows:

$$w_{d,q,L} = w_{d,q} + \omega_L \mathrm{L}(d) \tag{4.30}$$

where ω_L is the weight of the hyperlink structure analysis score L.

4.5.3 Evaluation of field retrieval with query-independent evidence

The current section evaluates the combination of field retrieval with three different query-independent sources of evidence. The first one is the URL path length, which is transformed to a score according to Equation (4.6). The second one is PageRank, where the damping factor is $prdf = 0.85$ (Brin & Page, 1998). The third source of query-independent evidence is the Absorbing Model with static priors, a novel hyperlink structure analysis algorithm described in Sections 4.5.2.4 and 4.5.2.5. The evaluation is performed for combinations of each field-based weighting model with one source of query-independent evidence, in order not to further increase the number of hyper-parameters for each retrieval approach.

Table A.8 in Appendix A presents the values of the parameters ω_u, k_u, ω_{pr}, k_{pr}, ω_{am} and k_{am} for the combination of the URL path length, PageRank and the Absorbing Model with the field-weighting models PL2F, PB2F, I(n_e)C2F, DLHF and BM25F, respectively. The parameter values are set in order to optimise MAP for each task. The setting of the parameters is based on a two-dimensional optimisation of the pairs of ω and k for each source of query-independent evidence. The optimisation is based on the same techniques that have been used to set the parameters of the weighting models, as described in Section 4.3.2.

The combination of a field-based weighting model with one of the URL path length, PageRank, or the Absorbing Model, is denoted by appending the letter U, P, A, respectively, to the name of the weighting model. For example BM25FU denotes the combination of the field-based weighting model BM25F with the URL path length, and PL2FA denotes the combination of the field-based weighting model PL2F with the Absorbing Model. The field-based weighting models employ the body, anchor text, and title fields of Web documents.

Table 4.8 contains the evaluation results of combining the weighting models PL2F, PB2F, I(n_e)C2, DLHF and BM25F with the evidence from the URL of documents (rows 12-22), PageRank (rows 23-33), and the Absorbing Model (rows 34-44). The entries in bold show the most effective combination of a weighting model with a query-independent source of evidence for a particular topic set. The baselines correspond to the field-based weighting models, which do not employ query-independent evidence. Their evaluation results are copied from Table 4.7 in the rows 1-11 of Table 4.8. Tables A.5, A.6, and A.7 in the Appendix A contain the precision at 10, the mean reciprocal rank of the first retrieved relevant document, and the number of retrieved relevant documents, respectively, for the evaluated retrieval approaches.

Row	Task	Mean Average Precision				
		PL2F	PB2F	I(n_e)C2F	DLHF	BM25F
1	tr2000	0.2047	0.1927	0.2066	0.1699	**0.2097**
2	tr2001	0.2144	0.2083	0.2199	0.1833	**0.2231**
3	td2002	**0.2155**	0.2115	0.2020	0.1764	0.2133
4	td2003	0.1745	0.1650	0.1577	0.1571	**0.1876**
5	td2004	0.1483	0.1316	0.1400	0.1343	**0.1497**
6	hp2001	0.6450	0.6252	0.6787	0.5534	**0.6874**

continued on next page

continued from previous page						
Row	Task	Mean Average Precision				
		PL2F	PB2F	I(n_e)C2F	DLHF	BM25F
7	hp2003	0.7281	0.6743	0.7201	0.6244	**0.7446**
8	hp2004	0.6559	0.5889	0.6519	0.5770	**0.6731**
9	np2002	0.7174	0.6888	**0.7302**	0.5829	0.7277
10	np2003	**0.7657**	0.7199	0.7068	0.5963	0.7138
11	np2004	**0.7437**	0.7189	0.7048	0.5354	0.7163
		PL2FU	PB2FU	I(n_e)C2FU	DLHFU	BM25FU
12	tr2000	0.2047	0.1927	0.2076	0.1705	**0.2122**
13	tr2001	0.2144	0.2083	0.2197	0.1848	**0.2231**
14	td2002	**0.2157**	0.2119	0.2020	0.1767	0.2133
15	td2003	0.2174	0.2036	0.2087	0.1793	**0.2338**
16	td2004	0.1869	0.1735	**0.2011**	0.1961	0.1981
17	hp2001	0.7946	0.7501	0.8148	0.7151	**0.8187**
18	hp2003	0.7803	0.7330	0.7958	0.7070	**0.8190**
19	hp2004	0.7032	0.6483	**0.7141**	0.6438	0.7100
20	np2002	0.7174	0.6904	**0.7302**	0.5829	0.7279
21	np2003	**0.7657**	0.7201	0.7068	0.5986	0.7138
22	np2004	**0.7458**	0.7331	0.7139	0.5380	0.7304
		PL2FP	PB2FP	I(n_e)C2FP	DLHFP	BM25FP
23	tr2000	0.2047	0.1927	0.2069	0.1704	**0.2102**
24	tr2001	0.2144	0.2083	0.2199	0.1868	**0.2231**
25	td2002	**0.2160**	0.2116	0.2006	0.1780	0.2138
26	td2003	0.1875	0.1700	0.1808	0.1642	**0.1966**
27	td2004	0.1525	0.1357	**0.1594**	0.1541	0.1549
28	hp2001	0.6452	0.6237	0.6839	0.5626	**0.6877**
29	hp2003	0.7403	0.7068	0.7746	0.7141	**0.8044**
30	hp2004	0.6763	0.6197	**0.7554**	0.6245	0.7461
31	np2002	0.7214	0.6901	**0.7400**	0.5865	0.7355
32	np2003	**0.7976**	0.7430	0.7629	0.6301	0.7916
33	np2004	**0.7552**	0.7284	0.7263	0.5365	0.7373
		PL2FA	PB2FA	I(n_e)C2FA	DLHFA	BM25FA
34	tr2000	0.1998	0.1927	0.2066	0.1736	**0.2096**
35	tr2001	0.2186	0.2115	0.2225	0.1848	**0.2231**
36	td2002	**0.2155**	0.2116	0.2022	0.1767	0.2137
37	td2003	0.1804	0.1660	0.1668	0.1573	**0.1871**
38	td2004	0.1506	0.1326	0.1454	0.1367	**0.1508**
39	hp2001	0.6435	0.6253	0.6827	0.5480	**0.6872**
40	hp2003	0.7363	0.7006	0.7461	0.6705	**0.7791**
41	hp2004	0.6681	0.5965	0.6972	0.5882	**0.7015**
42	np2002	0.7206	0.6897	**0.7348**	0.5830	0.7309
43	np2003	**0.7809**	0.7273	0.7159	0.6068	0.7522
44	np2004	**0.7612**	0.7310	0.7215	0.5354	0.7496

Table 4.8: Evaluation results of the combinations of field-based retrieval with the query-independent evidence from the URL path length, PageRank, and the Absorbing Model. The evaluation results of the baselines, which correspond to the field-based weighting models PL2F, PB2F, I(n_e)C2F, DLHF, and BM25F, are copied from Table 4.7.

Regarding the different sources of query-independent evidence, when the relevant

documents are restricted to be the home pages of relevant Web sites for the topic distillation tasks td2003 and td2004, the evidence from the URL of the Web documents (rows 15-16) results in important improvements over employing only field retrieval (rows 4-5), or a combination with PageRank (rows 26-27) or the Absorbing Model (rows 37-38). For example, the MAP achieved by BM25FU for td2003 is 0.2338 (row 15), while the MAP achieved by BM25FP for the same task is 0.1966 (row 26).

Regarding the home page finding tasks, using evidence from the URLs of Web documents (rows 17-19) is more effective than employing only field-based retrieval (rows 6-8), or its combinations with the Absorbing Model (rows 39-41). The combination of the field-based weighting models PL2F and PB2F with the URL path length is more effective than their combination with PageRank. However, when the weighting models I(n_e)2F, DLHF, and BM25F are employed, the differences are less marked.

For the named page finding tasks, there is no particular restriction to the type of the relevant documents. In this case, the combination of the most effective field-based weighting models (PL2F, I(n_e)C2F, and BM25F) with PageRank (rows 31-33), or with the Absorbing model (rows 42-44), outperforms the corresponding combination with evidence from the URL of Web documents (rows 20-22). Both PageRank and the Absorbing model result in comparable retrieval effectiveness for the named page finding topic sets. For example, the MAP of PL2FP for the tasks np2003 and np2004 is 0.7976, and 0.7552, respectively (rows 32-33). The MAP of PL2FA for the same tasks is 0.7809, and 0.7612, respectively (rows 43-44).

The combination of query-independent evidence with field retrieval does not yield any important improvements in retrieval effectiveness for the ad-hoc search tasks. For example, the MAP achieved by the field-based weighting model PL2F for the task tr2001 is 0.2144, while the MAP of the combination of PL2F with the Absorbing Model is 0.2186 (Table 4.8). This is due to the fact that the query-independent evidence are not necessarily useful for identifying relevant documents in ad-hoc search tasks (Craswell & Hawking, 2002).

4.5.4 Summary and conclusions

The combination of the field-based weighting models with query-independent sources of evidence, performs as well, or better than the best official submitted runs to the corresponding TREC Web tracks. For example, the obtained MAP from BM25F for

the task td2003 is 0.2338 (row 15 in Table 4.8), while the MAP of the best performing run submitted to the TREC 2003 Web track is 0.1543 (row 4 in Table 4.6). In addition, PL2FA achieves MAP of 0.7612 for the task np2004 (row 44 in Table 4.8), while the best performing run submitted to the TREC 2004 Web track achieved 0.7232 (row 11 in Table 4.6).

Overall, this section has investigated the use of query-independent sources of evidence for Web IR. Three sources of evidence have been employed: the URL path length; PageRank; and the Absorbing Model, a novel hyperlink structure analysis algorithm. The evaluation results have shown that the employed query-independent sources of evidence can be used effectively in order to enhance field-based retrieval.

The URL path length has been shown to be particularly effective for the topic distillation tasks. For the home page finding tasks, both the URL path length, and PageRank result in considerable improvements in retrieval effectiveness. Regarding the named page finding tasks, the most effective query-independent sources of evidence are PageRank and the Absorbing Model.

The next section investigates the performance of the described retrieval approaches in a setting which aims to reduce any overfitting effect of the applied optimisation process.

4.6 Obtaining a realistic parameter setting

In Sections 4.3 to 4.5, each retrieval approach has been optimised, and evaluated with a set of queries from a particular search task. This optimisation approach allows for the comparison of the retrieval approaches on the basis of their optimal retrieval performance. However, it may also result in the overfitting of a particular task. The aim of the current section is to introduce a more realistic setting for the optimisation of the proposed retrieval approaches. This setting involves the optimisation and the evaluation of the retrieval approaches with different mixed types of tasks, as well as a restriction of the optimisation process (Section 4.3.2), which is terminated early.

4.6.1 Using mixed tasks

The current section investigates the effectiveness of the retrieval approaches for a mixture of topic distillation, home page finding and named page finding topics. Two sets

of mixed tasks are used, as described in Section 4.2. The first one, denoted by mq2003, is a set of 350 topics from the TREC 2003 Web track. The second set of topics, denoted by mq2004, corresponds to 225 topics from the TREC 2004 Web track mixed query task. Due to a lack of test collections with various Web search tasks, only one test collection is used here, namely the .GOV collection. However, the tested tasks involve topic distillation, home page finding, and named page finding tasks. These three different types of tasks are specific to Web search, which is the focus of this thesis.

The mean average precision (MAP) of the employed retrieval approaches is optimised for one of the mixed tasks, and the obtained parameter values are used to evaluate the retrieval approach with a different set of mixed tasks. When the mixed task mq2003 is employed as a training set, the first 50 queries for each type of task are used, and this smaller set of queries is denoted by mq2003'. This choice is made in order not to bias the training towards a particular type of task (note that mq2003 consists of 50 topic distillation queries, 150 home page finding queries, and 150 named page finding queries).

The employed field-based retrieval models are PL2F, PB2F, I(n_e)C2F, DLHF, and BM25F (Section 4.4). The employed query-independent sources of evidence are the URL path length, PageRank, and the Absorbing Model (Section 4.5). The parameter values for the field-based retrieval models, and their combination with the query-independent sources of evidence are shown in Tables A.9 and A.10 of Appendix A, respectively.

The evaluation of the field-based weighting models and their combination with the query-independent evidence for the mixed-type query sets is shown in Table 4.9. The bold entries correspond to the most effective retrieval approach for each row of the table. In the column 'Task (train)', the task in brackets corresponds to the training task. For example, row 1 shows the evaluation results of the field-based weighting models for the mixed task mq2003, after optimising their MAP for the mixed task mq2004.

Regarding the mixed tasks mq2003 and mq2004, it is interesting to note that for the weighting models I(n_e)C2F and BM25F, the combination of field retrieval with PageRank (rows 17-18 in Table 4.9) is more effective than the combination of field retrieval with the URL path length (rows 9-10 in Table 4.9). This can be explained by the fact that the combination of field retrieval with PageRank improves the retrieval effectiveness for all three types of search tasks. The evidence from the URL path length

Mean Average Precision (MAP)						
Row	Task (train)	PL2F	PB2F	I(n_e)C2F	DLHF	BM25F
1	mq2003 (mq2004)	0.6132	0.5695	0.6174	0.5037	**0.6351**
2	mq2004 (mq2003')	**0.4916**	0.4638	0.4667	0.3914	0.4874
3	td2003 (mq2004)	0.1423	0.1366	0.1427	0.1357	**0.1449**
4	td2004 (mq2003')	**0.1454**	0.1307	0.1249	0.1177	0.1401
5	hp2003 (mq2004)	0.7032	0.6395	0.6889	0.6120	**0.7296**
6	hp2004 (mq2003')	**0.6410**	0.5893	0.5810	0.5602	0.6404
7	np2003 (mq2004)	0.6801	0.6437	**0.7041**	0.5182	0.7041
8	np2004 (mq2003')	0.6885	0.6716	**0.6940**	0.4963	0.6817
		PL2FU	PB2FU	I(n_e)C2FU	DLHFU	BM25FU
9	mq2003 (mq2004)	0.6317	0.5916	0.6461	0.5302	**0.6596**
10	mq2004 (mq2003')	**0.5363**	0.4954	0.5082	0.4321	0.5302
11	td2003 (mq2004)	**0.1942**	0.1810	0.1762	0.1586	0.1907
12	td2004 (mq2003')	**0.2015**	0.1577	0.1790	0.1630	0.1875
13	hp2003 (mq2004)	0.7732	0.7287	0.7632	0.6880	**0.7918**
14	hp2004 (mq2003')	0.7124	0.6450	0.6616	0.6325	**0.7172**
15	np2003 (mq2004)	0.6361	0.5913	**0.6856**	0.4962	0.6837
16	np2004 (mq2003')	**0.6950**	0.6834	0.6840	0.5007	0.6858
		PL2FP	PB2FP	I(n_e)C2FP	DLHFP	BM25FP
17	mq2003 (mq2004)	0.6308	0.5887	0.0.6644	0.5387	**0.6694**
18	mq2004 (mq2003')	0.4809	0.4639	0.5232	0.4155	**0.5322**
19	td2003 (mq2004)	0.1599	0.1558	**0.1791**	0.1524	0.1777
20	td2004 (mq2003')	0.1527	0.1307	0.1436	0.1418	**0.1656**
21	hp2003 (mq2004)	0.7300	0.6933	0.7657	0.6734	**0.7780**
22	hp2004 (mq2003')	0.6367	0.5895	0.7100	0.6123	**0.7487**
23	np2003 (mq2004)	0.6886	0.6285	**0.7249**	0.5330	0.7246
24	np2004 (mq2003')	0.6534	0.6716	**0.7160**	0.4925	0.6821
		PL2FA	PB2FA	I(n_e)C2FA	DLHFA	BM25FA
25	mq2003 (mq2004)	0.6225	0.5851	0.6334	0.5116	**0.6443**
26	mq2004 (mq2003')	**0.5016**	0.4638	0.4876	0.3947	0.4979
27	td2003 (mq2004)	0.1446	0.1330	**0.1449**	0.1336	**0.1449**
28	td2004 (mq2003')	**0.1448**	0.1307	0.1308	0.1232	0.1442
29	hp2003 (mq2004)	0.7212	0.6793	0.7385	0.6300	**0.7591**
30	hp2004 (mq2003')	0.6598	0.5893	0.6115	0.5697	**0.6675**
31	np2003 (mq2004)	0.6831	0.6417	0.6911	0.5192	**0.6959**
32	np2004 (mq2003')	0.7002	0.6716	**0.7206**	0.4913	0.6821

Table 4.9: The evaluation of the field retrieval weighting models and their combination with the query-independent evidence for the mixed-type query sets, and for the query-type specific topic subsets. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task.

is mostly beneficial for the topic distillation and home page finding tasks, where the relevant documents are home pages of Web sites.

On the other hand, the combination of the field-based models PL2F and PB2F, both of which employ a Poisson randomness model, with evidence from the URL path length for the tasks mq2003 and mq2004 (rows 9-10) seems to be more effective than

their combination with either Pagerank (rows 17-18), or the Absorbing Model (rows 25-26).

Overall, the training and evaluation of the retrieval approaches with different mixed tasks has a negative impact on MAP, compared to the results obtained from Table 4.8. This is explained in terms of the reduced effect of overfitting the data. However, some of the evaluated retrieval approaches still perform well compared to the best performing runs in the corresponding TREC Web tracks. For example, the MAP of the retrieval approach PL2FU for the task mq2004 is 0.5363 (row 10 in Table 4.9), while the highest MAP achieved by the submitted runs to the TREC 2004 is 0.5389 (row 12 in Table 4.6, page 67). The MAP of the same retrieval approach for the task td2004 is 0.2015 (row 12 in Table 4.9), while the highest MAP achieved for this task in TREC 2004 was 0.1791 (row 5 in Table 4.6).

4.6.2 Using mixed tasks and restricted optimisation

In addition to the evaluation of the retrieval approaches with mixed types of tasks, this section considers a setting, where the optimisation process is terminated after 20 iterations. The parameters are set to the values that resulted in the best retrieval effectiveness after 20 iterations of the optimisation process. This setting aims to further reduce any overfitting effect of the optimisation process. Tables A.11 and A.12 of Appendix A display the corresponding parameter values for the field-based weighting models, and their combination with the query-independent sources of evidence, respectively.

Table 4.10 shows the evaluation of the weighting models and their combination with query-independent sources of evidence when a restricted optimisation is performed. The bold entries correspond to the most effective retrieval approach for each tested topic set. In the column 'Task (train)', the task in brackets corresponds to the training task.

Compared to the retrieval effectiveness obtained from the full optimisation over a set of mixed types of queries (Table 4.9), it can be seen that, generally, the restricted optimisation results in lower MAP. This is expected because the optimisation process is stopped early.

In particular, the restricted optimisation has a negative effect on the retrieval effectiveness of BM25F. For example, the MAP of BM25F for the mixed task mq2003 with full optimisation is 0.6351 (row 1 in Table 4.9). However, in the case of the restricted

Mean Average Precision (MAP)						
Row	Task (train)	PL2F	PB2F	I(n_e)C2F	DLHF	BM25F
1	mq2003 (mq2004)	**0.6089**	0.5558	0.6071	0.4890	0.5533
2	mq2004 (mq2003')	**0.4444**	0.4114	0.4273	0.3792	0.4327
3	td2003 (mq2004)	**0.1474**	0.1401	0.1089	0.1410	0.1179
4	td2004 (mq2003')	**0.1299**	0.1137	0.1150	0.1138	0.1136
5	hp2003 (mq2004)	**0.7005**	0.5927	0.6872	0.5814	0.5629
6	hp2004 (mq2003')	0.4893	0.4262	0.4826	**0.5170**	0.5138
7	np2003 (mq2004)	0.6713	0.6575	**0.6930**	0.5127	0.6889
8	np2004 (mq2003')	**0.7141**	0.6944	0.6843	0.5069	0.6707
		PL2FU	PB2FU	I(n_e)C2FU	DLHFU	BM25FU
9	mq2003 (mq2004)	0.6206	0.5809	**0.6258**	0.5216	0.6237
10	mq2004 (mq2003')	**0.5254**	0.4723	0.4946	0.4273	0.4883
11	td2003 (mq2004)	**0.1939**	0.1520	0.1446	0.1600	0.1857
12	td2004 (mq2003')	**0.2092**	0.1404	0.1763	0.1565	0.1851
13	hp2003 (mq2004)	**0.7435**	0.6480	0.7343	0.6660	0.7365
14	hp2004 (mq2003')	**0.6674**	0.5523	0.6370	0.6278	0.6479
15	np2003 (mq2004)	0.6399	0.6568	**0.6778**	0.4978	0.6570
16	np2004 (mq2003')	**0.6997**	0.7241	0.6706	0.4978	0.6319
		PL2FP	PB2FP	I(n_e)C2FP	DLHFP	BM25FP
17	mq2003 (mq2004)	0.6238	0.5873	0.6453	0.5319	**0.6502**
18	mq2004 (mq2003')	0.4853	0.4723	**0.4983**	0.4156	0.4955
19	td2003 (mq2004)	0.1606	0.1445	0.1542	0.1455	**0.1640**
20	td2004 (mq2003')	**0.1459**	0.1402	0.1307	0.1371	0.1377
21	hp2003 (mq2004)	0.7174	0.6589	**0.7603**	0.6710	0.7516
22	hp2004 (mq2003')	0.6192	0.5677	**0.6632**	0.6149	0.6469
23	np2003 (mq2004)	0.6846	0.6634	0.6940	0.5216	**0.7108**
24	np2004 (mq2003')	0.6908	**0.7090**	0.7010	0.4947	0.7021
		PL2FA	PB2FA	I(n_e)C2FA	DLHFA	BM25FA
25	mq2003 (mq2004)	0.6164	0.5772	**0.6210**	0.5019	0.5894
26	mq2004 (mq2003')	**0.4717**	0.4462	0.4509	0.3959	0.4680
27	td2003 (mq2004)	**0.1558**	0.1417	0.1283	0.1284	0.1290
28	td2004 (mq2003')	0.1201	**0.1204**	0.0889	0.1167	0.1169
29	hp2003 (mq2004)	**0.7229**	0.6248	0.7227	0.6042	0.6498
30	hp2004 (mq2003')	0.5780	0.4965	0.5815	0.5555	**0.5975**
31	np2003 (mq2004)	0.6633	0.6748	**0.6836**	0.5241	0.6825
32	np2004 (mq2003')	0.7169	**0.7218**	0.6814	0.5154	0.6896

Table 4.10: The evaluation of the field retrieval weighting models and their combination with the query-independent evidence for the mixed-type query sets and for the query-type specific topic subsets, with restricted optimisation. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task.

optimisation, it drops to 0.5533 (row 1 in Table 4.10). This is explained because the optimisation of BM25F involves a higher number of two-dimensional optimisations, the restriction of which results in a setting further away from the optimum.

The DFR field-based weighting models are more robust, in the sense that they are less affected by the restricted optimisation. For example, the MAP of PL2F for mq2003

drops from 0.6132 (row 1 in Table 4.9) to 0.6089 (row 1 in Table 4.10).

It is worth noting that, despite the restricted optimisation, the combination of the field-based weighting model PL2F with the URL path length (PL2FU) achieves comparable MAP to that of the best performing run submitted to the TREC 2004 Web track (0.5254 from Table 4.10 with respect to 0.5389 from Table 4.6). This suggests that the retrieval approach PL2FU is robust with respect to setting its hyper-parameters.

4.6.3 Conclusions

This section has revisited the employed optimisation process from two perspectives in order to obtain a realistic setting for the hyper-parameters of the proposed retrieval approaches. First, the optimisation of precision, and the evaluation of the retrieval approaches has been performed with different sets of mixed tasks. The mixed tasks include topic distillation, home page finding, and named page finding tasks. Second, the two-step optimisation process described in Section 4.3.2 has been modified in order to terminate after 20 iterations.

The obtained parameter setting for the retrieval approaches does not always result in optimal retrieval performance. However, it represents a realistic setting, where the most effective parameter values are approximated. The setting, which employs the mixed tasks and the restricted optimisation, will be employed in the next section, in order to establish the potential of selective Web IR for improvements in retrieval effectiveness.

4.7 Potential improvements from selective Web information retrieval

The aim of this section is to investigate the potential improvements in retrieval effectiveness from selective Web IR. This investigation is performed in a setting where it is assumed that the most effective approach a_i is applied on a per-query basis.

The methodology to establish the potential for improvements from selective Web IR is the following. A set of retrieval approaches $a_1, a_2, \ldots$ is considered. It is assumed that there is a mechanism $\mathrm{MAX}(a_1, a_2, \ldots)$, which can identify and apply the most effective retrieval approach on a per-query basis. The retrieval effectiveness of the mechanism

MAX corresponds to the maximum retrieval effectiveness that can be obtained by selectively applying any of the approaches $a_1, a_2, \ldots$ on a per-query basis.

The employed retrieval approaches involve the field-based weighting models, and their combinations with query-independent evidence from the URLs of Web documents, PageRank, or the Absorbing Model. The parameters of the retrieval approaches have been set after a restricted optimisation with mixed tasks, as described in Section 4.6. The evaluation of the retrieval approaches has been shown in Table 4.10.

The described methodology is applied for pairs of retrieval approaches. Table 4.11 displays the pairs of retrieval approaches, for which the mechanism MAX results in the highest improvements over the most effective retrieval approach of the pair. The symbol * denotes that the difference between the MAP of the mechanism MAX and that of the most effective retrieval approach is statistically significant at $p = 0.05$ according to Wilcoxon's signed rank test. Rows 1-6 display the potential for improvements in retrieval effectiveness from the selective application of retrieval approaches that use the field-based weighting model PL2F. Row 1 of Table 4.11 refers to the following case. When the retrieval approach PL2F is applied for all queries of the task td2003, the achieved MAP is 0.1474. When the retrieval approach PL2FP is applied for all queries of the task td2003, the achieved MAP is 0.1606. When the mechanism MAX selects the most effective approach between PL2F and PL2FP for each query of the task td2003, the achieved MAP is 0.1726, which represents a relative improvement of +7.47% over the MAP of PL2FP (0.1606). According to Wilcoxon's signed rank test, the difference between the MAP of the decision mechanism MAX and that of PL2FP is statistically significant, as denoted by * in the table. For all the cases reported in Table 4.11, it can be seen that the improvements in MAP obtained by the mechanism MAX are statistically significant.

When the employed pairs of retrieval approaches use the same field-based weighting model (rows 1-30 in Table 4.11), then the highest potential for improvements in retrieval effectiveness is obtained when the field-based weighting model is I(n_e)C2F (rows 13-18 in Table 4.11). The lowest potential for improvements in retrieval effectiveness are obtained for the task np2003, when the employed retrieval approaches use the field-based weighting models PL2F or PB2F (+2.19% from rows 5 and 11 in Table 4.11).

If the available retrieval approaches employ different field-based weighting models (rows 31-36 in Table 4.11), the potential for improvements in retrieval effectiveness

increases considerably. For example, the maximum MAP achieved by the selective application of either PB2FU, or DLHFA, for the task hp2004, is 0.7025 (row 34 in Table 4.11). This corresponds to a relative increase of +26.46% from the MAP of DLHFA (0.5555).

In some cases, the maximum MAP achieved from the selective application of the pairs of retrieval approaches displayed in Table 4.11 is higher than the MAP of the best performing submitted run to the corresponding TREC Web track. For example, when either PB2F or I(n_e)C2FA are applied on a per-query basis for the task np2004, the mechanism MAX results in higher MAP than that of the best performing run in the same task of TREC 2004 Web track (0.8019 from row 36 in Table 4.11 vs. 0.7232 from row 11 in Table 4.6, page 67).

It is worth noting that the pairs of retrieval approaches that result in the highest potential for improvements, as shown in Table 4.11, do not necessarily involve the most effective retrieval approaches. For example, the maximum MAP obtained by the selective application of PL2F and PL2FP for the task td2003 is 0.1726. However, the MAP obtained by uniformly applying PL2FU, the most effective retrieval approach employing the field-based weighting model, is 0.1939 (row 11 from Table 4.10).

Overall, this section has shown that there is an important potential for statistically significant improvements in retrieval effectiveness from selective Web IR. The potential for improvements is higher when the applied retrieval approaches employ different field-based weighting models. Furthermore, there are important improvements from the selective application of retrieval approaches, even when the retrieval approaches are not the best performing ones.

4.8 Summary

This chapter has established the potential for improvements in selective Web IR, after a thorough evaluation of different retrieval approaches with a range of weighting models. The experimental setting has been described in Section 4.2. The evaluation of the retrieval approaches has been performed in three steps, where the mean average precision of each retrieval approach has been optimised with respect to each tested task.

First, Section 4.3 has examined the effectiveness of full text retrieval and retrieval from particular document representations, such as the title, the headings, and the

Row	Task	Mean Average Precision					
		First approach		Second approach		MAX	
1	td2003	PL2F	(0.1474)	PL2FP	(0.1606)	0.1726	(+ 7.47%)*
2	td2004	PL2F	(0.1299)	PL2FA	(0.1201)	0.1464	(+12.70%)*
3	hp2003	PL2FU	(0.7435)	PL2FA	(0.7229)	0.7633	(+ 2.66%)*
4	hp2004	PL2FU	(0.6674)	PL2FP	(0.6192)	0.7311	(+ 9.54%)*
5	np2003	PL2F	(0.6713)	PL2FA	(0.6633)	0.6860	(+ 2.19%)*
6	np2004	PL2F	(0.7141)	PL2FA	(0.7169)	0.7797	(+ 8.76%)*
7	td2003	PB2F	(0.1401)	PB2FA	(0.1417)	0.1490	(+ 5.15%)*
8	td2004	PB2FU	(0.1404)	PB2FP	(0.1402)	0.1614	(+14.96%)*
9	hp2003	PB2FU	(0.6480)	PB2FP	(0.6589)	0.6798	(+ 3.17%)*
10	hp2004	PB2FU	(0.5523)	PB2FP	(0.5677)	0.6340	(+11.68%)*
11	np2003	PB2F	(0.6575)	PB2FP	(0.6634)	0.6779	(+ 2.19%)*
12	np2004	PB2FU	(0.7241)	PB2FP	(0.7090)	0.7821	(+ 8.01%)*
13	td2003	I(n_e)C2F	(0.1089)	I(n_e)C2FA	(0.1283)	0.1574	(+22.68%)*
14	td2004	I(n_e)C2F	(0.1150)	I(n_e)C2FP	(0.1307)	0.1524	(+16.60%)*
15	hp2003	I(n_e)C2FU	(0.7343)	I(n_e)C2FA	(0.7227)	0.7600	(+ 3.50%)*
16	hp2004	I(n_e)C2FU	(0.6370)	I(n_e)C2FP	(0.6632)	0.7385	(+11.35%)*
17	np2003	I(n_e)C2F	(0.6930)	I(n_e)C2FP	(0.6940)	0.7262	(+ 4.64%)*
18	np2004	I(n_e)C2F	(0.6843)	I(n_e)C2FA	(0.6814)	0.7546	(+10.27%)*
19	td2003	DLHF	(0.1410)	DLHFP	(0.1455)	0.1578	(+ 8.45%)*
20	td2004	DLHF	(0.1138)	DLHFP	(0.1371)	0.1492	(+ 8.83%)*
21	hp2003	DLHFU	(0.6660)	DLHFP	(0.6710)	0.7018	(+ 4.59%)*
22	hp2004	DLHFU	(0.6278)	DLHFP	(0.6149)	0.6733	(+ 7.25%)*
23	np2003	DLHFP	(0.5216)	DLHFA	(0.5241)	0.5556	(+ 6.01%)*
24	np2004	DLHFU	(0.4978)	DLHFP	(0.4947)	0.5427	(+ 9.02%)*
25	td2003	BM25FU	(0.1857)	BM25FP	(0.1640)	0.2135	(+14.97%)*
26	td2004	BM25F	(0.1136)	BM25FA	(0.1169)	0.1255	(+ 7.35%)*
27	hp2003	BM25FU	(0.7365)	BM25FP	(0.7516)	0.8031	(+ 6.85%)*
28	hp2004	BM25FU	(0.6479)	BM25FP	(0.6469)	0.7062	(+ 9.00%)*
29	np2003	BM25F	(0.6889)	BM25FP	(0.7108)	0.7656	(+ 7.71%)*
30	np2004	BM25F	(0.6707)	BM25FU	(0.6319)	0.6966	(+ 3.86%)*
31	td2003	I(n_e)C2FU	(0.1446)	DLHFP	(0.1455)	0.1926	(+32.37%)*
32	td2004	PL2F	(0.1299)	I(n_e)C2FP	(0.1307)	0.1615	(+23.57%)*
33	hp2003	DLHFU	(0.6660)	BM25FA	(0.6498)	0.7658	(+14.98%)*
34	hp2004	PB2FU	(0.5523)	DLHFA	(0.5555)	0.7025	(+26.46%)*
35	np2003	PL2FP	(0.6846)	I(n_e)C2FA	(0.6836)	0.7827	(+14.33%)*
36	np2004	PB2F	(0.6944)	I(n_e)C2FA	(0.6814)	0.8019	(+16.52%)*

Table 4.11: Potential for improvements in retrieval effectiveness from the selective application of two retrieval approaches on a per-query basis. The retrieval approaches are based on a restricted optimisation, as reported in Table 4.10. The table displays the pairs of retrieval approaches that result in the highest improvements in MAP for the tested topic sets. The symbol * denotes that the difference between the MAP of MAX and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

anchor text of the incoming hyperlinks of Web documents. It has been shown that the effectiveness of each field depends on the search task (Table 4.4 on page 63). For

the ad-hoc retrieval tasks, the full text of documents is the most effective document representation. For the tasks, where there is a bias towards the home pages of Web sites, the anchor text representation is more effective than the full text of Web documents. The title representation of documents is less effective, but outperforms the headings representation for Web specific tasks.

Second, Section 4.4 has introduced per-field normalisation, a new normalisation technique for the DFR framework, which allows the term frequency normalisation and weighting of different document fields. The employed document fields are the body, the anchor text of incoming hyperlinks, and the title of Web documents. The field-based weighting models result in important improvements in retrieval effectiveness, compared to retrieval from the individual document representations, particularly for the named page finding tasks (Table 4.7 on page 72).

Third, Section 4.5 has enhanced the field-based weighting models with query-independent sources of evidence. In particular, the considered query-independent sources of evidence are the length in characters of the URL path of Web documents, PageRank (Brin & Page, 1998), as well as the Absorbing Model, a novel hyperlink structure analysis algorithm. The evaluation results have shown that the combination of field-based retrieval with the URLs of Web documents provides important improvements in MAP over field retrieval, for the topic distillation and home page finding tasks (Table 4.8 on page 90). This is due to the fact that there is a bias towards the home pages of Web sites for such search tasks. When employing PageRank, moderate improvements in retrieval effectiveness are obtained for all the Web specific search tasks, that is the topic distillation, home page finding and named page finding tasks. The Absorbing Model is particularly effective for the named page finding tasks.

Overall, the presented retrieval approaches achieve higher, or similar performance as the most effective official runs submitted to the corresponding tasks of the TREC Web tracks. Section 4.6 revisits the parameter setting of the retrieval approaches, by performing optimisation and evaluation on different sets of mixed tasks, as well as by terminating the optimisation process early. This allows for the reduction of the overfitting effects caused by the optimisation process, and shows that the retrieval approaches are robust (Table 4.10 on page 96).

This last setting is employed in Section 4.7, in order to demonstrate the potential for improvements in retrieval effectiveness from selective Web IR. The results show that

statistically significant improvements can be obtained with respect to the effectiveness of applying a retrieval approach uniformly for all queries (Table 4.11). After having established the potential for improvements in retrieval effectiveness from selective Web IR, Chapter 5 will introduce a framework for selective Web IR. The evaluation of the framework will be performed both in an optimal setting (Chapter 6), and in an operational setting with limited relevance information (Chapter 7).

Chapter 5

A framework for Selective Web Information Retrieval

5.1 Introduction

The previous chapter introduced a wide range of retrieval approaches for Web IR, and established that selective Web IR has the potential to result in improved retrieval effectiveness. This chapter proposes a novel framework for selective Web IR.

A central concept in this framework is the *decision mechanism*, which selects an appropriate retrieval approach to apply on a per-query basis. The selection of a retrieval approach is aided by an *experiment* $\mathcal{E}$, which extracts a feature from a sample of the set of retrieved documents. This is motivated by the following example. An informational query about a very broad topic may benefit from applying hyperlink structure analysis in order to detect the most authoritative Web sites and resources (Kleinberg, 1998). On the other hand, applying hyperlink analysis to an informational query about a topic which is not extensively represented in the document collection, may result in a topic drift (Bharat & Henzinger, 1998). In these two examples, the retrieval effectiveness of hyperlink analysis is related to the broadness of the topic. In terms of the characteristics of the set of retrieved documents for a particular query, the broadness of a topic can be seen as the proportion of the indexed documents that are being retrieved for a particular query. On the other hand, for navigational topics, if evidence from the URL of Web documents is uniformly used for all topics, then a relevant document with a relatively long URL would be penalised.

The remainder of this chapter is organised as follows. Section 5.2 describes the

framework for selective Web IR in terms of statistical decision theory, and discusses the differences between the selective Web IR and related work. The next two sections introduce a range of experiments $\mathcal{E}$. First, Section 5.3 defines the score-independent experiments $\mathcal{E}$, which are based on counting the occurrences of query terms in the set of retrieved documents, as well as in aggregates of related Web documents. The aggregates are defined as the Web documents that have the same domain name, or the Web documents that are stored in the same directory. Second, Section 5.4 introduces the score-dependent experiments $\mathcal{E}$, which employ evidence from the retrieval scores of Web documents, and from the hyperlink structure of the retrieved documents, in order to estimate the usefulness of the hyperlink structure. Finally, Section 5.5 introduces a Bayesian decision mechanism for the evaluation of selective Web IR.

5.2 Selective retrieval as a statistical decision problem

Selective Web IR can be seen as a statistical decision problem with a number of available actions a for a set of different states of nature s, a loss function l, and an experiment $\mathcal{E}$ (Lindgren, 1971; Wald, 1950). In the context of selective Web IR, the actions a correspond to the retrieval approaches that can be applied for a given query. Due to the inherent uncertainty of the retrieval process, a retrieval system can only guess which retrieval approach is most appropriate for a given query. The knowledge of which retrieval approach is most appropriate is modelled with the states of nature: when the *true state of nature* is s_i, the most appropriate retrieval approach for a given query is a_i. This formulation of the problem results in a one-to-one mapping between states of nature and actions. In this setting, a *decision mechanism* guesses the state of nature, or in other words, it aims to identify the most appropriate retrieval approach for a query.

The consequences of applying a retrieval approach a_i when the state of nature is s_j are modelled by a loss function $l(a_i, s_j)$, which expresses the loss of utility in each possible situation. The loss function can be defined in different ways, depending on the factors that affect utility. For example, it can be defined in terms of the effectiveness of the retrieval approaches. It is also reasonable to consider the computational cost of the retrieval approaches, especially in an operational setting where a retrieval system must process the users' queries in a timely manner. For example, the cost of

applying the HITS algorithm (Kleinberg, 1998) at query time is significantly higher than using PageRank (Page et al., 1998), because the hub and authority scores of the HITS algorithm need to be computed for each particular query. In contrast, PageRank scores are computed during indexing time. Hence, the overhead for combining them with content analysis scores is marginal. This thesis is focused on a TREC-like batch retrieval setting. It is also worth noting that the investigated retrieval approaches in Chapter 4 do not introduce any considerable overhead in the retrieval process. Therefore, it is assumed that the utility and the loss of applying a retrieval approach a when the state of nature is s depends only on the retrieval effectiveness of a, and not on its computational cost.

In the context of selective Web IR, the loss $l(a_i, s_j)$ of applying a retrieval approach a_i when the true state of nature is s_j can be defined with respect to a preference relationship among the retrieval approaches, as follows.

Definition 3 Suppose that there is a decision problem with n retrieval approaches, and n states of nature. The retrieval effectiveness of the retrieval approach a_i for the state s is denoted by $m(a_i, s)$. The n retrieval approaches are ranked in decreasing order of their retrieval effectiveness $m(a_i, s)$. In this way, the rank of the most effective retrieval approach is 1, the rank of the second most effective retrieval approach is 2, and so on. The rank of the least effective retrieval approach is n.

If the rank of the retrieval approach a_i is denoted by $r(a_i, s)$, then the loss function is defined as follows:

$$l(a_i, s) = \frac{r(a_i, s) - 1}{n - 1} \tag{5.1}$$

The definition of the loss function in Equation (5.1) does not consider the magnitude of the difference in retrieval effectiveness among the retrieval approaches, but only their ranking. Moreover, dividing with $n - 1$ only normalises the values of the loss function in the range $[0, 1]$, and it does not affect any further computations. □

Before continuing, an example is given in order to illustrate the formulation of selective Web IR in terms of a decision problem, as well as the definition of the loss function.

Example 4 Figure 5.1 describes selective Web IR as a decision problem with 3 states of nature (s_1, s_2, and s_3) and 3 retrieval approaches (a_1, a_2, and a_3). When the state

of nature is s_j, then the loss associated with applying retrieval approach a_i is denoted by $l(a_i, s_j)$. The loss $l(a_i, s_j)$ can be specified as follows.

	States of nature		
	s_1 = retrieval approach a_1 is appropriate	s_2 = retrieval approach a_2 is appropriate	s_3 = retrieval approach a_3 is appropriate
Apply retrieval approach a_1	$l(a_1, s_1)$	$l(a_1, s_2)$	$l(a_1, s_3)$
Apply retrieval approach a_2	$l(a_2, s_1)$	$l(a_2, s_2)$	$l(a_2, s_3)$
Apply retrieval approach a_3	$l(a_3, s_1)$	$l(a_3, s_2)$	$l(a_3, s_3)$

Figure 5.1: Selective application of retrieval approaches for three states of nature s_1, s_2, s_3 and three different retrieval approaches a_1, a_2, a_3. The loss associated with applying retrieval approach a_i when the true state of nature is s_j is denoted by $l(a_i, s_j)$.

When the state of nature is s_1, suppose that $m(a_1, s_1) > m(a_3, s_1) > m(a_2, s_1)$. In this case, $r(a_1, s_1) = 1$, $r(a_3, s_1) = 2$, and $r(a_2, s_1) = 3$. From Equation (5.1), the loss associated with a_1 when the true state of nature is s_1 is $l(a_1, s_1) = \frac{1-1}{3-1} = 0$. The loss for the retrieval approaches a_2 and a_3 is $l(a_2, s_1) = \frac{3-1}{3-1} = 1$, and $l(a_3, s_1) = \frac{2-1}{3-1} = 0.5$, respectively. The loss function $l(a_i, s_j)$ can be specified in the same way for all the possible pairs of retrieval approaches a_i and states of nature s_j. The decision problem can be formulated in the same way for any number of retrieval approaches and states of nature. In the case of a decision problem with two retrieval approaches, the output of the loss function is binary, i.e. 1 or 0. □

In order to identify the true state of nature, and to decide which retrieval approach to use, an experiment $\mathcal{E}$ is performed on a sample Ret_q of the set of retrieved documents for a query q. The sample Ret_q can be restricted to a number of top-ranked documents, ordered by a specific retrieval technique. This retrieval technique may correspond to any of the retrieval approaches presented in Chapter 4. The retrieval technique that generates the sample Ret_q is not necessarily used for the final ranking of documents. In other words, the experiment $\mathcal{E}$ does not depend on the retrieval approaches that the decision mechanism applies on a per-query basis. For example, suppose that an experiment $\mathcal{E}$ counts the number of documents that contain at least one query term in their title. A decision mechanism can employ this experiment in order to select on a per-query basis one of the field-based weighting models described in Section 4.4. The fact that the experiment employs only the documents with at least one query term in a particular field does not mean that the retrieval approaches cannot employ any other available document fields, such as the anchor text. This approach allows for more

flexibility in defining the experiment $\mathcal{E}$, as well as in selecting the retrieval approaches. In the remainder of this thesis, the defined experiments will be independent of the retrieval approaches used for the final ranking of documents.

The experiment $\mathcal{E}$ extracts a feature related to the query. This feature can be related to the statistical characteristics of the query terms, or the characteristics of the documents that are retrieved for this particular query. For example, the query performance pre-retrieval predictors (He & Ounis, 2004) can be seen as experiments $\mathcal{E}$ that use evidence only from the collection statistics of the query terms. A different experiment $\mathcal{E}$ may employ evidence from the hyperlink structure among the retrieved Web documents for a query.

The experiment $\mathcal{E}$ returns an outcome o, which can be either a categorical, or a numerical value. In the case of categorical values, the outcome o of an experiment, which detects how difficult the queries are, could be either `'Query is difficult'`, or `'Query is easy'`. In the case of numerical values, the outcome o of an experiment, which estimates the density of hyperlinks in the set of retrieved Web documents, could be a real number between 0 and 1. The decision mechanism needs to map the range of the possible outcome values of the employed experiment to particular retrieval approaches. According to the outcome of the experiment for a query, the decision mechanism selects an appropriate retrieval approach to apply.

When the outcome of the experiment $\mathcal{E}$ for a query predicts the true state of nature with some probability, it provides the decision mechanism with evidence to guess the true state of nature, and to apply an appropriate retrieval approach for the given query. Ideally, the probability distribution of the outcome of an experiment $\mathcal{E}$ when the state of nature is s_1 for a set of queries, should be different from the probability distribution of the outcome of $\mathcal{E}$ when the state of nature is s_2 for a different set of queries. In such a case, the experiment $\mathcal{E}$ would identify the true state of nature, without any error. Section 5.5 describes how the probability distribution of the outcome of $\mathcal{E}$ is empirically obtained from a set of training queries in the context of a Bayesian decision mechanism, which aims to minimise the expected loss from applying a retrieval approach.

For the remainder of this thesis, a particular experiment is referred to as $\mathcal{E}_x$, where x stands for the feature of the sample Ret_q that is quantified by the experiment. Next, the concept of an experiment $\mathcal{E}$ is further illustrated with an example.

Example 5 If the broadness of a topic is associated with the number of retrieved documents, then one experiment $\mathcal{E}_{broad}$ that estimates how broad a topic is, could be described as "Count the number of documents that contain at least one query term". The outcome of this experiment corresponds to the cardinality of the set Ret_q of retrieved documents containing at least one query term. □

In the context of Web retrieval, other types of experiments can exploit evidence from the hyperlink structure of the sample Ret_q of retrieved documents, or combinations of the hyperlink structure and retrieval from the text of documents. This thesis is focused on defining a range of experiments $\mathcal{E}$, and not on the definition of the loss function.

The remainder of the current section is organised as follows. Section 5.2.1 discusses the differences between selective Web IR and related work. Section 5.2.2 illustrates the terminology of the proposed framework, by describing the setting already used in Section 4.7 to establish the potential improvements in retrieval effectiveness from selective Web IR.

5.2.1 Selective Web information retrieval and related work

Selective Web IR is a different approach to optimising the retrieval effectiveness of a system from query type classification. Indeed, the selective application of different retrieval approaches differs from query type classification, as performed by Kang & Kim (2003), where the aim has been to identify whether a query is informational or navigational, and then to apply an appropriate retrieval approach for that particular query type. The selective application of retrieval approaches does not require knowing the type of a query. Instead, it selects a retrieval approach to apply for each query, irrespectively of its type. Therefore, the decision mechanism and the experiments $\mathcal{E}$ would not necessarily require modifications in the case of a new type of queries.

There is another difference between selective Web IR and the query-biased setting of weights and parameters for the combination of evidence. Amitay et al. (2002) adjusted the contribution of the hyperlink structure analysis on a per-query basis, according to the characteristics of the retrieved documents' hyperlink structure. Plachouras & Ounis (2005) adjusted the weights of content and hyperlink structure analysis with a Dempster-Shafer combination mechanism, according to the specificity of a query. On the other hand, selective Web IR applies a particular retrieval approach from a set of

available ones. In this context, the retrieval approach corresponds to a fixed combination of retrieval techniques. In other words, the retrieval approach corresponds to a description of all the steps followed in order to form the final ranking of retrieved documents. Therefore, the potential improvements in retrieval effectiveness from selective Web IR come from the relative difference in retrieval effectiveness between the different approaches, and not from the change in the weight of each source of evidence.

Selective Web IR is similar to query performance prediction, as discussed in Section 3.6.2, because it aims to predict how appropriate it is to apply a particular retrieval approach. However, query performance prediction is primarily focused on estimating the correlation of a predictor with the retrieval effectiveness of a particular retrieval approach. On the other hand, selective Web IR explicitly aims to predict the most effective retrieval approach from a set of at least two available retrieval approaches.

5.2.2 Decision mechanism with known states of nature

This section introduces a decision mechanism and an experiment $\mathcal{E}$ in order to describe the setting used for establishing the potential improvements in retrieval effectiveness from selective Web IR in Section 4.7.

Suppose that a decision mechanism can apply one of the retrieval approaches $a_1, a_2, \ldots, a_n$ on a per-query basis, and that the true state of nature s_i is known. In other words, it is assumed that the most effective retrieval approach a_i among $a_1, a_2, \ldots, a_n$ can be identified with certainty on a per-query basis. This setting corresponds to a situation where it is possible to design an experiment $\mathcal{E}_{max}$, so that its outcome is i when the true state of nature is s_i. Therefore, the mechanism MAX, described in Section 4.7, corresponds to a decision mechanism that would employ the outcome of $\mathcal{E}_{max}$, and select the retrieval approach a_i. In such a case, the retrieval effectiveness of the corresponding decision mechanism is the maximum that can be obtained by selectively applying one of the retrieval approaches $a_1, a_2, \ldots, a_n$ for each query.

The remainder of the chapter introduces a set of score-independent experiments (Section 5.3), and a set of score-dependent experiments (Section 5.4). Section 5.5 describes a Bayesian decision mechanism, and how it can be applied for selective Web IR.

5.3 Retrieval score-independent experiments

A wide range of experiments $\mathcal{E}$ can be defined, depending on the aim of the experiment and the employed sources of evidence. In the context of selective Web IR, the purpose of the experiment is to identify the queries for which a particular retrieval approach is more effective than other approaches. Since the different approaches may use evidence from the textual content of documents, as well as their structure and the hyperlink structure of the Web, it is reasonable to consider similar evidence in defining the experiments $\mathcal{E}$.

A first distinction of the possibilities for defining experiments $\mathcal{E}$ is whether scores or weights associated with ranking documents are used or not. In this context, the scores refer to either the scores assigned to documents by IR weighting models, such as the field-based weighting models described in Section 4.4. If such scores are not used, then a straight-forward way to define an experiment $\mathcal{E}$ is to consider whether query terms occur in documents. The current section investigates the latter approach in defining experiments $\mathcal{E}$. Section 5.3.1 introduces document-level experiments $\mathcal{E}$, which count the number of documents containing all or some of the query terms. Section 5.3.2 presents a refined set of experiments, where additional structural information from the distribution of documents in aggregates is considered.

5.3.1 Document-level experiments

Document-level experiments are based on whether query terms occur in documents. It is assumed that the broader topics are more widely covered in the collection. Therefore, there will be more documents that contain either all, or at least one of the query terms. For these topics, evidence from hyperlink analysis, or the URL of Web documents may be more useful in detecting high quality documents, or home pages of relevant Web sites.

For a given query, the outcome of the score-independent document-level experiments is related to the number of documents that satisfy a given condition. Several experiments can be defined for different conditions. For example, the condition $cond_{\forall}(d)$ that a document d in the sample Ret_q of the set of retrieved documents should contain all the terms of the query q is written as follows:

$$cond_{\forall}(d) : \forall t \in q \quad t \in d \quad d \in Ret_q \tag{5.2}$$

If at least one term of the query q is required to occur in the document d, then the condition $cond_{\exists}(d)$ is written as follows:

$$cond_{\exists}(d) : \exists t \in q \quad t \in d \quad d \in Ret_q \tag{5.3}$$

A range of more refined conditions can be defined when the fields of documents are considered. For example, a possible condition is that a document should contain at least one query term in its title field. In the case of documents with fields, the above conditions $cond_{\forall}(d)$ and $cond_{\exists}(d)$ are rewritten in order to distinguish between the occurrences of the same term in different fields. If $f(d)$ denotes the terms of d that appear in a particular field f, then the condition for checking whether all the query terms appear in any of the fields $f_1, \ldots, f_n$ of d is written as follows:

$$cond_{\forall}(d, f_1, \ldots, f_n) : \forall t \in q \quad t \in f_1(d) \vee \ldots \vee t \in f_n(d) \quad d \in Ret_q \tag{5.4}$$

The condition for checking whether at least one query term appears in any of the fields $f_1, \ldots, f_n$ of d is written as follows:

$$cond_{\exists}(d, f_1, \ldots, f_n) : \exists t \in q \quad t \in f_1(d) \vee \ldots \vee t \in f_n(d) \quad d \in Ret_q \tag{5.5}$$

The outcome o of the score-independent document-level experiments is computed as the number of documents for which a condition $cond_x(d)$ is true:

$$o = |\{d : cond_x(d) = true\}| \tag{5.6}$$

where x stands for either $\forall$ or $\exists$.

When documents with fields $f_1, \ldots, f_n$ are considered, the output of the experiments is computed as follows:

$$o = |\{d : cond_x(d, f_1, \ldots, f_n) = true\}| \tag{5.7}$$

For the rest of this thesis, the experiments that count the number of documents in Ret_q, with all the query terms, or at least one of them in a specific field f, will be denoted by $\mathcal{E}_{\forall(f)}$ or $\mathcal{E}_{\exists(f)}$, respectively. For example, the experiment that counts the number of documents with all the query terms in either the anchor text (*anchor*), or the title (*title*) fields is denoted by $\mathcal{E}_{\forall(at)}$. The outcome o of the experiment $\mathcal{E}_{\forall(at)}$ is computed as follows:

$$o = |\{d : cond_{\forall}(d, anchor, title) = true\}|$$

where

$$cond_{\forall}(d, anchor, title) : \forall t \in q \quad t \in anchor(d) \vee t \in title(d) \quad d \in Ret_q$$

The experiment that counts the number of documents with at least one query term in their body (*body*) is denoted by $\mathcal{E}_{\exists(b)}$, and its outcome o is computed as follows:

$$o = |\{d : cond_{\exists}(d, body) = true\}|$$

where

$$cond_{\exists}(d, body) : \forall t \in q \quad t \in body(d) \quad d \in Ret_q$$

The outcome o of the proposed experiments is an integer, ranging from 0 to the number of documents N in the collection. Plachouras, Ounis & Cacheda (2004) normalised the outcome values with $\min(\frac{o}{\alpha \cdot N}, 1)$. The current study primarily investigates the effectiveness of the different fields and the conditions for computing the outcome of the experiments. Therefore, the outcome of the score-independent document-level experiments is directly given by Equation (5.6), without any further normalisation.

Amitay et al. (2003) introduced a similar measure, the expected document frequency, which estimates the number of documents that contain all the query terms, by multiplying the probabilities of the query terms occurring in the collection. The underlying assumption is that the query terms are independent. In order to weaken the effect of this assumption, the expected document frequency was multiplied by the number of the query terms. The described experiments in this section compute the exact number of documents that satisfy a certain condition, allowing to consider or ignore the dependencies between the query terms. The computational cost of the document-level experiments is low, since the required information is available during retrieval.

5.3.2 Aggregate-level experiments

The proposed score-independent document-level experiments can be refined by considering additional structural information from the distribution of Web documents in *aggregates*. Indeed, the hypertext and the Web facilitate the organisation of related documents into aggregates. For example, in the case of the Web, the documents that belong to the same domain are likely to be about a particular topic, or a series of related topics. Therefore, they can be considered as an aggregate. This section investigates the

use of information from the distribution of Web documents in aggregates to define a range of experiments $\mathcal{E}$. This section introduces the aggregate-level experiments using abstract aggregates, and then it specifies how the aggregates are generated.

The underlying assumption for these experiments is that the distribution of documents in aggregates shows whether there exist large aggregates containing related documents, or whether the documents related to the topic are dispersed in different and unrelated aggregates. For example, evidence from the URL of Web documents or the hyperlink structure analysis may enhance the retrieval effectiveness, by identifying the entry points of large aggregates of documents.

The definition of the experiments $\mathcal{E}$ is based on the conditions introduced in Section 5.3.1. Indeed, by modifying the condition from Equation (5.3), the condition that at least one query term is required to appear in a document d from aggregate ag can be written as follows:

$$cond_{\exists}(d, ag) : \exists t \in q \quad d \in ag \quad t \in d \quad d \in Ret_q \tag{5.8}$$

The conditions (5.2), (5.3), (5.4), and (5.5) can be rewritten in the same way. The size of the aggregate ag is defined as follows:

$$|ag| = |\{d : cond_x(d, ag) = true\}| \tag{5.9}$$

where x corresponds to either $\exists$ or $\forall$.

Differently from the document-level experiments, the aggregate-level experiments are required to generate an outcome from the characteristics of the distribution of aggregate sizes. This work utilises three different characteristics of the distribution of aggregate sizes to generate the outcome of the experiments. The first two characteristics correspond to the average aggregate size $\overline{|ag|}$ and the standard deviation $\sigma_{|ag|}$ of the aggregate size distribution, respectively. The third characteristic of the aggregate size distribution is the number of large aggregates, which corresponds to the aggregates with size greater than $\overline{|ag|} + 2\sigma_{|ag|}$.

This work looks at two approaches to define aggregates. The first one is based on comparing the domain name of the URL of Web documents and aggregating the documents with the same domain name. This definition results in relatively broad aggregates, and it may not be appropriate to aggregate documents from Web sites that

Experiment	Aggregate Type	Distribution feature	Condition
$\mathcal{E}_{\exists(f),std(dom)}$	domain	$\sigma_{\lvert ag \rvert}$	$\exists t \in q \quad d \in ag \quad t \in f(d) \quad d \in Ret_q$
$\mathcal{E}_{\forall(f),avg(dom)}$	domain	$\overline{\lvert ag \rvert}$	$\forall t \in q \quad d \in ag \quad t \in f(d) \quad d \in Ret_q$
$\mathcal{E}_{\exists(f),lrg(dom)}$	domain	$\lvert\{ag : \lvert ag \rvert > \overline{\lvert ag \rvert} + 2\sigma_{\lvert ag \rvert}\}\rvert$	$\exists t \in q \quad d \in ag \quad t \in f(d) \quad d \in Ret_q$

Table 5.1: Notation examples for the aggregate-level experiments.

contain very diverse content. One such example is `http://www.geocities.com/`[1], which provides a free service for hosting Web sites. In this case, the fact that two Web documents appear in the same domain does not mean that they are about the same, or even a similar topic. On the other hand, aggregating Web documents by domain name is more appropriate when a different domain name is assigned to divisions, or departments of large organisations.

The second way to aggregate documents considers the directory under which the Web pages are stored. In this way, two Web documents, which are accessible through the URLs `http://a.b/d/e/y.html` and `http://a.b/d/e/z.html`, respectively, will be assigned to the same aggregate, but `http://a.b/d/x.html` will not. This approach partly overcomes the problem posed by Web sites such as `http://www.geocities.com/`, but it may result in large numbers of small aggregates. This approach is also more dependent on the way Web sites are organised. For example, Web sites with dynamically generated content by scripts may not have a useful directory structure related to the topics covered by documents.

Even though there are many ways to define aggregates of Web documents, such as clustering, the two introduced approaches provide a simple definition of aggregates. They also have the advantage that they identify aggregates by simply matching the string of the document URLs. This can take place during querying, by accessing a URL database for Web documents, or during indexing, by assigning an aggregate identifier to each document. The associated computational cost of the score-independent aggregate-level experiments is thus very low.

Table 5.1 summarises the notation that will be used for the aggregate-level experiments in the rest of this thesis. For example, the experiment that counts the average size of domain aggregates of documents that contain all query terms in the field f is denoted by $\mathcal{E}_{\forall(f),avg(dom)}$, as shown in the second row of the table.

[1] Visited on 11th August 2005.

5.4 Retrieval score-dependent experiments

Both the score-independent document-level and aggregate-level experiments depend solely on the occurrence of query terms in documents. Therefore, they are independent of any retrieval approach, or any score that is assigned to documents. However, not all the documents that contain a query term are relevant to a query. In addition, the outcome of the experiments that depend only on the occurrence of the query terms may be biased by frequent terms. The current section introduces experiments that employ the scores assigned to documents by a retrieval approach. This retrieval approach is not necessarily used for obtaining the final ranking of documents, as discussed in Section 5.2.

The introduced experiments employ a score distribution assigned to documents and transform it into a new score distribution, after a one-step propagation of the scores through the incoming hyperlinks of the documents, in order to favour the documents that point to other highly scored documents. The main underlying assumption of the experiments is that the difference between the two tested score distributions is related to the usefulness of evidence from the hyperlink structure of Web documents. For example, when there is a great difference between the two tested score distributions, employing additional evidence from the hyperlink structure analysis or the URL of documents may be more effective for retrieval.

The distribution of retrieval scores has been used in order to predict the effectiveness of a retrieval system. Manmatha et al. (2001) modelled the retrieval scores as a mixture of a Gaussian distribution for relevant documents, and an exponential distribution for non-relevant documents. The difference between the mean of the Gaussian distribution, and the point where the two distributions intersect indicates how well a system is expected to distinguish the relevant from the non-relevant documents. Cronen-Townsend et al. (2002) modelled the clarity of a query as the information theoretic divergence between the query language model and the collection language model. When the two language models are different, then the retrieval is expected to be effective. The defined experiments in this section are related to the approach of Cronen-Townsend et al. in the sense that both the clarity of a query and the introduced score-dependent experiments measure the difference or divergence between two probability distributions. Indeed, the introduced score-dependent experiments in this thesis focus on estimating

the difference between two score distributions (Section 5.4.1), after performing a one-step propagation of the document scores through the incoming hyperlinks of documents (Section 5.4.2).

5.4.1 Divergence between probability distributions

There are several different ways to estimate the divergence between probability distributions. A commonly used definition of information theoretic divergence between two probability distributions $P = \{p_i\}$ and $Q = \{q_i\}$ is the Kullback-Leibler divergence $I(P, Q)$ (Kullback, 1959):

$$I(P, Q) = \sum_i p_i \log_2 \frac{p_i}{q_i} \tag{5.10}$$

It easy to verify from Equation (5.10) that $I(P, Q) \neq I(Q, P)$, or in other words, that the Kullback-Leibler divergence is not symmetric. Following from this, the symmetric Kullback-Leibler divergence $J(P, Q)$ is defined as the sum of the divergences $I(P, Q)$ and $I(Q, P)$ (Kullback, 1959):

$$\begin{aligned} J(P, Q) &= I(P, Q) + I(Q, P) \\ &= \sum_i (p_i - q_i) \cdot \log_2 \frac{p_i}{q_i} \end{aligned} \tag{5.11}$$

From the above Equations (5.10) and (5.11), it can be seen that $I(P, Q) \geq 0$ and $J(P, Q) \geq 0$, respectively. Both $I(P, Q)$ and $J(P, Q)$ are equal to zero if and only if the distributions P and Q are equivalent. However, note that there is no upper bound for the values of $I(P, Q)$ and $J(P, Q)$.

This is addressed by the Jensen-Shannon divergence (Lin, 1991), which corresponds to the Kullback-Leibler divergence of the probability distribution P from the average of the probability distributions P and Q, as follows:

$$\begin{aligned} K(P, Q) &= I(P, \frac{1}{2} \cdot P + \frac{1}{2} \cdot Q) \\ &= \sum_i p_i \log_2 \frac{p_i}{\frac{1}{2} \cdot p_i + \frac{1}{2} \cdot q_i} \end{aligned} \tag{5.12}$$

The symmetric Jensen-Shannon divergence is defined as follows:

$$\begin{aligned} L(P, Q) &= K(P, Q) + K(Q, P) \\ &= \sum_i p_i \log_2 \frac{p_i}{\frac{1}{2} \cdot p_i + \frac{1}{2} \cdot q_i} + \sum_i q_i \log_2 \frac{q_i}{\frac{1}{2} \cdot p_i + \frac{1}{2} \cdot q_i} \end{aligned} \tag{5.13}$$

One of the properties of this measure of divergence is that there exists an upper bound for its value, $L(P,Q) \leq 2$ (Lin, 1991). The symmetric Jensen-Shannon divergence is also known as *total divergence from the average* (Pirolli & Pitkow, 1999), and it is a special case of the weighted Information Radius (Jardine & Sibson, 1971, page 13):

$$\sum_i \frac{w_p p_i}{w_p + w_q} \cdot \log_2 \frac{p_i \cdot (w_p + w_q)}{w_p p_i + w_q q_i} + \sum_i \frac{w_q q_i}{w_p + w_q} \cdot \log_2 \frac{q_i \cdot (w_p + w_q)}{w_p p_i + w_q q_i} \tag{5.14}$$

In the above formula, w_p and w_q are the weights of the probability distributions $\{p_i\}$ and $\{q_i\}$, respectively. Indeed, the information radius of two probability distributions with the same weights is equal to half their symmetric Jensen-Shannon divergence.

5.4.2 Usefulness of hyperlink structure

The current section defines an experiment based on measuring the divergence between the score distribution of a retrieval approach, and a modified score distribution, obtained after a one-step propagation of scores through the incoming hyperlinks of Web documents. The underlying assumption is that if there are non-random patterns of hyperlinks among the retrieved Web documents for a particular query, then the divergence between the original and the modified distributions of document scores will be higher. This suggests that the hyperlink structure is more useful, or, in other words, that the use of structural evidence may be more effective for retrieval.

The usefulness of the hyperlink structure of a sample Ret_q of the set of retrieved documents is defined as the information theoretic divergence between two probability distributions. The first one is the distribution S of the scores sc_i for the documents $d_i \in Ret_q$. The scores sc_i can be the relevance scores assigned to documents by any of the retrieval approaches introduced in Chapter 4, or a query-independent source of evidence, such as PageRank or the Absorbing Model. In the remainder of this thesis, the scores sc_i correspond to the relevance scores assigned by a particular retrieval approach. The second distribution is constructed so as to favour the highly scored documents that point to other highly scored documents in Ret_q. This is a desired property under the assumption that it is more useful for a user, who is browsing a highly scored document, to be able to access other highly scored documents by following hyperlinks. The new distribution U of scores u_i is defined as follows. The score u_i for document $d_i \in Ret_q$

depends on the score sc_i, as well as on the scores sc_j of all the retrieved documents d_j, which are pointed to by d_i:

$$u_i = sc_i + \sum_{d_i \to d_j} sc_j \quad d_i, d_j \in Ret_q \tag{5.15}$$

where $d_i \to d_j$ means that there exists a hyperlink from d_i to d_j.

The measures of divergence introduced in the previous section estimate the difference between two probability distributions. However, the score distributions S and U do not necessarily correspond to probability distributions. Indeed, the Divergence From Randomness (DFR) weighting models and their field-based extensions rank documents according to the divergence of the occurrences of a term in a document from a random distribution. The resulting scores are in the range $(0, +\infty)$. The scores assigned to documents by the weighting model BM25 and its extension BM25F also fall within the same range. Therefore, it is necessary to normalise the retrieval scores from $(0, +\infty)$ to $(0, 1]$. Nottelmann & Fuhr (2003) compared linear and sigmoid functions to transform the document scores into probabilities of relevance. For simplicity, and in order to reduce the number of the introduced parameters, the scores are normalised by dividing them with their sum. In this case, the normalised scores are in the range $(0, 1]$ and their sum is equal to 1:

$$sn_i = \frac{sc_i}{\sum_{d_j \in Ret_q} sc_j} \quad un_i = \frac{u_i}{\sum_{d_j \in Ret_q} u_j} \tag{5.16}$$

The distribution $U = \{u_i\}$ has been defined in order to favour the highly scored documents d_i that point to other highly scored documents. According to Equation (5.15), the score $u_i \geq sc_i > 0$. Therefore, highly scored documents, which do not point to any other documents, still have a high score. In order to favour only those documents that point to other highly scored documents, a new distribution $U' = \{u_i\}$ is defined as follows:

$$u'_i = \sum_{d_i \to d_j} sc_j \quad \text{and } un'_i = \frac{u'_i}{\sum_{d_j \in Ret_q} u'_j}. \tag{5.17}$$

The normalised distribution $\{un'_i\}$ is denoted by U'_n. The distribution $\{u'_i\}$ differs from $\{u_i\}$ in the sense that the dependence $u_i \geq sc_i$ is removed. For the distribution $\{u'_i\}$, it is easy to verify that if a document d_i does not have outgoing links, then $u'_i = 0$. If d_i points to documents with low scores, it may be the case that $0 < u'_i < sc_i$. Therefore,

the dependence of $\{u_i\}$ on $\{sc_i\}$ is stronger than the dependence of $\{u'_i\}$ on $\{sc_i\}$, and the hyperlink structure among the documents in Ret_q is expected to have a greater impact on the distribution $\{u'_i\}$.

Having defined the score distributions $S_n = \{sn_i\}$, $U_n = \{un_i\}$, and $U'_n = \{un'_i\}$, the usefulness of the hyperlink structure is estimated as the symmetric Jensen-Shannon divergence between the normalised distributions S_n and U_n:

$$L(S_n, U_n) = \sum_{d_i \in Ret_q} un_i \log_2 \frac{un_i}{\frac{un_i}{2} + \frac{sn_i}{2}} + \sum_{d_i \in Ret_q} sn_i \log_2 \frac{sn_i}{\frac{un_i}{2} + \frac{sn_i}{2}} \tag{5.18}$$

or as the symmetric Jensen-Shannon divergence between the normalised distributions S_n and U'_n

$$L(S_n, U'_n) = \sum_{d_i \in Ret_q} un'_i \log_2 \frac{un'_i}{\frac{un'_i}{2} + \frac{sn_i}{2}} + \sum_{d_i \in Ret_q} sn_i \log_2 \frac{sn_i}{\frac{un'_i}{2} + \frac{sn_i}{2}} \tag{5.19}$$

The usefulness of the hyperlink structure is defined using the symmetric Jensen-Shannon divergence, instead of the symmetric Kullback-Leibler divergence, because the values of the former are in the range $[0, 2]$. An additional reason for employing the symmetric Jensen-Shannon divergence is that the two probability distributions do not have to be mutually absolutely continuous, as it is the case for the Kullback-Leibler divergence. This means that the Kullback-Leibler divergence is defined only for probability distributions for which $sn_i = 0$, for all i, for which $un_i = 0$, and *vice versa*. In the case of the distributions S_n and U_n, this condition is satisfied, because the definition of the distribution U from Equation (5.15) suggests that $u_i \geq sc_i > 0$, and consequently $un_i > 0$ and $sn_i > 0$ for all $d_i \in Ret_q$. However, the Kullback-Leibler divergence cannot be defined for the distributions S_n and U'_n, because un'_i can be 0 even if $sn_i > 0$. Therefore, the symmetric Jensen-Shannon divergence is more appropriate to use in the context of selective Web IR. Note that, the Jensen-Shannon divergence has been used in the context of pattern recognition to measure the distance between random graphs (Wong & You, 1985).

Before continuing, an example is provided in order to illustrate the introduced experiments, and to show how the divergence between the distributions S_n and U_n or U'_n depends on the scores and the hyperlink structure of the retrieved documents.

Example 6 Suppose that for a particular query, six documents, numbered from 1 to 6, have been retrieved and ranked according to the scores $\{s_i\} = \{0.9, 0.4, 0.3, 0.2, 0.2, 0.1\}$,

i.e., the score of document 1 is 0.9, the score of document 2 is 0.4, and so on. The six documents are connected with hyperlinks, as shown in Figure 5.2. The divergences $L(S_n, U_n)$ and $L(S_n, U'_n)$ are computed for the three graphs of hyperlinks shown in the figure, as well as for a fourth case of a complete graph, where there is a hyperlink between any ordered pair of documents.

The first graph, shown in Figure 5.2(a), corresponds to a case, where there is no apparent pattern in the way the hyperlinks are distributed. After the distributions U_n and S_n are computed, $L(S_n, U_n) = 0.0728$ and $L(S_n, U'_n) = 0.6875$. The second graph of hyperlinks, in Figure 5.2(b), corresponds to a case, where the top three ranked documents are strongly connected. In the same way as before, it is easy to compute that $L(S_n, U_n) = 0.1226$ and $L(S_n, U'_n) = 0.4273$. For the third graph, shown in Figure 5.2(c), there is a group of documents that are strongly connected, without all of them being highly ranked. In this case, $L(S_n, U_n) = 0.2167$ and $L(S_n, U'_n) = 0.9386$. For the last case, suppose that the graph of the example is complete, in that it contains one hyperlink between each and every ordered pair of documents. In this case, $L(S_n, U_n) = 0.1675$ and $L(S_n, U'_n) = 0.2485$.

The divergence $L(S_n, U_n)$ has the lowest value when there is no apparent structure in the way hyperlinks are distributed, and increases its value when there is a connected group of documents. The increase is higher if the documents from the connected group are ranked lower in the list of documents. In this case, it is assumed that the information from the hyperlink structure is more useful. The computed values of $L(S_n, U'_n)$ are higher than those of $L(S_n, U_n)$, because the distribution U'_n is less dependent on S_n than the distribution U_n, as discussed above. The divergence $L(S_n, U_n)$ for the graph (a) is higher than $L(S_n, U_n)$ for the graph (b), while the divergence $L(S_n, U'_n)$ for the graph (a) is lower than $L(S_n, U'_n)$ for the graph (b). This fact indicates that $L(S_n, U_n)$ and $L(s_n, U'_n)$ can be used to define two experiments $\mathcal{E}$, which are not equivalent. Moreover, the fact that the complete graph does not result in the highest divergence for both $L(S_n, U_n)$ and $L(S_n, U'_n)$ indicates that the usefulness of the hyperlink structure does not depend only on the number of hyperlinks. □

The usefulness of the hyperlink structure has been defined by using a one-step propagation of scores through the incoming hyperlinks of documents. An n-step propagation would result in a weaker dependence between either of the score distributions

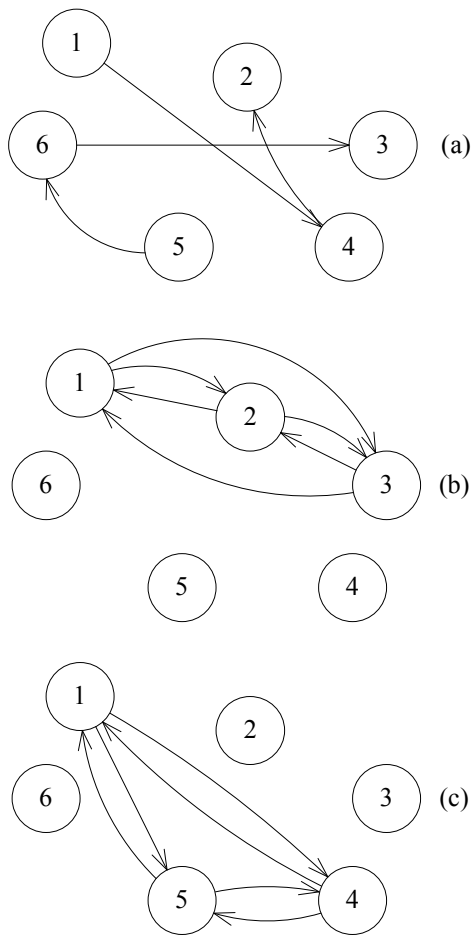

Figure 5.2: The hyperlink graphs of the ranked documents, corresponding to the first three cases described in the Example 6.

U_n or U'_n, and the initial distribution S_n. In addition, the computational overhead of computing U_n and U'_n would increase with every step.

In the remainder of this thesis, the experiments $\mathcal{E}$ that employ either $L(S_n, U_n)$ or $L(S_n, U'_n)$, when considering the documents with all the query terms in a combination of fields f, are denoted by $\mathcal{E}_{\forall(f),L(SU)_{wm}}$ and $\mathcal{E}_{\forall(f),L(SU')_{wm}}$, respectively. In this notation, wm stands for a scoring technique that assigns the score distribution S to documents. This scoring technique can be any of the retrieval approaches described in Chapter 4. When considering documents with at least one query term in a particular combination of fields, the experiments are denoted by $\mathcal{E}_{\exists(f),L(SU)_{wm}}$ and $\mathcal{E}_{\exists(f),L(SU')_{wm}}$, respectively.

5.5 Bayesian decision mechanism

A range of experiments $\mathcal{E}$ has been defined in the Sections 5.3 and 5.4, using different sources of evidence. In the context of a decision mechanism, the effectiveness of these experiments depends on how successful they are in detecting the true state of nature, and hence, in identifying the most appropriate retrieval approach to use for each given query. The current section defines a Bayesian decision mechanism (Section 5.5.1), and discusses how it can be applied for selective Web IR (Sections 5.5.2 and 5.5.3).

5.5.1 Definition of the Bayesian decision mechanism

The Bayesian decision mechanism is defined as follows. Suppose that there are k available retrieval approaches and r states of nature, where $k = r$. For each state of nature s_i, the retrieval approaches are ordered according to an evaluation measure m, as described in Section 5.2. In this way, the most effective retrieval approach a_i for the state of nature s_i ($m(a_i, s_i) > m(a_j, s_i)$) corresponds to a loss in utility equal to $l(a_i, s_i) = 0$, while the other retrieval approaches have a higher loss of utility $l(a_j, s_i)$. The least effective retrieval approach corresponds to a loss of utility equal to 1. Each state of nature s_i has a prior probability $P(s_i)$ of being the true state of nature for a particular query. This prior probability $P(s_i)$ is defined as the number of queries for which $m(a_i) > m(a_j), i \neq j$. In other words, the prior probability of a state of nature s_i depends on the number of queries for which the corresponding retrieval approach a_i is the most effective among the k retrieval approaches used in the decision mechanism. The Bayesian decision mechanism employs the experiment $\mathcal{E}$ in order to make an informed guess about the true state of nature. The conditional probability $P(o|s_i)$ denotes the probability of obtaining o as the outcome of the experiment $\mathcal{E}$, when the state of nature is s_i.

According to the Bayes decision rule, the posterior probability that s_i is the true state of nature depends on the prior probability $P(s_i)$ of s_i and the evidence from the outcome o of the experiment $\mathcal{E}$, as follows:

$$P(s_i|o) = \frac{P(s_i) \cdot P(o|s_i)}{P(o)} \tag{5.20}$$

where:

$$P(o) = \sum_{i=1}^{k} P(s_i) \cdot P(o|s_i) \tag{5.21}$$

Then, for each action a_i, the expected loss $E[l(a_i)]$ for all the states of nature is given by[1]:

$$E[l(a_i)] = \sum_{j=1}^{k} l(a_i, s_j) \cdot P(s_j|o) \tag{5.22}$$

The Bayesian decision mechanism selects the retrieval approach a_i with the minimum expected loss $E[l(a_i)]$. This mechanism is optimal in the sense that it minimises the average classification error (Duda & Hart, 1973), or, in other words, the expected loss. In the case of selective Web IR, this is desirable in order to evaluate the effectiveness of the employed experiment $\mathcal{E}$ in identifying the true state of nature, and applying an appropriate retrieval approach.

It is important to note that the denominator $P(o)$ in Equation (5.20) is a constant, and it is used as a normalisation factor in order to obtain probabilities. In the context of selective Web IR, the objective of the decision mechanism is to select a retrieval approach to apply. Therefore, the denominator $P(o)$ can be ignored, without affecting the selection of the retrieval approach to apply for a particular query. The use of the Bayesian decision mechanism is illustrated in the following example.

Example 7 Suppose that a decision mechanism selects one retrieval approach from three available ones: a_1, a_2, and a_3. The decision mechanism performs an experiment $\mathcal{E}$, for which the posterior likelihoods $P(s_j|o)$ are shown in the upper diagram of Figure 5.3. The loss $l(a_i, s_j)$ associated with applying a_i when the state of nature is s_j is specified in the matrix $[l_{ij}]$, where $l_{ij} = l(a_i, s_j)$:

$$[l_{ij}] = \begin{bmatrix} 0.0 & 1.0 & 0.5 \\ 1.0 & 0.0 & 1.0 \\ 0.5 & 0.5 & 0.0 \end{bmatrix}$$

The lower diagram in Figure 5.3 shows the expected loss from applying each of the three retrieval approaches a_1, a_2, and a_3, as computed from Equation (5.22). In this diagram, the intersections of the loss curves define the decision boundaries, that is the outcome values of the experiment $\mathcal{E}$, which serve as thresholds for selecting one of the retrieval approaches a_i. For example, if the experiment $\mathcal{E}$ results in outcome $o < o_1$, then the retrieval approach a_3 is applied, because it results in the lowest expected loss

[1]Note that $l(a_i, s_j)$ denotes the loss from selecting the action a_i when the true state of nature is s_j, while $E[l(a_i)]$ denotes the expected loss from selecting the action a_i.

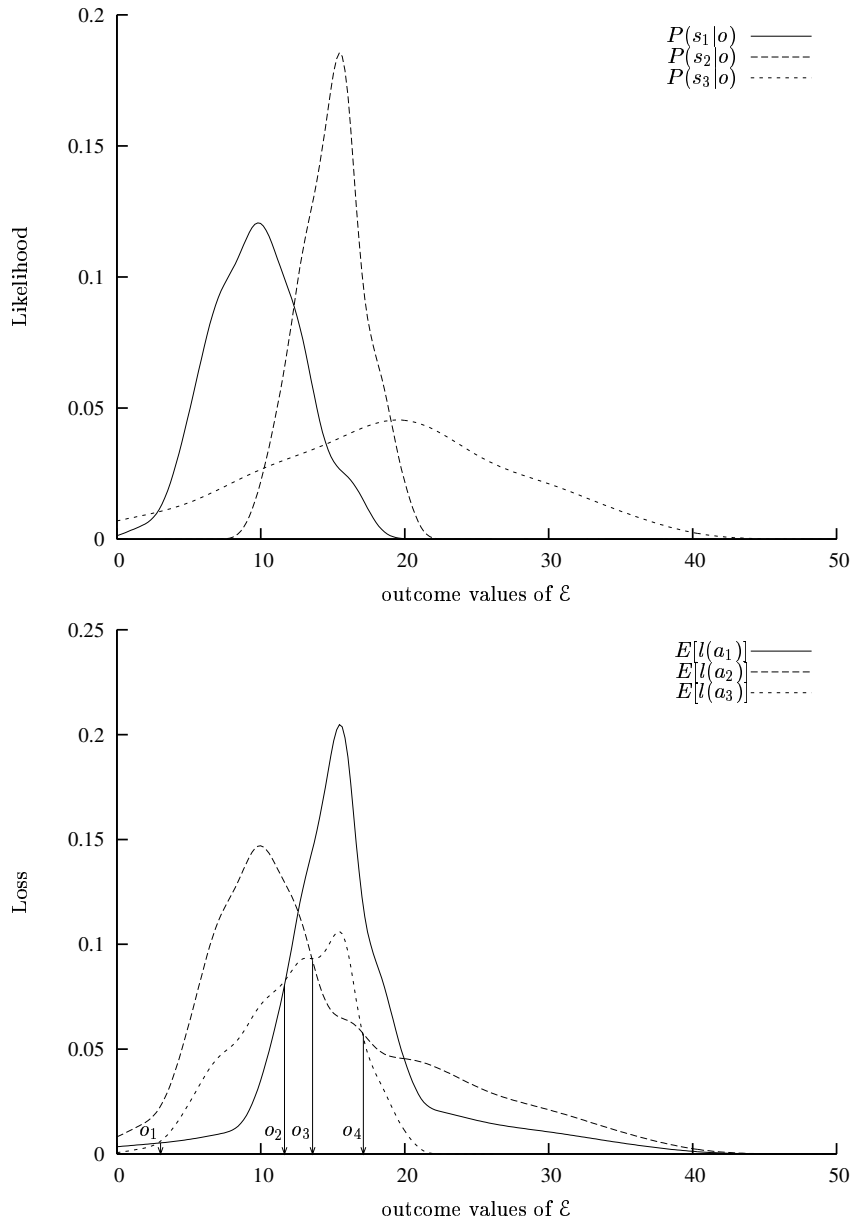

Figure 5.3: Example of a Bayesian Decision mechanism with 3 available retrieval approaches and three states of nature. The upper diagram shows the estimated densities of the posterior likelihoods for each state of nature. The lower diagram shows the corresponding loss curves $E[l(a_i)]$ for each retrieval approach. The outcome values o_1, o_2, o_3, and o_4 of $\mathcal{E}$ corresponding to the intersection points of the loss curves represent the decision boundaries of the decision mechanism.

$E[l(a_3)]$. If $o_1 < o < o_2$, then the expected loss $E[l(a_1)]$ is lower than both $E[l(a_2)]$ and $E[l(a_3)]$. Therefore, the retrieval approach a_1 is applied. In a similar way, the retrieval approach a_3 is applied when $o_2 < o < o_3$, the retrieval approach a_2 is applied when $o_3 < o < o_4$, and the retrieval approach a_3 is applied when $o_4 < o$. In this way, the decision mechanism selects a particular retrieval approach for every possible outcome of the experiment $\mathcal{E}$.

A decision mechanism that selects one out of two retrieval approaches is a special case, where the above description can be simplified. Indeed, in the case of two retrieval approaches, or equivalently two states of nature, the output of the loss function is binary:

$$l(a_i, s_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Therefore, $E[l(a_1) = l(a_1, s_2) \cdot P(s_2|o) = P(s_2|o)$ and $E[l(a_2)] = l(a_2, s_1) \cdot P(s_1|o) = P(s_1|o)$. The decision mechanism applies retrieval approach a_1 when $P(s_2|o) < P(s_1|o)$, or, in other words, when the posterior likelihood of s_1 is greater than that of s_2. Generally, selecting one out of k retrieval approaches can always be mapped to a series of $k-1$ decisions between two retrieval approaches.

5.5.2 Application of the Bayesian decision mechanism

This section discusses the application of the Bayesian decision mechanism in order to perform selective Web IR. It describes how to estimate the required quantities of the decision mechanism with respect to a set of training data.

The application of the Bayesian decision mechanism requires the estimation of three quantities:

- The prior probability $P(s_i)$ that a state of nature is the true state of nature. This corresponds to the prior probability that the retrieval approach a_i is the most effective. When performing selective Web IR with the Bayesian decision mechanism, the prior probability $P(s_i)$ is set equal to the proportion of the number of training queries, for which the retrieval approach a_i is the most effective.

- The loss $l(a_j, s_i)$ associated with the application of the retrieval approach a_j when the true state of nature is s_i. When a set of training queries is available, the loss $l(a_j, s_i)$ is defined as the difference between the retrieval effectiveness of a_i and

that of a_j, for the subset of training queries for which a_i is the most effective retrieval approach.

- The probability $P(o|s_i)$ that the outcome of an experiment $\mathcal{E}$ is o when the state of nature is s_i. This probability is computed by estimating the density of the outcome values of the experiment $\mathcal{E}$, for the subset of training queries, for which the retrieval approach a_i is the most effective. A more detailed discussion about the density estimation is given in Section 5.5.3

5.5.3 Density estimation

The last point of discussion with respect to the Bayesian decision mechanism is related to the density estimation of $P(o|s_j)$ from the outcome o of the experiment $\mathcal{E}$ for a number of queries. Bishop (1995, Chapter 2) identifies three main types of density estimation techniques: parametric methods that assume a certain functional form for the estimated density, non-parametric methods that allow the available data to completely specify the estimated density, and semi-parametric methods such as mixture models. A disadvantage of the parametric methods is that it may be difficult to find an appropriate functional form for the estimated density. Although non-parametric methods alleviate this disadvantage, the complexity of the estimated density depends on the number of available data points. Mixture models do not assume a particular functional form for the estimated density, and result in less complex models, but they are computationally expensive. In this thesis, the density estimation of $P(o|s_j)$ is performed using non-parametric methods, because of the relatively small amount of training data. In particular, a Gaussian kernel-based density estimation technique, with automatic setting of the bandwidth (Silverman, 1986), is employed[1].

Due to the limited amount of training data, it is necessary to pay special attention to the existence of outliers in the available experiment outcome values. Figure 5.4 shows the box-and-whisker plots (Chambers et al., 1983) of the outcome values for the score-independent document-level experiment, computed for the task td2003. In each plot, the ends of each box correspond to the first and third quartiles of the distribution of experiment outcome values. The bold line corresponds to the median of the distribution. The whiskers extend to the farthest points that are within 3/2 times the

[1] The density estimation was performed with the software package R: A language and environment for statistical computing (R Development Core Team, 2005).

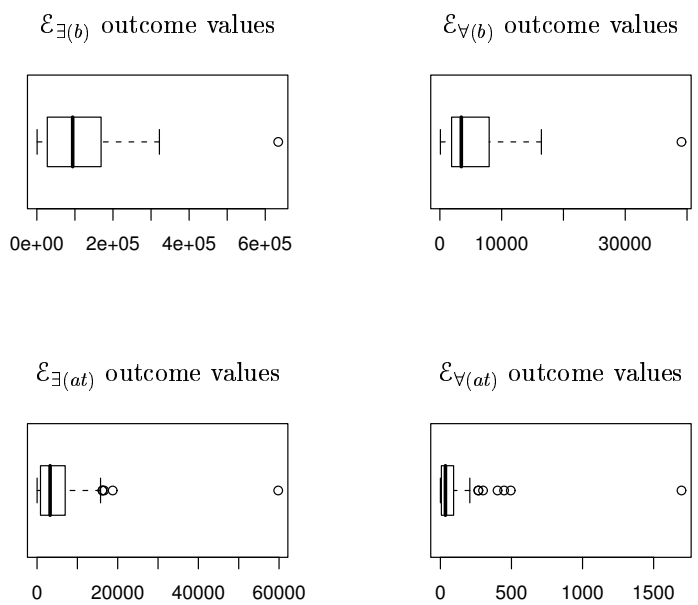

Figure 5.4: Box-and-whisker plots of the score-independent document-level experiment outcome values for the task td2003.

interquartile range of the first and third quartiles. Any points that are farther than the whiskers are considered to be outliers, and they are denoted with a circle.

The top left box-and-whisker plot shows that there is one outlier among the outcome values of the experiment $\mathcal{E}_{\exists(b)}$ for the task td2003, corresponding to the query TD39: `national public tv radio`. This is due to the very high document frequency of the query terms `national` and `public` in the .GOV test collection, resulting in 634,053 documents with at least one query term in their body. The outcome values for the experiments $\mathcal{E}_{\forall(b)}$, $\mathcal{E}_{\exists(at)}$, and $\mathcal{E}_{\exists(at)}$ are lower than those of $\mathcal{E}_{\exists(b)}$, but there exist outliers in all cases. More specifically, the experiments $\mathcal{E}_{\exists(at)}$ and $\mathcal{E}_{\forall(at)}$ result in more outliers that the experiments $\mathcal{E}_{\forall(b)}$ and $\mathcal{E}_{\exists(b)}$, because the obtained outcome values depend on the distribution of hyperlinks with the query terms in the associated anchor text: there are many distinct anchor texts associated with few hyperlinks, while there are only few anchor texts associated with many hyperlinks. The density estimation is performed for the range of obtained outcome values that lie within 3/2 times the interquartile range of the first and third quartiles.

In the next chapter, the Bayesian decision mechanism will be employed to evaluate the proposed score-independent (Section 5.3), and score-dependent experiments (Section 5.4) in a setting where relevance information is assumed to exist.

5.6 Summary

This chapter has introduced a novel framework for selective Web IR. The framework is formulated in terms of statistical decision theory (Section 5.2). One of its main concepts is the decision mechanism, which selects one retrieval approach from a set of available ones on a per-query basis. The selection of the applied retrieval approach is aided by the experiment $\mathcal{E}$, which extracts a feature from a sample of the set of retrieved documents.

The introduced framework for selective Web IR is different from the related work in several aspects (Section 5.2.1). First, it differs from query-type classification, because the aim is to apply an appropriate retrieval approach on a per-query basis, instead of a particular retrieval approach for each query-type. Second, it differs from the dynamic adjustment of the weights of each source of evidence, because each retrieval approach is assumed to be fixed. Third, the introduced framework is more general than query-performance prediction, which primarily estimates the correlation of a predictor with the effectiveness of a retrieval approach. Selective Web IR aims to predict the difference in the retrieval effectiveness between several retrieval approaches.

Several experiments $\mathcal{E}$ have been defined. Section 5.3 introduces a range of experiments based on counting the occurrences of query terms in documents, or in particular fields of documents. These documents are called score-independent, because they do not consider any score assigned to documents. The score-independent document-level experiments count the number of documents with at least one, or all query terms. The score-independent aggregate-level experiments consider the structural information from the distribution of documents in aggregates of related documents. The aggregates correspond to the documents that belong to the same domain, or the documents that are stored in the same directory.

Section 5.4 has presented a range of experiments based on estimating the usefulness of the hyperlink structure for a sample of the retrieved documents. These experiments

are called score-dependent, because they compute the information theoretical divergence between two score distributions. The first one is the score distribution assigned to documents by a particular retrieval approach, such as a field-based weighting model. The second score distribution is obtained after a one-step propagation of the document scores through their incoming hyperlinks.

The Bayesian decision mechanism defined in Section 5.5 provides a means for the evaluation of the proposed experiments $\mathcal{E}$, by applying a retrieval approach with the minimum expected loss. The Bayesian decision mechanism can be used to select one retrieval approach from any number of available ones. The estimation of the likelihoods that a particular retrieval approach is appropriate is carefully performed, by considering the fact that there may be outliers in the obtained outcomes of an experiment $\mathcal{E}$.

Overall, the introduced framework for selective Web IR represents a general approach to the problem of identifying appropriate retrieval approaches to apply on a per-query basis. The remainder of this thesis focuses on evaluating the effectiveness of the proposed experiments in identifying the most appropriate retrieval approaches to apply on a per-query basis. Chapter 6 evaluates the proposed experiments in a setting, where it is assumed that relevance information exists. Chapter 7 investigates the evaluation of the proposed experiments in a more realistic setting, where it is assumed that limited relevance information exists.

Chapter 6

Evaluation of Selective Web Information Retrieval

6.1 Introduction

The potential for improvements in retrieval effectiveness from selective Web IR has been established in Chapter 4. Furthermore, Chapter 5 has proposed a new framework for selective Web IR, which employs a range of experiments $\mathcal{E}$. The current chapter aims to evaluate the proposed framework, and to establish the effectiveness of the introduced experiments $\mathcal{E}$ in an setting, where relevance information is assumed to exist.

This chapter starts with Section 6.2, which introduces the evaluation methodology for the experiments $\mathcal{E}$. Each experiment $\mathcal{E}$ is evaluated in the context of a Bayesian decision mechanism, which selectively applies two retrieval approaches on a per-query basis, assuming that there exists relevance information. The two retrieval approaches are chosen according to their potential for improvements from selective Web IR, and employ different field-based weighting models, as described in Chapter 4. An example of a Bayesian decision mechanism, which selectively applies three retrieval approaches on a per-query basis is also provided later in this chapter.

Section 6.3 discusses the evaluation of the score-independent experiments $\mathcal{E}$, which were proposed in Section 5.3. These experiments include both the document-level, as well as the domain, and directory aggregate-level experiments $\mathcal{E}$. Next, Section 6.4 presents the evaluation of the score-dependent experiments, which estimate the usefulness of the hyperlink structure.

The chapter continues with Section 6.5, where the proposed experiments are computed from small samples of documents, in order to reduce the associated computational overhead, and to assess whether highly scored documents are more useful for computing the outcome of experiments. Section 6.6 discusses the evaluation of the experiments $\mathcal{E}$, when the decision mechanism selects between retrieval approaches, which employ the same field-based weighting models. Section 6.7 investigates an example of a Bayesian decision mechanism, which uses more than two retrieval approaches. The chapter closes with a discussion of the findings in Section 6.8.

6.2 Evaluation methodology

The aim of this section is to introduce the evaluation methodology that will be used for the remainder of the chapter. First, it describes how the effectiveness of an experiment $\mathcal{E}$ will be evaluated. Next, it defines the experimental setting, in which a Bayesian decision mechanism, as discussed in Section 5.5, employs the proposed score-independent and score-dependent experiments to perform selective retrieval. This section closes with a brief description of the presentation of the results in the remainder of the chapter.

6.2.1 Effectiveness of experiments $\mathcal{E}$

This section discusses issues related to the evaluation of the proposed experiments $\mathcal{E}$. The effectiveness of an experiment $\mathcal{E}$ is evaluated with respect to the number of decision boundaries used in a decision mechanism, the achieved mean average precision (MAP) by the decision mechanism, and whether the correct decision is made for a statistically significant number of queries.

As discussed in Section 5.5, in the context of a Bayesian decision mechanism, which employs an experiment $\mathcal{E}$, the decision boundaries correspond to the intersection points of the curves of expected loss for each of the employed retrieval approaches. An effective experiment $\mathcal{E}$ should result in a different distribution of the expected loss for each retrieval approach. In such a case, the number of intersection points between the curves of expected loss is likely to be low. However, if the experiment $\mathcal{E}$ does not result in a different distribution of the expected loss for each retrieval approach over the range of outcome values of $\mathcal{E}$, the number of intersection points between the curves is expected to be high. If there are no intersection points between the curves of expected

loss, because the loss of one retrieval approach is always lower than that of the other retrieval approaches, then the decision mechanism cannot selectively apply different retrieval approaches on a per-query basis. In such a case, the experiment $\mathcal{E}$ is considered to be less effective for selective Web IR. The same discussion applies to the case of a Bayesian decision mechanism, which employs two retrieval approaches, and selects the one with the higher posterior likelihood to be the most effective retrieval approach, as described in Section 5.5.

The application of the most appropriate retrieval approach by a decision mechanism on a per-query basis should have a positive impact on MAP, compared to the MAP of the individual retrieval approaches. Therefore, the effectiveness of an experiment should be reflected on the resulting MAP of the decision mechanism. The most effective experiments should result in improvements in MAP similar to that obtained by the hypothetical experiment $\mathcal{E}_{max}$, which always applies the most effective retrieval approach on a per-query basis (Section 5.2.2). The Wilcoxon's signed rank test is used to indicate whether the difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach is statistically significant.

The resulting MAP is not the only indication of the experiment's effectiveness. If the employed retrieval approaches have similar performance for a query, then applying the most effective retrieval approach for that particular query is not expected to have an important impact on the effectiveness of the decision mechanism. In order to take this issue into account, the sign test (Hoel, 1984; Siegel & Castellan, 1988) is used to denote whether the most appropriate retrieval approach is applied for a statistically significant number of queries.

6.2.2 Evaluation setting

This section provides an overview of the evaluation setting. It briefly describes the employed retrieval approaches, and their corresponding notations, as it has been introduced in Chapter 4. Next, it describes the setting of the Bayesian decision mechanism, which is used for the evaluation of the proposed experiments $\mathcal{E}$.

6.2.2.1 Description of retrieval approaches setting

Selective Web IR can be performed with any retrieval approach. In this thesis, the employed retrieval approaches correspond to either one of the field-based weighting

models (PL2F, PB2F, I(n_e)C2F, DLHF, and BM25F) (Section 4.4 on page 67), or their combination with query-independent sources of evidence (Section 4.5 on page 74). The employed fields are: the body; the anchor text of incoming hyperlinks; and the title. Compared to the original weighting models, the field-based weighting models are preferred, because they provide important gains in retrieval effectiveness for Web specific search tasks, as shown in Chapter 4.

The employed query-independent sources of evidence are the URL path length (Section 4.5.1 on page 74), PageRank (Brin & Page, 1998), and the novel Absorbing Model with static priors (Section 4.5.2.4 on page 83, and Section 4.5.2.5 on page 86, respectively). Their combination with a field-based weighting model is denoted by appending the letters U, P, and A, respectively, at the end of the weighting model's name. For example, the combination of PL2F with PageRank is denoted by PL2FP, and the combination of I(n_e)C2F with the Absorbing model is denoted by I(n_e)C2FA. Each field-based weighting model is combined with one source of query-independent evidence, in order not to further increase the number of hyper-parameters in the retrieval approaches, as described in Section 4.5.3.

The hyper-parameters of the employed retrieval approaches have been set, as described in Section 4.6, page 92. The mean average precision (MAP) of each retrieval approach is directly optimised for a mixed task. The optimisation process is terminated after 20 iterations, so that the hyper-parameters do not necessarily converge to their optimal values. The obtained hyper-parameter values from the above training process are used in order to evaluate the same retrieval approach with other tasks than the training ones. For example, a retrieval approach is optimised for the mixed task mq2004, and then it is evaluated for the tasks td2003, hp2003, or np2003. The evaluation results of the employed retrieval approaches are displayed in Table 4.10, page 96.

6.2.2.2 Description of Bayesian decision mechanism setting

The Bayesian decision mechanism, which has been described in Section 5.5, is used to perform the evaluation of the proposed experiments $\mathcal{E}$. The employed tasks are: td2003; td2004; hp2003; hp2004; np2003; and np2004. The training of the decision mechanism, and the evaluation of the experiments $\mathcal{E}$ are performed with the same task. This setting has been chosen in order to reduce any effect on the evaluation of the experiments from the differences among the employed tasks. Chapter 7 discusses the

evaluation of the experiments $\mathcal{E}$ in a setting with limited relevance information, where different mixed tasks are employed for the training of the Bayesian decision mechanism, and the evaluation of the experiments.

In order to obtain a clear indication about the effectiveness of the evaluated experiments $\mathcal{E}$, the employed retrieval approaches by the Bayesian decision mechanism correspond to the ones with the highest potential for improvements in retrieval effectiveness, as discussed in Section 4.7, and presented in Table 4.11. More specifically, the Bayesian decision mechanism employs pairs of retrieval approaches, which use different field-based weighting models. For ease of reference, Table 6.1 presents the evaluation of the selected pairs of retrieval approaches. This setting is chosen in order to provide a clear indication of the effectiveness of the proposed experiments $\mathcal{E}$.

		Mean Average Precision		
Row	Task	First approach	Second approach	MAX
1	td2003	I(n_e)C2FU (0.1446)	DLHFP (0.1455)	0.1926 (+32.37%)*
2	td2004	PL2F (0.1299)	I(n_e)C2FP (0.1307)	0.1615 (+23.57%)*
3	hp2003	DLHFU (0.6660)	BM25FA (0.6498)	0.7658 (+14.98%)*
4	hp2004	PB2FU (0.5523)	DLHFA (0.5555)	0.7025 (+26.46%)*
5	np2003	PL2FP (0.6846)	I(n_e)C2FA (0.6836)	0.7827 (+14.33%)*
6	np2004	PB2F (0.6944)	I(n_e)C2FA (0.6814)	0.8019 (+16.52%)*

Table 6.1: The pairs of retrieval approaches employed by the Bayesian decision mechanism in the evaluation of the proposed experiments $\mathcal{E}$. The columns 'First approach' and 'Second approach' show the employed retrieval approaches and their MAP for the corresponding task within brackets. The column 'MAX' shows the maximum MAP that can be obtained by selectively applying one of the two retrieval approaches on a per-query basis. The value within brackets is the relative increase in MAP from the most effective individual retrieval approach. The symbol * indicates that the difference in MAP between the mechanism MAX and the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test. The results are copied from Table 4.11.

The outcome of the evaluated experiments is computed from a sample Ret_q of the set of retrieved documents, as discussed in Section 5.2. This sample Ret_q is formed with documents that contain at least one query term in either their body, or their title. For example, the sample Ret_q contains documents with query terms in their anchor text, and at least one query term in either their body, or their title. However, documents that only contain query terms in their anchor text are not included in the sample Ret_q.

The experiments $\mathcal{E}$ have been defined for the different fields of documents, as discussed in Sections 5.3 and 5.4. From all the possible combinations of the three document

fields (body, anchor text, and title), the evaluated experiments employ either the body field (b), or a combination of the anchor text and title fields (at). The body field is selected, because it is similar to the full text of documents, while the combination of the anchor text and the title corresponds to fields that provide a concise description of the documents. Initial experiments have shown that other combinations of the body, anchor text, and title fields perform either similarly to the body field, or similarly to the combination of the anchor text and title fields.

6.2.3 Presentation and analysis of results

Here, a brief description of the presentation and the analysis of the results is given, before proceeding to the evaluation of the proposed experiments $\mathcal{E}$.

In the subsequent Sections 6.3 and 6.4, each row in the tables shows the following information: a row identifier for ease of reference ('Row'); the employed task ('Task'); the employed pair of retrieval approaches ('Retrieval approaches') and the mean average precision of the most effective one ('Baseline'); the experiment employed by the decision mechanism ('$\mathcal{E}$'); the achieved mean average precision by the Bayesian decision mechanism ('MAP'); the relative difference between the MAP of the most effective retrieval approach and the achieved MAP by the decision mechanism ('+/-%'); a †, which signifies that the number of times the decision mechanism applies the correct retrieval approach is statistically significant at level 0.05 according to the sign test; a *, which signifies that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant at level 0.05 according to Wilcoxon's singed rank test; and the number of decision boundaries in the decision mechanism ('Bnd').

The tables report the evaluation results for the experiments $\mathcal{E}$, which identify at least one decision boundary for each of the tested tasks. This choice is made in order to focus on the experiments $\mathcal{E}$ that are effective for all the three types of tasks (topic distillation, home page finding, and named page finding). The comparison of the effectiveness of the experiments $\mathcal{E}$ is performed with respect to their performance for all the tested tasks, in order to focus the analysis on the experiments that perform well for a range of different tasks. The results are mainly discussed from two perspectives: the impact of the particular fields used to compute the experiments $\mathcal{E}$, i.e., the body field, or a

combination of the anchor text and the title fields; and the particular characteristics of each experiment $\mathcal{E}$.

The remainder of the chapter is organised as follows. Sections 6.3 and 6.4 present the evaluation of the score-independent and the score-dependent experiments, in the described setting. The experimental setting and the presentation of the results are revisited in Sections 6.5, 6.6, and 6.7. More specifically, Section 6.5 introduces document sampling in order to reduce the computational overhead of the experiments, and to assess their effectiveness when using only highly scored documents. Section 6.6 discusses the effectiveness of a decision mechanism when the retrieval approaches employ the same field-based weighting model. Section 6.7 describes the results from an example of a Bayesian decision mechanism, which employs three retrieval approaches.

6.3 Evaluation of score-independent experiments

This section evaluates the effectiveness of the score-independent document-level and aggregate-level experiments, which were introduced in Sections 5.3.1 and 5.3.2, respectively. First, the evaluation of the document-level experiments is presented in Section 6.3.1. The evaluation of the domain and directory aggregate-level experiments are presented in Sections 6.3.2.1 and 6.3.2.2, respectively. For each type of experiment, the evaluation results are followed by an illustrative example for a particular task. The examples are intended to provide insight in using the experiments in the context of the Bayesian decision mechanism. Section 6.3.3 provides some concluding remarks about the evaluation of the score-independent experiments.

6.3.1 Document-level experiments

The current section presents the evaluation of the document-level experiments. The score-independent document-level experiments are based on counting the number of documents, which contain query terms in particular fields. The considered documents may contain all the query terms, or at least one of them, in a particular field. The considered fields are: the body; and a combination of the anchor text and the title.

The evaluation of the score-independent document-level experiments is performed in the context of a Bayesian decision mechanism, which employs a particular pair

of retrieval approaches for each of the tested tasks (Table 6.1). As described in Section 6.2.2.2, the tested tasks are: td2003; td2004; hp2003; hp2004; np2003; and np2004. The same task is used for training the Bayesian decision mechanism, and for evaluating each experiment. This setting has been chosen in order to reduce any effect from the differences between the employed tasks on the evaluation of the experiments.

6.3.1.1 Evaluation results for document-level experiments

Table 6.2 presents the evaluation of the document-level experiments, which identify at least one decision boundary for each of the tested tasks[1]. The results from column '+/- %' of Table 6.2 are presented in a histogram in Figure 6.1. Row 1 in the table shows the evaluation of the experiment $\mathcal{E}_{\exists(b)}$, which counts the number of documents that contain at least one query term in their body. For each query from the task td2003, the Bayesian decision mechanism selectively applies either the combination of the field-based weighting model I(n_e)C2F with evidence from the URL path length of documents (I(n_e)C2FU), or the combination of the field-based weighting model DLHF with PageRank (DLHFP). The achieved MAP of the Bayesian decision mechanism is 0.1483, which represents an improvement of +1.92% over the baseline MAP of the most effective individual approach (0.1455). Moreover, the decision mechanism applies the most effective retrieval approach for a statistically significant number of topics, as denoted by †.

From Table 6.2, it can be seen that the experiments $\mathcal{E}_{\forall(b)}$ (rows 7-12) and $\mathcal{E}_{\forall(at)}$ (rows 13-18) result, on average, in a lower number of decision boundaries, than the experiment $\mathcal{E}_{\exists(b)}$ (rows 1-6).

For all the tested cases, the Bayesian decision mechanism results in improved retrieval effectiveness, as indicated by the positive differences in the column '+/- %'. The most notable case is shown in row 4, where there is an improvement of 11.65% in MAP for the task hp2004. For the task np2004, the obtained MAP when the decision mechanism uses the experiment $\mathcal{E}_{\forall(b)}$ is 0.7341, which is higher than the MAP of the best performing run in the corresponding task of the TREC 2004 Web track (0.7232 from row 11 in Table 4.6, page 67).

[1] The evaluation results of the experiments, which do not identify at least one decision boundary for each of the tested tasks, are given in Table B.1 (page 239) of Appendix B.

The sign test shows that the decision mechanism has applied the most appropriate retrieval approach for a significant number of queries in 3 cases for the experiment $\mathcal{E}_{\exists(b)}$ (rows 1, 3, and 4), in 1 case for the experiment $\mathcal{E}_{\forall(b)}$ (row 9), and in 2 cases for the experiment $\mathcal{E}_{\forall(at)}$ (rows 13, 17). The Wilcoxon's signed rank test shows that the decision mechanism results in statistically significant improvements in MAP compared to the most effective retrieval approach, in 1 case for the experiment $\mathcal{E}_{\forall(b)}$ (row 9), and in 1 case for the experiment $\mathcal{E}_{\forall(at)}$ (row 13).

Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
1	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(b)}$	0.1483	+ 1.92†	1
2	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(b)}$	0.1313	+ 0.46	2
3	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(b)}$	0.6849	+ 2.84†	3
4	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(b)}$	0.6202	+11.65†	1
5	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(b)}$	0.7007	+ 2.35	1
6	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(b)}$	0.7220	+ 3.97	2
7	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b)}$	0.1476	+ 1.44	2
8	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b)}$	0.1402	+ 7.27	2
9	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b)}$	0.6942	+ 4.23$^{\dagger *}$	1
10	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b)}$	0.5635	+ 1.44	1
11	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b)}$	0.6940	+ 1.37	1
12	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b)}$	0.7341	+ 5.72	1
13	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(at)}$	0.1568	+ 7.77$^{\dagger *}$	1
14	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(at)}$	0.1322	+ 1.15	1
15	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(at)}$	0.6803	+ 2.15	1
16	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(at)}$	0.5871	+ 5.69	2
17	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(at)}$	0.7091	+ 3.58†	1
18	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(at)}$	0.7150	+ 2.97	1

Table 6.2: Evaluation of score-independent document-level experiments $\mathcal{E}_{\exists(f)}$ and $\mathcal{E}_{\forall(f)}$ for combination of fields f, which result in at least one decision boundary for each tested topic set. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test. The column 'Bnd' reports the number of decision boundaries for each case.

6.3.1.2 Example for document-level experiments

Figure 6.2 illustrates the decision boundaries of the Bayesian decision mechanism for the topic set hp2004, where the experiment $\mathcal{E}_{\exists(b)}$ performs very well (row 4 in Table 6.2). The decision mechanism selectively applies either a combination of the field-based weighting model PB2F with the URL path length (PB2FU), or a com-

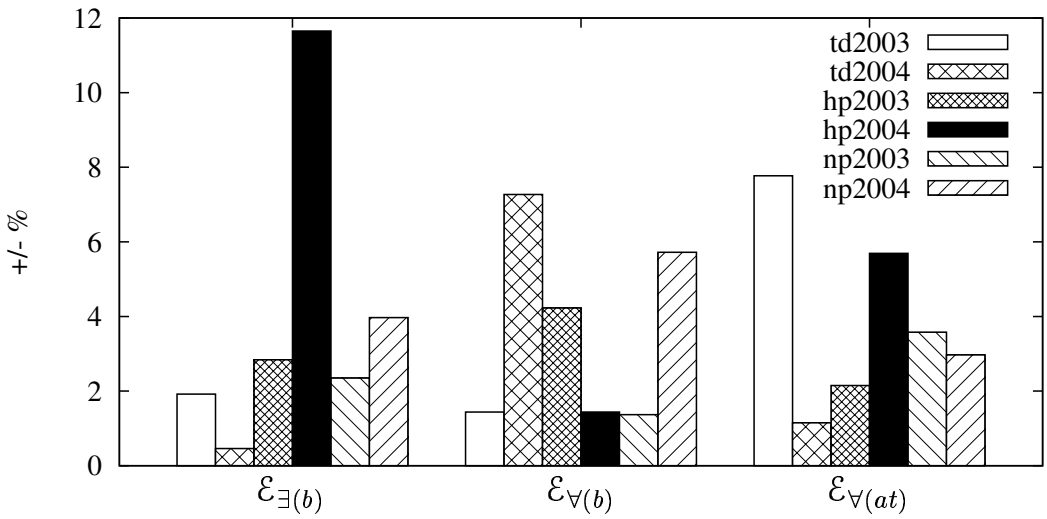

Figure 6.1: Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.2.

bination of the field-based model DLHF with the Absorbing model (DLHFA). The curves in the figure correspond to the estimated density of the posterior likelihoods P(PB2FU)·P($\mathcal{E}$|PB2FU) and P(DLHFA)·P($\mathcal{E}$|DLHFA) for the experiment $\mathcal{E}_{\exists(b)}$ (top diagram), and for the experiment $\mathcal{E}_{\forall(b)}$ (bottom diagram), respectively.

From the top diagram in Figure 6.2, it can be seen that the combination of field retrieval and evidence from the URL of documents (PB2FU) is more effective when the outcome of the experiment $\mathcal{E}_{\exists(b)}$, which considers documents with at least one query term in their body, is lower than 300551. When the outcome of the experiment $\mathcal{E}_{\exists(b)}$ is higher than 300551, the combination of the field-based weighting model DLHF with the Absorbing Model (DLHFA) is more effective. On the other hand, the bottom diagram indicates that DLHFA is more effective when the outcome of the experiment $\mathcal{E}_{\forall(b)}$ is lower than 3782.337, while PB2FU is more effective when the outcome of $\mathcal{E}_{\forall(b)}$ is higher than 3782.337. This suggests that, for this particular example, the type of the experiment, that is, whether the considered documents contain all or at least one of the query terms, has an important effect on the range of the outcome values for which a retrieval approach is effective.

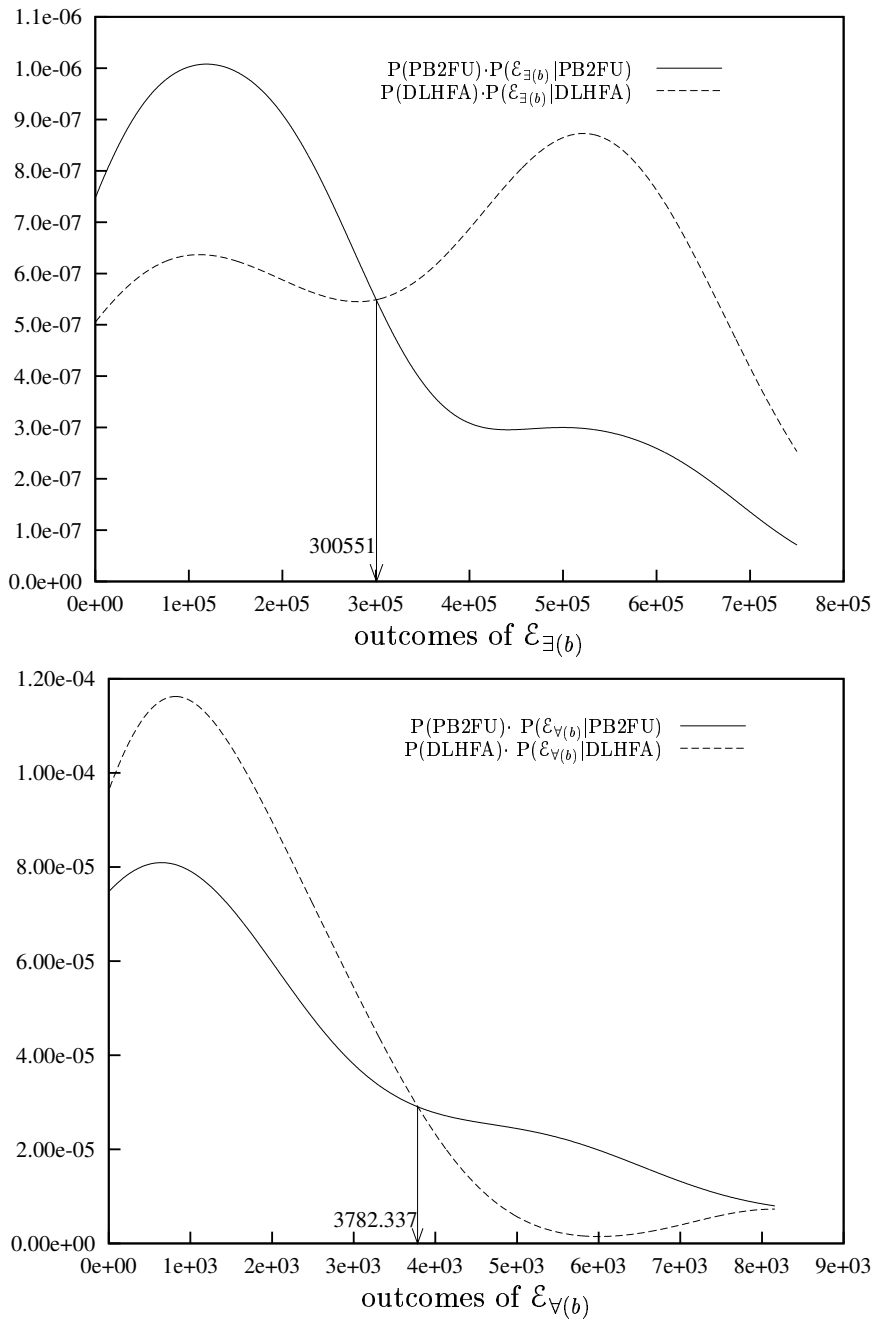

Figure 6.2: Posterior likelihoods of the experiments $\mathcal{E}_{\exists(b)}$ and $\mathcal{E}_{\forall(b)}$ for the topic set hp2004, where one of the retrieval approaches PB2FU or DLHFA is selected to be applied for each query.

6.3.1.3 Discussion

The evaluation results for the score-independent document-level experiments have shown that the experiments $\mathcal{E}_{\forall(b)}$ and $\mathcal{E}_{\forall(at)}$ result, on average, in a lower number of decision boundaries than $\mathcal{E}_{\exists(b)}$ (Table 6.2). For example, the experiment $\mathcal{E}_{\forall(at)}$ results in one decision boundary for all the tested tasks, apart from hp2004, for which there are two decision boundaries.

The experiments $\mathcal{E}_{\forall(b)}$ and $\mathcal{E}_{\forall(at)}$ count only the documents, in which all the query terms appear in the body, or in a combination of the anchor text and title fields. This provides a strong indication that the documents are related to the query. Therefore the used evidence by the experiments provides a better indication of how broad or specific a query is. When there are many documents with all the query terms in a particular field or a combination of fields, then evidence from the hyperlink structure or the URL of Web documents can be used to detect documents of higher quality, or documents that are likely to be home pages of relevant Web sites. On the other hand, it is not expected that there are many documents containing all the terms of a specific query. For this reason, the document-level experiments, which consider documents with all the query terms, are more appropriate for selective Web IR.

6.3.2 Aggregate-level experiments

This section discusses the evaluation of the decision mechanism that employs the score-independent aggregate-level experiments, described in Section 5.3.2. The aggregates are defined as the documents that belong to the same domain, or the documents that are stored in the same directory. The considered experiments first identify all the aggregates in the sample Ret_q of the set of retrieved documents. Next, they extract a feature of the size distribution of the aggregates. As discussed in Section 5.3.2, the employed features are: the average of the aggregates' size (avg); the standard deviation of the aggregates' size (std); and the number of large aggregates (lrg). The aggregate-level experiments consider documents with either at least one, or all the query terms in the body, or in a combination of the anchor text and the title of documents.

Sections 6.3.2.1 and 6.3.2.2 describes the evaluation results for the domain and directory aggregate-level experiments, respectively. Section 6.3.2.3 presents an illustrative example of applying the aggregate-level experiments for a particular task. Finally,

Section 6.3.2.4 discusses the evaluation of the aggregate-level experiments.

6.3.2.1 Evaluation results for domain aggregate-level experiments

Table 6.3 presents the evaluation of the experiments $\mathcal{E}$ that employ domain aggregates, and result in at least one decision boundary for all the tested tasks[1]. Figure 6.3 summarises the results from column '+/- %' of Table 6.3 in a histogram. The results indicate that only the experiments $\mathcal{E}_{\exists(b),avg(dom)}$ (rows 1-6), $\mathcal{E}_{\exists(at),avg(dom)}$ (rows 13-18), $\mathcal{E}_{\forall(b),std(dom)}$ (rows 19-24) and $\mathcal{E}_{\exists(b),lrg(dom)}$ (rows 31-36), result in improvements in MAP for all the tested tasks. On the other hand, the experiments $\mathcal{E}_{\forall(b),avg(dom)}$ (rows 7-12), $\mathcal{E}_{\exists(at),std(dom)}$ (rows 25-30), and $\mathcal{E}_{\forall(b),lrg(dom)}$ (rows 37-42), result in a decrease in MAP for some of the tested tasks.

Regarding the number of decision boundaries, only the experiment $\mathcal{E}_{\forall(b),std(dom)}$ (rows 19-24) results in a decision mechanism with only 1 decision boundary for all the tested tasks. The experiment $\mathcal{E}_{\exists(at),std(dom)}$ also identifies one decision boundary for four out of the six tested tasks (rows 25-30). The rest of the experiments shown in Table 6.3 result in a variable number of decision boundaries. For example, the experiment $\mathcal{E}_{\forall(b),avg(dom)}$ results in at least two decision boundaries for all the tested topic sets (rows 7-12), while the experiment $\mathcal{E}_{\exists(b),lrg(dom)}$ results in either one, two, or four decision boundaries (rows 31-36).

Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
1	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1482	+ 1.86†	1
2	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1347	+ 3.06	3
3	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6732	+ 1.08	3
4	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6202	+11.65†	1
5	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6929	+ 1.21	1
6	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(b),avg(dom)}$	0.7187	+ 3.50	2
7	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1429	- 1.79	3
8	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1386	+ 6.04†	3
9	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6593	- 1.01	2
10	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6054	+ 8.98	2
11	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7031	+ 2.70	2
12	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7005	+ 0.88	2
13	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1464	+ 0.62†	2
14	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1316	+ 0.69	1
15	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6895	+ 3.53†	4
16	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6215	+11.88$^{\dagger *}$	2

continued on next page

[1] The evaluation results of the experiments, which do not identify at least one decision boundary for all the tested tasks, are given in Tables B.2 (page 240), B.3 (page 242), and B.4 (page 244) of Appendix B.

continued from previous page								
Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
17	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6943	+1.42	1
18	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7298	+5.10	3
19	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b),std(dom)}$	0.1525	+4.81	1
20	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b),std(dom)}$	0.1353	+3.52	1
21	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b),std(dom)}$	0.6682	+0.33	1
22	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b),std(dom)}$	0.5622	+1.21	1
23	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b),std(dom)}$	0.7230	+5.61	1
24	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b),std(dom)}$	0.7184	+3.46	1
25	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(at),std(dom)}$	0.1426	-2.00†	1
26	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(at),std(dom)}$	0.1347	+3.06	1
27	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(at),std(dom)}$	0.6746	+1.29	2
28	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(at),std(dom)}$	0.5871	+5.69	2
29	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(at),std(dom)}$	0.7055	+3.05†	1
30	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(at),std(dom)}$	0.7088	+2.07	1
31	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1463	+0.55	4
32	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1399	+7.04†	2
33	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6719	+0.89	2
34	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6064	+9.16†	1
35	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6895	+0.72	1
36	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.7125	+2.61	2
37	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1534	+5.43	2
38	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1378	+5.43	5
39	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6881	+3.32$^{\dagger *}$	1
40	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.5483	-1.30	1
41	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6880	+0.50	2
42	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6959	+0.22	2

Table 6.3: Evaluation of score-independent aggregate-level experiments with domains, which result in at least one decision boundary for each tested topic set. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

The highest improvement in MAP is obtained for the task hp2004, where the Bayesian decision mechanism employs the experiment $\mathcal{E}_{\exists(at),avg(dom)}$, and selectively applies either PB2FU, or DLHFA (row 16 in Table 6.3). As denoted by *, this improvement in MAP is statistically significant according to Wilcoxon's signed rank test. For the task np2004, the MAP of the decision mechanism that employs the same experiment is 0.7298, which is higher than that obtained by the best performing run in the same task of the TREC 2004 Web track (0.7232 in row 11 of Table 4.6, page 67).

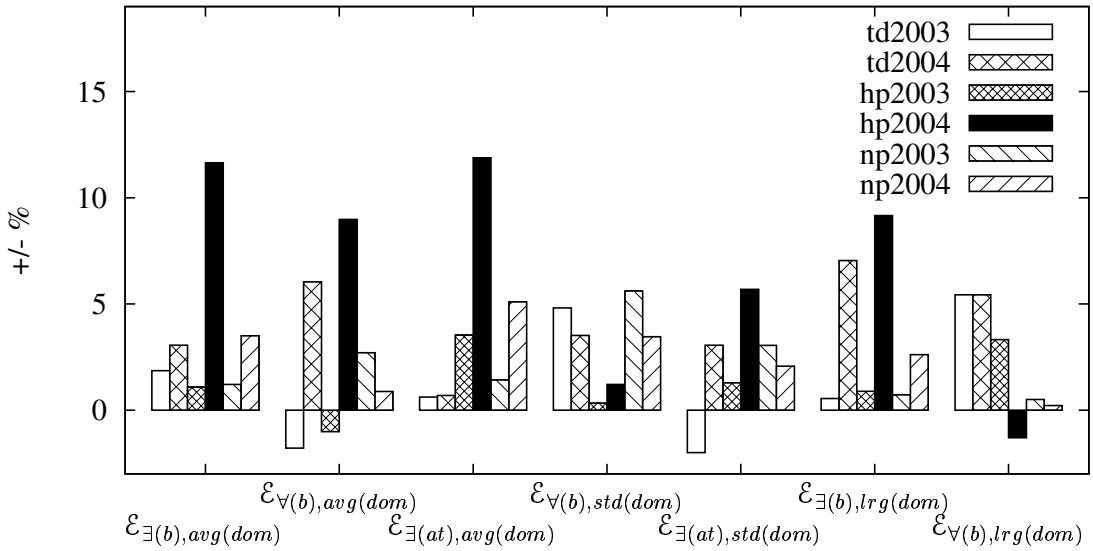

Figure 6.3: Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.3.

6.3.2.2 Evaluation results for directory aggregate-level experiments

The aggregates can also be defined in terms of documents that are stored in the same directory, as described in Section 5.3.2. Table 6.4 displays the evaluation results for the directory aggregate-level experiments, which identify at least on decision boundary for each tested task[1]. For example, row 1 in the table gives the evaluation results obtained when the Bayesian decision mechanism selectively applies either the combination of the field-based weighting model I(n_e)C2F with the URL path length (I(n_e)C2FU), or the combination of the field-based weighting model DLHF with PageRank (DLHFP), for the task td2003. The resulting MAP is 0.1483, which corresponds to a relative improvement of +1.92% over the MAP of the most effective individual approach (0.1455). Furthermore, the decision mechanism applies the most effective retrieval approach for a statistically significant number of queries from td2003, as indicated by †. Figure 6.4 provides an overview of the results from column '+/- %' of Table 6.4 in the form of a histogram.

[1] The evaluation results of the directory aggregate-level experiments, which do not identify at least one decision boundary for all the tested tasks, are given in Tables B.5 (page 245), B.6 (page 247), and B.7 (page 249) of Appendix B.

The evaluation results show that only the experiments, which compute the average size of aggregates, result in improvements for all the tested tasks (rows 1-18). The directory aggregate-level experiments, which compute either the standard deviation of the aggregates' sizes, or the number of large aggregates, do not always result in consistent improvements in retrieval effectiveness for all the tested tasks (rows 19-48).

The column 'Bnd' of Table 6.4 shows that the experiments, which estimate the average size of the directory aggregate size (rows 1-18), identify a variable number of decision boundaries for each task. For example, row 15 shows that the decision mechanism, which employs the experiment $\mathcal{E}_{\forall(at),avg(dir)}$ to select either DLHFU or BM25FA on a per-query basis, has seven decision boundaries. The experiments $\mathcal{E}_{\forall(b),std(dir)}$ (rows 25-30), and $\mathcal{E}_{\exists(at),std(dir)}$ (rows 31-36), result in either one or two decision boundaries in most of the tested cases. This suggests that the standard deviation is a robust feature of the aggregate size distribution, and it is in agreement with the obtained results for the domain aggregate-level experiments, as discussed in Section 6.3.2.1.

The most notable improvements in MAP are shown in row 4, where the MAP obtained by the Bayesian decision mechanism for the task hp2004 represents an improvement of 15.93% over the MAP of the most effective retrieval approach. This improvement is statistically significant according to Wilcoxon's signed rank test, as denoted by *. In the case of the task td2003, when the experiment $\mathcal{E}_{\forall(at),avg(dir)}$ is employed in order to select between I(n_e)C2FU, or DLHFP (row 13), the resulting MAP is 0.1613 (+10.86% relative improvement compared to the MAP of the most effective retrieval approach). Regarding the task np2004, the MAP achieved by the decision mechanism, which employs the experiment $\mathcal{E}_{\forall(b),std(dir)}$, is 0.7261 (row 30 in Table 6.4), which is slightly higher than that of the best performing run in the corresponding task of the TREC 2004 Web track (0.7232 from row 11 in Table 4.6 on page 67).

Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
1	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1483	+ 1.92†	1
2	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1336	+ 2.22	2
3	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6742	+ 1.23	4
4	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6440	+15.93$^{\dagger *}$	2
5	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(b),avg(dir)}$	0.7045	+ 2.91	4
6	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6975	+ 0.45	2
7	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1497	+ 2.89†	1

continued on next page

continued from previous page								
Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
8	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1411	+ 7.96†	3
9	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6855	+ 2.93†	4
10	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(at),avg(dir)}$	0.5903	+ 6.26	4
11	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7100	+ 3.71†	3
12	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7216	+ 3.92	4
13	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1613	+10.86	2
14	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1338	+ 2.37	1
15	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6836	+ 2.64	7
16	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6279	+13.03$^{\dagger *}$	3
17	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6861	+ 0.22	5
18	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7027	+ 1.20	1
19	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(b),std(dir)}$	0.1422	- 2.30	1
20	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(b),std(dir)}$	0.1318	+ 0.84	2
21	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(b),std(dir)}$	0.6699	+ 0.59†	3
22	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(b),std(dir)}$	0.6040	+ 8.73	2
23	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(b),std(dir)}$	0.6934	+ 1.29	2
24	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(b),std(dir)}$	0.7104	+ 2.30	4
25	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b),std(dir)}$	0.1565	+ 7.56	2
26	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b),std(dir)}$	0.1359	+ 3.98	3
27	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b),std(dir)}$	0.6710	+ 0.75	1
28	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b),std(dir)}$	0.5517	- 0.68	1
29	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b),std(dir)}$	0.7070	+ 3.27	1
30	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b),std(dir)}$	0.7261	+ 4.57	1
31	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(at),std(dir)}$	0.1284	-12.00	1
32	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(at),std(dir)}$	0.1321	+ 1.07	1
33	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(at),std(dir)}$	0.6849	+ 2.84	2
34	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(at),std(dir)}$	0.5898	+ 6.17	2
35	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(at),std(dir)}$	0.7111	+ 3.87†	2
36	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(at),std(dir)}$	0.7040	+ 1.38	1
37	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1547	+ 6.32†	2
38	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1295	- 0.92	2
39	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6712	+ 0.78	2
40	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6042	+ 8.77	1
41	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6872	+ 0.38	1
42	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7132	+ 2.71	2
43	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1469	+ 0.96	3
44	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1300	- 0.54	1
45	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6768	+ 1.62	6
46	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.5550	- 0.09	1
47	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6859	+ 0.19	1
48	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6910	- 0.49	1

Table 6.4: Evaluation of score-independent aggregate-level experiments with directories, which result in at least one decision boundary for each tested topic set. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

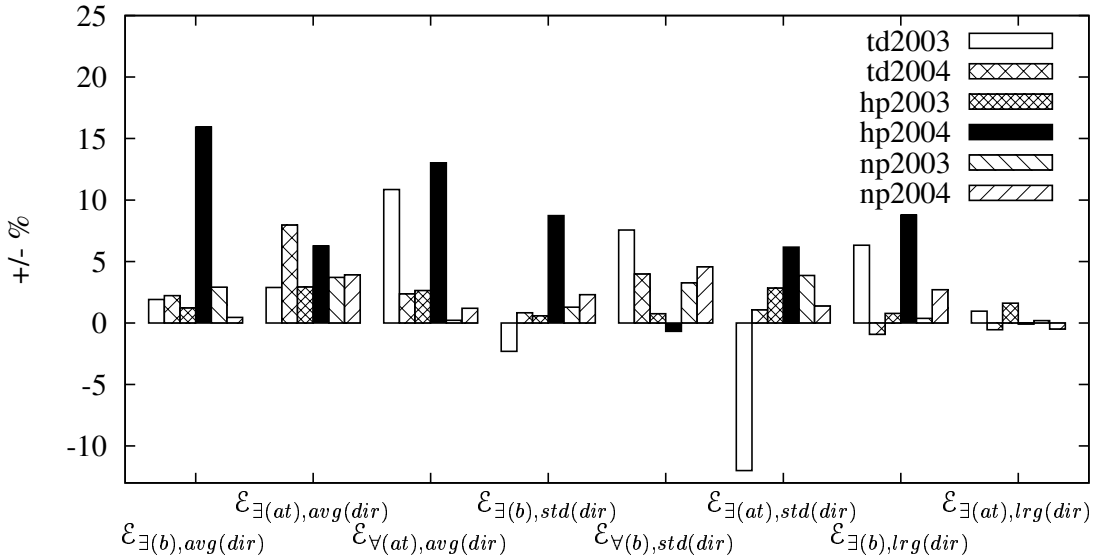

Figure 6.4: Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.4.

6.3.2.3 Example for domain and directory aggregate-level experiments

This section provides an illustrative example of applying the Bayesian decision mechanism with domain and directory aggregate-level experiments. The example is based on the experiments $\mathcal{E}_{\exists(b),avg(dom)}$ and $\mathcal{E}_{\exists(b),avg(dir)}$, which have resulted in considerable improvements in MAP for the task hp2004, with either one or two decision boundaries (see row 4 in Table 6.3, and row 4 in Table 6.4, respectively).

Figure 6.5 displays the posterior likelihoods P(PB2FU)·P($\mathcal{E}_{\exists b,avg(dom)}$|PB2FU) (top diagram) and P(DLHFA)·P($\mathcal{E}_{\exists b,avg(dir)}$|DLHFA) (bottom diagram), which have been estimated during the application of the Bayesian decision mechanism for the topic set hp2004. The decision mechanism selectively applies either PB2FU, or DLHFA on a per-query basis. PB2FU denotes the combination of the field-based weighting model PB2F with evidence from the URL path length. DLHFA denotes the combination of the field-based weighting model DLHF with the Absorbing Model.

The top diagram shows that the retrieval approach PB2FU is more effective for the lower outcomes of the experiment $\mathcal{E}_{\exists b,avg(dom)}$, while DLHFA is more effective for the higher outcomes of $\mathcal{E}_{\exists b,avg(dom)}$. There is one decision boundary, and the diagram

suggests that there is a clear separation between the curves corresponding to the two posterior likelihoods. The outcome values of the experiment $\mathcal{E}_{\exists b,avg(dom)}$ approximately fall between 0 and 120.

When the aggregates are based on directories, the bottom diagram shows that there are two decision boundaries, and the separation between the two posterior likelihoods is less clear than in the case of the experiment $\mathcal{E}_{\exists(b),avg(dom)}$. However, the outcome values of the experiment $\mathcal{E}_{\exists(b),avg(dir)}$ are considerably lower than those of the experiment $\mathcal{E}_{\exists b,avg(dom)}$, and they approximately fall between 1.5 and 6. This range is two orders of magnitude smaller than the range of outcome values of the experiment $\mathcal{E}_{\exists(b),avg(dom)}$. As a consequence, the estimated densities of the posterior likelihoods are more likely to overlap, resulting in a higher number of decision boundaries than the one resulting from domain aggregate-level experiments. In this particular example, it is preferable to employ the domain aggregates, because the corresponding experiment results in a lower number of decision boundaries.

6.3.2.4 Discussion

This section provides a discussion related to the evaluation of the domain and directory aggregate-level experiments, which have been evaluated in Sections 6.3.2.1 and 6.3.2.2, respectively. There are three main points of discussion related to: the differences between the domain and directory aggregates; the effectiveness of the three employed features of the aggregate size distribution (average size, standard deviation, and number of large aggregates); the fields used to compute the aggregate-level experiments.

Domain versus directory aggregates The evaluation results for the domain aggregates indicate that out of the seven experiments, which identify at least one decision boundary for all the tested tasks, there are four experiments that result in improvements for all the tested tasks (rows 1-6, 13-18, 19-24, and 31-36 in Table 6.3). Regarding the directory aggregates, there are three experiments that result in improvements in MAP for all the tested tasks (rows 1-6, 7-12, and 13-18 in Table 6.4), out of the eight experiments that identify at least one decision boundary for all the tested tasks. This suggests that the domain aggregates are more robust, and provide a better indication of the distribution of related documents in aggregates. On the other hand, the directory aggregates are expected to be smaller, and their distribution depends on the particular

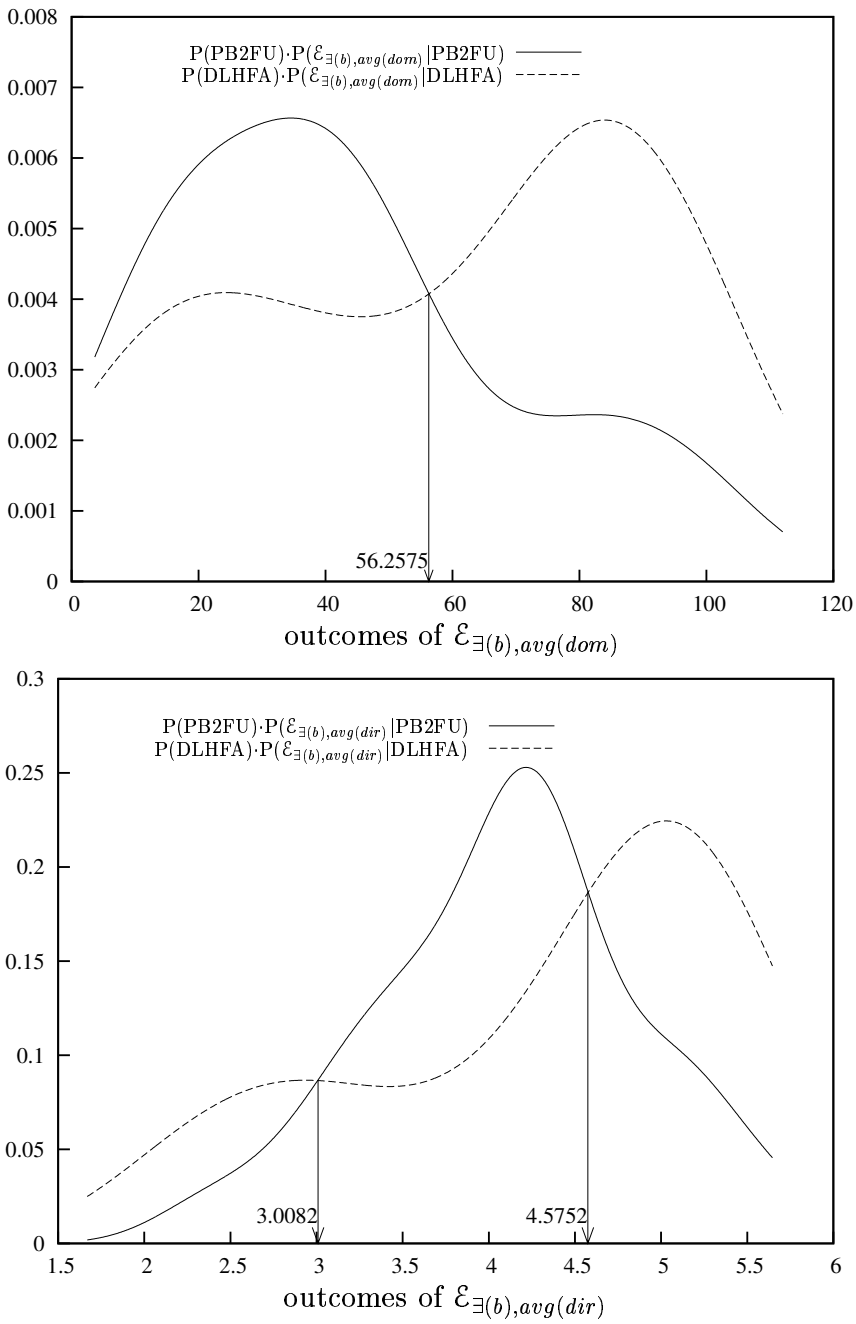

Figure 6.5: Posterior likelihoods of the score-independent aggregate-level experiments $\mathcal{E}_{\exists b,avg(dom)}$ and $\mathcal{E}_{\exists b,avg(dir)}$, for the topic set hp2004, where one of the retrieval approaches PB2FU or DLHFA is selectively applied for each query. The posterior likelihoods for the domain and the directory based aggregates are presented on top, and the bottom diagram, respectively.

structure of Web sites. Therefore, the domain aggregates provide a better indication than the directory aggregates that a query is broad, and that the retrieval effectiveness may be enhanced by employing evidence from the hyperlink structure of documents, or the URLs.

Features of the aggregate size distribution The aggregate-level experiments, which compute the standard deviation of the aggregate size distribution result in a relatively low number of decision boundaries (rows 19-30 in Table 6.3 for domain aggregates, and rows 25-36 in Table 6.4 for directory aggregates). In particular, the experiment $\mathcal{E}_{\forall(b),std(dom)}$ is the only aggregate-level experiment, which identifies only one decision boundary and results in improvements in MAP for each of the tested tasks (rows 19-24 in Table 6.3). This suggests that the standard deviation of the aggregate size distribution is effective in separating the queries for which each of the employed retrieval approaches is more effective. On the other hand, the experiments that compute the average size, or the number of large aggregates tend to identify a higher and more variable number of decision boundaries for the different topic sets (see rows 7-12 in Table 6.3, and rows 43-48 in Table 6.4). For this reason, estimating the standard deviation of the aggregate size distribution provides a better indication about which retrieval approach is more effective, and hence it is more appropriate for defining experiments $\mathcal{E}$ for selective Web IR.

Document fields for aggregate-level experiments Regarding the domain aggregate-level experiments, the evaluation results from Table 6.3 show that improvements in MAP for all the tested topics are obtained with the experiments $\mathcal{E}_{\exists(b),avg(dom)}$ (rows 1-6), $\mathcal{E}_{\exists(at),avg(dom)}$ (rows 13-18), $\mathcal{E}_{\forall(b),std(dom)}$ (rows 19-24), and $\mathcal{E}_{\exists(b),lrg(dom)}$ (rows 31-36). Among these experiments, only $\mathcal{E}_{\exists(at),avg(dom)}$ considers the documents for which the query terms appear in either the anchor text, or the title field. This can be explained because the number of documents containing a particular term in their body is likely to be higher than the number of documents that contain the same term in either their anchor text or their title. Employing the body of documents provides more documents from which to generate the domain aggregates, and, therefore, a more representative distribution of domain aggregate sizes. Similarly, only the experiment $\mathcal{E}_{\forall(b),std(dom)}$ considers documents that contain all the query terms. Considering documents with

all the query terms in a particular field may result in a less representative distribution of domain aggregate sizes. The results for the directory aggregate-level experiments (Table 6.4) do not exhibit any particular trend regarding the document fields.

6.3.3 Conclusions

Overall, this section has evaluated in the context of a Bayesian decision mechanism the score-independent experiments, which have been introduced in Section 5.3. The results suggest that the proposed score-independent experiments $\mathcal{E}$ allow the decision mechanism to distinguish and apply appropriate retrieval approaches on a per-query basis.

Both the document-level experiments (Section 6.3.1), as well as the aggregate-level experiments, which compute the standard deviation of the aggregate sizes (Sections 6.3.2.1 and 6.3.2.2), result in a low number of decision boundaries. This suggests that they can capture a simple relation between the effectiveness of the different retrieval approaches.

The document-level experiments that consider documents with all the query terms tend to result in a lower number of decision boundaries or thresholds, because the occurrence of all the query terms in a particular part of a document provides stronger evidence about the topic of the document. Therefore, the outcome of the experiment is computed from a more cohesive set of documents (Section 6.3.1.3)

The domain aggregates are more stable than the directory aggregates, because the size distribution of the directory aggregates is more dependent on the structure of Web sites (Section 6.3.2.4).

6.4 Evaluation of score-dependent experiments

This section focuses on the evaluation of the score-dependent experiments that estimate the usefulness of the hyperlink structure, by computing the divergence between two score distributions, as described in Section 5.4. The first score distribution, S_n, corresponds to the scores of documents assigned by a weighting model. The second score distribution is formed in order to favour documents that point to other highly scored documents. Two different definitions for the second distribution are tested: U_n, where the scores of documents pointed to by a document are added to its original score;

and U'_n, where the sum of the scores of documents pointed to by a document replaces its original score. The scores in both distributions are normalised between 0 and 1.

The score distribution S_n can be defined with respect to any of the retrieval approaches described in Chapter 4. In the context of the evaluation of the experiments, two field-based weighting models are employed, namely PL2F and I(n_e)C2F, in order to test the impact of different weighting models on the effectiveness of the experiments. These two field-based models are statistically independent, as shown in Chapter 4. The weighting models PL2F and I(n_e)C2F are used independently of the retrieval approaches employed for the final document ranking. In this way, the definition of the experiments does not depend on which retrieval approaches are considered by the decision mechanism.

When the weighting model PL2F is employed, then the score-dependent experiments, which define the usefulness of the hyperlink structure as $L(S_n, U_n)$, are denoted by $\mathcal{E}_{\exists(f),L(SU)_{pl}}$ and $\mathcal{E}_{\forall(f),L(SU)_{pl}}$, depending on whether documents with at least one or all the query terms in the field f are considered, respectively. In the same way, when the weighting model I(n_e)C2F is used to define S_n, the score-dependent experiments, which define the usefulness of the hyperlink structure as $L(S_n, U'_n)$, are denoted by $\mathcal{E}_{\exists(f),L(SU')_{in}}$ and $\mathcal{E}_{\forall(f),L(SU')_{in}}$.

After describing the setting of the distribution S_n, this section presents the evaluation results for the score-dependent experiments, and closes with a discussion and some concluding remarks.

6.4.1 Setting the score distribution S_n

The score distribution S_n is defined by using the field-based weighting models PL2F, and I(n_e)C2F. Each of the weighting models has six hyper-parameters: the term frequency normalisation parameters c_b, c_a, and c_t, for the body, anchor text, and title fields, respectively, and the three corresponding field weights w_b, w_a, and w_t.

In order to define the score distribution S_n independently of the retrieval approaches used for the final ranking of documents, the hyper-parameters are set in the following way. For both weighting models, the parameters related to the length normalisation of the fields are set $c_b = c_a = c_t = 1$. The weights of the body and title fields are set equal to $w_b = w_t = 1$. The weight of the anchor text field is set equal to $w_a = 0$. Indeed, if $w_a > 0$, then the anchor text would contribute to the score of both the

source and the destination documents. Therefore, the estimated distribution U_n would incorporate the effect of the anchor text twice. For the evaluation of the score-dependent experiments, the described setting of the parameters will be referred to as the *default setting*. The remainder of this section investigates the impact of the parameter setting on the distribution of the experiment outcome values.

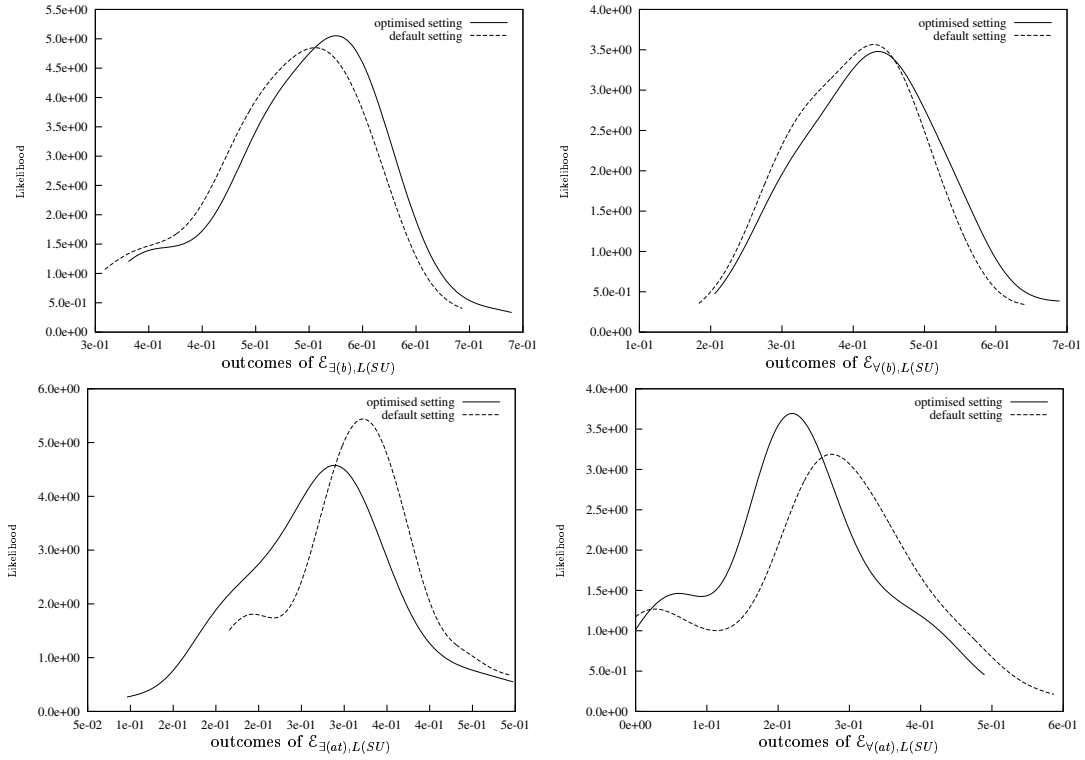

Figure 6.6: Density estimates of the usefulness of the hyperlink structure experiments, according to whether an optimised or the default parameter setting is used. The shown density estimates are obtained for the topic set td2003 and the distribution S_n is based on the weighting model PL2F.

Figure 6.6 displays the density estimates for the experiment outcome values, which are based on $L(S_n, U_n)$, when the score distribution S_n is generated with the weighting model PL2F, and either the above described default setting of the parameters, or the parameter setting used for retrieval with the task td2003, as discussed in Section 4.6.2 (first row of Table A.11 on page 235).

The differences in the estimated density curves suggest that the parameter setting does not have a strong impact on the obtained outcomes, when the experiment considers

the documents with at least one (top left diagram), or all the query terms in their body (top right diagram). On the other hand, the difference is greater when the considered documents contain the query terms in their anchor text, or their title fields (bottom left and right diagrams). This is explained by the fact that the optimised setting weights the anchor text of documents, while the default setting uses a zero weight for the anchor text.

The bottom diagrams in Figure 6.6 show that the outcome values of the experiments $\mathcal{E}_{\exists(at),L(SU)_{pl}}$ and $\mathcal{E}_{\forall(at),L(SU)_{pl}}$ are lower for the optimised setting of the parameters, than the ones obtained for the default setting. This is due to the fact that, in the optimised setting, the effect of the hyperlinks is already incorporated in the score distribution S_n, and it justifies setting the weight of the anchor text equal to zero for computing the usefulness of the hyperlink structure. Similar results are obtained when the usefulness of the hyperlink structure is represented with the divergence $L(S_n, U'_n)$, as well as when the score distribution S_n is based on the weighting model I(n_e)C2F. Therefore, the default setting of the hyper-parameters of the field-based weighting models is appropriate for computing the outcome of the score-dependent experiments.

6.4.2 Evaluation results of experiments based on the usefulness of hyperlink structure $L(S_n, U_n)$

This section presents the evaluation results of the experiments $\mathcal{E}_{\exists(f),L(SU)}$ and $\mathcal{E}_{\forall(f),L(SU)}$, which estimate the usefulness of the hyperlink structure $L(S_n, U_n)$. The score distribution S_n is formed by using either the field-based model PL2F or I(n_e)C2F, with the default parameter setting described in Section 6.4.1. The score distribution U_n is generated by adding the scores of documents pointed to by a document to its original score (Equations (5.15) and (5.16) on page 118). As in the case of the score-independent experiments (Section 6.3), the experiments are evaluated for either the body field of documents (b), or a combination of the anchor text and title fields (at). For example, the experiment $\mathcal{E}_{\exists(b),L(SU)_{in}}$ employs the field-based weighting model I(n_e)C2F to form the score distribution S_n, and considers documents with at least one query term in their body, in order to compute the symmetric Jensen-Shannon divergence $L(S_n, U_n)$. All the different combinations of options (using either PL2F or I(n_e)C2F to define S_n; using either $\exists$ or $\forall$; and using either the body (b) or the anchor text and title fields (at)) result in eight different experiments.

The evaluation results in Table 6.5 show that all the eight experiments, which estimate the usefulness of the hyperlink structure $L(S_n, U_n)$, identify at least one decision boundary for all the tested tasks. However, the number of identified decision boundaries varies. For example, the experiment $\mathcal{E}_{\forall(at),L(SU)_{pl}}$ identifies one decision boundary for the tasks hp2004, np2003, and np2004 (rows 22-24), and at least three decision boundaries for the tasks td2003, td2004, and hp2003 rows (19-21).

Regarding the obtained MAP by the decision mechanism, it can be seen that only the experiments $\mathcal{E}_{\exists(at),L(SU)_{pl}}$ (rows 13-18) and $\mathcal{E}_{\exists(at),L(SU)_{in}}$ (37-42) result in improvements for all the tested tasks. Furthermore, when the decision mechanism selectively applies either PB2F or I(n_e)C2FA for the task np2004 (row 48), the obtained MAP is 0.7468, which is higher than the MAP of the best performing run in the corresponding task of the TREC 2004 Web track (0.7232 from row 11 in Table 4.6 on page 67). In this case, the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, as denoted by †. When the decision mechanism employs the experiments $\mathcal{E}_{\forall(b),L(SU)_{pl}}$ or $\mathcal{E}_{\forall(at),L(SU)_{pl}}$ for the task np2004, or the experiment $\mathcal{E}_{\forall(b),L(SU)_{in}}$ for the task hp2003, the difference between the MAP of the decision mechanism and that of the most effective baseline is statistically significant, as denoted by * (rows 12, 24, and 33 of Table 6.5, respectively). Figure 6.7 provides an overview of the differences between the MAP of the decision mechanism and that of the most effective retrieval approach, as reported in column '+/- %' of Table 6.5.

Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
1	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1432	- 1.58	2
2	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1433	+ 9.64†	3
3	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6670	+ 0.15	3
4	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.5612	+ 1.03	3
5	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6899	+ 0.77	1
6	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.7312	+ 5.30	2
7	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1607	+10.45	3
8	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1287	- 1.50	1
9	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6943	+ 4.25†	5
10	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6132	+10.39	3
11	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7213	+ 5.36†	4
12	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7373	+ 6.18*	3
13	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1484	+ 1.99†	3
14	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1337	+ 2.30	3
15	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6796	+ 2.04	5

continued on next page

continued from previous page								
Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
16	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.5877	+5.80	1
17	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6946	+1.46	3
18	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7382	+6.31	2
19	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1571	+7.97†	4
20	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1322	+1.15	5
21	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6589	-1.10	3
22	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.5707	+2.74	1
23	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7013	+2.44	1
24	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7460	+7.43$^{\dagger *}$	1
25	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1404	-3.51	2
26	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1405	+7.50	2
27	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6600	-0.90	1
28	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.5628	+1.31	3
29	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6944	+1.43†	3
30	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7432	+7.03	4
31	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1561	+7.29	1
32	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1289	-1.40	1
33	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7038	+5.68$^{\dagger *}$	5
34	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6038	+8.69	4
35	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7149	+4.43†	5
36	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7368	+6.11	3
37	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1500	+3.09†	3
38	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1367	+4.59†	2
39	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6772	+1.68	4
40	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.5933	+6.80	2
41	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7017	+2.50	3
42	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7231	+4.13	3
43	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1445	-0.69	2
44	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1350	+3.29	3
45	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6608	-0.78	4
46	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.5607	+0.94	1
47	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6904	+0.85	3
48	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7468	+7.55†	1

Table 6.5: Evaluation of score-dependent experiments based on estimating the usefulness of the hyperlink structure $L(S_n, U_n)$, which result in at least one decision boundary for each tested topic set. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

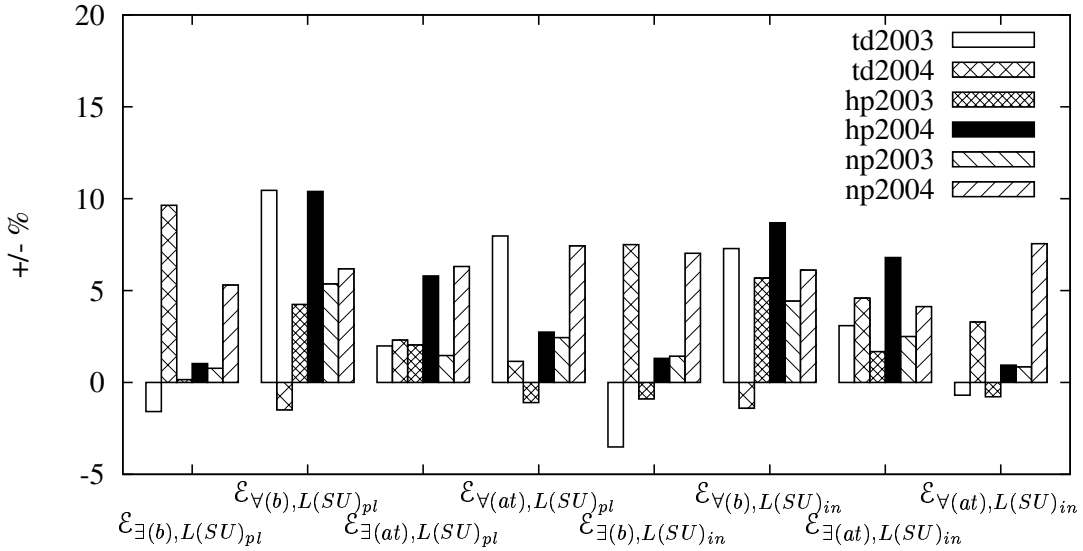

Figure 6.7: Histogram summarising the relative differences between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.5.

6.4.3 Evaluation results of experiments based on the usefulness of hyperlink structure $L(S_n, U'_n)$

The current section presents the evaluation results for the score-dependent experiments that compute the usefulness of the hyperlink structure $L(S_n, U'_n)$. The score distribution S_n corresponds to the scores assigned to documents by the field-based weighting models PL2F or I(n$_e$)C2F. According to the distribution U'_n, the score of a document corresponds to the sum of the scores of the documents it points to (Equation (5.17) on page 118). Table 6.6 presents the evaluation of the experiments that depend on the divergence $L(S_n, U'_n)$[1]. Figure 6.8 presents a histogram of the relative differences between the MAP of the decision mechanism and the most effective retrieval approach, as reported in column '+/- %' of Table 6.6.

The evaluation results show that there is no experiment that results in a consistently low number of decision boundaries for the different tasks (column 'Bnd' in Table 6.6). For example, when the decision mechanism employs the experiment $\mathcal{E}_{\exists(at),L(SU')_{in}}$, there

[1] The evaluation results of the experiments, which do not identify at least one decision boundary for all the tested tasks, are given in Tables B.10 (page 254), and B.11 (page 255) of Appendix B.

is one decision boundary for the named page finding tasks (rows 23-24), but there are five decision boundaries for the hp2004 task (row 22).

Regarding the achieved MAP by the decision mechanism, the experiments $\mathcal{E}_{\forall(b),L(SU')_{pl}}$ (rows 1-6), $\mathcal{E}_{\forall(at),L(SU')_{pl}}$ (rows 7-12), and $\mathcal{E}_{\forall(b),L(SU')_{in}}$ (rows 13-18) result in improvements for all the tested tasks. In particular, when the experiment $\mathcal{E}_{\forall(b),L(SU')_{pl}}$ is used to selectively apply either I(n_e)C2FU or DLHFP, for the td2003 task, the obtained MAP is 0.1655 (row 1). The obtained MAP corresponds to a relative increase of 13.75% over the MAP of the most effective individual approach (0.1455). This increase is statistically significant according to Wilcoxon's signed rank test, as denoted by *. The same decision mechanism also applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test, as denoted by †. When the decision mechanism uses the experiment $\mathcal{E}_{\exists(at),L(SU')_{in}}$, the obtained MAP for the np2004 task is 0.7269 (row 24). This is slightly higher than the MAP of the best performing run in the corresponding task of TREC 2004 Web track (0.7232 from row 11 in Table 4.6 on page 67).

Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
1	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1655	+13.75$^{\dagger *}$	3
2	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1343	+ 2.75	3
3	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6829	+ 2.54	1
4	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.5939	+ 6.91*	3
5	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6914	+ 0.99	2
6	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7173	+ 3.30	2
7	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1470	+ 1.03	1
8	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1355	+ 3.67	2
9	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6806	+ 2.19	4
10	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.5751	+ 3.53	1
11	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6925	+ 1.15	2
12	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.7155	+ 3.04	1
13	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1570	+ 7.90†	3
14	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1327	+ 1.53	2
15	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6816	+ 2.34	1
16	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6001	+ 8.03	3
17	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7059	+ 3.11	4
18	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7217	+ 3.93	2
19	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1437	- 1.20	1
20	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1357	+ 3.83	3
21	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6807	+ 2.21	4
22	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.5949	+ 7.09†	5

continued on next page

continued from previous page								
Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
23	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7131	+4.16	1
24	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7269	+4.68	1
25	td2003	I(n_e)C2FU	DLHFP	0.1455	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1503	+3.30	1
26	td2004	PL2F	I(n_e)C2FP	0.1307	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1376	+5.28	4
27	hp2003	DLHFU	BM25FA	0.6660	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6719	+0.89	2
28	hp2004	PB2FU	DLHFA	0.5555	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.5640	+1.53	1
29	np2003	PL2FP	I(n_e)C2FA	0.6846	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.7009	+2.38	2
30	np2004	PB2F	I(n_e)C2FA	0.6944	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6914	-0.43	1

Table 6.6: Evaluation of score-dependent experiments based on estimating the usefulness of the hyperlink structure $L(S_n, U'_n)$, which result in at least one decision boundary for each tested topic set. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

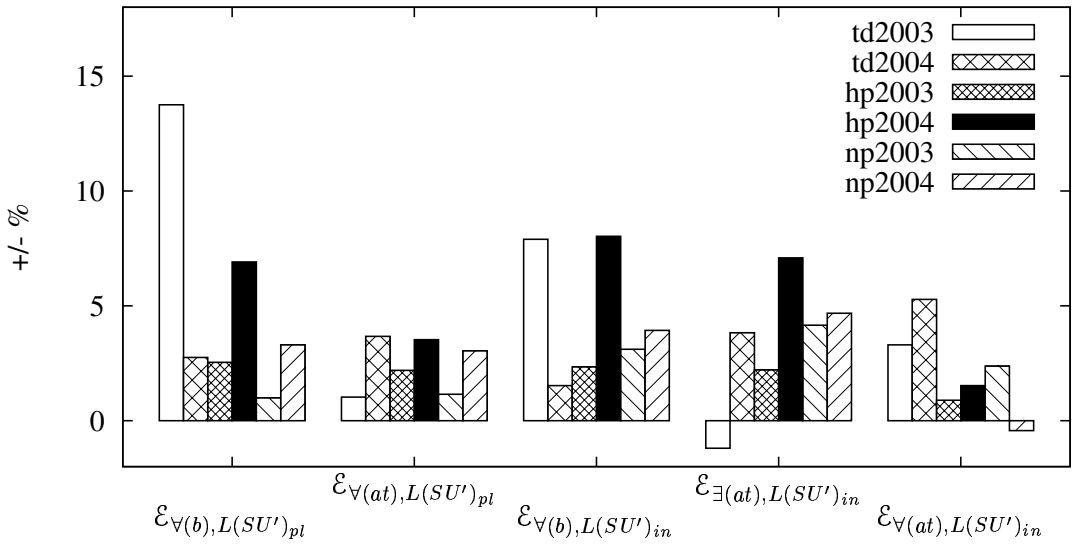

Figure 6.8: Histogram summarising the relative differences between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.6.

6.4.4 Example of the usefulness of hyperlink structure experiments

This section presents an illustrative example of using the experiments based on the usefulness of the hyperlink structure. Figure 6.9 displays the posterior likelihoods

P(I(n_e)C2FU)·P($\mathcal{E}$|P(I(n_e)C2FU) and P(DLHFP)·P($\mathcal{E}$|P(DLHFP) for the score-dependent experiments $\mathcal{E}_{\forall(b),L(SU)_{pl}}$ (top left diagram), $\mathcal{E}_{\forall(b),L(SU')_{pl}}$ (top right diagram), $\mathcal{E}_{\forall(b),L(SU)_{in}}$ (bottom left diagram), and $\mathcal{E}_{\forall(b),L(SU')_{in}}$ (bottom right diagram). The employed task is td2003, where all four experiments achieved an improvement of at least +7.29% over the baseline (rows 7 and 31 from Table 6.5, and rows 1 and 13 from Table 6.6).

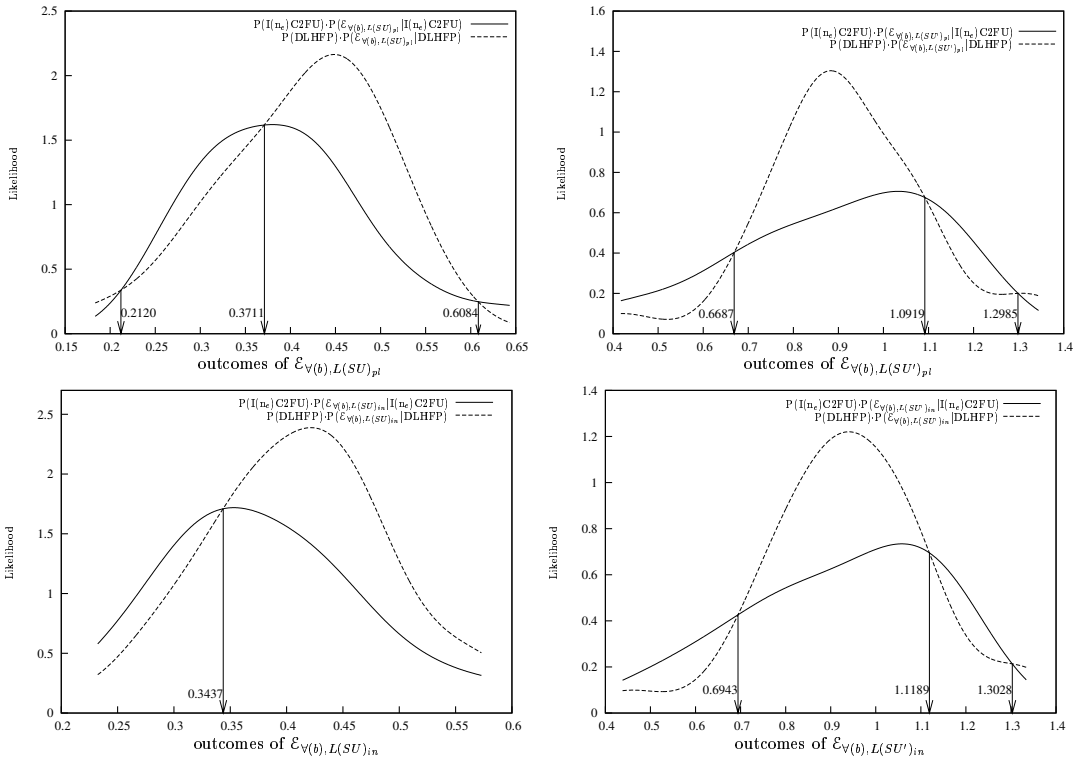

Figure 6.9: Posterior likelihoods of the score-dependent experiments for the topic set td2003, where one of the retrieval approaches I(n_e)C2FU or DLHFP is selectively applied on a per-query basis. The posterior likelihoods for the experiments that estimate the usefulness of the hyperlink structure $L(S_n, U_n)$ and $L(S_n, U'_n)$ are shown on the left hand and the right hand side of the figure, respectively. The score distribution S_n is generated with either PL2F (top diagrams), or I(n_e)C2F (bottom diagrams).

For this particular example, a comparison between the top diagrams and the bottom diagrams suggests that using a different weighting model, such as PL2F, or I(n_e)C2F, has a small effect on the outcome values of the experiments. The shapes of the posterior likelihoods for the experiments $\mathcal{E}_{\forall(b),L(SU')_{pl}}$ and $\mathcal{E}_{\forall(b),L(SU')_{in}}$ (left hand side of Figure 6.9) are very similar. The shapes of the posterior likelihoods for the experiments $\mathcal{E}_{\forall(b),L(SU)_{pl}}$ and $\mathcal{E}_{\forall(b),L(SU)_{in}}$ are also similar. However, the number of intersection

points, or in other words, decision boundaries, is different (3 decision boundaries for the experiment $\mathcal{E}_{\forall(b),L(SU)_{pl}}$ vs. 1 for the experiment $\mathcal{E}_{\forall(b),L(SU)_{in}}$). This variability is explained by the fact that the estimated posterior likelihoods for the lower and the higher outcome values of the experiments are relatively low and less reliable, because there are only few training queries that result in such outcome values.

It is also worth noting that the divergence values obtained with $L(S_n, U'_n)$ (right hand side of Figure 6.9) are considerably higher than those obtained with $L(S_n, U_n)$ (left hand side of Figure 6.9). This confirms that the distribution U'_n is less dependent on S_n, because the original score of a document in S_n is replaced in U'_n by the sum of the scores of the documents it points to, as discussed in Section 5.4.2 (page 117). On the other hand, the distribution U_n is more similar to S_n, because the score of a document in U_n depends on its original score in S_n.

6.4.5 Discussion

This section provides a discussion of issues related to the score-dependent experiments, which have been evaluated in Sections 6.4.2 and 6.4.3. The discussion is focused on the perspectives of the different definitions of S_n, and the effectiveness of using documents with at least one, or all the query terms in particular fields.

Defining S_n with either PL2F or I(n_e)C2F The score distribution S_n can be defined in several different ways. In the context of the evaluation of the experiments, the distribution S_n has been defined with respect to the field-based weighting models PL2F and I(n_e)C2F. The two employed weighting models affect the Bayesian decision mechanism. For example, when the experiment $\mathcal{E}_{\forall(b),L(SU)_{pl}}$, which employs the symmetric Jensen-Shannon divergence $L(S_n, U_n)$, is used to selectively apply PB2FU or DLHFA for the task hp2004, the MAP and the number of decision boundaries are 0.6132 and 3, respectively (row 10 in Table 6.5). When the experiment $\mathcal{E}_{\forall(b),L(SU)_{in}}$ is used in the same setting, the obtained MAP is 0.6038, and there are 4 decision boundaries (row 34 in Table 6.5). However, the obtained results are consistent to some extent. For example, only the experiments $\mathcal{E}_{\exists(at),L(SU)_{pl}}$ (rows 13-18 from Table 6.5) and $\mathcal{E}_{\exists(at),L(SU)_{in}}$ (rows 37-42 from Table 6.5) result in improvements in MAP for all the tested tasks. The experiments that estimate the usefulness of the hyperlink structure $L(S_n, U'_n)$ follow similar trends. For example, both $\mathcal{E}_{\forall(b),L(SU')_{pl}}$ and $\mathcal{E}_{\forall(b),L(SU')_{in}}$

(rows 1-6 and 13-18 from Table 6.6, respectively) result in improvements in MAP for all the tested tasks. Therefore, the score-dependent experiments are robust with respect to the different weighting models that can be used to define the score distribution S_n.

Using documents with all, or at least one of the query terms Table 6.5 shows that all the experiments, which estimate the usefulness of the hyperlink structure $L(S_n, U_n)$, identify at least one decision boundary for all the tested tasks. This is not the case for the experiments, which estimate the usefulness of the hyperlink structure $L(S_n, U'_n)$. As shown in Table 6.6, out of the five different experiments that identify at least one decision boundary for each of the tested tasks, four of them consider only the documents with all the query terms in a particular field, or a combination of fields (rows 1-6, 7-12, 13-18, and 25-30 from Table 6.6). This is due to the fact that the score distribution U'_n is less dependent on S_n, than the score distribution U_n. Therefore, considering only the documents with all the query terms allows to compute the divergence $L(S_n, U'_n)$ from a more cohesive set of documents, which are more likely to be about the topic of the query.

6.4.6 Conclusions

Overall, this section has presented the evaluation of the document score-dependent experiments, which compute the usefulness of the hyperlink structure as the divergence between two score distributions (Section 5.4 on page 115). The first score distribution S_n corresponds to the scores assigned to documents by a retrieval approach. In the context of the evaluation of the experiments, two field-based weighting models, PL2F and I(n_e)C2F, are employed to form S_n. The second score distribution is defined in two ways, in order to favour documents that point to other highly-scored documents: the distribution U_n, where the score of a document corresponds to its original score plus the scores of documents that it points to; and the distribution U'_n, where the score of a document is equal to the sum of the scores of documents that it points to. The usefulness of the hyperlink structure corresponds to the symmetric Jensen-Shannon divergence between S_n and U_n, or between S_n and U'_n.

The evaluation results have shown that the experiments that employ the divergence $L(S_n, U_n)$ result in identifying at least one decision boundary for all the tested topic sets (Section 6.4.2). Therefore, they are very robust for applying selective Web IR. The

experiments that use the divergence $L(S_n, U'_n)$ are robust when the documents with all the query terms are considered (Section 6.4.5). However, both the experiments that use either $L(S_n, U_n)$ or $L(S_n, U'_n)$ result in a variable number of decision boundaries.

The outcome values of the score-dependent experiments that estimate the usefulness of the hyperlink structure $L(S_n, U_n)$ or $L(S_n, U'_n)$ depend on the definition of the distribution S_n. However, the effectiveness of the experiments, which define S_n in terms of two statistically independent field-based weighting models, namely PL2F and I(n_e)C2F, is consistent to an extent (Section 6.4.5).

6.5 Document sampling

The evaluation methodology, which has been described in Section 6.2 (page 131), has stated that the outcome of an experiment is computed from the sample of retrieved documents Ret_q, which contain at least one query term in either their body, or their title fields. For example, Ret_q contains documents for which the query terms occur in the anchor text and the body or the title of the document, but it does not contain documents for which the query terms occur only in the anchor text. Depending on the document frequency of the query terms, the size of Ret_q can be anything between few documents to a large proportion of the document collection. The aim of the current section is to evaluate the proposed experiments, when their outcome is computed from small samples of a fixed number of documents $TopRet_q \subset Ret_q$.

The advantage of using a subset of the set of retrieved documents is mainly the reduced time for computing the outcome of the experiment $\mathcal{E}$. Employing small samples of a fixed number of documents can potentially indicate whether the documents retrieved at higher ranks are more useful for computing the outcome of experiments, and whether there is an optimal number of documents that should used.

The sample of documents $TopRet_q$ is obtained by ranking documents with respect to the score assigned by a retrieval approach. For this purpose, any of the field-based weighting models (Section 4.4 on page 67), or their combination with the query-independent sources of evidence (Section 4.5 on page 74), can be used. In this section, the evaluation of document sampling is performed with two statistically independent field-based weighting models, namely PL2F and I(n_e)C2F. The use of two different

weighting models allows for evaluating the robustness of document sampling. The default setting described in Section 6.4 is used to set the associated hyper-parameters: $c_b = c_a = c_t = 1$, $w_b = w_t = 1$, and $w_a = 0$.

In the context of evaluating document sampling, the effectiveness of computing the outcome of an experiment is tested with two sizes of samples. First, the top 5000 ranked documents are used to form a sample of moderate size. Second, the top 500 ranked documents are used to form a sample of small size. Regarding the queries, which retrieve several tens of thousands of documents, both 5000 and 500 document samples are relatively small.

The remainder of this section is organised as follows. First, the definition of the experiments $\mathcal{E}$ is revisited in order to employ the sample of documents $TopRet_q$. Next, a brief description of the experimental setting and the presentation of results is given. The current section continues with the evaluation of the score-independent and score-dependent experiments, and closes with a discussion, and some concluding remarks.

6.5.1 Revisiting the definition of experiments $\mathcal{E}$

This section revisits the definition of the score-independent (Section 5.3 on page 110) and the score-dependent experiments (Section 5.4 on page 115), in order to use document sampling, and compute their outcome from a sample of top ranked documents $TopRet_q$.

The definitions of the score-independent experiments are updated by replacing Ret_q with $TopRet_q$ in the Equations (5.2)-(5.5), and (5.8). For example, Equation (5.2):

$$cond_{\forall}(d) : \forall t \in q \quad t \in d \quad d \in Ret_q$$

is rewritten as follows:

$$cond_{\forall}(d) : \forall t \in q \quad t \in d \quad d \in TopRet_q \tag{6.1}$$

For the score-dependent experiments that compute the usefulness of the hyper-link structure, the definitions of the score distributions $U = \{u_i\}$ (Equation (5.15) on page 118) and $U' = \{u'_i\}$ (Equation (5.17) on page 118), are updated as follows:

$$u_i = sc_i + \sum_{d_i \to d_j} sc_j \quad d_i \in TopRet_q, d_j \in Ret_q \tag{6.2}$$

$$u'_i = \sum_{d_i \to d_j} sc_j \quad d_i \in TopRet_q, d_j \in Ret_q \tag{6.3}$$

so that all the outgoing hyperlinks from the documents in $TopRet_q$ to the documents in Ret_q are used. In this way, the number of employed hyperlinks is greater than in the case where only the hyperlinks within the set $TopRet_q$ would be used. Therefore, more information from the hyperlink structure is employed.

6.5.2 Description of experimental setting and presentation of results

This section provides a brief description of the experimental setting and the presentation of the results for the evaluation of the experiments $\mathcal{E}$ with document sampling.

As described in the previous sections, the Bayesian decision mechanism selectively applies one retrieval approach from the pair of retrieval approaches that results in the highest potential for improvements in retrieval effectiveness. The evaluation is performed for six different tasks: td2003; td2004; hp2003; hp2004; np2003; and np2004. The employed pairs of retrieval approaches for each task are given in Table 6.1 (page 134).

Each of the tables used for the evaluation of document sampling provides the following information: a row identifier ('Row'); the tested task ('Task'); the mean average precision (MAP) of the most effective individual retrieval approach ('Baseline'); the relative difference in MAP from the baseline, and the number of decision boundaries obtained by the decision mechanism without document sampling (column denoted with the symbol of an experiment, i.e., $\mathcal{E}_{\forall(b)}$ for the experiment that counts the number of documents with all the query terms in their body); the relative difference in MAP from the baseline, and the number of decision boundaries obtained by the decision mechanism with document sampling (columns '$pl5000$', '$pl500$', '$in5000$', '$in500$', where, for example, $pl5000$ corresponds to the sample $TopRet_q$ formed by the top 5000 documents ranked by PL2F, and $in500$ corresponds to the sample $TopRet_q$ formed by the top 500 documents ranked by I(n_e)C2F). The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

The subsequent tables present the evaluation results for the score-independent experiments, which identify at least one decision boundary for all the tested topic sets, when the outcomes are computed for the set Ret_q, and for both $pl5000$ and $pl500$, or

both $in5000$ and $in500$. The reported evaluation results for the score-dependent experiments correspond to the experiments, which identify at least one decision boundary for all the tested topic sets, when the outcomes are computed for the set Ret_q, and for all four samples $pl5000$, $pl500$, $in5000$, and $in500$. This choice is made in order to focus the evaluation on the experiments that can effectively be used for selective Web IR, and to avoid situations that result in the application of one retrieval approach for all queries. The cases for which the decision mechanism does not have at least one decision boundary for a particular task are denoted by '– –' in the tables.

6.5.3 Document sampling for score-independent document-level experiments

Table 6.7 shows the evaluation results of the document-level experiments $\mathcal{E}_{\forall(b)}$ and $\mathcal{E}_{\forall(at)}$ for sampling with either PL2F or I(n_e)C2F. The results from Table 6.7 are summarised in the form of a histogram in Figure 6.10. When the experiment $\mathcal{E}_{\forall(b)}$ is used with a sample of 5000 documents, the number of identified decision boundaries is low (rows 1-6 in columns '$pl5000$' and '$in5000$'). The obtained mean average precision (MAP) is higher than that of the baseline, with the exception of sampling 5000 documents with PL2F for the task hp2004 (row 4).

Compared to the MAP obtained from the decision mechanism without document sampling (column '$\mathcal{E}_{\forall(b)}$'), there are some fluctuations resulting from the document sampling. For example, the MAP achieved by using the experiment $\mathcal{E}_{\forall(b)}$ without document sampling is +7.27% above the baseline for the task td2004 (row 2 of column '$\mathcal{E}_{\forall(b)}$'). However, when sampling the top 5000 ranked documents with PL2F, the relative difference in MAP between the baseline and the decision mechanism drops to +2.52% (row 2 and column '$pl5000$').

When sampling the top 5000 ranked documents with either PL2F or I(n_e)C2F, the experiment $\mathcal{E}_{\forall(at)}$ performs well in retaining the improvements in MAP of the decision mechanism without sampling, and identifying only 1 decision boundary for each of the tested tasks. For example, the performance of the decision mechanism without sampling, or with sampling 5000 documents for the task td2003 remains the same (row 7 and columns '$\mathcal{E}_{\forall(at)}$', '$pl5000$', and '$in5000$').

When the sample is reduced to 500 documents, then the choice of the weighting model to use for sampling has a more considerable effect on the performance of the

decision mechanism. For example, sampling the top 500 documents with PL2F, and using the experiment $\mathcal{E}_{\forall(at)}$ for the task np2003, does not result in identifying any decision boundary (row 11 of column '*pl*500' in Table 6.7). However, sampling with I(n_e)C2F results in identifying one decision boundary (row 11 and column '*in*500' in Table 6.7). This is related to the fact that for small sample sizes, the top ranked documents are more likely to depend on the employed weighting model.

From the obtained results, it can be seen that the experiment $\mathcal{E}_{\forall(at)}$ results in improvements over the baseline for all tasks when used with a sample of the top 5000 documents, ranked with either PL2F or I(n_e)C2F (rows 7-12 and columns *pl*5000 and *in*5000).

Row	Task	Baseline	$\mathcal{E}_{\forall(b)}$		*pl*5000		*pl*500		*in*5000		*in*500	
1	td2003	0.1455	+1.44	2	+6.67	1	+2.68	1	+7.29	1	+10.10	3
2	td2004	0.1307	+7.27	2	+2.52	1	+0.15	2	+1.22	1	-1.80	2
3	hp2003	0.6660	$+4.23^{\dagger *}$	1	+2.69	1	+0.47	1	+2.70	1	+2.82	3
4	hp2004	0.5555	+1.44	1	-0.63	1	+0.34	1	+1.78	2	+6.41	2
5	np2003	0.6846	+1.37	1	+2.40	1	+2.82	1	+1.34	1	–	–
6	np2004	0.6944	+5.72	1	+1.14	1	+3.11	1	+1.93	1	–	–
Row	Task	Baseline	$\mathcal{E}_{\forall(at)}$		*pl*5000		*pl*500		*in*5000		*in*500	
7	td2003	0.1455	$+7.77^{\dagger *}$	1	$+7.77^{\dagger *}$	1	+6.74	1	$+7.77^{\dagger *}$	1	$+7.49^{\dagger *}$	1
8	td2004	0.1307	+1.15	1	+1.22	1	$+5.43^{\dagger *}$	2	+1.76	1	+4.28	1
9	hp2003	0.6660	+2.15	1	+1.34	1	+1.50	1	+1.38	1	+1.26	1
10	hp2004	0.5555	+5.69	2	+4.79	1	+3.91	2	+1.57	1	+5.11	2
11	np2003	0.6846	$+3.58^{\dagger}$	1	+2.40	1	–	–	+2.40	1	$+2.38^{\dagger}$	1
12	np2004	0.6944	+2.97	1	+0.75	1	+1.21	1	+0.75	1	+3.05	1

Table 6.7: The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs score-independent document level experiments with document sampling of 5000 and 500 top ranked documents with PL2F (*pl*5000 and *pl*500), and I(n_e)C2F (*in*5000 and *in*500), using the default parameter setting. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

6.5.4 Document sampling for score-independent aggregate-level experiments

This section is focused on the evaluation of the score-independent aggregate-level experiments. It is organised in three parts. Each of the parts evaluates the application

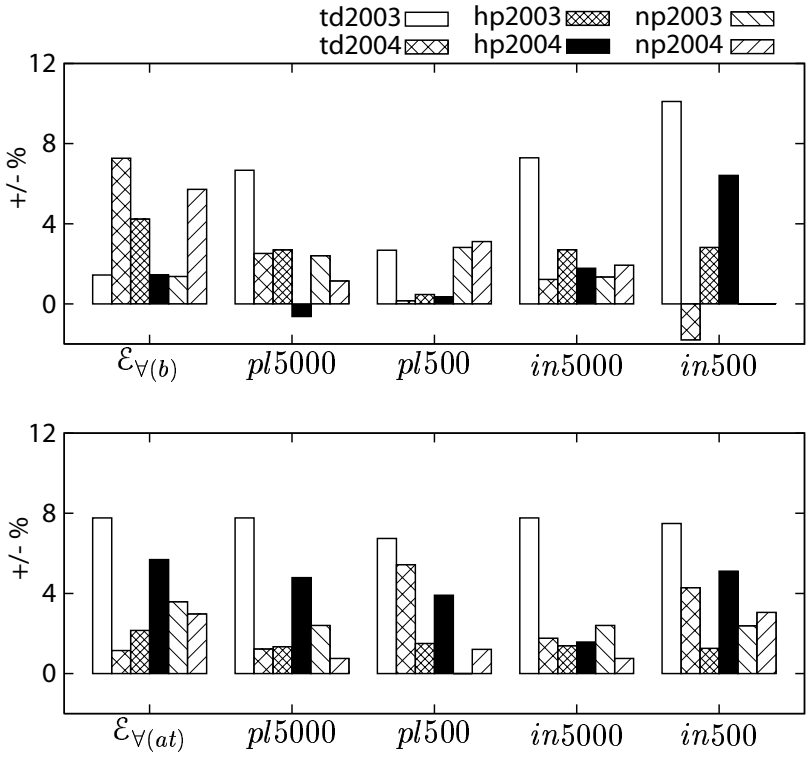

Figure 6.10: Histogram summarising the relative differences between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.7.

of sampling with the experiments that compute the average of the aggregate size distribution, its standard deviation, and the number of large aggregates, respectively, as described in Section 5.3.2 (page 112).

6.5.4.1 Average of the aggregate size distribution

Table 6.8 displays the results from document sampling with the experiments that compute the average size of aggregates. Figure 6.11 presents an overview of the results from Table 6.8 in the form of a histogram.

When the experiment $\mathcal{E}_{\exists(b),avg(dom)}$ is used with sampling of documents, then all sampling methods result in a relatively high number of decision boundaries (rows 1-6). For example, when sampling the top 5000 documents with I(n_e)C2F, there are 4 identified decision boundaries for each of the tasks td2004, hp2003, hp2004, and np2003 (rows 2-5).

In the case of the experiment $\mathcal{E}_{\forall(b),avg(dom)}$, sampling 500 documents with I(n_e)C2F results in improvements in retrieval effectiveness and two decision boundaries for each of the tested tasks (rows 7-12 and column '*in*500').

In the case of directory based aggregates, the experiment $\mathcal{E}_{\exists(at),avg(dir)}$ performs well when used with document sampling, and it results in a relatively low number of thresholds (rows 19-24 in Table 6.8). In particular, sampling 500 documents with I(n_e)C2F results in one decision boundary and improvements in retrieval effectiveness for all tested topic sets (rows 19-24 and column '*in*500').

When the decision mechanism employs the experiment $\mathcal{E}_{\forall(at),avg(dir)}$, sampling with either the weighting models PL2F or I(n_e)C2F produces a variable number of decision boundaries and has a mixed effect in the retrieval effectiveness of the decision mechanism (rows 25-30).

6.5.4.2 Standard deviation of the aggregate size distribution

This section discusses the effect of document sampling on the performance of the aggregate-level experiments that compute the standard deviation of the aggregate size distribution. Table 6.9 displays the obtained results, and Figure 6.12 provides an overview of the results in the form of a histogram.

Using the experiment $\mathcal{E}_{\forall(b),std(dom)}$ with document sampling results in a variable number of decision boundaries (rows 1-6 in the last 4 columns in Table 6.9). In particular, the decision mechanism does not detect any decision boundary for the task td2004 (row 2 and column '*pl*5000'), and the named page finding tasks (rows 5-6 in column '*pl*500').

In the case of directory aggregates, the experiments $\mathcal{E}_{\exists(b),std(dir)}$, $\mathcal{E}_{\forall(b),std(dir)}$, and $\mathcal{E}_{\exists(at),std(dir)}$ result in improvements over the baseline (rows 7-12 and column '*in*5000', rows 13-18 and column '*pl*500', rows 19-24 and column '*pl*5000', respectively, from Table 6.9). However, the results depend on the employed field-based weighting model for sampling. For example, the experiment $\mathcal{E}_{\forall(b),std(dir)}$ results in an improvement of +1.86% for td2003, when sampling the top 5000 documents with PL2F. The same experiments results in an improvement of +9.14% for the same task, when sampling the top 5000 documents with I(n_e)C2F.

Row	Task	Baseline	$\mathcal{E}_{\exists(b),avg(dom)}$		$pl5000$		$pl500$		$in5000$		$in500$	
1	td2003	0.1455	+1.86†	1	+0.82	1	+3.44	3	-2.80	1	+7.63†	2
2	td2004	0.1307	+3.08	3	+10.18†	4	+0.31	3	+7.57	4	+1.91	2
3	hp2003	0.6660	+1.08	3	+3.36	2	+3.17	3	+1.55	4	+2.06	1
4	hp2004	0.5555	+11.65†	1	+6.23	2	+5.76	3	+4.46	4	+5.72	3
5	np2003	0.6846	+1.21	1	+0.22	3	+1.31	3	+0.67	4	-0.28	1
6	np2004	0.6944	+3.50	2	+1.41	2	-1.40	3	+3.31	3	+0.95	1
Row	Task	Baseline	$\mathcal{E}_{\forall(b),avg(dom)}$		$pl5000$		$pl500$		$in5000$		$in500$	
7	td2003	0.1455	-1.79	3	+2.34	3	+15.53†	3	+1.92	2	+3.37	2
8	td2004	0.1307	+6.04†	3	+4.82	1	-2.40	2	+5.74	1	+2.14	2
9	hp2003	0.6660	-1.01	2	+2.21†	5	+4.83†	2	+0.80	5	+3.18†	2
10	hp2004	0.5555	+8.98	2	+6.08	2	+9.38	4	+7.18	4	+9.85	2
11	np2003	0.6846	+2.70	2	+3.72	1	+2.06	2	+5.02	1	+2.22	2
12	np2004	0.6944	+0.88	2	-1.60	4	+3.96	2	-0.58	2	+2.64	2
Row	Task	Baseline	$\mathcal{E}_{\exists(b),avg(dir)}$		$pl5000$		$pl500$		$in5000$		$in500$	
13	td2003	0.1455	+1.92†	1	+13.06$^{\dagger *}$	2	+5.70	2	+1.58	1	+3.16	1
14	td2004	0.1307	+2.22	2	+7.50†	4	+0.77	2	+7.88	4	+5.74†	5
15	hp2003	0.6660	+1.23	4	+2.39	4	+0.41	2	-1.13†	3	+2.09†	2
16	hp2004	0.5555	+15.93$^{\dagger *}$	2	+6.34	3	+10.62†	2	+3.89	2	+6.16†	3
17	np2003	0.6846	+2.91	4	+3.80†	5	+3.17†	4	+5.57$^{\dagger *}$	3	+2.42	2
18	np2004	0.6944	+0.45	2	+1.56	4	+0.85	2	+5.72$^{\dagger *}$	2	+0.53	3
Row	Task	Baseline	$\mathcal{E}_{\exists(at),avg(dir)}$		$pl5000$		$pl500$		$in5000$		$in500$	
19	td2003	0.1455	+2.89†	1	+2.13	1	+5.29	2	+1.51	3	+3.99	1
20	td2004	0.1307	+7.96†	3	+3.06	2	-1.80	3	+1.91	2	+1.07	1
21	hp2003	0.6660	+2.93†	4	+1.34	2	-1.80	2	-0.27	3	+1.98†	1
22	hp2004	0.5555	+6.26	4	+4.37	2	+3.20	2	+5.99	1	+1.76	1
23	np2003	0.6846	+3.71†	3	+0.66	1	+1.50	2	+1.31	2	+0.53	1
24	np2004	0.6944	+3.92	4	+4.65	2	+4.94	2	-1.20	2	+1.21	1
Row	Task	Baseline	$\mathcal{E}_{\forall(at),avg(dir)}$		$pl5000$		$pl500$		$in5000$		$in500$	
25	td2003	0.1455	+10.86	2	+2.47	2	+6.80	2	+4.26	2	+5.50	2
26	td2004	0.1307	+2.37	1	+0.46	3	0.00	2	0.00	2	0.00	2
27	hp2003	0.6660	+2.64	7	-1.70	4	-0.03	3	-0.29	3	-0.92	4
28	hp2004	0.5555	+13.03$^{\dagger *}$	3	+10.66$^{\dagger *}$	1	+9.45†	3	+10.66$^{\dagger *}$	1	+9.45†	3
29	np2003	0.6846	+0.22	5	+1.99	1	+1.74	1	+1.49	1	+0.74	1
30	np2004	0.6944	+1.20	1	+0.49	1	+1.97	3	+0.49	1	+2.85	1

Table 6.8: The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. The experiments compute the average domain or directory aggregate sizes. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

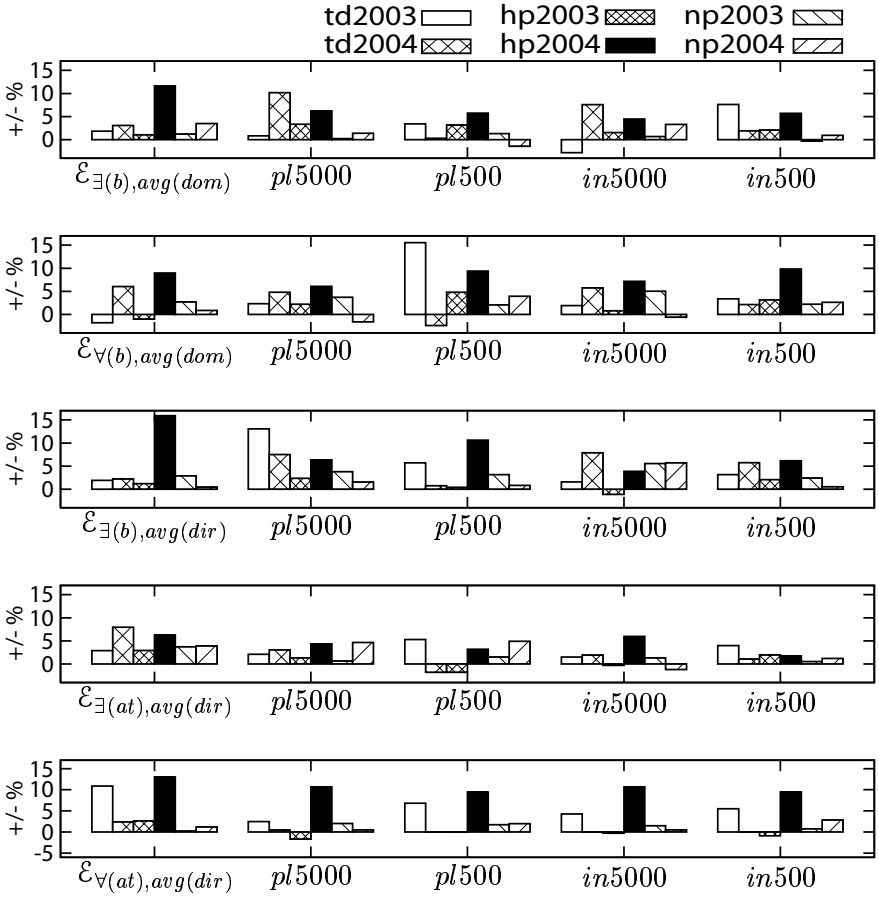

Figure 6.11: Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.8.

6.5.4.3 Number of large aggregates

Table 6.10 displays the results obtained by a decision mechanism, which employs the experiments that count the number of large aggregates with document sampling. Figure 6.13 provides an overview of the results from Table 6.10 in the form of a histogram.

When domain aggregates are used, the experiment $\mathcal{E}_{\exists(b),lrg(dom)}$ identifies at least one decision boundary for all the tested topic sets, and both sample sizes of 5000 and 500 documents, respectively (rows 1-6 and last 4 columns in Table 6.10). However, the number of decision boundaries varies for each task. On the other hand, when the employed experiment is $\mathcal{E}_{\forall(b),lrg(dom)}$, then sampling 5000 documents with either the weighting models PL2F or I(n_e)C2F results in a less variable number of decision

Row	Task	Baseline	$\mathcal{E}_{\forall(b),std(dom)}$		$pl5000$		$pl500$		$in5000$		$in500$	
1	td2003	0.1455	+4.81	1	-5.09	1	+10.38	2	-5.09	1	+1.79	2
2	td2004	0.1307	+3.52	1	–	–	+2.68	1	-1.84	1	+0.84	1
3	hp2003	0.6660	+0.33	1	+1.98	3	+2.91	1	+3.59	3	-1.23	1
4	hp2004	0.5555	+1.21	1	+11.95†	4	+5.11	2	+7.81	4	+5.58	2
5	np2003	0.6846	+5.61	1	+4.83†	1	–	–	+5.84†	1	+2.86	1
6	np2004	0.6944	+3.46	1	+1.31	1	–	–	+1.11	1	+2.68	1
Row	Task	Baseline	$\mathcal{E}_{\exists(b),std(dir)}$		$pl5000$		$pl500$		$in5000$		$in500$	
7	td2003	0.1455	-2.30	1	-2.27	2	+6.12	1	+12.44	3	+2.68	1
8	td2004	0.1307	+0.84	2	+8.72†	4	-0.61	2	+7.88†	2	+0.23	2
9	hp2003	0.6660	+0.59†	3	+0.83†	3	+0.62	1	+0.21	2	+2.24†	1
10	hp2004	0.5555	+8.73	2	+6.97	3	+8.01	2	+9.04	1	+4.77	2
11	np2003	0.6846	+1.29	2	+2.16	2	+4.31†	2	+2.25	1	-0.13	1
12	np2004	0.6944	+2.30	4	+1.87	4	-1.90	4	+0.84	3	+6.74$^{\dagger *}$	3
Row	Task	Baseline	$\mathcal{E}_{\forall(b),std(dir)}$		$pl5000$		$pl500$		$in5000$		$in500$	
13	td2003	0.1455	+7.56	2	+1.86	2	+7.77	1	+9.14	2	+4.12	1
14	td2004	0.1307	+3.98	3	+3.60	1	+1.61	1	+3.52	2	+2.30	3
15	hp2003	0.6660	+0.75	1	0.00	1	+2.03	1	-1.90	2	+3.21†	1
16	hp2004	0.5555	-0.68	1	+11.38	3	+5.51	2	+2.21	1	+6.21	2
17	np2003	0.6846	+3.27	1	+3.16	1	+1.80	1	-0.83	2	–	–
18	np2004	0.6944	+4.57	1	+0.23	1	+2.89	2	+2.07	1	+4.33	1
Row	Task	Baseline	$\mathcal{E}_{\exists(at),std(dir)}$		$pl5000$		$pl500$		$in5000$		$in500$	
19	td2003	0.1455	-12.00	1	+2.68	3	+4.67	2	+4.54	2	+5.91	2
20	td2004	0.1307	+1.07	1	+7.19	2	-1.30	1	+2.07	2	–	–
21	hp2003	0.6660	+2.84	2	+1.05	1	-0.83	1	-0.20	1	-0.20	2
22	hp2004	0.5555	+6.17	2	+4.55	2	+5.83†	4	+4.07	2	+4.46	1
23	np2003	0.6846	+3.87†	2	+3.07	2	+1.87	2	+1.55	1	+1.68	1
24	np2004	0.6944	+1.38	1	+6.01	1	+4.62	1	+6.06	3	+3.60	2

Table 6.9: The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. The experiments compute the standard deviation of the domain or directory aggregate sizes. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

boundaries for each of the tested tasks (rows 7-12 and columns '$pl5000$', '$in5000$', respectively).

Regarding the directory aggregates, the combination of the experiment $\mathcal{E}_{\exists(at),lrg(dir)}$ and sampling of the top 5000 ranked documents according to PL2F results in a low number of decision boundaries (rows 19-24 and column '$pl5000$' in Table 6.10). However, it results in reduced precision for the task td2003, compared to the baseline (row

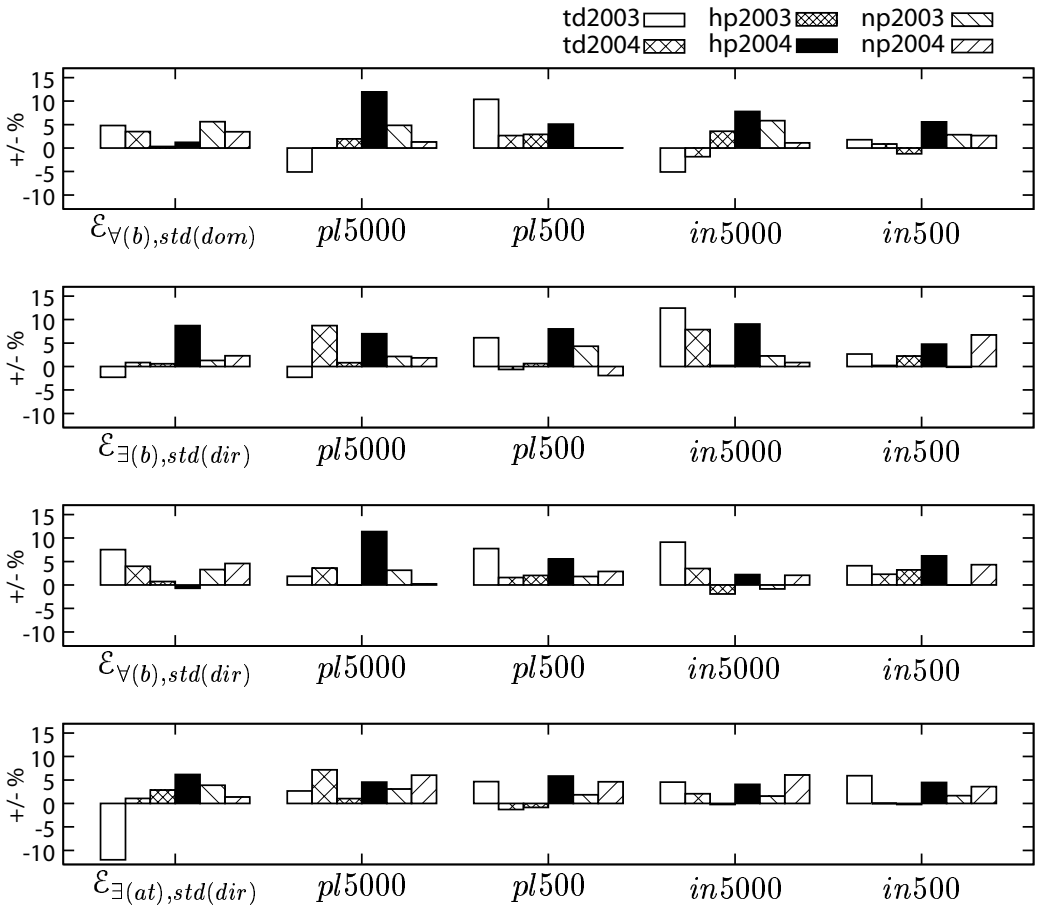

Figure 6.12: Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.9.

19 and column '$pl5000$').

6.5.5 Document sampling for score-dependent experiments

Table 6.11 presents the evaluation of the decision mechanism, which employs the experiments that estimate the usefulness of the hyperlink structure $L(S_n, U_n)$ (rows 1-12) and $L(S_n, U'_n)$ (rows 13-24). An overview of the results is also provided in the form of a histogram in Figure 6.14. The experiments $\mathcal{E}_{\exists(at),L(SU)_{pl}}$ and $\mathcal{E}_{\exists(at),L(SU)_{in}}$ can be effectively used with document sampling, with the exception of the task hp2003 and sampling 500 documents with PL2F (row 3 and column '$pl500$' from Table 6.11). In most of the cases, the decision mechanism results in more than one decision boundaries,

Row	Task	Baseline	$\mathcal{E}_{\exists(b),lrg(dom)}$		$pl5000$		$pl500$		$in5000$		$in500$	
1	td2003	0.1455	+0.55	4	+9.28	2	+9.62	2	+9.90	3	+2.27	3
2	td2004	0.1307	$+7.04^{\dagger}$	2	-1.10	3	+2.60	2	+0.38	2	+1.61	1
3	hp2003	0.6660	+0.89	2	0.00	2	-1.40	2	-0.59	3	+2.15	2
4	hp2004	0.5555	$+9.16^{\dagger}$	1	+7.27	2	+5.29	1	+2.07	2	+4.77	1
5	np2003	0.6846	+0.72	1	+2.21	2	$+1.91^{\dagger}$	2	+0.72	4	+1.33	2
6	np2004	0.6944	+2.61	2	-0.99	3	+0.91	2	+1.07	2	+0.12	2
Row	Task	Baseline	$\mathcal{E}_{\forall(b),lrg(dom)}$		$pl5000$		$pl500$		$in5000$		$in500$	
7	td2003	0.1455	+5.43	2	+3.23	1	$+7.01^{\dagger}$	1	+6.12	1	+0.55	1
8	td2004	0.1307	+5.43	5	$+5.05^{\dagger}$	1	+2.07	1	+5.66	1	+1.61	2
9	hp2003	0.6660	$+3.32^{\dagger *}$	1	+2.03	1	+0.74	5	$+2.63^{\dagger}$	2	+2.87	1
10	hp2004	0.5555	-1.30	1	+4.81	3	+4.57	4	+3.01	2	+5.71	5
11	np2003	0.6846	+0.50	2	-0.20	1	+0.79	3	+1.93	2	+4.46	1
12	np2004	0.6944	+0.22	2	-0.69	1	-2.40	1	+0.86	1	+4.95	2
Row	Task	Baseline	$\mathcal{E}_{\exists(b),lrg(dir)}$		$pl5000$		$pl500$		$in5000$		$in500$	
13	td2003	0.1455	$+6.32^{\dagger}$	2	+6.12	2	$+7.29^{\dagger}$	2	$+16.36^{\dagger *}$	3	+5.57	2
14	td2004	0.1307	-0.92	2	$+10.02^{\dagger}$	3	+3.14	1	+6.89	3	-0.23	1
15	hp2003	0.6660	+0.78	2	+0.36	2	-0.41	1	+0.57	2	+2.13	1
16	hp2004	0.5555	+8.77	1	+6.26	4	+4.19	2	$+12.10^{\dagger *}$	3	+2.90	2
17	np2003	0.6846	+0.38	1	-0.16	2	+0.34	1	+1.88	2	-0.66	1
18	np2004	0.6944	+2.71	2	+4.59	3	+0.26	2	$+10.21^{\dagger *}$	2	$+6.05^{*}$	2
Row	Task	Baseline	$\mathcal{E}_{\exists(at),lrg(dir)}$		$pl5000$		$pl500$		$in5000$		$in500$	
19	td2003	0.1455	+0.96	3	$-5.70^{\dagger}$	1	$+14.50^{\dagger}$	3	$-0.62^{\dagger}$	1	-2.50	1
20	td2004	0.1307	-0.54	1	+0.31	1	+1.37	2	+0.69	1	+1.30	2
21	hp2003	0.6660	+1.62	6	+1.10	1	$+2.82^{\dagger}$	2	+3.27	4	$+4.19^{\dagger *}$	1
22	hp2004	0.5555	-0.09	1	+4.34	2	-0.99	2	+0.94	2	+3.92	2
23	np2003	0.6846	+6.67	1	+2.95	1	–	–	+0.76	3	+5.62	2
24	np2004	0.6944	+4.97	1	+1.57	1	$+5.99^{\dagger *}$	2	+1.92	2	+3.07	2

Table 6.10: The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. The experiments compute the number of large domain or directory aggregates. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

similarly to the case where no sampling is used (Section 6.4.2 on page 154)

The experiments $\mathcal{E}_{\forall(b),L(SU')_{pl}}$ and $\mathcal{E}_{\forall(b),L(SU')_{in}}$ also result in improvements over the baseline (rows 13-18 and 19-24 in Table 6.11, respectively). In particular, the sampling of 500 documents with either PL2F, or I(n_e)C2F, to compute the outcome of $\mathcal{E}_{\forall(b),L(SU')_{pl}}$ and $\mathcal{E}_{\forall(b),L(SU')_{in}}$, respectively, results in one decision boundary for the topic distillation and the home page finding tasks (rows 13-16 in column '$pl500$', and

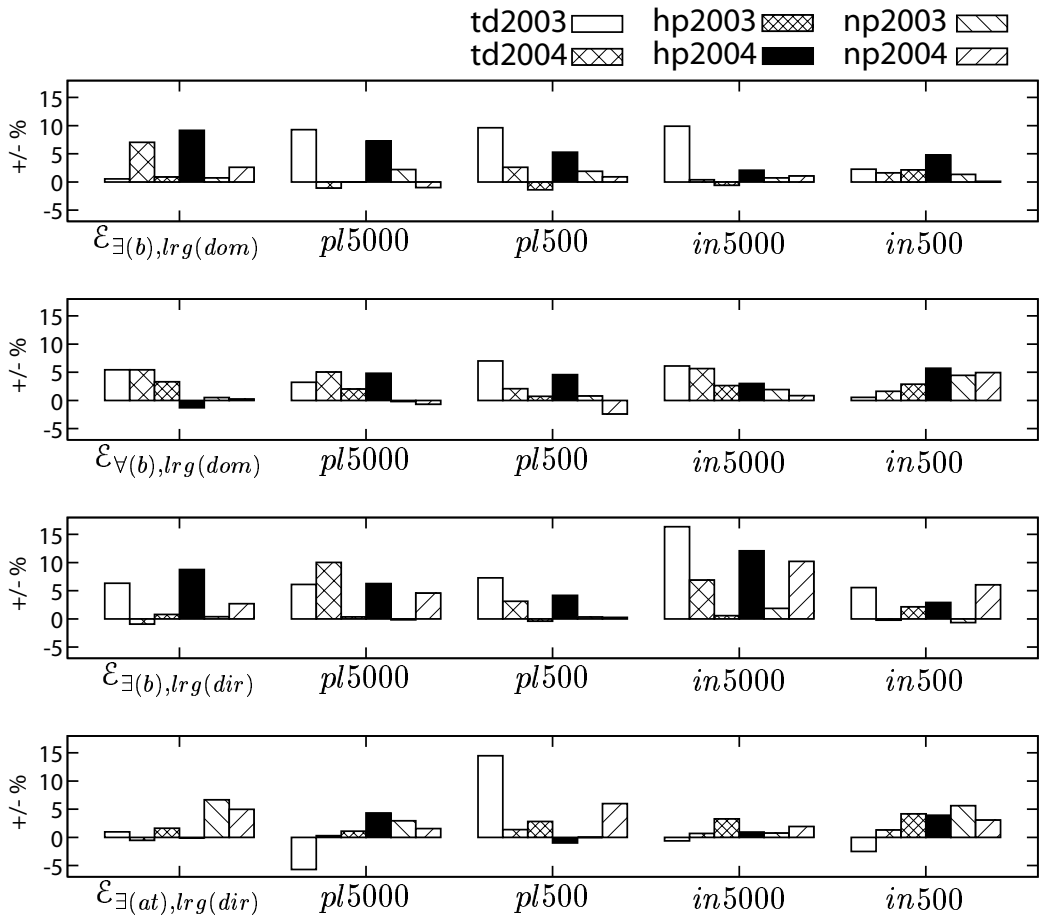

Figure 6.13: Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.10.

rows 19-22 in column '$in500$', respectively).

6.5.6 Discussion

This section presents a discussion related to document sampling from the perspectives of: the effectiveness of experiments $\mathcal{E}$ with document sampling; the size of document samples; and generating document samples with different retrieval approaches.

Effective experiments with document sampling The score-independent document-level experiments $\mathcal{E}_{\forall(b)}$ and $\mathcal{E}_{\forall(at)}$ are particularly effective with document sampling, as discussed in Section 6.5.3 and shown in Table 6.7. The score-dependent experi-

Row	Task	Baseline	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$		$pl5000$		$pl500$	
1	td2003	0.1455	$+1.99^{\dagger}$	3	$+7.90^{\dagger *}$	2	$+2.75^{\dagger}$	2
2	td2004	0.1307	+2.30	3	+1.15	2	$+5.89^{\dagger}$	2
3	hp2003	0.6660	+2.04	5	+0.24	2	$-0.30^{\dagger}$	3
4	hp2004	0.5555	+5.80	1	+2.72	3	+5.18	3
5	np2003	0.6846	+1.46	3	+2.06	3	+1.17	2
6	np2004	0.6944	+6.31	2	$+6.98^{\dagger *}$	5	+2.36	2
Row	Task	Baseline	$\mathcal{E}_{\exists(at),L(SU)_{in}}$		$in5000$		$in500$	
7	td2003	0.1455	$+3.09^{\dagger}$	3	$+3.99^{\dagger}$	2	$+5.70^{\dagger *}$	2
8	td2004	0.1307	$+4.59^{\dagger}$	2	+2.91	2	$+10.33^{\dagger *}$	3
9	hp2003	0.6660	+1.68	4	$+1.97^{\dagger}$	2	$+2.57^{\dagger}$	1
10	hp2004	0.5555	+6.80	2	+6.91	2	+4.18	2
11	np2003	0.6846	+2.50	3	$+4.73^{\dagger}$	3	+3.18	3
12	np2004	0.6944	+4.13	3	+4.57	2	+0.75	2
Row	Task	Baseline	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$		$pl5000$		$pl500$	
13	td2003	0.1455	$+13.75^{\dagger *}$	3	$+10.45^{\dagger *}$	3	+7.22	1
14	td2004	0.1307	+2.75	3	+1.61	2	$+1.07^{\dagger}$	1
15	hp2003	0.6660	+2.54	1	+1.44	1	$+3.32^{\dagger}$	1
16	hp2004	0.5555	$+6.91^{*}$	3	+5.33	2	+3.01	1
17	np2003	0.6846	+0.99	2	$+4.46^{\dagger}$	3	+2.41	2
18	np2004	0.6944	+3.30	2	+1.47	2	+4.49	2
Row	Task	Baseline	$\mathcal{E}_{\forall(b),L(SU')_{in}}$		$in5000$		$in500$	
19	td2003	0.1455	$+7.90^{\dagger}$	3	$+11.68^{\dagger *}$	3	+4.26	1
20	td2004	0.1307	+1.53	2	+2.07	1	$+2.07^{\dagger}$	1
21	hp2003	0.6660	+2.34	1	$+3.02^{\dagger}$	1	+1.77	1
22	hp2004	0.5555	+8.03	3	+4.82	2	+2.29	1
23	np2003	0.6846	+3.11	4	+2.07	3	+1.80	3
24	np2004	0.6944	+3.93	2	+4.75	2	+1.71	2

Table 6.11: The relative difference between the MAP of a decision mechanism and that of the most effective individual retrieval approach, and the corresponding number of decision boundaries. The decision mechanism employs the score-dependent experiments and document sampling of 5000 and 500 top ranked documents with PL2F ($pl5000$ and $pl500$), and I(n_e)C2F ($in5000$ and $in500$), using the default parameter setting. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

ments that estimate the usefulness of the hyperlink structure $L(S_n, U_n)$ or $L(S_n, U'_n)$ are also effectively applied with document sampling (Section 6.5.5 and Table 6.11). A reason that can explain these facts is the following: when the outcome of either the score-independent document-level experiments, or the score-dependent experiments is computed from the same number of documents, then a similar amount of information is considered for each query, making the comparison between the outcome values of

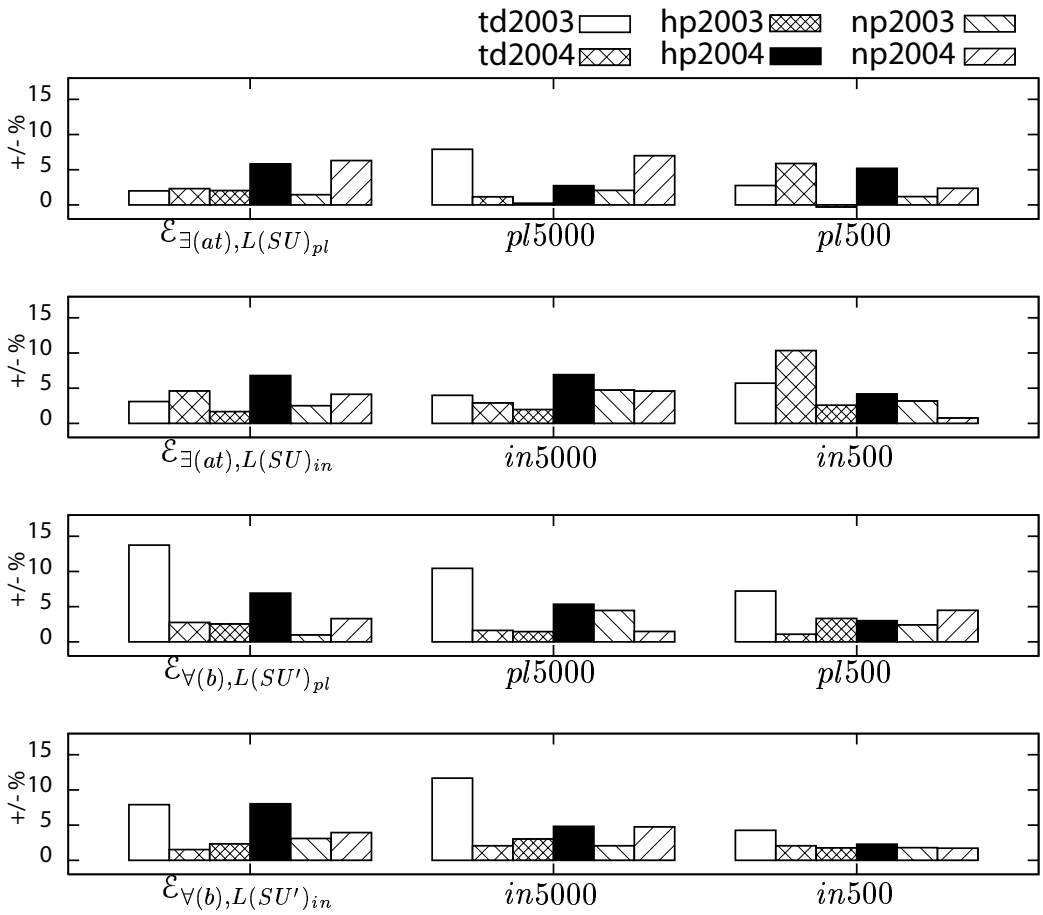

Figure 6.14: Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from Table 6.11.

the experiments easier. The score-independent aggregate-level experiments have not been shown to be particularly effective with document sampling (Section 6.5.4 and Tables 6.8, 6.9, and 6.10). This may be explained by the fact that since the outcome of the aggregate-level experiments is based on the distribution of aggregates, and not on the distribution of documents, then more documents are required in order to obtain a representative distribution of aggregates.

Size of document samples In the evaluation of the experiments $\mathcal{E}$ with document sampling, the document samples consisted of 5000 or 500 documents. The score-independent document-level experiments, which count the number of documents in

which the query terms occur, resulted in improvements in retrieval effectiveness, and a low number of thresholds for the samples of 5000 documents. However, their performance was harmed for the samples of 500 documents (Section 6.5.3). This is because smaller document samples reduce the available information to compute the outcome of the experiments. On the other hand, the score-dependent experiments have been shown to be robust for samples of 500 documents (Section 6.5.5). This is because the score-dependent experiments, which estimate the usefulness of the hyperlink structure, employ all the outgoing links from the sample of documents to the whole set of retrieved documents, as described in Section 6.5.1. Therefore, even for small samples of documents, the experiments consider more information from the hyperlink structure.

Generating document samples with different retrieval approaches The document samples have been generated with two different field-based weighting models, namely PL2F and I(n_e)C2F. The experiments that exhibited a weak dependence on the particular weighting model used for sampling were the score-independent document-level $\mathcal{E}_{\forall(b)}$ and $\mathcal{E}_{\forall(at)}$ (Section 6.5.3). This is explained, because these experiments simply count the number of documents in which query terms occur. On the other hand, the score-dependent experiments that compute the usefulness of the hyperlink structure explicitly employ the score of documents (Section 6.5.5). Consequently, their performance is more dependent on the employed weighting model, especially for small document samples.

6.5.7 Conclusions

This section has evaluated the proposed experiments $\mathcal{E}$ when their outcomes are computed from small sets of documents. The evaluation results have shown that document sampling can be effectively used to reduce the computational cost of the experiments, while still retaining the improvements in retrieval effectiveness.

Document sampling is used more effectively with either the score-independent document-level experiments (Section 6.5.3), or the score-dependent experiments that compute the usefulness of the hyperlink structure (Section 6.5.5). The score-independent aggregate-level experiments do not perform as well as when no sampling is used (Section 6.5.6).

When the document sample is considerably reduced, then the document-level experiment are less effective. The score-dependent experiments perform well, but also exhibit

a stronger dependence on the particular method used for performing the sampling of documents (Section 6.5.6).

6.6 Using retrieval approaches based on the same weighting model

The evaluation results presented in Sections 6.3 and 6.4 refer to a Bayesian decision mechanism that selectively applies retrieval approaches, which use different field-based weighting models. This section discusses the evaluation of the experiments $\mathcal{E}$, when the decision mechanism employs retrieval approaches that use the same field-based weighting models. For example, the Bayesian decision mechanism can selectively apply the field-based weighting model PL2F, or its combination with PageRank (PL2FP) for the task td2003. Hence, the employed experiment $\mathcal{E}$ is required to identify the most effective retrieval approach, based on differences due to the used sources of query-independent evidence. The remainder of this section aims to identify which of the field-based weighting models can be used more effectively in the context of a Bayesian decision mechanism, which employs the proposed experiments $\mathcal{E}$ to selectively apply combinations of a particular field-based weighting model and query-independent sources of evidence on a per-query basis.

The experimental setting is the following. A Bayesian decision mechanism employs pairs of retrieval approaches, which are restricted to use the same field-based weighting model. For each of the tested tasks (td2003, td2004, hp2003, hp2004, np2003, and np2004), and each of the field-based weighting models (PL2F, PB2F, I(n_e)C2F, DLHF, and BM25F), the decision mechanism employs the pairs of retrieval approaches that result in the highest potential for improvements in retrieval effectiveness (rows 1-30 in Table 4.11, page 100). For example, in the case of the task td2003 and the weighting model BM25F, the Bayesian decision mechanism selectively applies either the combination of BM25F with evidence from the URL path length (BM25FU), or the combination of BM25F with PageRank (BM25FP) (row 25 in Table 4.11). Overall there are 11 different experiments: 1 score-independent document-level experiment; 6 score-independent domain and directory aggregate-level experiments, which compute the average, the standard deviation, and the number of large aggregates; and 4 score-dependent experiments which compute the divergences $L(S_n, U_n)$ and $L(S_n, U'_n)$ by

setting the distribution S_n with either PL2F or I(n_e)C2F. By considering the 6 tested tasks, the body (b) and the combination of the anchor text and title fields (at), and the conditions $\exists$ and $\forall$, each type of experiment has 24 different configurations. The total number of configurations for the 11 different experiments is $24 \cdot 11 = 264$.

Table 6.12 provides an overview of the evaluation results, with respect to: the number of times a particular type of experiments identifies at least one decision boundary (column 'B>0'); and the number of times a particular type of experiments results in improvements in mean average precision, compared to the most effective individual retrieval approach (column '+'). These numbers are given for each of the five field-based weighting models. For each type of experiments, the column 'Table' indicates the table in Appendix B, which contains the evaluation results for the corresponding experiment $\mathcal{E}$. The row 'Total' of the table displays the sum of the corresponding columns for all the experiments $\mathcal{E}$. The row 'Ratio + / B>0' corresponds to the ratio of the number of times when there is an improvement in MAP from selective Web IR, over the number of times when there is at least one decision boundary.

		PL2F		PB2F		I(n_e)C2F		DLHF		BM25F	
Experiment $\mathcal{E}$	Table	B>0	+	B>0	+	B>0	+	B>0	+	B>0	+
Score-independent experiments											
$\mathcal{E}_{\exists(f)}$, $\mathcal{E}_{\forall(f)}$	B.1	18	11	20	13	12	7	14	6	20	10
$\mathcal{E}_{\exists(f),avg(dom)}$, $\mathcal{E}_{\forall(f),avg(dom)}$	B.2	18	10	19	15	16	12	16	11	24	14
$\mathcal{E}_{\exists(f),std(dom)}$, $\mathcal{E}_{\forall(f),std(dom)}$	B.3	17	8	17	11	15	8	18	13	23	13
$\mathcal{E}_{\exists(f),lrg(dom)}$, $\mathcal{E}_{\forall(f),lrg(dom)}$	B.4	16	11	17	8	12	10	13	8	19	11
$\mathcal{E}_{\exists(f),avg(dir)}$, $\mathcal{E}_{\forall(f),avg(dir)}$	B.5	17	13	20	11	16	9	22	15	24	11
$\mathcal{E}_{\exists(f),std(dir)}$, $\mathcal{E}_{\forall(f),std(dir)}$	B.6	18	12	19	14	16	11	20	14	23	12
$\mathcal{E}_{\exists(f),lrg(dir)}$, $\mathcal{E}_{\forall(f),lrg(dir)}$	B.7	15	8	17	9	11	7	15	7	21	12
Score-dependent experiments											
$\mathcal{E}_{\exists(f),L(SU)_{pl}}$, $\mathcal{E}_{\forall(f),L(SU)_{pl}}$	B.8	19	14	22	16	21	19	21	15	24	20
$\mathcal{E}_{\exists(f),L(SU)_{in}}$, $\mathcal{E}_{\forall(f),L(SU)_{in}}$	B.9	18	13	22	15	20	18	17	11	24	18
$\mathcal{E}_{\exists(f),L(SU')_{pl}}$, $\mathcal{E}_{\forall(f),L(SU')_{pl}}$	B.10	18	14	22	18	15	14	17	10	23	19
$\mathcal{E}_{\exists(f),L(SU')_{in}}$, $\mathcal{E}_{\forall(f),L(SU')_{in}}$	B.11	15	10	23	18	15	10	19	13	23	19
Total		189	124	218	148	169	125	192	123	248	159
Ratio + /B>0		0.66		0.68		0.74		0.64		0.64	

Table 6.12: The number of times for which there is at least one decision boundary ('B>0'), or improvements in retrieval effectiveness ('+'), when the Bayesian decision mechanism selectively applies retrieval approaches, which use the same field-based weighting model.

The results from Table 6.12 indicate that when the Bayesian decision mechanism employs retrieval approaches, which use the weighting model BM25F, there is at least

one decision boundary identified for 248 out of the 264 experiment configurations, and improvements in retrieval effectiveness for 159 experiment configurations (row 'Total'). On the other hand, when the decision mechanism employs retrieval approaches, which use the weighting model I(n_e)C2F, there are only 169 out of 264 configurations of the experiments, which result in at least one decision boundary (row 'Total'). The 125 out of these 169 configurations (0.74%) result in improvements in retrieval effectiveness (rows 'Total' and '+ / B>0'). Therefore, the field-based weighting model BM25F is more appropriate to be used in selective Web IR than I(n_e)C2F. This can be explained by the fact that the restricted optimisation, which has been described in Section 4.6.2, harmed the retrieval effectiveness of BM25F more than that of the Divergence From Randomness (DFR) field-based weighting models. Therefore, the benefit from selective Web IR is greater for the less robust field-based weighting model BM25F.

Table 6.12 also suggests that the score-dependent experiments are particularly robust when they are used to selectively apply retrieval approaches based on the weighting models PB2F and BM25F (rows '$\mathcal{E}_{\exists(f),L(SU)_{pl}}$, $\mathcal{E}_{\forall(f),L(SU)_{pl}}$' to '$\mathcal{E}_{\exists(f),L(SU')_{in}}$, $\mathcal{E}_{\forall(f),L(SU')_{in}}$').

Overall, this section has provided an overview of the evaluation of the proposed experiments, when the Bayesian decision mechanism selectively applies retrieval approaches, which employ the same weighting model. The results suggest that there are improvements in retrieval effectiveness in most of the cases. When both the applied retrieval approaches use the field-based weighting model BM25F, there is at least one identified decision boundary for most of the tested cases (row 'Total' in Table 6.12). The score-dependent experiments are also robust, and they result in improvements in retrieval effectiveness for most of the tested cases.

6.7 Decision mechanism with more than two retrieval approaches

The evaluation of the proposed experiments in this chapter has, so far, been performed with a Bayesian decision mechanism, which uses two retrieval approaches. However, the Bayesian decision mechanism can selectively apply from any number of retrieval approaches, as it has been described in Example 7 of Section 5.5 (page 122). In such

a case, the decisions depend on the expected loss of each retrieval approach, instead of only the posterior likelihood that a given retrieval approach is the most effective.

The current section presents an illustrative example of a Bayesian decision mechanism, which employs 3 retrieval approaches. Chapter 4 has described and evaluated 20 different retrieval approaches (5 field-based weighting models, and their combinations with 3 different sources of query-independent evidence), there are $20 \cdot 19 \cdot 18 = 6840$ ways to select a set of three distinct retrieval approaches. This section presents the evaluation of a particular set of retrieval approaches, which have been selected for being diverse, and using all three different sources of query-independent evidence. The selected approaches are: the combination of the field-based weighting model PL2F with the Absorbing Model (PL2FA); the combination of I(n_e)C2F with evidence from the URL path length (I(n_e)C2FU); and the combination of BM25F with PageRank (BM25FP). The evaluation is performed for each of the tasks: td2003; td2004; hp2003; hp2004; np2003; and np2004.

Table 6.13 displays the evaluation of the decision mechanism that employs the above mentioned retrieval approaches, for the cases when at least one decision boundary is identified for all tasks, and there are improvements in retrieval effectiveness for at least three of the tested tasks. This choice is made in order to focus the analysis on the most effective experiments. Figure 6.15 provides an overview of the results from column '+/- %' of Table 6.13 in the form of a histogram. The results suggest that the decision mechanism can lead to small improvements in retrieval effectiveness over the baseline in some of the cases. For example, the MAP achieved by the decision mechanism with the experiment $\mathcal{E}_{\forall(b),L(SU)_{pl}}$ is 0.1726 (row 19 in Table 6.13). This represents an improvement of +5.24% over the MAP of the most effective individual retrieval approach for the task td2003 (0.1640). When the decision mechanism uses the experiment $\mathcal{E}_{\exists(b),std(dir)}$ for the task hp2004, there is a statistically significant improvement in MAP, and the most appropriate retrieval approach is applied for a statistically significant number of queries (row 16 in Table 6.13). However, there is no experiment that results in improvements in MAP for all the tested tasks (column '+/-%' in Table 6.13). For example, none of the experiments results in improvements for the task td2004. The number of decision boundaries also varies for each of the tested experiments and tasks (column 'Bnd' in Table 6.13). It is worth noting that 6 out of the 9 experiments that identify at least one decision boundary for all tasks, and result in improvements for at least three

of the tested tasks, are score-dependent experiments, which estimate the usefulness of the hyperlink structure (rows 19-54), while there are only 3 score-independent directory aggregate-level experiments (rows 1-18). The results indicate that the score-dependent experiments are more robust than the score-independent experiments in the described setting.

The unstable performance of the Bayesian decision mechanism in the employed setting can be attributed to the fact that the higher number of retrieval approaches require more queries for training the decision mechanism. As described in Section 5.5.2, the estimation of the prior probability that a particular retrieval approach is effective, the estimation of the loss function, and the density estimation of the likelihoods of obtaining a particular experiment outcome, are performed from subsets of the training queries. These subsets correspond to the queries for which a particular retrieval approach is the most effective one. Therefore, as the number of retrieval approaches increases, the size of the training subsets of queries decreases, providing less evidence for setting the Bayesian decision mechanism.

Row	Task	Retrieval Approaches	Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
1	td2003	PL2FA I(n_ϵ)C2FU BM25FP	0.1640	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1623	-1.04	3
2	td2004	PL2FA I(n_ϵ)C2FU BM25FP	0.1763	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1704	-3.35	4
3	hp2003	PL2FA I(n_ϵ)C2FU BM25FP	0.7516	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7574	+0.77	1
4	hp2004	PL2FA I(n_ϵ)C2FU BM25FP	0.6469	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6500	+0.48	1
5	np2003	PL2FA I(n_ϵ)C2FU BM25FP	0.7108	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7180	+1.01	1
6	np2004	PL2FA I(n_ϵ)C2FU BM25FP	0.7169	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7024	-2.02	2
7	td2003	PL2FA I(n_ϵ)C2FU BM25FP	0.1640	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1558	-5.00	2
8	td2004	PL2FA I(n_ϵ)C2FU BM25FP	0.1763	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1714	-2.78	1
9	hp2003	PL2FA I(n_ϵ)C2FU BM25FP	0.7516	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7662	+1.94	3
10	hp2004	PL2FA I(n_ϵ)C2FU BM25FP	0.6469	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6539	+1.08	7
11	np2003	PL2FA I(n_ϵ)C2FU BM25FP	0.7108	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7079	-0.41	3
12	np2004	PL2FA I(n_ϵ)C2FU BM25FP	0.7169	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7393	+3.12	1
13	td2003	PL2FA I(n_ϵ)C2FU BM25FP	0.1640	$\mathcal{E}_{\exists(b),std(dir)}$	0.1558	-5.00	3
14	td2004	PL2FA I(n_ϵ)C2FU BM25FP	0.1763	$\mathcal{E}_{\exists(b),std(dir)}$	0.1711	-2.94	1
15	hp2003	PL2FA I(n_ϵ)C2FU BM25FP	0.7516	$\mathcal{E}_{\exists(b),std(dir)}$	0.7517	+0.01	1
16	hp2004	PL2FA I(n_ϵ)C2FU BM25FP	0.6469	$\mathcal{E}_{\exists(b),std(dir)}$	0.6713	$+3.77^{\dagger *}$	1
17	np2003	PL2FA I(n_ϵ)C2FU BM25FP	0.7108	$\mathcal{E}_{\exists(b),std(dir)}$	0.7063	-0.63	1
18	np2004	PL2FA I(n_ϵ)C2FU BM25FP	0.7169	$\mathcal{E}_{\exists(b),std(dir)}$	0.7316	+2.05	4
19	td2003	PL2FA I(n_ϵ)C2FU BM25FP	0.1640	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1726	+5.24	3
20	td2004	PL2FA I(n_ϵ)C2FU BM25FP	0.1763	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1676	-4.93	4
21	hp2003	PL2FA I(n_ϵ)C2FU BM25FP	0.7516	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7597	+1.08	2
22	hp2004	PL2FA I(n_ϵ)C2FU BM25FP	0.6469	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6261	-3.22	2

continued on next page

continued from previous page							
Row	Task	Retrieval Approaches	Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
23	np2003	PL2FA I(n_e)C2FU BM25FP	0.7108	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7131	+0.32	2
24	np2004	PL2FA I(n_e)C2FU BM25FP	0.7169	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7082	-1.21	3
25	td2003	PL2FA I(n_e)C2FU BM25FP	0.1640	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1614	-1.59	3
26	td2004	PL2FA I(n_e)C2FU BM25FP	0.1763	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1699	-3.63	1
27	hp2003	PL2FA I(n_e)C2FU BM25FP	0.7516	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7558	+0.56	1
28	hp2004	PL2FA I(n_e)C2FU BM25FP	0.6469	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6347	-1.89	4
29	np2003	PL2FA I(n_e)C2FU BM25FP	0.7108	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7140	+0.45	1
30	np2004	PL2FA I(n_e)C2FU BM25FP	0.7169	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7314	+2.02	2
31	td2003	PL2FA I(n_e)C2FU BM25FP	0.1640	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1674	+2.07	1
32	td2004	PL2FA I(n_e)C2FU BM25FP	0.1763	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1693	-3.97	3
33	hp2003	PL2FA I(n_e)C2FU BM25FP	0.7516	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7563	+0.63	2
34	hp2004	PL2FA I(n_e)C2FU BM25FP	0.6469	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6338	-2.03	2
35	np2003	PL2FA I(n_e)C2FU BM25FP	0.7108	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7183	+1.06	3
36	np2004	PL2FA I(n_e)C2FU BM25FP	0.7169	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7148	-0.29	4
37	td2003	PL2FA I(n_e)C2FU BM25FP	0.1640	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1558	-5.00	1
38	td2004	PL2FA I(n_e)C2FU BM25FP	0.1763	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1645	-6.69	3
39	hp2003	PL2FA I(n_e)C2FU BM25FP	0.7516	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7637	+1.61	1
40	hp2004	PL2FA I(n_e)C2FU BM25FP	0.6469	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6508	+0.60	1
41	np2003	PL2FA I(n_e)C2FU BM25FP	0.7108	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7125	+0.24	1
42	np2004	PL2FA I(n_e)C2FU BM25FP	0.7169	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7319	+2.09	1
43	td2003	PL2FA I(n_e)C2FU BM25FP	0.1640	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1570	-4.27	2
44	td2004	PL2FA I(n_e)C2FU BM25FP	0.1763	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1629	-7.60	3
45	hp2003	PL2FA I(n_e)C2FU BM25FP	0.7516	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7639	+1.64	1
46	hp2004	PL2FA I(n_e)C2FU BM25FP	0.6469	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6295	-2.69	2
47	np2003	PL2FA I(n_e)C2FU BM25FP	0.7108	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7125	+0.24	1
48	np2004	PL2FA I(n_e)C2FU BM25FP	0.7169	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7253	+1.17	1
49	td2003	PL2FA I(n_e)C2FU BM25FP	0.1640	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1578	-3.78	1
50	td2004	PL2FA I(n_e)C2FU BM25FP	0.1763	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1630	-7.54	4
51	hp2003	PL2FA I(n_e)C2FU BM25FP	0.7516	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7524	+0.11	1
52	hp2004	PL2FA I(n_e)C2FU BM25FP	0.6469	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6487	+0.28	4
53	np2003	PL2FA I(n_e)C2FU BM25FP	0.7108	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7022	-1.21	2
54	np2004	PL2FA I(n_e)C2FU BM25FP	0.7169	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7346	+2.47	1

Table 6.13: Evaluation of the decision mechanism, which employs the retrieval approaches PL2FA, I(n_e)C2FU, and BM25FP, for the experiments that identify at least one decision boundary for all the tested tasks, and result in improvements in retrieval effectiveness for at least three tested tasks. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

Overall, this section has presented an example of a Bayesian decision mechanism, which employs three different retrieval approaches. In this example, the Bayesian decision mechanism can lead to small improvements in retrieval effectiveness. However, the increased number of retrieval approaches requires a higher number of training queries

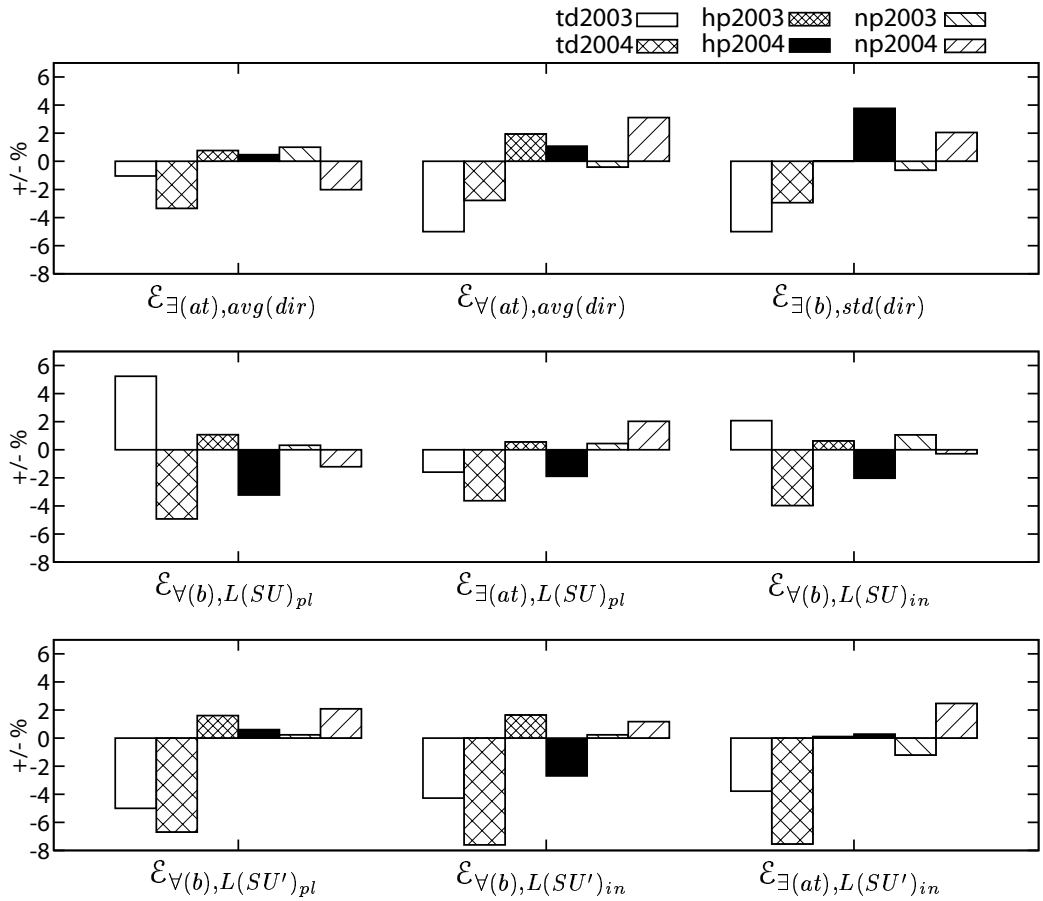

Figure 6.15: Histogram summarising the relative difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach from column '+/- %' of Table 6.13.

in order to reliably set the Bayesian decision mechanism. In order to alleviate the need for a higher number of training queries, a different approach can be taken. For example, the problem of selecting one retrieval approach among k available ones can always be transformed to a series of $k-1$ selections of one among two retrieval approaches.

6.8 Discussion

Overall, the evaluation of the proposed experiments in the context of the Bayesian decision mechanism has shown that the introduced framework for selective Web IR is a promising approach, which can lead to improvements in retrieval effectiveness. This section further discusses the obtained results from a range of additional perspectives.

Range of experiment outcome values The outcome of the proposed experiments fall within different ranges. For example, the outcome of the score-independent document-level experiments can be any number within the range $[0, N]$, where N is the number of documents in the employed collection. On the other hand, the outcome of the experiments that compute the usefulness of the hyperlink structure fall within the range of the symmetric Jensen-Shannon divergence values $[0, 2]$ (Section 5.4.1). In addition, the average size of domain aggregates is expected to be higher than the average size of the directory aggregates, as discussed in Section 6.3.2.3 (page 147). The illustrative examples in Sections 6.3.1.2 (page 138) for the document-level experiments, 6.3.2.3 (page 147) for the aggregate-level experiments, and 6.4.4 (page 159) for the score-dependent experiments, suggest that the estimated posterior likelihoods are higher when the outcome of an experiment falls within a smaller range of values. The higher posterior likelihoods correspond to stronger evidence for the appropriateness of a particular retrieval approach. However, the smaller range of outcome values of an experiment $\mathcal{E}$ is likely to result to overlapping densities for the posterior likelihoods, and hence, a higher number of decision boundaries. Therefore, there is a tradeoff between the range of the outcome values of an experiment and the expected number of decision boundaries. This tradeoff explains the fact that all the score-dependent experiments, which compute the symmetric Jensen-Shannon divergence $L(S_n, U_n)$ identify at least one decision boundary for all tested tasks (Table 6.5 on page 156).

Applying appropriate retrieval approaches and improvements in retrieval effectiveness The effectiveness of the proposed experiments $\mathcal{E}$ is shown by the number of decision boundaries, the improvements in retrieval effectiveness, and the number of topics for which the Bayesian decision mechanism applies the most appropriate retrieval approach (Section 6.2.1 on page 131). For example, when the Bayesian decision mechanism uses $\mathcal{E}_{\exists(b),avg(dir)}$ to selectively apply either PB2FU or DLHFA for the task hp2004, there are two decision boundaries, there is a statistically significant improvement in MAP of +13.03%, and the most appropriate retrieval approach is applied for a statistically significant number of queries (row 16 in Table 6.4, page 146). However, in some of the tested settings, the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, but there are no important improvements in retrieval effectiveness. For example, row 21 in Table 6.4

shows that the decision mechanism results in a relative improvement of only 0.59% in MAP for the task hp2003, even though the most appropriate retrieval approach is applied for a statistically significant number of queries. This fact can be explained in the following way. The decision mechanism may apply the most appropriate retrieval approach for a statistically significant number of queries, but a small number of wrong decisions may cancel the positive effect in the overall retrieval effectiveness. Indeed, when the decision mechanism is trained, it does not consider the magnitude of the difference in retrieval effectiveness between the most effective and the less effective retrieval approaches. Future work can address this issue by investigating different definitions for the loss function described in Section 5.2, which would consider the magnitude of differences in retrieval effectiveness.

Potential for improvements and obtained improvements from selective Web IR The evaluation of the proposed experiments in this chapter has been performed in the context of a Bayesian decision mechanism, which selectively applies a pair of retrieval approaches. These retrieval approaches have been selected on the basis of their potential for improvements in retrieval effectiveness. Table 6.1 (page 134) shows the potential for improvements in retrieval effectiveness for each pair of retrieval approaches used in the tested tasks. For example, when selectively applying the retrieval approaches I(n_e)C2FU and DLHFP for the task td2003, the maximum mean average precision (MAP) can be 0.1926 (row 1 in Table 6.1). This corresponds to a relative improvement of 32.37% from the MAP of the most effective retrieval approach (0.1455). When employing the Bayesian decision mechanism, the highest obtained MAP is 0.1655, when the experiment $\mathcal{E}_{\forall(b),L(SU')_{pl}}$ is employed (row 1 from Table 6.6 on page 159). This corresponds to a relative improvement of 13.75% from 0.1455, which is statistically significant.

The difference between the potential for improvements and the obtained improvements can be interpreted in two ways. First, this difference is due to the fact that the maximum MAP is obtained in a hypothetical setting, where a decision mechanism makes perfect decisions for all the queries. Second, the difference between the maximum and the obtained retrieval effectiveness suggests that there is more room for improvements by introducing more effective experiments $\mathcal{E}$ for selective Web IR.

Generalising findings from selective Web IR The evaluation of the proposed experiments $\mathcal{E}$ has been performed in a particular setting, with a broad set of retrieval approaches and a range of different tasks, as described in Section 6.2. The obtained results depend on several factors. One such factor is the particular field-based weighting models and their combination with the query-independent evidence. The improvements in retrieval effectiveness from selective Web IR can be higher if less robust retrieval approaches are employed. For example, the field-based weighting model BM25F has been shown to be more appropriate for selective Web IR, because it is less robust than the DFR field-based weighting models with respect to the setting of the hyper-parameters, as discussed in Section 6.6. Another factor that affects the results is whether the employed tasks are good representatives of particular types of tasks. If they are good representatives of a type of tasks, then the Bayesian decision mechanism and the proposed experiments should have a similar performance for the tasks of the same type.

In order to alleviate the effect on the obtained results from the above mentioned factors, the evaluation has focused on the experiments $\mathcal{E}$, which are effective across different types of tasks. Therefore, it is expected that the main findings from the evaluation of the experiments $\mathcal{E}$ would hold for different experimental settings. Furthermore, Chapter 7 will perform an evaluation of the experiments $\mathcal{E}$ in a different setting, where there exists limited relevance information, if any.

6.9 Summary

This chapter has presented the evaluation of the framework for selective Web IR, which has been proposed in Chapter 5. The evaluation has been performed in the context of a Bayesian decision mechanism, which employs pairs of retrieval approaches to apply on a per-query basis. These pairs of retrieval approaches have been selected with respect to their potential for improvements in retrieval effectiveness, and they employ different field-based weighting models (Section 6.2). In order to focus the evaluation on the effectiveness of the employed experiments $\mathcal{E}$, the training and testing of the decision mechanism have been performed with the same task, assuming that there exists relevance information.

The score-independent document-level experiments perform well when their outcome is computed from the set of documents that contain all the query terms in their anchor text. This is because these documents are more likely to be about the query topic, and therefore, the resulting set of documents is more cohesive (Section 6.3.1).

The score-independent aggregate-level experiments perform well when the considered documents contain at least one or all the query terms in their body, because a larger set of documents is required in order to obtain a representative distribution of aggregate sizes (Section 6.3.2.4).

The score-dependent experiments are robust, and they result in improvements for most of the tested settings, when the usefulness of the hyperlink structure is estimated by the symmetric Jensen-Shannon divergence $L(S_n, U_n)$ (Section 6.4.2).

The current chapter has also investigated document sampling, in order to reduce the computational cost of the introduced experiments $\mathcal{E}$, as well as to test whether the experiments $\mathcal{E}$ are more effective with documents, which have been highly scored by a weighting model. The results show that document sampling can be effectively employed for the score-independent document-level experiment $\mathcal{E}_{\forall(at)}$ (Section 6.5.3), as well as for the score-dependent experiments $\mathcal{E}_{\exists(at),L(SU)_{pl}}$, $\mathcal{E}_{\exists(at),L(SU)_{in}}$, $\mathcal{E}_{\forall(at),L(SU')_{pl}}$ and $\mathcal{E}_{\forall(at),L(SU')_{in}}$ (Section 6.5.5).

When the Bayesian decision mechanism selectively applies retrieval approaches, which employ the same weighting model, the introduced experiments can also be used to improve the retrieval effectiveness of the individual retrieval approaches (Section 6.6). An example of a Bayesian decision mechanism, which selectively applies three retrieval approaches, has also been presented in Section 6.7. In the illustrated example, the performance of the Bayesian decision mechanism is unstable, because the higher number of retrieval approaches requires more training queries to appropriately set the decision mechanism.

Overall, the evaluation has shown that selective Web IR can lead to improvements in retrieval effectiveness, which are statistically significant in some of the tested cases, and that the introduced experiments allow the Bayesian decision mechanism to apply an appropriate retrieval approach for a statistically significant number of queries. The evaluation has primarily focused on the experiments, which perform well across all the tested tasks. The following chapter will evaluate selective Web IR in a setting, where only limited relevance information exists.

Chapter 7

Selective Web IR with limited relevance information

7.1 Introduction

The framework for selective Web IR and the proposed experiments $\mathcal{E}$ have been so far evaluated with a Bayesian decision mechanism, which was trained and tested with the same task, as presented in Chapter 6. In this way, the evaluation has been focused on the effectiveness of the proposed experiments $\mathcal{E}$ to identify appropriate retrieval approaches, assuming that relevance information does exist. The objective of this chapter is to investigate the effectiveness of the decision mechanism and the experiments $\mathcal{E}$ in a more realistic and operational setting, where a decision mechanism is trained with limited relevance information. This setting is represented by training the decision mechanism with a known set of queries, and performing the evaluation with a different set of queries. Moreover, each set of queries represents a mixed set of different search tasks, such as topic distillation, home page finding, and named page finding.

This chapter also proposes an ad-hoc decision mechanism that can be used when only limited relevance information exists. This ad-hoc decision mechanism approximates the decision boundaries by automatically generating samples of representative queries. The automatically generated queries can be single-term queries, or more realistic multiple term queries. The multiple term queries are generated by either applying automatic query expansion to a random seed term, or sampling anchor text from the document collection.

The remainder of this chapter is organised as follows. Section 7.2 defines how limited

relevance information is modelled, and describes the experimental setting that is used in the remainder of this chapter. Next, Section 7.3 evaluates the proposed experiments $\mathcal{E}$ in the experimental setting of this chapter. It evaluates both the score-independent document and aggregate-level experiments, as well as the score-dependent experiments, which have been introduced in Chapter 5. Section 7.4 proposes an ad-hoc decision mechanism, which can be applied, when the available relevance information is limited. The ad-hoc decision mechanism sets its decision boundaries by using novel techniques to automatically generate samples of queries. These samples of queries correspond to single term queries, or queries with multiple terms, which are generated by applying automatic query expansion, or by sampling the anchor text of documents.

7.2 Limited relevance information

The proposed retrieval approaches in Chapter 4 have been optimised and evaluated with different sets of mixed tasks, in order to obtain a realistic setting of the hyper-parameters. However, the evaluation of selective Web IR and the proposed experiments $\mathcal{E}$ in Chapter 6 has been performed by training and testing a Bayesian decision mechanism with the same task, assuming that relevance information does exist. This choice was made in order to focus the evaluation on the effectiveness of the experiments, and to reduce any effect from using different training and testing tasks. The current chapter aims to evaluate selective Web IR in a setting, where a decision mechanism has only limited relevance information. This section explains the concept of limited relevance information in the context of selective Web IR, and it describes the experimental setting used for evaluating the proposed experiments $\mathcal{E}$ in the remainder of this chapter.

7.2.1 Modelling limited relevance information

A retrieval system will almost certainly not have complete relevance information for the search requests that it processes. However, some limited relevance information may be available for the queries that have been already processed. In the context of selective Web IR, the concept of limited relevance information is defined with respect to: (a) the type of the queries, which are processed by the decision mechanism; (b) the training and evaluation of the decision mechanism with different sets of queries.

Type of queries In an operational setting, a retrieval system processes a stream of queries, which are submitted by users. The queries are not associated with explicit evidence about the aim of the user. For example, the retrieval system is not aware whether a particular query is an informational, or a navigational query, unless further analysis of the queries is performed (Beitzel et al., 2004; Bomhoff et al., 2005; Rose & Levinson, 2004). This means that the type of the relevant documents, or the number of relevant documents is unknown. For example, if a system is not aware whether a query is related to a navigational or an informational task, it does not know whether there is one or few relevant documents. Similarly, a system does not know that the relevant document for a home page finding task is indeed a home page of a Web site. Craswell & Hawking (2004) suggested that effective retrieval can be performed without knowing the type of the queries. However, in the context of selective Web IR, queries from mixed tasks may have an impact on the training of the decision mechanism. This is because different types of queries are likely to result in different distributions of outcome values for an experiment. For example, a query related to a home page finding task is likely to retrieve more documents from a particular Web site, resulting in a small number of large domain aggregates. On the other hand, a query related to a topic distillation task is likely to retrieve many documents from several Web sites, resulting in a high number of large domain aggregates. A very specific query related to a named page finding task is likely to retrieve few documents with all the query terms. Therefore, using mixed tasks intends to test whether the setting of the decision mechanism is affected from processing different types of queries.

Using different training and testing tasks In addition to processing queries from mixed tasks, a retrieval system is usually trained and evaluated with different sets of queries. The retrieval effectiveness of a system is optimised with respect to a set of training queries. Then, the retrieval system is required to process both previously unseen queries, and possibly queries, which have been used for training. If the set of training queries is representative of a particular type of search task, then the performance of the retrieval system is likely to be close to that obtained during training. In the context of selective Web IR, using different query sets for training and evaluating the decision mechanism aims to test whether the decision mechanism can apply appropriate retrieval approaches for previously unseen queries.

7.2.2 Experimental setting for limited relevance information

This section describes the experimental setting, which will be employed in the remainder of this chapter to evaluate selective Web IR when limited relevance information is available. The experimental setting is defined as follows.

1. As described in Section 4.6, two mixed tasks are selected to be used for training and testing the decision mechanism, respectively. Both mixed tasks correspond to a mix of queries from three different tasks: topic distillation; home page finding; and named page finding. The first mixed task is denoted by mq2003, and corresponds to the queries from the tasks td2003, hp2003, and np2003. When the mixed task mq2003 is employed as a training set, the first 50 topics for each type of task are used, and this smaller set of queries is denoted by mq2003'. The mixed task mq2003' is used for training in order not to bias the results towards a particular type of task. The second mixed task corresponds to the queries used in the mixed query task of TREC 2004 Web track (Craswell & Hawking, 2004). When the mixed task mq2003' is used for training, the mixed task mq2004 is employed for the evaluation, and when the mixed task mq2004 is employed for training, the mixed task mq2003 is employed for the evaluation. Details about the employed mixed tasks have also been given in Section 4.2 (page 52).

2. As described in Section 4.6.2, the hyper-parameters of the employed retrieval approaches are set in order to optimise mean average precision for the training mixed task. In order not to overfit the training mixed task, the optimisation process is terminated after 20 iterations (see Section 4.6.2 on page 95). The employed retrieval approaches correspond to the field-based weighting models PL2F, PB2F, I(n_e)C2F, DLHF, and BM25F (Section 4.4 on page 67), as well as to their combinations with evidence from the URL path length, PageRank, and the Absorbing Model (Section 4.5 on page 74)[1]. The document fields are the body, the anchor text of incoming hyperlinks, and the title of documents.

3. The Bayesian decision mechanism employs one of the proposed experiments $\mathcal{E}$ and sets the decision boundaries for one of the training mixed tasks, i.e., mq2003' or

[1] The values of the hyper-parameters associated with the field-based weighting models are displayed in Table A.11 (page 235) of Appendix A. The values of the hyper-parameters associated with the query-independent sources of evidence are displayed in Table A.12 (page 235) of Appendix A.

mq2004. Then, the Bayesian decision mechanism is tested with the corresponding evaluation mixed task. For example, if it has been trained with mq2003', then the evaluation employs the mixed task mq2004.

In each case, the Bayesian decision mechanism selectively applies the two retrieval approaches, which have the highest potential for improvements from selective Web IR. For each of the evaluation mixed tasks, the potential for improvements in retrieval effectiveness is shown in Table 7.1. It is computed by assuming that a decision mechanism MAX employs two retrieval approaches and selectively applies the most appropriate one on a per-query basis, as described in Sections 4.7 and 5.2.2. The resulting retrieval effectiveness is the maximum that can be obtained from selectively applying the two retrieval approaches, and it is statistically significantly higher than that of the most effective individual retrieval approach, as denoted by *. The Bayesian decision mechanism employs the pairs of retrieval approaches that result in the highest potential for improvements in retrieval effectiveness. The employed pairs of retrieval approaches correspond to: DLHFP and BM25F for the evaluation task mq2003 (row 11); DLHFP and PB2F for the evaluation task mq2004 (row 12).

Row	Task	Mean Average Precision				
		First approach		Second approach		MAX
1	mq2003	PL2FU	(0.6206)	PL2FP	(0.6238)	0.6529 (+ 4.66%)*
2	mq2004	PL2F	(0.4444)	PL2FA	(0.4717)	0.5094 (+ 7.99%)*
3	mq2003	PB2FU	(0.5809)	PB2FP	(0.5873)	0.6029 (+ 2.66%)*
4	mq2004	PB2FU	(0.4723)	PB2FP	(0.4723)	0.5258 (+11.33%)*
5	mq2003	I(n_e)C2FU	(0.6258)	I(n_e)C2FA	(0.6210)	0.6511 (+ 4.04%)*
6	mq2004	I(n_e)C2FU	(0.4946)	I(n_e)C2FP	(0.4983)	0.5561 (+11.60%)*
7	mq2003	DLHFU	(0.5216)	DLHFP	(0.5319)	0.5577 (+ 4.85%)*
8	mq2004	DLHFU	(0.4273)	DLHFP	(0.4156)	0.4618 (+ 8.07%)*
9	mq2003	BM25FU	(0.6237)	BM25FP	(0.6502)	0.6921 (+ 6.44%)*
10	mq2004	BM25FU	(0.4883)	BM25FA	(0.4680)	0.5284 (+ 8.21%)*
11	mq2003	DLHFP	(0.5319)	BM25F	(0.5533)	0.6582 (+18.96%)*
12	mq2004	DLHFP	(0.4156)	PB2F	(0.4114)	0.5304 (+27.62%)*

Table 7.1: Evaluation of a decision mechanism MAX, which selectively applies the most effective retrieval approach on a per-query basis. The retrieval approaches are based on a restricted optimisation, as reported in Table 4.10 (page 96). The table displays the pairs of retrieval approaches that result in the highest improvements in MAP for the tested mixed tasks mq2003 and mq2004. The symbol * denotes that the difference between the MAP of the decision mechanism MAX and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

Overall, the described experimental setting allows to investigate the effectiveness of the proposed framework for Web IR in a setting where the decision mechanism is trained and evaluated with different sets of mixed tasks. The next section evaluates the proposed experiments $\mathcal{E}$ in the described experimental setting.

7.3 Evaluation of experiments $\mathcal{E}$ with limited relevance information

This section presents the evaluation of the proposed experiments $\mathcal{E}$ with limited relevance information, as described in Section 7.2. The evaluated experiments are the score-independent document-level and aggregate-level experiments (Section 5.3 on page 110), as well as the score-dependent experiments that estimate the usefulness of the hyperlink structure (Section 5.4 on page 115). This section closes with a discussion and conclusions from the evaluation of the Bayesian decision mechanism and the experiments with limited relevance information (Section 7.3.3).

7.3.1 Score-independent experiments with limited relevance information

Table 7.2 displays the evaluation results for those score-independent experiments, which result in improvements in MAP, compared to the most effective retrieval approach, for both tasks mq2003 and mq2004[1]. This choice is made in order to focus the analysis on the experiments that allow the decision mechanism to obtain improved retrieval effectiveness. For example, row 1 in Table 7.2 corresponds to a decision mechanism, which has been trained for the mixed task mq2004 and it is evaluated for the task mq2003. This decision mechanism, selectively applies either the combination of the field-based weighting model DLHF with PageRank (DLHFP), or the field-based weighting model BM25F, on a per-query basis. The employed experiment is $\mathcal{E}_{\forall(at)}$, which counts the number of documents with all the query terms in the anchor text. The MAP of the decision mechanism is 0.5775, which represents a relative improvement of +4.37% over the MAP of the most effective individual approach (0.5533). This improvement in MAP is statistically significant according to Wilcoxon's signed rank test, as denoted by *, and the corresponding decision mechanism applies the most appropriate retrieval

[1] The evaluation results for all the score-independent experiments in the same setting appear in Table B.12 (page 257) of Appendix B.

approach for a statistically significant number of queries according to the sign test, as denoted by †.

From the results, it can be seen that the document-level experiment $\mathcal{E}_{\forall(at)}$ results in improved MAP over the baseline, and it identifies only 1 decision boundary for both tasks mq2003 and mq2004 (rows 1-2). Moreover, when the decision mechanism uses $\mathcal{E}_{\forall(at)}$, it applies the most appropriate retrieval approach for a significant number of queries, and there is only one decision boundary.

Regarding the aggregate-level experiments, there are four experiments that employ the domain aggregates (rows 3-10). Three of these experiments, namely $\mathcal{E}_{\forall(b),std(dom)}$, $\mathcal{E}_{\exists(at),std(dom)}$, and $\mathcal{E}_{\forall(at),std(dom)}$ compute the standard deviation of the domain aggregate size distribution. The experiment $\mathcal{E}_{\forall(b),std(dom)}$ results in the highest improvement over the most effective retrieval approach for the task mq2003 (+4.41% from row 5 of Table 7.2). The experiment $\mathcal{E}_{\forall(at),avg(dom)}$ results in the highest improvement for the task mq2004 (+7.12% from row 4 in Table 7.2). The directory aggregate-level experiments achieve lower improvements in retrieval effectiveness, but they result in a decision mechanism, which applies the most appropriate retrieval approach for a statistically significant number of queries (rows 11-13 and 15-16 from Table 7.2).

It is worth noting from Table 7.2 that six out of the eight experiments compute their outcome from the documents that contain all the query terms in their anchor text (rows 1-4 and 9-16). In the context of processing queries from mixed tasks for selective Web IR, this can be explained by the fact that the terms of either broad queries or queries about the home page of a particular Web site, are more likely to appear in the anchor text of documents. Therefore, the experiments, which count the number of documents with all the query terms in their anchor text, aid the decision mechanism to identify the queries for which the relevant documents are likely to be home pages, and therefore, to apply more evidence from the hyperlink structure.

7.3.2 Score-dependent experiments with limited relevance information

Table 7.3 presents the evaluation results for those score-dependent experiments, which result in retrieval effectiveness improvements, compared to the most effective individual

Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
1	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at)}$	0.5775	$+4.37^{\dagger *}$	1
2	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at)}$	0.4381	$+5.41^{\dagger}$	1
3	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),avg(dom)}$	0.5626	+1.68	2
4	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),avg(dom)}$	0.4452	$+7.12^{\dagger}$	2
5	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),std(dom)}$	0.5777	$+4.41^{\dagger *}$	1
6	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),std(dom)}$	0.4212	+1.34	2
7	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),std(dom)}$	0.5554	+0.38	3
8	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),std(dom)}$	0.4265	+2.62	1
9	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),std(dom)}$	0.5622	+1.61	2
10	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),std(dom)}$	0.4233	+1.85	2
11	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),avg(dir)}$	0.5626	$+1.68^{\dagger}$	1
12	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),avg(dir)}$	0.4395	$+5.75^{\dagger}$	2
13	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),std(dir)}$	0.5648	$+2.08^{\dagger *}$	2
14	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),std(dir)}$	0.4374	+5.25	1
15	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.5613	$+1.45^{\dagger}$	3
16	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.4421	$+6.37^{\dagger}$	1

Table 7.2: Evaluation of the score-independent document-level and aggregate-level experiments with limited relevance information. The table displays the evaluation results of a decision mechanism, which is trained and evaluated with different mixed tasks. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

retrieval approach, for both tested tasks mq2003 and mq2004[1]. Overall, there are seven experiments, which result in improved performance for both tested mixed tasks. All the seven experiments compute their outcome from the documents that contain all the query terms in their body (rows 1-2, 5-6, and 9-12), or in the combination of their anchor text and title fields (rows 3-4, 7-8, and 13-14). This indicates that documents with all the query terms provide more useful evidence in order to compute an experiment $\mathcal{E}$.

The experiment $\mathcal{E}_{\forall(at),L(SU)_{in}}$, which employs the field-based weighting model I(n_e)C2F to assign the score distribution S_n, results in the highest improvements in retrieval effectiveness for both tasks mq2003 and mq2004 (+3.80% and +2.60% from rows 7 and 8, respectively). The experiment $\mathcal{E}_{\forall(at),L(SU)_{pl}}$, which employs the field-based weighting model PL2F to assign the score distribution S_n (rows 3-4), achieves lower improvements than the experiment $\mathcal{E}_{\forall(at),L(SU)_{in}}$.

The experiment $\mathcal{E}_{\forall(at),L(SU')_{in}}$ results in one decision boundary for both tested tasks

[1] The evaluation results for all the score-dependent experiments in the same setting appear in Table B.13 (page 258) of Appendix B.

(rows 13-14). Furthermore, when the Bayesian decision mechanism employs the experiment $\mathcal{E}_{\forall(b),L(SU')_{pl}}$, the most appropriate retrieval approach is applied for a statistically significant number of queries for both tasks mq2003 and mq2004 (rows 9-10). When the Bayesian decision mechanism employs the experiments $\mathcal{E}_{\forall(at),L(SU)_{in}}$ or $\mathcal{E}_{\forall(at),L(SU')_{in}}$ for the task mq2003, there is a statistically significant improvement in MAP, and the most appropriate retrieval approach is applied for a statistically significant number of queries. However, there is no particular experiment that results in both statistically significant improvements in MAP over the baseline, as well as in applying the most appropriate retrieval approach for a statistically significant number of queries, for both mq2003 and mq2004.

Row	Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	Bnd
1	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.5539	+0.11	4
2	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.4215	$+1.42^{\dagger}$	2
3	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.5685	$+2.75^{\dagger}$	2
4	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.4215	+1.42	1
5	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.5702	+3.05	4
6	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.4207	$+1.23^{\dagger}$	2
7	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.5743	$+3.80^{\dagger *}$	2
8	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.4264	+2.60	2
9	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.5698	$+2.98^{\dagger}$	3
10	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.4201	$+1.08^{\dagger}$	1
11	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.5742	$+3.78^{\dagger}$	3
12	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.4194	+0.91	1
13	mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.5733	$+3.61^{\dagger *}$	1
14	mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.4157	+0.02	1

Table 7.3: Evaluation of score-dependent experiments with limited relevance information. The table displays the evaluation results of a decision mechanism, which is trained and evaluated with different mixed tasks. The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test.

7.3.3 Discussion and conclusions

This section has evaluated the Bayesian decision mechanism in a setting with limited relevance information. This setting corresponds to training and evaluating the Bayesian decision mechanism with different sets of mixed tasks. The evaluation results for both the score-independent (Section 7.3.1) and the score-dependent experiments

(Section 7.3.2) show that the proposed decision mechanism and the experiments $\mathcal{E}$ result in improved retrieval effectiveness, even when limited relevance information exists. This suggests that selective Web IR can be effectively applied in a realistic setting.

The score-independent document-level experiment $\mathcal{E}_{\forall(at)}$ performs well for both the tested mixed tasks (rows 1-2 from Table 7.2). The domain aggregate-level experiments $\mathcal{E}_{\forall(at),avg(dom)}$ and $\mathcal{E}_{\forall(b),std(dom)}$ also perform well for the tasks mq2004 and mq2003, respectively (rows 4 and 5 from Table 7.2, respectively). Moreover, four out of the seven aggregate-level experiments, which are shown in Table 7.2, estimate the standard deviation of the aggregates' size. This indicates that estimating the standard deviation of the aggregates' size results in robust experiments, and it is in agreement with the results from Sections 6.3.2.1 and 6.3.2.2.

The score-dependent experiments, which estimate the usefulness of the hyperlink structure also result in improvements in retrieval effectiveness. However, no particular trend has been observed from the results in Table 7.3. Moreover, the score-independent experiments outperform the score-dependent ones with respect to the obtained improvements in MAP by the decision mechanism.

An observation related to both the score-independent (Section 7.3.1) and the score-dependent experiments (Section 7.3.2) is that the documents that contain all the query terms in the anchor text or the title fields provide more robust evidence to compute the outcome of the experiments. This can be explained because a high number of documents with the all the query terms in the anchor text or the title fields indicates that there are Web sites related to the query. Hence, applying hyperlink analysis is likely to be effective in order to detect the home pages of those Web sites.

7.4 Ad-hoc decision mechanism and query sampling

When setting the Bayesian decision mechanism, as described in Section 5.5.2 (page 125), a training set of queries and the corresponding relevance assessments are employed in the following way. First, they are used to set the prior probability that a particular retrieval approach is effective. Second, they are used to estimate the associated loss with applying a particular retrieval approach. Third, they are employed to estimate the density of the likelihood that a particular retrieval approach is the most effective one for a range of the experiment outcome values. A drawback of this process is that

the setting of the decision mechanism primarily depends on the availability of training queries.

The current section aims to reduce the dependence on the training queries by introducing a simple ad-hoc decision mechanism, which employs the distribution of the outcome values of an experiment $\mathcal{E}$ to set its decision boundary. The distribution of the outcome values of $\mathcal{E}$ is obtained from a sample of queries, which is automatically generated.

The automatic generation of query samples is performed with three different techniques. The first one involves the random sampling of single terms from the vocabulary of the collection. The second technique applies automatic query expansion to randomly selected terms from the vocabulary and generates queries with more than one terms. The third one is a novel technique, which samples the anchor text of documents and generates queries with more than one terms.

The query sampling is evaluated with respect to the similarity of the outcome values of an experiment $\mathcal{E}$ for the sampled queries, and the outcome values of the same experiment $\mathcal{E}$ for the TREC Web track queries. The ad-hoc decision mechanism is evaluated with respect to the obtained improvements in retrieval effectiveness from selective Web IR. For both the evaluation of query sampling and the ad-hoc decision mechanism, the employed experiments are $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$, which have been shown to be effective when limited relevance information is available (rows 1-2, 4, and 5 in Table 7.2).

The remainder of this section is organised as follows. The ad-hoc decision mechanism is introduced in Section 7.4.1. Section 7.4.2 introduces the three query sampling techniques. Section 7.4.3 evaluates the similarity of the sampled queries to the queries used in the TREC 2003 and 2004 Web tracks (Craswell & Hawking, 2004; Craswell et al., 2003). Next, 7.4.4 presents the evaluation of the ad-hoc decision mechanism.

7.4.1 Ad-hoc decision mechanism

In order to alleviate the dependence of the Bayesian decision mechanism on the training queries and the corresponding relevance assessments, a new ad-hoc decision mechanism is proposed. This section describes how this ad-hoc mechanism is set and applied for selective Web IR.

The ad-hoc decision mechanism is used to select one out of two retrieval approaches a_1 and a_2. It is set in two steps. First, it estimates the distribution of the outcome values of an experiment $\mathcal{E}$. Second, it sets a decision boundary Bnd, so that the outcome o of the experiment $\mathcal{E}$ is lower than Bnd with a given probability $P(o < Bnd)$.

During retrieval, if the outcome of the experiment $\mathcal{E}$ is $o < Bnd$, then the decision mechanism applies the retrieval approach a_1. Otherwise, it applies the retrieval approach a_2.

The probability $P(o < Bnd)$ can be set equal to the prior probability that the retrieval approach a_1 is more effective than the retrieval approach a_2. In this case, the prior probability that the retrieval approach a_2 is more effective than the retrieval approach a_1 would be $1 - P(o < Bnd)$. If the prior probabilities cannot be determined, then a uniform prior probability can be assigned to the retrieval approaches by setting $P(o < Bnd) = 0.5$. Generally, the probability $P(o < Bnd)$ can be set equal to any value within the range $[0, 1]$.

The defined ad-hoc decision mechanism employs two retrieval approaches. The selection of one out of k retrieval approaches can always be modelled by $k - 1$ ad-hoc decision mechanisms, which select one out of two retrieval approaches. The remainder of this chapter will illustrate the ad-hoc decision mechanism with two retrieval approaches.

The next section focuses on automatically estimating the distribution of outcome values of the experiment $\mathcal{E}$, by generating samples of queries.

7.4.2 Query sampling

In this thesis, query sampling can be seen as the automatic generation of a sample of queries from a document collection. It aims to approximate real queries and provide meaningful queries that users could have formed in order to satisfy an information need.

In the context of selective Web IR, query sampling is used to approximate the distribution of outcome values of an experiment $\mathcal{E}$, in order to set the decision boundary of an ad-hoc decision mechanism. In this case, even if the sampled queries are not meaningful, they should at least result in a similar distribution of outcome values to the one obtained from real queries.

The remainder of this section introduces three techniques for query sampling. The first one generates single-term queries by randomly sampling terms from the vocabulary of a document collection. The second technique generates queries with more than one

terms, by applying automatic query expansion to randomly selected terms from the vocabulary. The third one is a novel query sampling technique, which generates queries from the anchor text of Web documents. This section is followed by the evaluation of the three query sampling techniques with respect to the TREC 2003 and 2004 Web track queries.

7.4.2.1 Single-term query sampling

Sampling of terms and their statistics has been employed in the context of distributed IR, in order to obtain a representation of remote databases (Callan & Connell, 2001). Cronen-Townsend et al. (2002) employed single-term query sampling in order to obtain a distribution of the clarity scores of the terms in a collection. Plachouras & Ounis (2004) used single-term sampling to obtain a distribution of values for the usefulness of the hyperlink structure.

Here, in order to sample meaningful terms and avoid generating queries with either very frequent or rare terms, the terms in the vocabulary are ranked according to their frequency in the collection. Then, terms with a rank in the range $[r_{lo}, r_{hi}]$, are randomly sampled. The thresholds r_{lo} and r_{hi} correspond to the low and the high rank, respectively. This technique is referred to as *Single-Term Sampling* (STS).

Single-term sampling is only a simple approximation of the querying process, because real queries are likely to have more than one term (Jansen & Pooch, 2001). In addition, the query terms tend to be correlated and to co-occur in the documents (Silverstein et al., 1999). For these reasons, the two following query sampling techniques generate queries with more than one terms.

7.4.2.2 Multiple term query sampling

The second technique is based on using the Divergence From Randomness (DFR) framework (Amati & Van Rijsbergen, 2002) to find informative terms from the set of retrieved documents for a query (He & Ounis, 2005*b*). The informative terms from the set of top ranked documents are used to form the sampled queries.

The sampled queries are generated in the following way. The terms of the vocabulary are ranked according to their term frequency in the collection, as described in Section 7.4.2.1. A term with rank in the range $[r_{lo}, r_{hi}]$ is randomly selected to be used as a single-term query. The thresholds r_{lo} and r_{hi} correspond to the low and the

high rank, respectively. From the top ranked documents, the most informative term is extracted and it is used as an intermediate seed term to perform retrieval again. This intermediate seed term is employed in order to reduce the effect of initially selecting a random term. From the new set of retrieved documents, the most informative terms are extracted from the top-ranked documents in order to form a sampled query.

The number of extracted terms depends on the required length of the sampled queries. He & Ounis (2005*b*) used a uniform query length distribution. In this thesis, this technique is refined by using a Gaussian distribution with mean μ and standard deviation σ for the query length distribution of the sampled queries. This technique is referred to as *Multiple Term Sampling* (MTS).

The advantage of MTS is that it can be applied to any document collection, irrespectively of the type of documents, because it does not make use of any particular feature of Web documents. However, the length of the generated queries should be set in an appropriate way. This issue is further investigated in Section 7.4.3.

7.4.2.3 Anchor text query sampling

This section introduces a novel technique for sampling queries. It is based on the observation that user queries are similar to the anchor text of Web documents (Eiron & McCurley, 2003*b*). The user queries have similar length and term frequency distributions with the anchor text of the incoming hyperlinks of a Web document, which can be seen as a concise textual description of that document. Therefore, it is reasonable to employ the anchor text for query sampling.

The sampled queries are generated as follows. The frequency of the anchor text is computed by counting the number of times each anchor text appears in the collection. The anchor texts with a frequency less than the threshold af_{lo} are discarded. This restriction aims to reduce the bias of the sampled queries towards anchor texts that appear very few times. From the remaining anchor texts, one is randomly selected as the sampled query. This technique is referred to as *Anchor Text Sampling* (ATS).

The advantage of this query sampling technique is that the number of terms in the anchor text of hyperlinks is similar to the length of queries. In addition, these terms are correlated and they are highly likely to co-occur in the text of Web documents.

7.4.3 Evaluation of query sampling

The query sampling techniques described in Sections 7.4.2.1, 7.4.2.2, and 7.4.2.3 will be used in the context of an ad-hoc decision mechanism to approximate the outcome distribution of an experiment $\mathcal{E}$ for real queries. In order to do so, it is necessary to assess the quality of the query sampling techniques. The evaluation of the query sampling techniques is based on measuring the difference between the distribution of the experiment outcomes for the sampled and the TREC 2003 and 2004 Web track queries (tasks mq2003 and mq2004). The difference between the two distributions is estimated in terms of the symmetric Jensen-Shannon divergence (Equation (5.13) on page 116). When the divergence between the distributions is low, the sampled queries and the real TREC Web track queries are considered to be similar with regard to the outcome values of the experiment. The employed experiments are $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$, which have been shown to be effective when limited relevance information is available (Section 7.3.3).

The remainder of this section introduces the experimental setting for the evaluation of the query sampling techniques, and presents the evaluation results.

7.4.3.1 Experimental setting for evaluation of query sampling

This section describes the employed experimental setting for the evaluation of the query sampling techniques.

For each of the sampling techniques STS, MTS, and ATS, 500 queries are generated. Regarding the used thresholds r_{lo} and r_{hi} to randomly select single terms for the STS and MTS techniques, the employed values for r_{lo} are 20, 200, 2000 and 20000, and the employed values for r_{hi} are 200, 2000 and 20000. Because $r_{lo} < r_{hi}$, only some combinations of threshold values are used: $[20, 200]$, $[20, 2000]$, $[20, 20000]$, $[200, 2000]$, $[200, 20000]$ and $[2000, 20000]$. The different values for the thresholds r_{lo} and r_{hi} are selected in order to test the effect of randomly selecting terms with high frequency (low rank), or lower frequency (and higher rank). The threshold af_{lo} for the ATS technique is set equal to 5, 20, and 50, respectively. This choice is made in order to test the effect of discarding the infrequent anchor texts from the sampling process. For example, when $af_{lo} = 5$, then ATS considers only the anchor texts that appear at least 5 times in the collection to generate queries.

For the MTS technique in particular, the intermediate query and the final sampled query are generated by using the field-based weighting model PL2F, where the hyper-parameters c_b, c_a, c_t are set equal to 1.0 and the weights w_b, w_a, w_t are set equal to 1.0, 0.0, and 1.0, respectively. The informativeness of terms is estimated using the term weighting model Bo1 from the Divergence From Randomness framework (Amati, 2003). The terms from the top x retrieved documents for a query are ranked according to the weight:

$$w(t) = tf_x \cdot \log_2 \frac{1 + P_n}{P_n} + \log_2(1 + P_n) \tag{7.1}$$

where tf_x is the frequency of the term t in the top x documents and $P_n = \frac{F}{N}$. F is the frequency of a term in the collection, and N is the number of documents in the collection. The parameter x is set equal to 3. The number of extracted terms for the final sampled query depends on its length.

The MTS technique has two parameters μ and σ related to the average and the standard deviation of the length of the generated queries. These parameters are set in order to match the average and the standard deviation of the length of the TREC 2003 and 2004 Web track queries (Craswell & Hawking, 2004; Craswell et al., 2003). Table 7.4 displays the average and the standard deviation of the query length distribution of the TREC 2003 and 2004 Web track queries, after removing stop words. From the table, it can be seen that the topic distillation queries (row 1) are shorter than the home page finding, or the named page finding queries (rows 2 and 3, respectively). The length distribution of all the queries from the TREC 2003 and 2004 Web tracks is close to that of the home page finding and the named page finding (rows 5 and 4, respectively). This is partly because there are more home page finding and named page finding queries, than topic distillation queries in mq2003. For the evaluation of the MTS technique, two different settings of the parameters μ and σ are tested. The first one corresponds to the query length distribution of the topic distillation tasks ($\mu = 2.1$ and $\sigma = 0.78$ from row 1 in Table 7.4). The second setting corresponds to the query length distribution of all the TREC 2003 and 2004 Web track tasks ($\mu = 3.2$ and $\sigma = 1.31$ from row 5 in Table 7.4).

The next section employs the described experimental setting in order to evaluate the three proposed query sampling techniques.

Row	Task	Query length Average	Standard Dev.
1	td2003 & td2004	2.1	0.78
2	hp2003 & hp2004	3.5	1.23
3	np2003 & np2004	3.6	1.30
4	hp2003, np2004, hp2003 & np2004	3.5	1.26
5	mq2003 & mq2004	3.2	1.31

Table 7.4: Average and standard deviation for the length of the TREC 2003 and 2004 Web track queries.

7.4.3.2 Evaluation results for query sampling techniques

This section evaluates the three proposed query sampling techniques in the setting described in Section 7.4.3.1. The evaluation is performed by measuring the similarity between the outcome values of three experiments for the sampled queries and the TREC 2003 and 2004 Web track queries. As mentioned before, the employed experiments $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$, have been selected because they perform well in a setting with limited relevance information (Section 7.3.3).

The similarity corresponds to the symmetric Jensen-Shannon divergence between the distributions of outcome values of the experiments for the sampled and the TREC queries. The range of values of the symmetric Jensen-Shannon divergence is $[0, 2]$. Because the divergence measures dissimilarity, the higher values of the symmetric Jensen-Shannon divergence suggest that there are important differences between the sampled and the real TREC queries, with respect to the outcomes of the employed experiment $\mathcal{E}$. When the divergence approaches zero, then the distributions of the outcome values of the experiment are very similar.

Table 7.5 displays the symmetric Jensen-Shannon divergence between the sets of sampled queries with any of the three sampling techniques, and the queries from the TREC 2003 and 2004 Web track mixed tasks. The bold values indicate the set of query samples that results in the lowest divergence for each experiment $\mathcal{E}$ and each sampling technique. For the MTS technique, two different query length distributions are evaluated in order to test whether generating shorter or longer queries is more effective. These length distributions are chosen to match the length distribution of the queries from the tasks td2003 and td2004 (row 1 of Table 7.4), as well as that of the queries from the tasks mq2003 and mq2004 (row 5 from Table 7.4).

The sampled queries with STS are more similar to the real TREC queries, when the ranks of the sampled terms are within the range $[20, 20000]$ (rows 3, and 5-6 in Table 7.5). This is because, the experiment $\mathcal{E}_{\forall(at)}$ is expected to be sensitive to the frequency of the sampled terms. When the ranks of the sampled terms are in the range $[200, 20000]$, the lowest divergence is obtained between the sampled queries and the TREC queries for the outcome values of the experiment $\mathcal{E}_{\forall(at)}$ (row 5). Regarding the experiments $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$, the lowest divergence value is obtained when the ranks of the sampled terms are in the range $[20, 20000]$ (row 3). When the ranks of the sampled terms are very low (row 1), the resulting divergence value is high. This suggests that the sampled terms are very frequent and the outcome of the experiments is very different from that of the TREC queries.

Symmetric J-S divergence between query samples and mq2003 & mq2004					
Row			$\mathcal{E}_{\forall(at)}$	$\mathcal{E}_{\forall(at),avg(dom)}$	$\mathcal{E}_{\forall(b),std(dom)}$
	r_{lo}	r_{hi}	STS		
1	20	200	1.9813	1.0894	1.9753
2	20	2000	1.8176	0.5347	1.2302
3	20	20000	0.4912	**0.0836**	**0.0431**
4	200	2000	1.7964	0.4662	1.2671
5	200	20000	**0.3488**	0.0869	0.0574
6	2000	20000	0.5842	0.2003	0.1158
	r_{lo}	r_{hi}	MTS $\mu = 2.1$ $\sigma = 0.78$		
7	20	200	0.5014	0.0705	0.3492
8	20	2000	0.1299	0.0393	0.1804
9	20	20000	**0.0815**	0.0471	**0.0604**
10	200	2000	0.1186	**0.0309**	0.1921
11	200	20000	0.1425	0.0577	0.1114
12	2000	20000	0.0957	0.0793	0.0619
	r_{lo}	r_{hi}	MTS $\mu = 3.2$ $\sigma = 1.31$		
13	20	200	**0.2482**	**0.1515**	0.1309
14	20	2000	0.5850	0.3403	0.0490
15	20	20000	0.9428	0.5085	0.1052
16	200	2000	0.6409	0.3054	**0.0070**
17	200	20000	0.9452	0.5610	0.0943
18	2000	20000	0.9287	0.4902	0.0887
		af_{lo}	ATS		
19		5	**0.0105**	0.1747	**0.0477**
20		20	0.3926	**0.0329**	0.2842
21		50	0.6961	0.0467	0.4147

Table 7.5: Symmetric Jensen-Shannon (J-S) divergence between the distribution of experiment outcome values for the generated queries with STS, MTS, and ATS and the TREC 2003 and 2004 Web track queries (mq2003 and mq2004). The experiments are $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$. The mean and standard deviation of the query length distribution in MTS are denoted by μ and σ.

Table 7.5 displays the divergence between the sampled queries generated with MTS and the real TREC queries in rows 7-12 for the short queries ($\mu = 2.1$ and $\sigma = 0.78$), and in rows 13-18 for the longer queries ($\mu = 3.2$ and $\sigma = 1.31$), respectively. Regarding the shorter queries, sampling a random term with rank within $[20, 20000]$ provides the lowest divergence for the experiments $\mathcal{E}_{\forall(at)}$ and $\mathcal{E}_{\forall(b),std(dom)}$ (row 9 in Table 7.5). The lowest divergence for the experiment $\mathcal{E}_{\forall(at),avg(dom)}$, which computes the average size of domain aggregates, is obtained when the random term has rank within $[200, 2000]$ (row 10). It should be noted that the divergence values obtained for the experiment $\mathcal{E}_{\forall(at),avg(dom)}$ are lower than 0.08, regardless of the range of ranks. This suggests that this experiment is robust and the distribution of its outcome values is not affected by the rank of the random term, which is used during the first step of the MTS technique.

Regarding the longer queries generated with MTS (rows 13-18 from Table 7.5), the divergence values are lower when the randomly sampled terms have low ranks, or in other words, when the randomly sampled terms have high frequency. For example, the divergence between the outcome values of the experiments $\mathcal{E}_{\forall(at)}$ and $\mathcal{E}_{\forall(at),avg(dom)}$ for the sampled and the TREC queries is the lowest when the rank of the randomly sampled term is within the range $[20, 200]$ (row 13 in Table 7.5). The obtained divergence value for the experiment $\mathcal{E}_{\forall(b),std(dom)}$, which estimates the standard deviation of the size of the domain aggregates, is the lowest for the range of ranks $[200, 2000]$ (0.0070 from row 16 in Table 7.5).

The evaluation results for the generated queries with the ATS technique are displayed in rows 19-21 in Table 7.5. Regarding the experiments $\mathcal{E}_{\forall(at)}$ and $\mathcal{E}_{\forall(b),std(dom)}$, the lowest divergence is obtained when the anchor texts, which appear less than $af_{lo} = 5$ times are discarded during the sampling process (row 19). The divergence increases as the value of the threshold af_{lo} increases accordingly (rows 20-21). Sampling queries with the ATS technique performs better for the experiment $\mathcal{E}_{\forall(at),avg(dom)}$ when the threshold $af_{lo} = 20$ (row 20 in Table 7.5).

7.4.3.3 Discussion

Overall, query sampling is an effective method for generating queries with a similar distribution of outcome values for an experiment to the one obtained from real TREC queries. Query sampling can be seen as an approximation of the query generation process.

The first query sampling technique, STS, provides only a rough approximation of real queries through sampling of single terms. Queries with more than one terms are generated with either MTS, which is based on extracting the most informative terms from a set of documents, or ATS, which samples the anchor text of hyperlinks between Web documents. The former uses only statistical evidence, while the latter takes advantage of the similarity between real queries and the anchor text (Eiron & McCurley, 2003*b*).

The results from Table 7.5 show that the MTS and ATS techniques are more effective than STS in generating queries with a distribution of outcome values for the tested experiments similar to that of the TREC Web track queries (rows 19, 10, and 16 in Table 7.5 for the experiments $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$, respectively).

Regarding the MTS technique, the results suggest that when employing an experiment, which considers documents with all the query terms in a combination of the anchor text and title fields, sampling shorter queries is more effective than sampling longer queries (rows 7-12 vs. rows 13-18 in Table 7.5). However, the experiment $\mathcal{E}_{\forall(b),std(dom)}$, which considers documents with all the query terms in their body, benefits from sampling longer queries. This indicates that the type of the experiment should be considered when setting the characteristics of the length distribution for the sampled queries.

The most effective sampling technique for the experiment $\mathcal{E}_{\forall(at)}$, which counts the number of documents with all the query terms in their anchor text or title, is ATS (row 19 from Table 7.5). This fact confirms the correspondence between queries and the anchor text of Web documents suggested by Eiron & McCurley (2003*b*). An advantage of ATS over MTS is that sampling the anchor text provides queries with a representative length distribution. On the other hand, the MTS technique requires to specify the distribution of the query length.

Overall, query sampling can be employed to automatically generate queries, in order to approximate the distribution of outcome values of an experiment for the real TREC Web track queries. In the next section, query sampling is employed in the context of an ad-hoc decision mechanism, in order to set the value of a decision boundary. This will allow to reduce the dependence of the decision mechanism on training data. Therefore, it facilitates the application of selective Web IR in a setting, where relevance information is hardly used to set the decision mechanism.

7.4.4 Evaluation of ad-hoc decision mechanism

This section investigates the effectiveness of an ad-hoc decision mechanism for selective Web IR with limited relevance information, as described in Section 7.4.1. The ad-hoc decision mechanism employs query sampling in order to set a decision boundary.

The ad-hoc decision mechanism employs two retrieval approaches a_1 and a_2, and it has only one decision boundary Bnd. The decision boundary Bnd is set so that the outcome o of an experiment $\mathcal{E}$ is lower than Bnd with a given probability $P(o < Bnd)$. This probability is obtained from the distribution of the outcome o of the experiment $\mathcal{E}$ using the query sampling techniques described in Section 7.4.2. When the outcome of an experiment $\mathcal{E}$ for a query is lower than the decision boundary Bnd, then the retrieval approach a_1 is applied, otherwise, the retrieval approach a_2 is applied.

The remainder of this section describes the experimental setting for the evaluation of the ad-hoc decision mechanism, presents the evaluation results, and closes with a discussion of the findings.

7.4.4.1 Experimental setting for the ad-hoc decision mechanism

This section briefly describes the used experimental setting for the evaluation of the ad-hoc decision mechanism.

The employed experiments $\mathcal{E}$ are the ones that performed well when a Bayesian decision mechanism has been employed in a setting with limited relevance information: $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$ (Section 7.3 on page 195).

For each of the employed experiments $\mathcal{E}$, the decision boundary Bnd of the ad-hoc decision mechanism is set so that $P(o < Bnd) = 0.5$. In other words, the decision boundary Bnd is set so that the outcome o of $\mathcal{E}$ is less than the decision boundary for 50% of the queries. The value of Bnd is set with respect to the distribution of o obtained with the most effective query sampling technique for each experiment $\mathcal{E}$ (Section 7.4.3.3 and Table 7.5).

The evaluation is performed with two mixed tasks, namely mq2003 and mq2004. For each task, the ad-hoc decision mechanism employs the two retrieval approaches with the highest potential for improvements in retrieval effectiveness, as described in Table 7.1: the field-based weighting model BM25F and the combination of the field-based weighting model DLHF with PageRank (DLHFP) for evaluating with mq2003

(row 11 in Table 7.1); the field-based weighting model PB2F and DLHFP for evaluating with mq2004 (row 12 in Table 7.1).

The combination of the field-based weighting model DLHF with PageRank is applied for the queries which result in an experiment outcome $o > Bnd$, in order to favour the broader queries, which are more likely to retrieve many documents with all the query terms, or many aggregates of documents. When the decision mechanism is used for the mixed task mq2003, if the outcome o of an experiment $\mathcal{E}$ is lower than the decision boundary Bnd, then BM25F is applied, otherwise DLHFP is applied. Similarly, when the decision mechanism is evaluated for the mixed task mq2004, if the outcome o of an experiment $\mathcal{E}$ is lower than the decision boundary Bnd, then PB2F is applied, otherwise DLHFP is applied.

7.4.4.2 Evaluation results for the ad-hoc decision mechanism

Table 7.6 displays the evaluation results for the ad-hoc decision mechanism that employs the experiments $\mathcal{E}_{\forall(at)}$ (rows 1-2), $\mathcal{E}_{\forall(at),avg(dom)}$ (rows 3-4), and $\mathcal{E}_{\forall(b),std(dom)}$ (rows 5-6). The rows preceding the evaluation results in the table describe the setting of the decision mechanism, that is the query sampling technique used to obtain the distribution of outcome values for each experiment, and the value of the decision boundary Bnd, as explained in Section 7.4.4.1. The column 'Baseline' in the table displays the mean average precision (MAP) of the most effective individual retrieval approach, and the column 'Bayesian' displays the obtained MAP by the Bayesian decision mechanism, which is applied with limited relevance information in the same setting, as presented in Table 7.2.

The results indicate that the ad-hoc decision mechanism is effective in the considered setting. When employing the experiment $\mathcal{E}_{\forall(at)}$ (rows 1-2), the obtained mean average precision (MAP) for the task mq2003 by the decision mechanism is 0.5903, which represents an improvement of 6.69% over the most effective retrieval approach (0.5533). The ad-hoc decision mechanism also performs better than the corresponding Bayesian decision mechanism, which is applied with limited relevance information in the same setting (0.5903 vs. 0.5775 from row 1 and columns 'MAP' and 'Bayesian', respectively). The ad-hoc decision mechanism performs well when applied to the task mq2004, and results in +5.65% improvement over the most effective retrieval approach (row 2). For both tested tasks, the ad-hoc decision mechanism applies the most effective retrieval

approach for a statistically significant number of queries, as indicated by †. In particular for the task mq2003, the improvement in MAP is statistically significant according to Wilcoxon's signed rank, as denoted by *.

Row	Task	Retrieval approaches		Baseline	Bayesian	$\mathcal{E}$	MAP	+/- %
ATS with $af_{lo} = 5$, $P(o < Bnd) = 0.5$, and $Bnd = 11.9956$								
1	mq2003	BM25F	DLHFP	0.5533	0.5775	$\mathcal{E}_{\forall(at)}$	0.5903	$+6.69^{\dagger *}$
2	mq2004	PB2F	DLHFP	0.4145	0.4381	$\mathcal{E}_{\forall(at)}$	0.4391	$+5.65^{\dagger}$
MTS with $r_{lo} = 200$, $r_{hi} = 2000$, $\mu = 2.1$, $\sigma = 0.78$, $P(o < Bnd) = 0.5$, and $Bnd = 1.5129$								
3	mq2003	BM25F	DLHFP	0.5533	0.5626	$\mathcal{E}_{\forall(at),avg(dom)}$	0.5728	$+3.52^{\dagger *}$
4	mq2004	PB2F	DLHFP	0.4145	0.4452	$\mathcal{E}_{\forall(at),avg(dom)}$	0.4431	+6.62
MTS with $r_{lo} = 200$, $r_{hi} = 2000$, $\mu = 3.2$, $\sigma = 1.31$, $P(o < Bnd) = 0.5$, and $Bnd = 24.6724$								
5	mq2003	BM25F	DLHFP	0.5533	0.5777	$\mathcal{E}_{\forall(b),std(dom)}$	0.5799	$+4.81^{\dagger *}$
6	mq2004	PB2F	DLHFP	0.4145	0.4212	$\mathcal{E}_{\forall(b),std(dom)}$	0.4349	$+4.64^{\dagger}$

Table 7.6: Evaluation of the ad-hoc decision mechanism with the experiments $\mathcal{E}_{\forall(at)}$, $\mathcal{E}_{\forall(at),avg(dom)}$, and $\mathcal{E}_{\forall(b),std(dom)}$. The table displays the evaluation task ('Task'), the employed retrieval approaches ('Retrieval approaches'), the mean average precision of the most effective retrieval approach ('Baseline'), the mean average precision obtained from a Bayesian decision mechanism, applied with limited relevance information for the same setting ('Bayesian'), the employed experiment ('$\mathcal{E}$'), the mean average precision of the ad-hoc decision mechanism ('MAP'), and the relative improvement over the baseline ('+/- %'). The symbol † denotes that the decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries, according to the sign test. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective retrieval approach is statistically significant, according to Wilcoxon's signed rank test. The rows preceeding the results describe the setting of the ad-hoc decision mechanism.

The ad-hoc decision mechanism, which employs the experiment $\mathcal{E}_{\forall(at),avg(dom)}$ (rows 3-4), also results in improvements over the baseline for both tested tasks mq2003 and mq2004 (0.5728 vs. 0.5533 and 0.4431 vs. 0.4145, respectively). In addition, it applies the most effective retrieval approach for a statistically significant number of queries from the mixed task mq2003, and it outperforms the corresponding Bayesian decision mechanism (0.5728 vs. 0.5626 from row 3). When the decision mechanism employs the experiment $\mathcal{E}_{\forall(b),std(dom)}$, the obtained MAP is higher than that of the corresponding decision mechanism for both tested tasks (0.5799 vs. 0.5777 and 0.4349 vs. 0.4212 from rows 5-6, respectively). The decision mechanism also selectively applies the most effective retrieval approach for a statistically significant number of queries for both tested tasks, as indicated by †. In the case of mq2003, there is a statistically significant improvement in MAP, according to Wilcoxon's signed rank test, as denoted by *.

7.4.4.3 Discussion

Overall, it has been shown that query sampling can be effectively used to set the decision boundary of the proposed ad-hoc decision mechanism in the tested setting (Table 7.6). The ad-hoc decision mechanism performs as well as the corresponding Bayesian decision mechanism with limited relevance information. This section further discusses two issues related to the ad-hoc decision mechanism.

Selecting the retrieval approach to apply when $o < Bnd$ In the case of the task mq2004, the evaluated decision mechanism in Section 7.4.4.2 applies the field-based weighting model PB2F when the outcome o of an experiment is lower than the decision boundary Bnd, otherwise, it applies the combination of the field-based weighting model DLHF with PageRank (DLHFP). This setting has been based on expecting that employing PageRank performs better for the most broad queries, which result in higher outcome values for the experiments. However, if some training data is available, then they can be used to suggest which retrieval approach to apply when $o < Bnd$. For example, the training data can be used to estimate the likelihood of obtaining particular outcome values when a retrieval approach is effective. This likelihood can be employed to indicate whether a retrieval approach is expected to be effective for the low or the high outcome values of an experiment, and hence, to select which retrieval approach to apply when the outcome o of the employed experiment is lower than the decision boundary Bnd.

Setting the decision boundary Bnd In the described experiments in Section 7.4.4.2, the decision boundary of the ad-hoc decision mechanism has been set so that $P(o < Bnd) = 0.5$. In other words, the probability that the outcome of an experiment $\mathcal{E}$ is lower than Bnd is 0.5. If further evidence exist to suggest the prior probability that a retrieval approach is effective, then the probability $P(o < Bnd)$ could be set accordingly. For example, when the ad-hoc decision mechanism is applied to the task mq2003, the retrieval approach BM25F outperforms DLHFP for 122 out of the 350 queries. DLHFP outperforms BM25F for 115 queries, while both retrieval approaches result in the same MAP for 113 queries. In this case, the prior probability that BM25F is effective is $\frac{122}{122+115} = 0.515$, which suggests that $P(o < Bnd) = 0.5$ is appropriate for the tested setting.

7.4.5 Conclusions

This section has introduced an ad-hoc decision mechanism and novel techniques for automatically generating samples of queries. The ad-hoc decision mechanism aims to reduce the dependence of applying selective Web IR on training data, by setting its decision boundary with respect to the distribution of the outcome values of an experiment (Section 7.4.1). This distribution of outcome values is obtained from a sample of automatically generated queries (Section 7.4.2).

Three techniques have been proposed for automatically generating queries. In the first one, STS, a generated query corresponds to a randomly sampled term from the vocabulary of the collection (Section 7.4.2.1). In the second technique, MTS, a generated query corresponds to a number of the most informative terms from a set of documents, and the length of the generated query follows a Gaussian distribution (Section 7.4.2.2). The third one, ATS, is a novel technique, where a generated query corresponds to a randomly sampled anchor text from the collection (Section 7.4.2.3). The evaluation of the three proposed techniques has shown that generating queries with either MTS or ATS is more effective than STS in approximating the distribution of the outcome values of an experiment (Section 7.4.3). Moreover, ATS is more effective than MTS in approximating the distribution of the outcome values of the experiment $\mathcal{E}_{\forall(at)}$. On the other hand, MTS is more effective than ATS in approximating the distribution of the outcome values of the experiments $\mathcal{E}_{\forall(at),avg(dom)}$ and $\mathcal{E}_{\forall(b),std(dom)}$. Sampling short queries with MTS is more effective for approximating the distribution of outcome values of $\mathcal{E}_{\forall(at),avg(dom)}$, while sampling long queries with MTS is more effective for approximating the distribution of outcome values of $\mathcal{E}_{\forall(b),std(dom)}$ (Section 7.4.3.3).

The ad-hoc decision mechanism has employed the most effective query sampling technique in order to set its decision boundary for a given experiment (Section 7.4.4.1). The evaluation results in Section 7.4.4.2 have indicated that the ad-hoc mechanism can be effectively used with the experiments $\mathcal{E}$, which have been shown to perform well in the context of a Bayesian decision mechanism with limited relevance information.

In particular, the ad-hoc decision mechanism results in statistically significant improvements in MAP, and it applies the most appropriate retrieval approach for a statistically significant number of queries in the case of the task mq2003.

Overall, this section has proposed an alternative way to using the Bayesian decision mechanism, in order to apply selective Web IR and reduce the dependence on available relevance information.

7.5 Summary

This chapter has investigated the application of selective Web IR in a setting, where limited relevance information exists in order to set a decision mechanism. The concept of limited relevance information corresponds to processing queries from mixed tasks, as well as to training and evaluating the Bayesian decision mechanism with different sets of queries (Section 7.2). This definition of limited relevance information provides a realistic setting for evaluating the effectiveness of selective Web IR.

The evaluation of the Bayesian decision mechanism with limited relevance information has shown that selective Web IR can be effectively applied (Section 7.3). Both the score-independent and the score-dependent experiments resulted in improvements in retrieval effectiveness (Table 7.2 on page 197, and Table 7.3 on page 198). For example, the experiment $\mathcal{E}_{\forall(at)}$, which counts the number of documents with all the query terms in the anchor text, has been shown to be particularly effective (rows 1-2 in Table 7.2). Indeed, the decision mechanism that employs $\mathcal{E}_{\forall(at)}$ applies the most appropriate retrieval approach for a statistically significant number of queries. In particular for the task mq2003, the Bayesian decision mechanism results in statistically significant improvements in MAP.

The introduction of a simple ad-hoc decision mechanism has further reduced the dependence of the Bayesian decision mechanism on training queries. The ad-hoc decision mechanism sets its decision boundary with respect to the distribution of outcome values for a given experiment. This distribution is obtained with three novel query sampling techniques, which generate either single-term queries (Section 7.4.2.1), or queries with more than one term (Sections 7.4.2.2 and 7.4.2.3). The latter generate a distribution of outcome values, which are closer to those corresponding to real TREC Web track queries (Section 7.4.3.3). The evaluation of the ad-hoc decision mechanism has shown that it can be applied effectively in conjunction with the experiments that perform well in the context of a Bayesian decision mechanism with limited relevance information (Section 7.4.4.2). Indeed, the ad-hoc decision mechanism can lead

to statistically significant improvements in retrieval effectiveness, as well as in applying the most appropriate retrieval approach for a statistically significant number of queries (Table 7.6 on page 212). Therefore, the ad-hoc decision mechanism can be useful in the context of an operational Web retrieval setting.

Overall, this chapter has shown that selective Web IR can be applied when limited relevance information exists. This complements the evaluation of selective Web IR in a setting where it has been assumed that relevance information is available, as discussed in Chapter 6.

Chapter 8

Conclusions and Future Work

8.1 Contributions and conclusions

This thesis has investigated selective Web information retrieval, a technique by means of which appropriate retrieval approaches are applied on a per-query basis. This section discusses the contributions and conclusions of this thesis.

8.1.1 Contributions

The main contributions of this thesis are the following:

- A general framework for selective Web information retrieval (IR) has been proposed, and a range of experiments $\mathcal{E}$ has been defined.
- The proposed experiments have been thoroughly evaluated in the context of a Bayesian decision mechanism with a range of Web specific search tasks and a standard TREC Web test collection. The evaluation has been performed in a setting, where relevance information is assumed to exist, as well as in a more realistic and operational setting, where only limited relevance information is available.
- Techniques for the automatic generation of query samples, including a novel technique based on sampling the anchor text of documents, have been introduced and evaluated in the context of setting the decision boundary of a proposed ad-hoc decision mechanism.
- A range of different retrieval approaches for Web IR have been introduced and thoroughly evaluated for a number of different types of search tasks, including

both ad-hoc and Web specific search tasks from standard TREC Web test collections. The introduced retrieval approaches include an extension of the Divergence From Randomness framework to perform per-field normalisation, as well as the Absorbing Model, a novel hyperlink structure analysis algorithm.

8.1.2 Conclusions

This section discusses the achievements and conclusions of this work.

Effectiveness of selective Web information retrieval The work in this thesis has been motivated by the wealth of different retrieval approaches that can be used for Web information retrieval (IR), as well as the fact that there are different types of search tasks performed by users. Most of the related works in the literature have described retrieval approaches that are applied for all the queries uniformly. Other related works have considered the prediction of the performance of a retrieval approach, or the classification of the query type.

The aim of this thesis has been to introduce a general framework for selective Web IR, in order to apply an effective retrieval approach on a per-query basis. Selective Web IR is different from the related work, in the sense that: different retrieval approaches can be applied to queries of the same type of task; and the selection of the most effective retrieval approach considers the retrieval effectiveness of at least two retrieval approaches. The obtained experimental results suggest that selective Web IR can lead to statistically significant improvements in retrieval effectiveness, and that the most effective retrieval approach can be applied for a statistically significant number of queries.

Potential for improvements in retrieval effectiveness from selective Web information retrieval Chapter 4 has established the potential for improvements from selective Web IR, by investigating several retrieval approaches. The proposed retrieval approaches range from field-based weighting models, which perform term frequency normalisation and weighting of each document field independently (Section 4.4), to combinations of the field-based weighting models with query-independent evidence from the URL of Web documents, PageRank, and the Absorbing Model, a novel algorithm for the analysis of the hyperlink structure (Section 4.5). The employed documents

fields are the body, the anchor text of the incoming hyperlinks, and the title of Web documents.

First the proposed retrieval approaches have been compared on the basis of their optimal performance, by setting their hyper-parameters in order to optimise the retrieval effectiveness for each tested task independently (Section 4.3.2). A more realistic setting of the hyper-parameters involved training with mixed types of tasks, and terminating the optimisation process early (Section 4.6). This setting has been employed in order to establish the potential for improvements when the most effective retrieval approach is applied on a per-query basis (Section 4.7). The obtained results have shown that selective Web IR has the potential for statistically significant improvements in retrieval effectiveness.

Decision theoretical framework for selective Web information retrieval Chapter 5 has introduced a new framework for selective Web IR, based on statistical decision theory. One of the main concepts of this framework is the decision mechanism, which employs an experiment $\mathcal{E}$ in order to guess the state of nature, or in other words, in order to select the most effective retrieval approach to apply on a per-query basis (Section 5.2). The consequences of applying a particular retrieval approach are modelled by the loss function, which corresponds to a preference relationship among the retrieval approaches with respect to their retrieval effectiveness.

A range of score-independent and score-dependent experiments $\mathcal{E}$ has been defined. The score-independent experiments consider the occurrence of query terms in documents (Section 5.4). The score-independent document-level experiments count the number of documents with all, or at least one query term in a particular combination of fields of a document (Section 5.3.1). The score-independent aggregate-level experiments consider aggregates of related documents from the same domain, or directory, and estimate features of the aggregate size distribution (Section 5.3.2). Three features of the aggregate size distribution are considered: the average size; the standard deviation of the size; and the number of large aggregates.

The score-dependent experiments $\mathcal{E}$ estimate the usefulness of the hyperlink structure, an indication of whether there are non-random patterns of hyperlinks among the retrieved documents for a query (Section 5.4). The usefulness of the hyperlink structure

is defined in terms of the information theoretic divergence between two score distributions. The first distribution S_n corresponds to the scores assigned to documents by a weighting model. The second distribution is formed in order to favour the documents that point to other highly scored documents. The score of each document corresponds to either the sum of the original score of the document and the scores of all the documents that it points to (U_n), or only the sum of the scores of all the documents that a document points to (U'_n). The usefulness of the hyperlink structure is defined as the symmetric Jensen-Shannon divergence $L(S_n, U_n)$ or $L(S_n, U'_n)$.

Selective Web IR has been primarily investigated in the context of a Bayesian decision mechanism, which has been introduced in Section 5.5. The Bayesian decision mechanism is trained with a set of queries. According to the outcome of an experiment for a given query, the Bayesian decision mechanism selects the retrieval approach, which results in the lowest expected loss. Overall, the proposed framework for selective Web IR is general in the sense that it does not depend on any particular retrieval approach, and it can be applied to select one out of any number of retrieval approaches.

Evaluation of Selective Web information retrieval with relevance information The evaluation in Chapter 6 has been performed with different types of Web search tasks, including the topic distillation, home page finding, and named page finding tasks from the TREC 2003 and 2004 Web tracks (Craswell & Hawking, 2004; Craswell et al., 2003). The evaluation of the proposed experiments has been performed by training and testing the Bayesian decision mechanism with the same task, in order to focus the analysis on the effectiveness of the experiments, and reduce the effect of using different training and testing tasks on the Bayesian decision mechanism. The evaluation results have shown that the Bayesian decision mechanism can selectively apply the most effective retrieval approach on a per-query basis for a statistically significant number of queries, and improve the achieved retrieval effectiveness for both the score-dependent (Section 6.3.3) and the score-independent experiments (Section 6.4.6).

The score-independent document-level experiments, as well as the score-independent aggregate-level experiments, which estimate the standard deviation of the aggregate size distribution, result in a low number of decision boundaries for all the tested tasks (Section 6.3.3). This suggests that they are more effective in identifying the queries for which a retrieval approach is more appropriate. The document-level experiments also

perform well when the considered documents contain all the query terms in a particular field, or a combination of fields. This is explained by the fact that the outcome of the experiment is computed from a cohesive set of documents, which are likely to be about the query (Section 6.3.1.3). The aggregate-level experiments perform well when the considered documents contain the query terms in their body, because the distribution of aggregate sizes is generated from a higher number of documents, compared to when the experiments consider only the documents with query terms in their anchor text or title fields (Section 6.3.2.4). The domain aggregates are also more effective than the directory aggregates, because they provide a better indication that there are groups of related documents about a query (Section 6.3.2.4). As discussed in Section 5.3.2, this may depend on the characteristics of the Web sites that appear in the collection.

The score-dependent experiments that estimate the usefulness of the hyperlink structure $L(S_n, U_n)$ have been shown to be robust in identifying at least one decision boundary in all the tested cases (Section 6.4.2). The experiments that compute the usefulness of the hyperlink structure $L(S_n, U'_n)$ have been more effective when considering documents with all the query terms in a particular field, or in a combination of fields (Section 6.4.5).

Evaluation of Selective Web information retrieval with limited relevance information Selective Web IR has been also evaluated in a setting, where it is assumed that only limited relevance information does exist. The concept of limited relevance information corresponds to employing different mixed tasks for the training and the evaluation of the decision mechanism (Section 7.2).

The application of the Bayesian decision mechanism in this setting has resulted in improvements in retrieval effectiveness (Section 7.3). The score-independent document-level experiment $\mathcal{E}_{\forall(at)}$, which counts the number of documents with all the query terms in their anchor text, has performed particularly well for both tested mixed tasks. The obtained results suggest that the Bayesian decision mechanism can be trained with one set of queries, and selectively apply appropriate retrieval approaches to other previously unseen queries (Section 7.3.3).

Ad-hoc decision mechanism and query sampling techniques In order to reduce the dependence of the Bayesian decision mechanism on training data, an ad-hoc

decision mechanism has been introduced in Section 7.4. The ad-hoc decision mechanism sets its decision boundary according to the distribution of outcome values of an experiment $\mathcal{E}$. This distribution is obtained using three techniques for the automatic generation of queries.

The first technique, STS, generates single-term queries by randomly sampling the vocabulary of the collection (Section 7.4.2.1). The second technique, MTS, generates queries with more than one term by extracting the most informative terms from a set of documents (Section 7.4.2.2). The third technique, ATS, generates queries from the anchor text of the documents in the collection (Section 7.4.2.3). This is based on the observation that the anchor text of Web documents resembles queries (Eiron & McCurley, 2003*b*). The three proposed query sampling techniques have been evaluated with respect to the similarity of the distribution of outcome values of an experiment $\mathcal{E}$ obtained from the sampled queries and the real TREC 2003 and 2004 Web track queries. The evaluation results of the proposed query sampling techniques suggest that using MTS or ATS to sample queries with more than one term performs better than using STS to sample single-term queries (Section 7.4.3.3).

The evaluation of the ad-hoc decision mechanism, which uses query sampling to set its decision boundary, has shown that it can be effectively applied and achieve similar, or better performance than that of the Bayesian decision mechanism, when limited relevance information exists (Section 7.4.4.2).

8.2 Future work

This section discusses several directions for future work related to, or stemming from this thesis.

Selective Web information retrieval with more than two retrieval approaches, or experiments $\mathcal{E}$ The evaluation of selective Web IR in this thesis has been mainly focused on using one of the proposed experiments $\mathcal{E}$ to selectively apply one out of two retrieval approaches. An extension of this work may consider using combinations of experiments in order to obtain more evidence to select the appropriate retrieval approach to apply. For example, using evidence from both score-independent and score-dependent experiments $\mathcal{E}$ may lead to improvements in the performance of the

decision mechanism, because the different experiments may capture diverse and different features of the set of retrieved documents.

Considering more than two retrieval approaches to select from provides another direction for future work. One issue related to employing several retrieval approaches is the investigation of how the Bayesian decision mechanism can be effectively trained in order to apply an appropriate retrieval approach out of several ones. Another interesting direction for future work is the investigation of automatic techniques to select the retrieval approaches that can be effectively applied for selective Web IR.

Definition of the loss function and Games against nature This thesis has focused on the evaluation of selective Web IR from the perspective of retrieval effectiveness. The loss associated with the application of a retrieval approach has been defined in terms of the retrieval effectiveness, as described in Section 5.2. However, other factors can also be incorporated in defining the loss of a retrieval approach.

The computational overhead and the efficiency of each retrieval approach can be considered in defining the loss function. For example, a retrieval approach, which is very effective, but also has a significant computational overhead, may not be appropriate for a Web search task, where fast response times are required from an IR system. Information from a user profile can also be used to define the loss associated with the application of the retrieval approaches. For example, a loss function can bias the decision mechanism towards retrieval approaches that are effective in finding entry points, or detailed information, according to the preferences of a particular user. Such a technique can lead to the application of the proposed approaches in this thesis for performing IR in context, or adaptive IR.

It is also interesting to consider a different formulation of selective Web IR, in terms of *games against nature*. Luce & Raiffa (1957) showed that a statistical decision problem can be transformed into a game against nature, where the reasoning is made in terms of selecting a decision mechanism, instead of the actions to perform for each state of nature. Each of the decision mechanisms has a different loss function, which may correspond to different types of search tasks. Regarding selective Web IR, such a setting can be used to aid the selection of a retrieval approach, by weighting its retrieval effectiveness in the different tasks.

Updating the decision mechanism with relevance feedback The investigation of selective Web IR has been focused on a TREC-like experimental setting, where all the queries are processed in a batch mode. An interesting direction of future work is related to employing selective Web IR in an interactive setting, where the decision mechanism is updated while processing queries. The updating of the decision mechanism can be performed with explicit, or implicit feedback from users. For example, clickthrough data can be used as an indication of the relevance of documents for a given query (Joachims et al., 2005). In this way, the decision mechanism can be refined with more accurate information, and it can adapt the loss function to the search behaviour of the users.

Sampling distribution of experiments $\mathcal{E}$ The proposed query sampling techniques have been used in conjunction with an ad-hoc decision mechanism, in order to set its decision boundary. Another application of the query sampling techniques is related to assigning a probability to the outcome of an experiment $\mathcal{E}$ for a given query. This probability may be used in order to adjust the belief of a decision mechanism in the obtained evidence from the experiment $\mathcal{E}$. For example, the low probability of obtaining a particular outcome for an experiment $\mathcal{E}$ may provide a stronger indication about the structure of the corresponding set of documents, and hence, about applying a particular retrieval approach.

Appendix A

Parameter settings and evaluation of retrieval approaches

This appendix presents the parameter settings used for the evaluation of the retrieval approaches described in Chapter 4. In addition, it presents the precision at 10 retrieved documents (P10), the mean reciprocal rank of the first retrieved relevant document (MRR1), and the number of retrieved relevant documents for the field-based weighting models, and their combination with the query-independent sources of evidence.

	PL2	PB2	$I(n_e)C2$	BM25	
Task	c	c	c	b	k_1
Full text					
tr2000	11.9420	52.9390	6.0645	0.4424	0.47
tr2001	12.3985	10.6837	11.4277	0.4221	0.98
td2002	1.2712	1.1485	0.9024	0.6788	3.29
td2003	0.4134	0.2613	0.1040	0.8827	11.70
td2004	0.1536	0.1417	0.1059	0.9524	15.05
hp2001	0.3456	0.3410	0.5781	0.8349	2.67
hp2003	0.4973	0.5064	0.4320	0.9233	1.01
hp2004	0.2642	0.2881	0.2650	0.8433	11.36
np2002	2.0354	1.4713	2.4378	0.8072	2.99
np2003	0.9420	1.0364	0.9790	0.6975	1.13
np2004	1.7393	1.4253	2.0450	0.5548	3.13
Title					
tr2000	99.3880	193.2049	142.2381	0.0610	0.47
tr2001	7.1904	6.1322	12.2877	0.7527	0.12
td2002	3.1646	5.4616	4.6198	0.4519	0.98
td2003	1.1635	1.2124	1.7706	0.8069	0.86
td2004	1.3719	1.3898	4.1331	0.5006	1.16

continued on next page

continued from previous page					
	PL2	PB2	I(n_e)C2	BM25	
Task	c	c	c	b	k_1
Title					
hp2001	2.0843	2.4594	4.2034	0.5787	0.92
hp2003	2.9547	4.8307	2.6783	0.9991	0.34
hp2004	10.7558	1.2479	4.0749	0.5039	2.71
np2002	5.4129	6.0953	3.9854	0.2840	0.87
np2003	8.0238	9.1423	9.1528	0.6329	0.37
np2004	16.1657	12.1322	10.5189	0.4488	0.73
Headings					
tr2000	23.4745	21.0648	18.4154	0.4134	0.42
tr2001	7.8859	5.7164	8.3176	0.3931	0.17
td2002	10.6788	10.0058	7.8139	0.2905	0.71
td2003	0.9540	0.9499	0.9435	0.8588	0.77
td2004	10.7798	13.1460	14.1948	0.2486	0.64
hp2001	10.2345	4.3676	5.6029	0.4189	0.42
hp2003	3.0215	3.6548	4.8310	0.5268	0.80
hp2004	1.1069	1.0981	1.0772	0.5410	1.11
np2002	9.0942	17.3411	12.4019	0.9039	0.09
np2003	5.8931	10.9804	5.7167	0.9131	0.09
np2004	4.4467	2.8013	3.4802	0.6609	0.52
Anchor text					
tr2000	72.3341	0.7031	0.6915	0.9859	0.12
tr2001	1.0301	1.1675	2.7026	0.8271	0.26
td2002	5.1115	1.4353	1.2275	0.7752	2.40
td2003	1.1242	1.0912	0.9769	0.3439	2.09
td2004	526.7705	605.4442	83.1691	0.1076	1.38
hp2001	805.5639	930.4702	915.9571	0.0175	0.15
hp2003	322.8153	452.3235	931.4237	0.0093	0.99
hp2004	81.1350	71.2218	94.2461	0.0174	0.35
np2002	14.6921	2.0410	11.4034	0.6172	0.50
np2003	13.4140	9.5270	35.6099	0.6643	0.36
np2004	10.7286	6.3414	209.9415	0.5743	0.59

Table A.1: Parameter values for retrieval from the full text, title, headings, and anchor text of documents, with the DFR weighting models PL2, PB2 and I(n_e)C2, and the weighting model BM25.

Task	c_b	c_a	c_t	w_a	w_t
PL2F					
tr2000	4.0855	6.7723	142.5767	0.3284	0.1014
tr2001	10.7273	2.6661	5.8433	6.5407	0.7304
td2002	1.3457	1.0278	26.6474	2.3710	0.8517
td2003	0.3360	4.0263	4.2368	2.2810	2.1520
td2004	0.1293	324.2448	4.5425	0.3886	0.7046
hp2001	0.7721	908.9378	5.2074	0.5038	15.2169
hp2003	0.4573	306.6216	26.9822	2.2132	9.6546
hp2004	0.4129	87.0932	100.4131	6.1152	7.9996
np2002	1.4210	27.9542	16.7559	1.4820	3.8587
np2003	1.0380	25.7228	10.9078	0.4325	3.0119
np2004	1.2840	15.9861	7.2411	3.5909	33.0245
PB2F					
tr2000	16.0796	5.9412	42.1003	0.0791	0.0166
tr2001	8.9545	1.7225	2.3878	2.2264	1.7090
td2002	0.9980	3.5936	20.0154	0.6134	2.7686
td2003	0.2794	5.3132	10.4665	3.5192	3.7206
td2004	0.1032	49.5182	7.0637	2.3072	6.9674
hp2001	0.3838	324.6800	3.4474	0.3079	8.0844
hp2003	0.5326	43.1579	2.7982	5.6396	31.8623
hp2004	0.2873	64.3357	42.3266	30.5284	31.3110
np2002	1.1296	5.1085	35.5495	0.6452	3.2041
np2003	0.5159	7.6558	46.3645	0.7124	2.1435
np2004	1.0491	5.9074	4.2198	3.8822	15.7021
I(n_e)C2F					
tr2000	3.4647	9.1803	21.2702	0.2040	0.5960
tr2001	8.9296	1.6934	1.7729	5.8078	5.2829
td2002	0.7487	1.0332	8.1612	0.6465	0.5532
td2003	0.0750	1.2520	6.0006	1.0876	0.6437
td2004	0.0556	13.7654	1.6926	0.1441	0.6937
hp2001	0.5287	291.4738	2.8953	0.4801	6.9953
hp2003	0.2289	936.6813	3.4725	0.3949	2.1620
hp2004	0.1933	971.5727	51.9629	0.3274	0.6562
np2002	0.4348	14.4389	4.9082	0.6905	2.5353
np2003	0.5690	10.8766	8.9510	0.9375	6.7246
np2004	1.3135	28.2666	9.4366	2.4891	22.7825

Table A.2: The values of the c parameters and the weights of the fields for the weighting models PL2F, PB2F and I(n_e)C2.

DLHF		
Task	w_a	w_t
tr2000	0.0058	0.2247
tr2001	0.0428	0.0057
td2002	0.1757	0.7122
td2003	6.3925	1.2955
td2004	1.7329	0.2073
hp2001	2.1965	0.4854
hp2003	54.2242	9.4886
hp2004	6.4182	3.0871
np2002	0.9546	0.7899
np2003	1.1574	0.3274
np2004	59.8354	2.8585

Table A.3: The weights of the anchor text and title fields for the weighting model DLHF.

BM25F						
Task	b_b	b_a	b_t	k	w_a	w_t
tr2000	0.2850	0.9984	0.1926	0.52	0.2648	0.0416
tr2001	0.3605	0.8214	0.4723	0.60	1.8536	0.2763
td2002	0.6836	0.8437	0.5245	3.89	1.7451	2.7512
td2003	0.9198	0.3766	0.9910	21.00	7.1014	14.4419
td2004	0.9402	0.0499	0.4612	36.69	3.2380	15.8051
hp2001	0.8474	0.0079	0.5912	2.83	4.7198	20.3408
hp2003	0.9493	0.0185	0.7335	3.23	11.1898	28.1010
hp2004	0.8808	0.0031	0.8621	13.13	24.5700	38.9098
np2002	0.8315	0.4384	0.4660	5.04	6.0283	4.5403
np2003	0.8831	0.6641	0.3600	1.46	4.7054	9.7318
np2004	0.6246	0.6236	0.6737	5.20	7.6192	19.2977

Table A.4: The values of the parameters for the weighting model BM25F.

	P10				
Task	PL2F	PB2F	I(n_e)C2F	DLHF	BM25F
tr2000	0.2620	0.2540	0.2640	0.2260	0.2740
tr2001	0.3620	0.3440	0.3800	0.3220	0.3760
td2002	0.2680	0.2700	0.2440	0.2280	0.2640
td2003	0.1320	0.1200	0.1320	0.1200	0.1500
td2004	0.1960	0.1627	0.1813	0.1893	0.2053
hp2001	0.1207	0.1179	0.1248	0.1069	0.1234
hp2003	0.0987	0.0933	0.0980	0.0920	0.1000
hp2004	0.0853	0.0800	0.0893	0.0787	0.0893
np2002	0.0993	0.0973	0.0987	0.0853	0.0993
np2003	0.0953	0.0913	0.0947	0.0840	0.0940
np2004	0.0960	0.0960	0.0907	0.0800	0.0933
Task	PL2FU	PB2FU	I(n_e)C2FU	DLHFU	BM25FU
tr2000	0.2640	0.2540	0.2640	0.2280	0.2760
tr2001	0.3620	0.3440	0.3820	0.3200	0.3780
td2002	0.2680	0.2720	0.2440	0.2400	0.2640
td2003	0.1960	0.1620	0.1780	0.1520	0.2000
td2004	0.2413	0.2187	0.2733	0.2787	0.2613
hp2001	0.1255	0.1241	0.1303	0.1145	0.1297
hp2003	0.1073	0.1053	0.1100	0.0973	0.1113
hp2004	0.0893	0.0880	0.0947	0.0853	0.0960
np2002	0.0993	0.0980	0.0987	0.0853	0.1000
np2003	0.0953	0.0913	0.0947	0.0840	0.0940
np2004	0.0960	0.0960	0.0920	0.0800	0.0920
Task	PL2FP	PB2FP	I(n_e)C2FP	DLHFP	BM25FP
tr2000	0.2640	0.2540	0.2640	0.2280	0.2740
tr2001	0.3620	0.3440	0.3800	0.3300	0.3740
td2002	0.2680	0.2720	0.2440	0.2340	0.2640
td2003	0.1380	0.1200	0.1320	0.1340	0.1360
td2004	0.2000	0.1773	0.2200	0.2293	0.2120
hp2001	0.1207	0.1172	0.1248	0.1069	0.1248
hp2003	0.1033	0.1000	0.1047	0.0967	0.1040
hp2004	0.0880	0.0840	0.0960	0.0867	0.0960
np2002	0.0993	0.0973	0.0993	0.0887	0.1000
np2003	0.0980	0.0933	0.0967	0.0853	0.0960
np2004	0.0973	0.0973	0.0920	0.0800	0.0960
Task	PL2FA	PB2FA	I(n_e)C2FA	DLHFA	BM25FA
tr2000	0.2620	0.2540	0.2640	0.2120	0.2700
tr2001	0.3500	0.3420	0.3720	0.3200	0.3760
td2002	0.2680	0.2720	0.2420	0.2300	0.2620
td2003	0.1340	0.1200	0.1340	0.1220	0.1480
td2004	0.1987	0.1733	0.1987	0.1960	0.2080
hp2001	0.1207	0.1193	0.1241	0.1021	0.1228
hp2003	0.1033	0.0953	0.1013	0.0933	0.1033
hp2004	0.0893	0.0840	0.0933	0.0800	0.0907
np2002	0.0987	0.0973	0.0980	0.0853	0.1000
np2003	0.0980	0.0953	0.0947	0.0820	0.0960
np2004	0.0960	0.0960	0.0907	0.0800	0.0947

continued on next page

continued from previous page

	P10				
Task	PL2FA	PB2FA	I(n_e)C2FA	DLHFA	BM25FA

Table A.5: Precision at 10 retrieved documents (P10) for field retrieval and combination with query-independent evidence.

	MRR1				
Task	PL2F	PB2F	I(n_e)C2F	DLHF	BM25F
tr2000	0.5524	0.4800	0.5130	0.4658	0.5136
tr2001	0.7107	0.6581	0.6855	0.5565	0.6753
td2002	0.5711	0.5413	0.5479	0.5098	0.5423
td2003	0.3907	0.3915	0.3874	0.3670	0.4252
td2004	0.4511	0.4184	0.3858	0.4132	0.4420
hp2001	0.6797	0.6626	0.7138	0.5895	0.7231
hp2003	0.7879	0.7285	0.7746	0.6771	0.7940
hp2004	0.6711	0.6033	0.6666	0.5909	0.6868
np2002	0.7289	0.6971	0.7368	0.5893	0.7333
np2003	0.7669	0.7200	0.7083	0.5961	0.7134
np2004	0.7531	0.7285	0.7137	0.5453	0.7245
Task	PL2FU	PB2FU	I(n_e)C2FU	DLHFU	BM25FU
tr2000	0.5523	0.4800	0.5470	0.4693	0.5462
tr2001	0.7107	0.6581	0.6736	0.5596	0.6755
td2002	0.5709	0.5412	0.5479	0.5054	0.5423
td2003	0.4662	0.4694	0.4434	0.4461	0.5306
td2004	0.6156	0.6021	0.5917	0.5931	0.6388
hp2001	0.8270	0.7890	0.8363	0.7522	0.8415
hp2003	0.8273	0.7774	0.8348	0.7666	0.8522
hp2004	0.7206	0.6648	0.7298	0.6574	0.7311
np2002	0.7289	0.6971	0.7368	0.5893	0.7336
np2003	0.7669	0.7202	0.7083	0.5984	0.7134
np2004	0.7561	0.7455	0.7220	0.5480	0.7396
Task	PL2FP	PB2FP	I(n_e)C2FP	DLHFP	BM25FP
tr2000	0.5524	0.4800	0.5127	0.4662	0.5136
tr2001	0.7107	0.6581	0.6855	0.5695	0.6822
td2002	0.5859	0.5413	0.5693	0.5071	0.5425
td2003	0.4240	0.4033	0.4835	0.3936	0.4576
td2004	0.4717	0.4527	0.4667	0.4612	0.4839
hp2001	0.6777	0.6591	0.7176	0.5988	0.7146
hp2003	0.7965	0.7535	0.8240	0.7758	0.8474
hp2004	0.6943	0.6387	0.7812	0.6474	0.7671
np2002	0.7324	0.6989	0.7492	0.5909	0.7439
np2003	0.8011	0.7447	0.7687	0.6306	0.7925
np2004	0.7697	0.7415	0.7372	0.5464	0.7479
Task	PL2FA	PB2FA	I(n_e)C2FA	DLHFA	BM25FA
tr2000	0.5424	0.4800	0.5122	0.4601	0.5121

continued on next page

continued from previous page					
	MRR1				
Task	PL2FA	PB2FA	I(n_e)C2FA	DLHFA	BM25FA
tr2001	0.7200	0.6775	0.6800	0.5547	0.6753
td2002	0.5711	0.5414	0.5364	0.4919	0.5381
td2003	0.4092	0.3955	0.4176	0.3677	0.4340
td2004	0.4602	0.4421	0.4187	0.4165	0.4558
hp2001	0.6763	0.6592	0.7191	0.5851	0.7195
hp2003	0.7902	0.7564	0.7966	0.7292	0.8273
hp2004	0.6858	0.6146	0.7107	0.6085	0.7141
np2002	0.7321	0.6999	0.7439	0.5894	0.7370
np2003	0.7826	0.7290	0.7178	0.6072	0.7523
np2004	0.7707	0.7439	0.7302	0.5453	0.7574

Table A.6: Mean reciprocal rank of the first retrieved relevant document (MRR1) for field retrieval and combination with query-independent evidence.

Task	Relevant docs.	Retrieved relevant documents				
		PL2F	PB2F	I(n_e)C2F	DLHF	BM25F
tr2000	2590	1538	1412	1562	1458	1641
tr2001	3363	2466	2423	2409	2308	2400
td2002	1574	1197	1137	1160	1093	1183
td2003	516	403	383	403	401	398
td2004	1600	1133	979	1151	1094	1138
hp2001	252	246	241	249	238	249
hp2003	194	191	189	191	182	192
hp2004	83	81	80	82	82	82
np2002	170	169	169	168	167	169
np2003	158	157	157	157	155	157
np2004	80	80	80	80	77	80
		PL2FU	PB2FU	I(n_e)C2FU	DLHFU	BM25FU
tr2000	2590	1536	1413	1565	1458	1637
tr2001	3363	2466	2423	2409	2312	2399
td2002	1574	1196	1145	1160	1100	1183
td2003	516	424	402	422	417	410
td2004	1600	1208	1100	1214	1224	1182
hp2001	252	243	242	249	245	249
hp2003	194	191	190	194	188	194
hp2004	83	81	81	82	82	82
np2002	170	169	169	168	167	169
np2003	158	157	157	157	155	157
np2004	80	80	80	80	77	80
		PL2FPR	PB2FPR	I(n_e)C2FPR	DLHFPR	BM25FPR
tr2000	2590	1538	1412	1560	1451	1641
tr2001	3363	2466	2423	2409	2323	2399
td2002	1574	1203	1141	1169	1113	1186
td2003	516	411	392	403	404	409

continued on next page

continued from previous page

Task	Relevant docs.	Retrieved relevant documents				
		PL2FP	PB2FP	I(n_e)C2FP	DLHFP	BM25FP
td2004	1600	1145	1020	1193	1191	1173
hp2001	252	246	241	249	240	249
hp2003	194	191	190	192	187	193
hp2004	83	82	82	83	82	83
np2002	170	169	169	168	166	169
np2003	158	157	157	157	155	157
np2004	80	80	80	80	77	80
		PL2FA	PB2FA	I(n_e)C2FA	DLHFA	BM25FA
tr2000	2590	1540	1412	1562	1431	1641
tr2001	3363	2465	2414	2416	2303	2400
td2002	1574	1197	1141	1160	1104	1184
td2003	516	403	382	404	400	404
td2004	1600	1133	993	1165	1122	1142
hp2001	252	246	241	249	239	249
hp2003	194	192	191	194	185	193
hp2004	83	82	80	83	82	82
np2002	170	169	169	168	167	169
np2003	158	157	157	157	155	157
np2004	80	80	80	80	77	80

Table A.7: Number of retrieved relevant documents for field retrieval and combination with query-independent evidence.

	ω_u	k_u	ω_{pr}	k_{pr}	ω_{am}	k_{am}
Task	PL2FU		PL2FP		PL2FA	
tr2000	0.3093	1.5751	0.0375	23.9525	0.0375	23.9525
tr2001	0.0929	0.5403	0.0002	0.7134	0.0002	0.7134
td2002	1.5717	1.8468	3.3667	47.1483	3.3667	47.1483
td2003	7.8659	8.0463	5.1552	10.3136	5.1552	10.3136
td2004	7.6821	12.4129	1.8550	0.1420	1.8550	0.1420
hp2001	15.7055	19.3671	1.3039	35.1833	1.3039	35.1833
hp2003	14.3775	18.3087	6.2034	1.0129	6.2034	1.0129
hp2004	9.0759	14.6394	5.5842	0.8826	5.5842	0.8826
np2002	0.1013	0.1658	0.9174	0.1168	0.9174	0.1168
np2003	0.0046	5.5921	4.1131	0.1685	4.1131	0.1685
np2004	1.9834	19.4505	5.7268	0.8798	5.7268	0.8798
Task	PB2FU		PB2FP		PB2FA	
tr2000	0.0635	4.9097	0.0009	1.2607	0.0009	1.2607
tr2001	0.9795	0.1521	0.0089	4.0126	0.0089	4.0126
td2002	13.9630	0.3353	0.5162	6.0564	0.5162	6.0564
td2003	29.5918	5.1464	9.7372	3.7462	9.7372	3.7462
td2004	34.5830	9.7946	14.4232	0.7003	14.4232	0.7003
hp2001	25.5420	3.9264	0.6829	0.4400	0.6829	0.4400
hp2003	23.5538	14.0256	19.2171	4.0532	19.2171	4.0532
hp2004	50.8372	21.0273	31.6911	1.0226	31.6911	1.0226

continued on next page

continued from previous page						
	ω_u	k_u	ω_{pr}	k_{pr}	ω_{am}	k_{am}
Task	PB2FU		PB2FP		PB2FA	
np2002	2.9455	6.8139	1.0794	0.5002	1.0794	0.5002
np2003	0.9358	5.8195	16.8391	0.3043	16.8391	0.3043
np2004	14.1126	20.0929	9.9938	1.0405	9.9938	1.0405
Task	I(n_e)C2FU		I(n_e)C2FP		I(n_e)C2FA	
tr2000	0.4539	52.8343	0.2380	40.1671	0.2380	40.1671
tr2001	0.4432	0.2414	0.1169	322.0717	0.1169	322.0717
td2002	0.2309	0.1100	0.6187	3.8294	0.6187	3.8294
td2003	13.3819	12.7960	9.6733	1.6143	9.6733	1.6143
td2004	9.0207	14.1381	5.4772	0.9171	5.4772	0.9171
hp2001	5.7255	12.3012	1.1897	160.9115	1.1897	160.9115
hp2003	4.9946	12.6104	2.6812	11.0089	2.6812	11.0089
hp2004	2.1851	6.8189	9.1281	0.5099	9.1281	0.5099
np2002	0.2132	0.0152	1.9003	0.0842	1.9003	0.0842
np2003	0.0177	0.1168	4.0721	0.2500	4.0721	0.2500
np2004	0.7658	3.7596	0.5165	0.8709	0.5165	0.8709
Task	DLHFU		DLHFP		DLHFA	
tr2000	0.4859	23.1299	1.5792	18.0661	1.5792	18.0661
tr2001	1.2604	0.7566	0.5686	1.0517	0.5686	1.0517
td2002	2.4051	6.7482	1.4771	4.8690	1.4771	4.8690
td2003	7.4755	6.8915	1.8798	0.5452	1.8798	0.5452
td2004	9.1663	11.9887	4.2494	0.6156	4.2494	0.6156
hp2001	17.6257	28.4453	3.3613	1.1164	3.3613	1.1164
hp2003	16.4713	12.0029	13.9537	5.2911	13.9537	5.2911
hp2004	7.5731	10.0293	9.3553	1.0605	9.3553	1.0605
np2002	1.0460	0.0585	1.8805	0.1457	1.8805	0.1457
np2003	0.9707	1.2028	3.5503	0.4976	3.5503	0.4976
np2004	0.7591	45.0101	2.5503	1.2275	2.5503	1.2275
Task	BM25FU		BM25FP		BM25FA	
tr2000	0.8037	86.3178	0.0189	10.0952	0.0189	10.0952
tr2001	0.4702	0.1774	0.2245	53.1777	0.2245	53.1777
td2002	0.1301	0.0343	0.0677	10.3742	0.0677	10.3742
td2003	6.8095	16.3036	2.3923	4.7923	2.3923	4.7923
td2004	6.1798	37.1125	1.7851	3.5860	1.7851	3.5860
hp2001	3.8562	14.5958	0.3967	77.0965	0.3967	77.0965
hp2003	2.0919	13.4126	1.7147	7.9865	1.7147	7.9865
hp2004	1.6581	5.4128	7.3362	0.4661	7.3362	0.4661
np2002	0.1761	3.9992	2.0513	0.0784	2.0513	0.0784
np2003	0.0125	13.2079	2.7349	0.0784	2.7349	0.0784
np2004	2.1986	30.4512	2.3428	2.2427	2.3428	2.2427

Table A.8: The parameter values for the combination of the weighting models with the query-independent evidence.

Task (train)	c_b	c_a	c_t	w_a	w_t
	PL2F				
mq2003 (mq2004)	0.9319	78.1036	8.4729	10.2889	37.6963
mq2004 (mq2003')	0.6986	73.2827	26.7106	1.0094	2.8912
	PB2F				
mq2003 (mq2004)	0.6296	62.0220	22.0046	1.1658	12.3601
mq2004 (mq2003')	0.4608	21.7563	7.0632	1.4040	6.6747
	I(n_e)C2F				
mq2003 (mq2004)	0.2628	158.0428	9.2125	0.4006	3.3339
mq2004 (mq2003')	0.4487	12.4619	3.2772	1.5102	7.5873
	DLHF				
mq2003 (mq2004)	–	–	–	96.0313	37.3645
mq2004 (mq2003')	–	–	–	14.2996	3.6735

	BM25F					
Task (train)	b_b	b_a	b_t	k	w_a	w_t
mq2003 (mq2004)	0.8211	0.0093	0.4580	4.39	8.6475	33.7770
mq2004 (mq2003')	0.8896	0.0487	0.8453	5.62	8.1673	20.0292

Table A.9: The values of the parameters and the weights of the fields for the weighting models PL2F, PB2F, I(n_e)C2, DLHF and BM25F for training and evaluating with different mixed tasks. The parameter values used for the mixed tasks are the ones used for their corresponding subsets of tasks.

	ω_u	k_u	ω_{pr}	k_{pr}	ω_{am}	k_{am}
Task (train)	PL2FU		PL2FP		PL2FA	
mq2003 (mq2004)	9.4083	11.0740	9.1928	0.2024	2.4205	0.4538
mq2004 (mq2003')	9.7826	28.6998	4.7340	35.7684	1.8934	0.6473
	PB2FU		PB2FP		PB2FA	
mq2003 (mq2004)	21.6786	15.4118	17.9247	0.5128	5.0749	2.3209
mq2004 (mq2003')	8.0324	8.2846	0.0562	5.3236	0.0036	8.5315
	I(n_e)C2FU		I(n_e)C2FP		I(n_e)C2FA	
mq2003 (mq2004)	1.5936	6.5864	9.3379	0.2262	1.1061	1.3728
mq2004 (mq2003')	1.9588	7.6559	7.1016	0.2680	1.1468	17.5904
	DLHFU		DLHFP		DLHFA	
mq2003 (mq2004)	8.3892	13.8659	14.7347	0.2093	3.0739	14.1576
mq2004 (mq2003')	15.9754	71.0661	6.7616	1.1263	3.4603	34.2182
	BM25FU		BM25FP		BM25FA	
mq2003 (mq2004)	2.8241	79.7968	5.7691	0.1833	0.6453	1.0578
mq2004 (mq2003')	1.7968	12.7021	2.1615	2.0311	1.1634	20.6925

Table A.10: The values of the parameters for the combination of each field retrieval weighting model and the query-independent evidence for training and evaluating with different mixed tasks. The parameter values used for the mixed tasks are the ones used for their corresponding subsets of tasks. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task.

Task (train)	c_b	c_a	c_t	w_a	w_t
	PL2F				
mq2003 (mq2004)	0.9572	64.9514	8.0774	5.3941	8.8938
mq2004 (mq2003')	0.6607	1.2557	1.2172	5.2971	7.1962
	PB2F				
mq2003 (mq2004)	0.9905	4.9990	15.4419	8.2748	13.7280
mq2004 (mq2003')	1.0161	1.3790	3.0255	2.9781	5.0802
	I(n_e)C2F				
mq2003 (mq2004)	0.8462	113.3989	1.2494	1.3378	5.3631
mq2004 (mq2003')	1.1734	1.0854	2.7481	12.5489	26.0581
	DLHF				
mq2003 (mq2004)	–	–	–	9.3857	9.8862
mq2004 (mq2003')	–	–	–	4.9184	5.7829

	BM25F					
Task (train)	b_b	b_a	b_t	k	w_a	w_t
mq2003 (mq2004)	0.5804	0.4794	0.5462	2.92	13.2098	13.9637
mq2004 (mq2003')	0.4866	0.5291	0.5663	2.20	18.2071	9.4071

Table A.11: The values of the parameters and the weights of the fields for the weighting models PL2F, PB2F, I(n_e)C2, DLHF and BM25F for training and evaluating with mixed tasks, and restricted optimisation. The parameter values used for the mixed tasks are the ones used for their corresponding subsets of tasks. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task.

	ω_u	k_u	ω_{pr}	k_{pr}	ω_{am}	k_{am}
Task (train)	PL2FU		PL2FP		PL2FA	
mq2003 (mq2004)	6.8737	8.0400	5.5400	0.1651	2.2822	2.7487
mq2004 (mq2003')	9.0689	8.7801	6.5624	18.2335	5.7708	1.6061
	PB2FU		PB2FP		PB2FA	
mq2003 (mq2004)	5.8988	3.3266	14.9628	46.2044	4.9506	3.3182
mq2004 (mq2003')	10.1833	5.2613	20.4207	0.8000	8.1035	2.2948
	I(n_e)C2FU		I(n_e)C2FP		I(n_e)C2FA	
mq2003 (mq2004)	4.5683	0.1295	1.5623	16.6362	1.0130	5.1615
mq2004 (mq2003')	2.0438	2.4396	3.6810	0.4796	10.7205	0.0155
	DLHFU		DLHFP		DLHFA	
mq2003 (mq2004)	7.7638	2.7803	7.7489	8.4433	5.6169	0.0993
mq2004 (mq2003')	6.9989	4.2361	10.9063	0.4846	3.0052	0.8618
	BM25FU		BM25FP		BM25FA	
mq2003 (mq2004)	3.1348	3.0927	2.7073	3.7313	1.4702	41.2509
mq2004 (mq2003')	3.7348	3.5144	1.2952	2.8207	0.9682	13.3225

Table A.12: The values of the parameters for the combination of each field retrieval weighting model and the query-independent evidence for training and evaluating with mixed tasks, and restricted optimisation. The parameter values used for the mixed tasks are the ones used for their corresponding subsets of tasks. The task mq2003' corresponds to a subset of mq2003, which consists of the first 50 topics for each type of task.

Appendix B

Evaluation of experiments $\mathcal{E}$

This appendix presents the evaluation results from all the introduced experiments $\mathcal{E}$, in the context of a Bayesian decision mechanism, which employs two retrieval approaches based on PL2F, PB2F, I(n_e)C2F, DLHF, BM25F, or two different weighting models, respectively.

Tables B.1 to B.11 present the evaluation results for a Bayesian decision mechanism, which is trained and tested with the same search task, as described in Chapter 6. In these tables, the first column displays the name of the tested topic set, the two retrieval approaches employed by the decision mechanism, and the mean average precision of the most effective one. The second column displays the evaluation results for the experiments that consider documents with at least one query term in a particular combination of fields. Similarly, the third column displays the evaluation results for the experiments that consider documents with all the query terms in a particular combination of fields. For each evaluated decision mechanism, the tables report the employed experiment ('$\mathcal{E}$'), the obtained mean average precision ('MAP'), the relative difference between the MAP of the most effective retrieval approach and the obtained MAP by the decision mechanism ('+/-%'), and the number of decision boundaries ('B'). The symbol † denotes that the Bayesian decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach is statistically significant according to Wilcoxon's singed rank test. If an experiment $\mathcal{E}$ does not identify at least one decision boundary for a particular task, because the posterior likelihood of one retrieval approach is always

higher than the posterior likelihood of the other retrieval approach, then this is denoted by – in the tables. For example, when the Bayesian decision mechanism employs the experiment $\mathcal{E}_{\exists(at)}$ in order to selectively apply either PL2FA or I(n_e)C2FA for the task np2003, there is no decision boundary identified (Table B.1). For this reason, the results of the experiment $\mathcal{E}_{\exists(at)}$ are only reported in Table B.1, and not in Table 6.2 (page 138).

Tables B.12 and B.13 present the evaluation results for a Bayesian decision mechanism, which is trained and evaluated with different sets of mixed tasks, as described in Chapter 7. The columns of these tables display: the name of the evaluation task ('Task'); the two retrieval approaches employed by the decision mechanism ('Retrieval approaches'); the mean average precision of the most effective individual retrieval approach ('Baseline'); the employed experiment ('$\mathcal{E}$'); the obtained mean average precision by the Bayesian decision mechanism ('MAP'); the relative difference between the MAP of the most effective retrieval approach and the obtained MAP by the decision mechanism ('+/-%'); and the number of decision boundaries ('B'). The symbol † denotes that the Bayesian decision mechanism applies the most appropriate retrieval approach for a statistically significant number of queries. The symbol * denotes that the difference between the MAP of the decision mechanism and that of the most effective individual retrieval approach is statistically significant according to Wilcoxon's singed rank test.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	–	–	–
0.1606 PL2FP	$\mathcal{E}_{\exists(at)}$	0.1520	- 5.40†	1	$\mathcal{E}_{\forall(at)}$	0.1619	+ 0.81†	1
td2004 PL2F	$\mathcal{E}_{\exists(b)}$	0.1355	+ 4.31	1	$\mathcal{E}_{\forall(b)}$	0.1358	+ 4.54†	3
0.1299 PL2FA	$\mathcal{E}_{\exists(at)}$	0.1343	+ 3.39†	2	$\mathcal{E}_{\forall(at)}$	0.1331	+ 2.46	1
hp2003 PL2FU	$\mathcal{E}_{\exists(b)}$	0.7435	0.00†	1	$\mathcal{E}_{\forall(b)}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at)}$	0.7409	- 0.35†	1	$\mathcal{E}_{\forall(at)}$	0.7380	- 0.74	1
hp2004 PL2FU	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	–	–	–
0.6674 PL2FP	$\mathcal{E}_{\exists(at)}$	0.6674	0.00	1	$\mathcal{E}_{\forall(at)}$	0.6816	+ 2.13	1
np2003 PL2F	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	0.6742	+ 0.43†	1
0.6713 PL2FA	$\mathcal{E}_{\exists(at)}$	0.6687	- 0.39	2	$\mathcal{E}_{\forall(at)}$	0.6692	- 0.31	1
np2004 PL2F	$\mathcal{E}_{\exists(b)}$	0.7480	+ 4.34$^{\dagger *}$	2	$\mathcal{E}_{\forall(b)}$	0.7174	+ 0.07	1
0.7169 PL2FA	$\mathcal{E}_{\exists(at)}$	0.7501	+ 4.63†	2	$\mathcal{E}_{\forall(at)}$	0.7296	+ 1.77	1
td2003 PB2F	$\mathcal{E}_{\exists(b)}$	0.1389	- 2.00	2	$\mathcal{E}_{\forall(b)}$	0.1366	- 3.60	1
0.1417 PB2FA	$\mathcal{E}_{\exists(at)}$	0.1393	- 1.70	2	$\mathcal{E}_{\forall(at)}$	0.1377	- 2.80	1
td2004 PB2FU	$\mathcal{E}_{\exists(b)}$	0.1451	+ 3.35$^{\dagger *}$	1	$\mathcal{E}_{\forall(b)}$	0.1417	+ 0.93	1
0.1404 PB2FP	$\mathcal{E}_{\exists(at)}$	0.1440	+ 2.56	1	$\mathcal{E}_{\forall(at)}$	0.1441	+ 2.64*	1

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
hp2003 PB2FU	$\mathcal{E}_{\exists(b)}$	0.6621	+ 0.49	1	$\mathcal{E}_{\forall(b)}$	0.6696	+ 1.62†	1
0.6589 PB2FP	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.6658	+ 1.05	1
hp2004 PB2FU	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	0.5762	+ 1.50†	1
0.5677 PB2FP	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.5766	+ 1.57†	2
np2003 PB2F	$\mathcal{E}_{\exists(b)}$	0.6728	+ 1.42$^{\dagger *}$	3	$\mathcal{E}_{\forall(b)}$	–	–	–
0.6634 PB2FP	$\mathcal{E}_{\exists(at)}$	0.6662	+ 0.42	1	$\mathcal{E}_{\forall(at)}$	0.6692	+ 0.87†	1
np2004 PB2FU	$\mathcal{E}_{\exists(b)}$	0.7091	- 2.10	1	$\mathcal{E}_{\forall(b)}$	0.7318	+ 1.06	1
0.7241 PB2FP	$\mathcal{E}_{\exists(at)}$	0.7189	- 0.72	1	$\mathcal{E}_{\forall(at)}$	0.7174	- 0.93	1
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	–	–	–
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.1341	+ 4.52†	1
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	–	–	–
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.1271	- 2.80†	1
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b)}$	0.7320	- 0.31	1	$\mathcal{E}_{\forall(b)}$	–	–	–
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at)}$	0.7323	- 0.27	1	$\mathcal{E}_{\forall(at)}$	0.7220	- 1.70	1
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	0.6939	+ 4.63†	1
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.7031	+ 6.02†	3
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b)}$	0.6978	+ 0.55	1	$\mathcal{E}_{\forall(b)}$	–	–	–
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at)}$	0.7022	+ 1.18	1	$\mathcal{E}_{\forall(at)}$	–	–	–
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b)}$	0.6923	+ 1.17	1	$\mathcal{E}_{\forall(b)}$	–	–	–
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at)}$	0.6819	- 0.35	1	$\mathcal{E}_{\forall(at)}$	0.7079	+ 3.45	1
td2003 DLHF	$\mathcal{E}_{\exists(b)}$	0.1495	+ 2.75	2	$\mathcal{E}_{\forall(b)}$	0.1466	+ 0.76	2
0.1455 DLHFP	$\mathcal{E}_{\exists(at)}$	0.1453	- 0.14	2	$\mathcal{E}_{\forall(at)}$	0.1434	- 1.40	1
td2004 DLHF	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	–	–	–
0.1371 DLHFP	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.1312	- 4.30	1
hp2003 DLHFU	$\mathcal{E}_{\exists(b)}$	0.6747	+ 0.55	2	$\mathcal{E}_{\forall(b)}$	0.6644	- 0.98	1
0.6710 DLHFP	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	–	–	–
hp2004 DLHFU	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	0.6135	- 2.30	3
0.6278 DLHFP	$\mathcal{E}_{\exists(at)}$	0.6173	- 1.70	2	$\mathcal{E}_{\forall(at)}$	0.6399	+ 1.93†	1
np2003 DLHFP	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	–	–	–
0.5241 DLHFA	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.5377	+ 2.59$^{\dagger *}$	1
np2004 DLHFU	$\mathcal{E}_{\exists(b)}$	0.4973	- 0.10	1	$\mathcal{E}_{\forall(b)}$	0.5128	+ 3.01	1
0.4978 DLHFP	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.4871	- 2.10†	2
td2003 BM25FU	$\mathcal{E}_{\exists(b)}$	–	–	–	$\mathcal{E}_{\forall(b)}$	0.1861	+ 0.22	2
0.1857 BM25FP	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.1701	- 8.40	2
td2004 BM25F	$\mathcal{E}_{\exists(b)}$	0.1173	+ 0.34	3	$\mathcal{E}_{\forall(b)}$	0.1166	- 0.26	3
0.1169 BM25FA	$\mathcal{E}_{\exists(at)}$	0.1119	- 4.30	1	$\mathcal{E}_{\forall(at)}$	0.1170	+ 0.09	1
hp2003 BM25FU	$\mathcal{E}_{\exists(b)}$	0.7516	0.00	1	$\mathcal{E}_{\forall(b)}$	0.7481	- 0.47	1
0.7516 BM25FP	$\mathcal{E}_{\exists(at)}$	0.7602	+ 1.14	2	$\mathcal{E}_{\forall(at)}$	0.7516	0.00	2
hp2004 BM25FU	$\mathcal{E}_{\exists(b)}$	0.6681	+ 3.12	1	$\mathcal{E}_{\forall(b)}$	0.6635	+ 2.41	1
0.6479 BM25FP	$\mathcal{E}_{\exists(at)}$	0.6653	+ 2.69	2	$\mathcal{E}_{\forall(at)}$	0.6823	+ 5.31$^{\dagger *}$	3
np2003 BM25F	$\mathcal{E}_{\exists(b)}$	0.7031	- 1.10	1	$\mathcal{E}_{\forall(b)}$	0.7182	+ 1.04	1
0.7108 BM25FP	$\mathcal{E}_{\exists(at)}$	0.7068	- 0.56	1	$\mathcal{E}_{\forall(at)}$	–	–	–
np2004 BM25F	$\mathcal{E}_{\exists(b)}$	0.6658	- 0.73†	2	$\mathcal{E}_{\forall(b)}$	–	–	–
0.6707 BM25FU	$\mathcal{E}_{\exists(at)}$	0.6794	+ 1.30†	2	$\mathcal{E}_{\forall(at)}$	0.6698	- 0.13†	2
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b)}$	0.1483	+ 1.92†	1	$\mathcal{E}_{\forall(b)}$	0.1476	+ 1.44	2
0.1455 DLHFP	$\mathcal{E}_{\exists(at)}$	0.1319	- 9.35	1	$\mathcal{E}_{\forall(at)}$	0.1568	+ 7.77$^{\dagger *}$	1

continued on next page

continued from previous page								
Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2004 PL2F	$\mathcal{E}_{\exists(b)}$	0.1313	+ 0.46	2	$\mathcal{E}_{\forall(b)}$	0.1402	+ 7.27	2
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at)}$	0.1330	+ 1.76	2	$\mathcal{E}_{\forall(at)}$	0.1322	+ 1.15	1
hp2003 DLHFU	$\mathcal{E}_{\exists(b)}$	0.6849	+ 2.84†	3	$\mathcal{E}_{\forall(b)}$	0.6942	+ 4.23$^{\dagger *}$	1
0.6660 BM25FA	$\mathcal{E}_{\exists(at)}$	0.6809	+ 2.24	3	$\mathcal{E}_{\forall(at)}$	0.6803	+ 2.15	1
hp2004 PB2FU	$\mathcal{E}_{\exists(b)}$	0.6202	+11.65†	1	$\mathcal{E}_{\forall(b)}$	0.5635	+ 1.44	1
0.5555 DLHFA	$\mathcal{E}_{\exists(at)}$	0.5935	+ 6.84	1	$\mathcal{E}_{\forall(at)}$	0.5871	+ 5.69	2
np2003 PL2FP	$\mathcal{E}_{\exists(b)}$	0.7007	+ 2.35	1	$\mathcal{E}_{\forall(b)}$	0.6940	+ 1.37	1
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at)}$	–	–	–	$\mathcal{E}_{\forall(at)}$	0.7091	+ 3.58†	1
np2004 PB2F	$\mathcal{E}_{\exists(b)}$	0.7220	+ 3.97	2	$\mathcal{E}_{\forall(b)}$	0.7341	+ 5.72	1
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at)}$	0.7154	+ 3.02	1	$\mathcal{E}_{\forall(at)}$	0.7150	+ 2.97	1

Table B.1: Evaluation of score-independent document-level experiments $\mathcal{E}_{\exists(f)}$ and $\mathcal{E}_{\forall(f)}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1522	- 5.20†	1	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
0.1606 PL2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1614	+ 0.50†	1	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1599	- 0.44†	1
td2004 PL2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1355	+ 4.31	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1326	+ 2.08†	1
0.1299 PL2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1316	+ 1.31	1	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1334	+ 2.69	1
hp2003 PL2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.7435	0.00†	1	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7443	+ 0.11†	1	$\mathcal{E}_{\forall(at),avg(dom)}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
0.6674 PL2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6713	+ 0.58†	1	$\mathcal{E}_{\forall(at),avg(dom)}$	–	–	–
np2003 PL2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6681	- 0.48	3	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6704	- 0.13	3
0.6713 PL2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6703	- 0.15	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6664	- 0.73	1
np2004 PL2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.7555	+ 5.38$^{\dagger *}$	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7160	- 0.13	2
0.7169 PL2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7370	+ 2.80†	3	$\mathcal{E}_{\forall(at),avg(dom)}$	0.7289	+ 1.67	1
td2003 PB2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1402	- 1.06	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1419	+ 0.14	1
0.1417 PB2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1399	- 1.27	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1422	+ 0.35	1
td2004 PB2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1427	+ 1.64	1	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
0.1404 PB2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1416	+ 0.85	1
hp2003 PB2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6649	+ 0.91†	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6708	+ 1.81†	2
0.6589 PB2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6667	+ 1.18†	1
hp2004 PB2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dom)}$	0.5704	+ 0.48†	2
0.5677 PB2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dom)}$	0.5908	+ 4.07†	1
np2003 PB2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6650	+ 0.24†	3	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6686	+ 0.78	2
0.6634 PB2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6686	+ 0.78	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6694	+ 0.90†	1
np2004 PB2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.7090	- 2.09	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7246	+ 0.07	2
0.7241 PB2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7247	+ 0.08	3	$\mathcal{E}_{\forall(at),avg(dom)}$	0.7227	- 0.19	2
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1349	+ 5.14†	2
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dom)}$	–	–	–
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1270	- 2.83†	1
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.7320	- 0.31	3	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7423	+ 1.09	1	$\mathcal{E}_{\forall(at),avg(dom)}$	0.7355	+ 0.16	1

continued on next page

continued from previous page								
Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6919	+ 4.33†	1
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6811	+ 2.70†	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6807	+ 2.64†	4
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6993	+ 0.76	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6920	- 0.29	1
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7086	+ 2.10	1	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6962	+ 0.32	1
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6846	+ 0.04	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6755	- 1.23	1
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7004	+ 2.35	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.7019	+ 2.57	1
td2003 DLHF	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1485	+ 2.06	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1461	+ 0.41	2
0.1455 DLHFP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1464	+ 0.62	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1468	+ 0.89	2
td2004 DLHF	$\mathcal{E}_{\exists(b),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
0.1371 DLHFP	$\mathcal{E}_{\exists(at),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dom)}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6721	+ 0.16	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6710	0.00	3
0.6710 DLHFP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6736	+ 0.39†	3	$\mathcal{E}_{\forall(at),avg(dom)}$	–	–	–
hp2004 DLHFU	$\mathcal{E}_{\exists(b),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6209	- 1.10†	1
0.6278 DLHFP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6271	- 0.11	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6524	+ 3.92†	1
np2003 DLHFP	$\mathcal{E}_{\exists(b),avg(dom)}$	0.5242	+ 0.02†	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.5310	+ 1.32	1
0.5241 DLHFA	$\mathcal{E}_{\exists(at),avg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dom)}$	0.5301	+ 1.14†	2
np2004 DLHFU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.4973	- 0.10	1	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
0.4978 DLHFP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.4905	- 1.47	4	$\mathcal{E}_{\forall(at),avg(dom)}$	0.5076	+ 1.97†	2
td2003 BM25FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1889	+ 1.72	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1920	+ 3.39	5
0.1857 BM25FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1795	- 3.34	1	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1696	- 8.67	2
td2004 BM25F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1148	- 1.80†	4	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1203	+ 2.91	1
0.1169 BM25FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1193	+ 2.05†	4	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1156	- 1.11†	3
hp2003 BM25FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.7523	+ 0.09	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7476	- 0.53	2
0.7516 BM25FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7607	+ 1.21†	4	$\mathcal{E}_{\forall(at),avg(dom)}$	0.7516	0.00	1
hp2004 BM25FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6681	+ 3.12	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6717	+ 3.67	1
0.6479 BM25FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6815	+ 5.19$^{\dagger *}$	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6737	+ 3.98	1
np2003 BM25F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.7090	- 0.25	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7148	+ 0.56	1
0.7108 BM25FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7073	- 0.49	1	$\mathcal{E}_{\forall(at),avg(dom)}$	0.7154	+ 0.65	1
np2004 BM25F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6691	- 0.24	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6733	+ 0.39†	2
0.6707 BM25FU	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6769	+ 0.92	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6563	- 2.15	2
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1482	+ 1.86†	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1429	- 1.79	3
0.1455 DLHFP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1464	+ 0.62†	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.1573	+ 8.11†	2
td2004 PL2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.1347	+ 3.06	3	$\mathcal{E}_{\forall(b),avg(dom)}$	0.1386	+ 6.04†	3
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),avg(dom)}$	0.1316	+ 0.69	1	$\mathcal{E}_{\forall(at),avg(dom)}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6732	+ 1.08	3	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6593	- 1.01	2
0.6660 BM25FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6895	+ 3.53†	4	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6592	- 1.02	4
hp2004 PB2FU	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6202	+11.65†	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.6054	+ 8.98	2
0.5555 DLHFA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6215	+11.88$^{\dagger *}$	2	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6279	+13.03$^{\dagger *}$	1
np2003 PL2FP	$\mathcal{E}_{\exists(b),avg(dom)}$	0.6929	+ 1.21	1	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7031	+ 2.70	2
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.6943	+ 1.42	1	$\mathcal{E}_{\forall(at),avg(dom)}$	0.6972	+ 1.84	1
np2004 PB2F	$\mathcal{E}_{\exists(b),avg(dom)}$	0.7187	+ 3.50	2	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7005	+ 0.88	2
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dom)}$	0.7298	+ 5.10	3	$\mathcal{E}_{\forall(at),avg(dom)}$	0.7040	+ 1.38	3

Table B.2: Evaluation of score-independent domain aggregate-level experiments $\mathcal{E}_{\exists(f),avg(dom)}$ and $\mathcal{E}_{\forall(f),avg(dom)}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.1523	- 5.17†	1	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.1606 PL2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.1515	- 5.67†	1	$\mathcal{E}_{\forall(at),std(dom)}$	0.1606	0.00†	1
td2004 PL2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.1325	+ 2.00	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.1302	+ 0.23	1
0.1299 PL2FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.1324	+ 1.92	4	$\mathcal{E}_{\forall(at),std(dom)}$	0.1299	0.00	2
hp2003 PL2FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.7435	0.00†	1	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.7463	+ 0.38†	1	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.6674 PL2FP	$\mathcal{E}_{\exists(at),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dom)}$	0.6827	+ 2.29	1
np2003 PL2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.6638	- 1.12	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.6620	- 1.39	3
0.6713 PL2FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.6678	- 0.52	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.6725	+ 0.18†	1
np2004 PL2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.7348	+ 2.50	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.6998	- 2.39	2
0.7169 PL2FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.7341	+ 2.40	3	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–
td2003 PB2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.1404	- 0.92†	3	$\mathcal{E}_{\forall(b),std(dom)}$	0.1395	- 1.55	3
0.1417 PB2FA	$\mathcal{E}_{\exists(at),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dom)}$	0.1440	+ 1.62	2
td2004 PB2FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.1395	- 0.64	1	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.1404 PB2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.1455	+ 3.63†	1	$\mathcal{E}_{\forall(at),std(dom)}$	0.1402	- 0.14	2
hp2003 PB2FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.6635	+ 0.70	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.6706	+ 1.78†	3
0.6589 PB2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.6622	+ 0.50	2	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–
hp2004 PB2FU	$\mathcal{E}_{\exists(b),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dom)}$	0.5762	+ 1.50†	2
0.5677 PB2FP	$\mathcal{E}_{\exists(at),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–
np2003 PB2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.6597	- 0.56	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.6682	+ 0.72	1
0.6634 PB2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.6716	+ 1.24$^{\dagger *}$	4	$\mathcal{E}_{\forall(at),std(dom)}$	0.6722	+ 1.33$^{\dagger *}$	3
np2004 PB2FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.7076	- 2.28	1	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.7241 PB2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.7353	+ 1.55†	4	$\mathcal{E}_{\forall(at),std(dom)}$	0.73	+ 0.81	1
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dom)}$	0.1283	0.00†	2
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.1333	+ 1.99†	1	$\mathcal{E}_{\forall(at),std(dom)}$	0.1305	- 0.15†	2
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.7336	- 0.10	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.7343	0.00	2
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.7482	+ 1.89†	3	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dom)}$	0.6876	+ 3.68†	1
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.6914	+ 4.25†	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.7053	+ 6.35$^{\dagger *}$	3
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.6955	+ 0.22	4	$\mathcal{E}_{\forall(b),std(dom)}$	0.6921	- 0.27	1
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.7028	+ 1.27	1	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.6839	- 0.06	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.7155	+ 4.56†	2
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.6814	- 0.42	2	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–
td2003 DLHF	$\mathcal{E}_{\exists(b),std(dom)}$	0.1495	+ 2.75	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.1467	+ 0.82	1
0.1455 DLHFP	$\mathcal{E}_{\exists(at),std(dom)}$	0.1463	+ 0.55	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.1416	- 2.68	1
td2004 DLHF	$\mathcal{E}_{\exists(b),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.1371 DLHFP	$\mathcal{E}_{\exists(at),std(dom)}$	0.1396	+ 1.82†	1	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),std(dom)}$	0.6718	+ 0.12	2	$\mathcal{E}_{\forall(b),std(dom)}$	–	–	–
0.6710 DLHFP	$\mathcal{E}_{\exists(at),std(dom)}$	0.6744	+ 0.51	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.6709	- 0.02	1
hp2004 DLHFU	$\mathcal{E}_{\exists(b),std(dom)}$	0.6255	- 0.37	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.6342	+ 1.02†	1
0.6278 DLHFP	$\mathcal{E}_{\exists(at),std(dom)}$	0.6358	+ 1.27	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.6471	+ 3.07†	7
np2003 DLHFP	$\mathcal{E}_{\exists(b),std(dom)}$	0.5242	+ 0.02†	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.54	+ 3.03†	1
0.5241 DLHFA	$\mathcal{E}_{\exists(at),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dom)}$	0.5408	+ 3.19$^{\dagger *}$	1
np2004 DLHFU	$\mathcal{E}_{\exists(b),std(dom)}$	0.4963	- 0.30	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.4932	- 0.92	3
0.4978 DLHFP	$\mathcal{E}_{\exists(at),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dom)}$	0.5118	+ 2.81	3

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 BM25FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.1977	+ 6.46	3	$\mathcal{E}_{\forall(b),std(dom)}$	0.1902	+ 2.42	3
0.1857 BM25FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.1965	+ 5.82†	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.1808	- 2.64	1
td2004 BM25F	$\mathcal{E}_{\exists(b),std(dom)}$	0.1148	- 1.80†	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.1165	- 0.34	1
0.1169 BM25FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.1193	+ 2.05	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.1148	- 1.80	2
hp2003 BM25FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.7499	- 0.23	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.7532	+ 0.21	4
0.7516 BM25FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.7627	+ 1.48†	4	$\mathcal{E}_{\forall(at),std(dom)}$	0.7529	+ 0.17	1
hp2004 BM25FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.6607	+ 1.98	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.6623	+ 2.22	1
0.6479 BM25FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.6539	+ 0.93	1	$\mathcal{E}_{\forall(at),std(dom)}$	0.6822	+ 5.29$^{\dagger *}$	3
np2003 BM25F	$\mathcal{E}_{\exists(b),std(dom)}$	0.7050	- 0.82	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.7137	+ 0.41	1
0.7108 BM25FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.7108	0.00	3	$\mathcal{E}_{\forall(at),std(dom)}$	0.7023	- 1.20	1
np2004 BM25F	$\mathcal{E}_{\exists(b),std(dom)}$	0.6680	- 0.40†	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.6709	+ 0.03	2
0.6707 BM25FU	$\mathcal{E}_{\exists(at),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dom)}$	0.6459	- 3.70	1
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.1404	- 3.51	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.1525	+ 4.81	1
0.1455 DLHFP	$\mathcal{E}_{\exists(at),std(dom)}$	0.1426	- 2.00†	1	$\mathcal{E}_{\forall(at),std(dom)}$	0.1517	+ 4.26	1
td2004 PL2F	$\mathcal{E}_{\exists(b),std(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dom)}$	0.1353	+ 3.52	1
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),std(dom)}$	0.1347	+ 3.06	1	$\mathcal{E}_{\forall(at),std(dom)}$	0.1357	+ 3.83	2
hp2003 DLHFU	$\mathcal{E}_{\exists(b),std(dom)}$	0.6815	+ 2.33	3	$\mathcal{E}_{\forall(b),std(dom)}$	0.6682	+ 0.33	1
0.6660 BM25FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.6746	+ 1.29	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.6707	+ 0.71	1
hp2004 PB2FU	$\mathcal{E}_{\exists(b),std(dom)}$	0.6342	+14.17$^{\dagger *}$	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.5622	+ 1.21	1
0.5555 DLHFA	$\mathcal{E}_{\exists(at),std(dom)}$	0.5871	+ 5.69	2	$\mathcal{E}_{\forall(at),std(dom)}$	0.5869	+ 5.65	2
np2003 PL2FP	$\mathcal{E}_{\exists(b),std(dom)}$	0.7052	+ 3.01†	1	$\mathcal{E}_{\forall(b),std(dom)}$	0.7230	+ 5.61	1
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.7055	+ 3.05†	1	$\mathcal{E}_{\forall(at),std(dom)}$	0.7131	+ 4.16†	1
np2004 PB2F	$\mathcal{E}_{\exists(b),std(dom)}$	0.7054	+ 1.58	2	$\mathcal{E}_{\forall(b),std(dom)}$	0.7184	+ 3.46	1
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dom)}$	0.7088	+ 2.07	1	$\mathcal{E}_{\forall(at),std(dom)}$	–	–	–

Table B.3: Evaluation of score-independent domain aggregate-level experiments $\mathcal{E}_{\exists(f),std(dom)}$ and $\mathcal{E}_{\forall(f),std(dom)}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1608	+ 0.12†	1
0.1606 PL2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1609	+ 0.19†	1
td2004 PL2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1359	+ 4.62†	1	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1386	+ 6.70$^{\dagger *}$	3
0.1299 PL2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.1335	+ 2.77	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1318	+ 1.46	2
hp2003 PL2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7444	+ 0.12†	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6609	- 0.97	1	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6524	- 2.25	1
0.6674 PL2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6752	+ 1.17	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2003 PL2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6702	- 0.16	1	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6712	- 0.02	1
0.6713 PL2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6674	- 0.58	5	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2004 PL2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.7563	+ 5.50	3	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.7252	+ 1.16	1
0.7169 PL2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7310	+ 1.97	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
td2003 PB2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1412	- 0.35†	3	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1385	- 2.26	2
0.1417 PB2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.1399	- 1.27	3	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1393	- 1.69	2
td2004 PB2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1471	+ 4.77$^{\dagger *}$	1	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1403	- 0.07	2
0.1404 PB2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.1482	+ 5.56$^{\dagger *}$	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1469	+ 4.63$^{\dagger *}$	3

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
hp2003 PB2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6649	+ 0.91†	1	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6580	- 0.14	1
0.6589 PB2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6615	+ 0.39	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.6677	+ 1.34†	1
hp2004 PB2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.5677 PB2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2003 PB2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6705	+ 1.07†	1	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6714	+ 1.21	2
0.6634 PB2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6608	- 0.39	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2004 PB2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.7172	- 0.95	2
0.7241 PB2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7094	- 2.03	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1347	+ 4.99†	1
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1331	+ 3.74†	1
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1314	+ 0.54	1
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.7246	- 1.32	1
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7371	+ 0.38	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.74	+ 0.78	1
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6790	+ 2.38†	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.6632	0.00	2
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.7091	+ 2.18	2	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.7049	+ 1.57	1
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7049	+ 1.57	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6930	+ 1.27†	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
td2003 DLHF	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1454	- 0.07	3	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1405	- 3.44	1
0.1455 DLHFP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.1471	+ 1.10	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1457	+ 0.14†	2
td2004 DLHF	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.1371 DLHFP	$\mathcal{E}_{\exists(at),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6784	+ 1.10	3	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6771	+ 0.91	4
0.6710 DLHFP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6781	+ 1.06	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.6656	- 0.80	3
hp2004 DLHFU	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6483	+ 3.27	2
0.6278 DLHFP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6292	+ 0.22	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.6149	- 2.05	2
np2003 DLHFP	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.5301	+ 1.14†	3	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.5241 DLHFA	$\mathcal{E}_{\exists(at),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2004 DLHFU	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.4978 DLHFP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.4961	- 0.34†	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
td2003 BM25FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1937	+ 4.31†	3	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1990	+ 7.16†	1
0.1857 BM25FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.1759	- 5.28	3	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1976	+ 6.41	1
td2004 BM25F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1173	+ 0.34	2	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1172	+ 0.26	3
0.1169 BM25FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.1178	+ 0.77†	4	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1155	- 1.20	2
hp2003 BM25FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.7474	- 0.56	2	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.7492	- 0.32	4
0.7516 BM25FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7693	+ 2.35$^{\dagger *}$	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.7478	- 0.51	1
hp2004 BM25FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6656	+ 2.73	2	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.6479 BM25FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6852	+ 5.76$^{\dagger *}$	3	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.6469	- 0.15	2
np2003 BM25F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.7370	+ 3.69†	3	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.7117	+ 0.13	1
0.7108 BM25FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7074	- 0.48	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2004 BM25F	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dom)}$	–	–	–
0.6707 BM25FU	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6687	- 0.30†	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1463	+ 0.55	4	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1534	+ 5.43	2
0.1455 DLHFP	$\mathcal{E}_{\exists(at),lrg(dom)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1536	+ 5.57	1

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2004 PL2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.1399	+ 7.04†	2	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.1378	+ 5.43	5
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.1353	+ 3.52	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.1377	+ 5.36	7
hp2003 DLHFU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6719	+ 0.89	2	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6881	+ 3.32$^{\dagger *}$	1
0.6660 BM25FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.6758	+ 1.47	4	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.6855	+ 2.93*	3
hp2004 PB2FU	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6064	+ 9.16†	1	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.5483	- 1.30	1
0.5555 DLHFA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.5550	- 0.09	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2003 PL2FP	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.6895	+ 0.72	1	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6880	+ 0.50	2
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7028	+ 2.66	2	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
np2004 PB2F	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.7125	+ 2.61	2	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.6959	+ 0.22	2
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.7352	+ 5.88†	1	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–

Table B.4: Evaluation of score-independent domain aggregate-level experiments $\mathcal{E}_{\exists(f),lrg(dom)}$ and $\mathcal{E}_{\forall(f),lrg(dom)}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1515	- 5.67†	1	$\mathcal{E}_{\forall(b),avg(dir)}$	–	–	–
0.1606 PL2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1616	+ 0.62†	1	$\mathcal{E}_{\forall(at),avg(dir)}$	–	–	–
td2004 PL2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1357	+ 4.46	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.1337	+ 2.93†	1
0.1299 PL2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1303	+ 0.31	3	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1331	+ 2.46	1
hp2003 PL2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dir)}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7488	+ 0.71†	3	$\mathcal{E}_{\forall(at),avg(dir)}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6674	0.00	1	$\mathcal{E}_{\forall(b),avg(dir)}$	–	–	–
0.6674 PL2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7060	+ 5.78	1	$\mathcal{E}_{\forall(at),avg(dir)}$	–	–	–
np2003 PL2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6713	0.00	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6678	- 0.52	3
0.6713 PL2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6774	+ 0.91†	3	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6745	+ 0.48	2
np2004 PL2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.7510	+ 4.76†	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.72	+ 0.43	2
0.7169 PL2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7265	+ 1.34†	1	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7234	+ 0.91	1
td2003 PB2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1399	- 1.27	3	$\mathcal{E}_{\forall(b),avg(dir)}$	0.1417	0.00	1
0.1417 PB2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1421	+ 0.28†	3	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1443	+ 1.83	1
td2004 PB2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1421	+ 1.21†	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.1390	- 1.00	1
0.1404 PB2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1402	- 0.14	2	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1417	+ 0.93†	2
hp2003 PB2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6648	+ 0.90†	3	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6708	+ 1.81†	2
0.6589 PB2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6698	+ 1.65†	3	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6587	- 0.03	1
hp2004 PB2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.5807	+ 2.29†	1	$\mathcal{E}_{\forall(b),avg(dir)}$	–	–	–
0.5677 PB2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dir)}$	0.5973	+ 5.21$^{\dagger *}$	2
np2003 PB2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6576	- 0.87	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6664	+ 0.45	2
0.6634 PB2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6628	- 0.09	3	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6670	+ 0.54†	1
np2004 PB2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.7018	- 3.08	2	$\mathcal{E}_{\forall(b),avg(dir)}$	–	–	–
0.7241 PB2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7146	- 1.31	1	$\mathcal{E}_{\forall(at),avg(dir)}$	–	–	–
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),avg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dir)}$	–	–	–
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dir)}$	–	–	–
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1327	+ 1.53†	1	$\mathcal{E}_{\forall(b),avg(dir)}$	–	–	–
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dir)}$	–	–	–
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.7305	- 0.52	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.7322	- 0.29	2
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7434	+ 1.24	4	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7283	- 0.82	2

continued on next page

continued from previous page								
Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6632	0.00	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6758	+ 1.90†	3
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6774	+ 2.14	2	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6945	+ 4.72†	2
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6918	- 0.32	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.7001	+ 0.88	3
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6955	+ 0.22	2	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7007	+ 0.97	5
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),avg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6980	+ 2.00	2
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6828	- 0.22	1	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6756	- 1.27	1
td2003 DLHF	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1486	+ 2.13	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.1521	+ 4.54	2
0.1455 DLHFP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1378	- 5.29	1	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1476	+ 1.44	2
td2004 DLHF	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1355	- 1.17†	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.1330	- 2.99†	2
0.1371 DLHFP	$\mathcal{E}_{\exists(at),avg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),avg(dir)}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6775	+ 0.97	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6778	+ 1.01	2
0.6710 DLHFP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6740	+ 0.45	1	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6672	- 0.57	1
hp2004 DLHFU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6158	- 1.91	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6249	- 0.46†	3
0.6278 DLHFP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6547	+ 4.28$^{\dagger *}$	2	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6326	+ 0.76†	2
np2003 DLHFP	$\mathcal{E}_{\exists(b),avg(dir)}$	0.5260	+ 0.36†	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.5250	+ 0.17†	2
0.5241 DLHFA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.5262	+ 0.40†	4	$\mathcal{E}_{\forall(at),avg(dir)}$	0.5360	+ 2.27†	5
np2004 DLHFU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.4947	- 0.62	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.5071	+ 1.87†	2
0.4978 DLHFP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.5011	+ 0.66	1	$\mathcal{E}_{\forall(at),avg(dir)}$	0.5006	+ 0.56	4
td2003 BM25FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1934	+ 4.15	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.1768	- 4.79	3
0.1857 BM25FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1873	+ 0.86†	1	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1765	- 4.95	2
td2004 BM25F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1145	- 2.05	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.1192	+ 1.97	2
0.1169 BM25FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1172	+ 0.26†	4	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1156	- 1.11	3
hp2003 BM25FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.7653	+ 1.82	4	$\mathcal{E}_{\forall(b),avg(dir)}$	0.7507	- 0.12	3
0.7516 BM25FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7498	- 0.24	3	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7622	+ 1.41	2
hp2004 BM25FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6494	+ 0.23	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6544	+ 1.00	2
0.6479 BM25FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6951	+ 7.29$^{\dagger *}$	2	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6852	+ 5.76†	10
np2003 BM25F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.7093	- 0.21	1	$\mathcal{E}_{\forall(b),avg(dir)}$	0.7182	+ 1.04	1
0.7108 BM25FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7076	- 0.45	1	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7070	- 0.53	3
np2004 BM25F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6606	- 1.51	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6538	- 2.52†	3
0.6707 BM25FU	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6702	- 0.08†	2	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6275	- 6.44	1
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1483	+ 1.92†	1	$\mathcal{E}_{\forall(b),avg(dir)}$	–	–	–
0.1455 DLHFP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1497	+ 2.89†	1	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1613	+10.86	2
td2004 PL2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.1336	+ 2.22	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.1387	+ 6.12†	2
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),avg(dir)}$	0.1411	+ 7.96†	3	$\mathcal{E}_{\forall(at),avg(dir)}$	0.1338	+ 2.37	1
hp2003 DLHFU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6742	+ 1.23	4	$\mathcal{E}_{\forall(b),avg(dir)}$	0.7021	+ 5.42	4
0.6660 BM25FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.6855	+ 2.93†	4	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6836	+ 2.64	7
hp2004 PB2FU	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6440	+15.93$^{\dagger *}$	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.6164	+10.96	2
0.5555 DLHFA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.5903	+ 6.26	4	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6279	+13.03$^{\dagger *}$	3
np2003 PL2FP	$\mathcal{E}_{\exists(b),avg(dir)}$	0.7045	+ 2.91	4	$\mathcal{E}_{\forall(b),avg(dir)}$	0.7195	+ 5.10†	2
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.71	+ 3.71†	3	$\mathcal{E}_{\forall(at),avg(dir)}$	0.6861	+ 0.22	5
np2004 PB2F	$\mathcal{E}_{\exists(b),avg(dir)}$	0.6975	+ 0.45	2	$\mathcal{E}_{\forall(b),avg(dir)}$	0.7295	+ 5.05	2
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),avg(dir)}$	0.7216	+ 3.92	4	$\mathcal{E}_{\forall(at),avg(dir)}$	0.7027	+ 1.20	1

Table B.5: Evaluation of score-independent directory aggregate-level experiments $\mathcal{E}_{\exists(f),avg(dir)}$ and $\mathcal{E}_{\forall(f),avg(dir)}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.1599	- 0.44†	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.1693	+ 5.42†	2
0.1606 PL2FP	$\mathcal{E}_{\exists(at),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dir)}$	0.1603	- 0.19†	1
td2004 PL2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.1344	+ 3.46†	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.1295	- 0.31	2
0.1299 PL2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.1335	+ 2.77	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.1316	+ 1.31†	2
hp2003 PL2FU	$\mathcal{E}_{\exists(b),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dir)}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.7463	+ 0.38†	1	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.6747	+ 1.09	1	$\mathcal{E}_{\forall(b),std(dir)}$	0.6769	+ 1.42	1
0.6674 PL2FP	$\mathcal{E}_{\exists(at),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
np2003 PL2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.6684	- 0.43	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.6694	- 0.28	4
0.6713 PL2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.6662	- 0.76	1	$\mathcal{E}_{\forall(at),std(dir)}$	0.6757	+ 0.66†	5
np2004 PL2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.7447	+ 3.88†	3	$\mathcal{E}_{\forall(b),std(dir)}$	0.7388	+ 3.05	2
0.7169 PL2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.7548	+ 5.29†	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.7195	+ 0.36	1
td2003 PB2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.1418	+ 0.07	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.1410	- 0.49	1
0.1417 PB2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.1403	- 0.99	3	$\mathcal{E}_{\forall(at),std(dir)}$	0.1448	+ 2.19$^{\dagger *}$	2
td2004 PB2FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.1428	+ 1.71†	1	$\mathcal{E}_{\forall(b),std(dir)}$	0.1453	+ 3.49$^{\dagger *}$	4
0.1404 PB2FP	$\mathcal{E}_{\exists(at),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dir)}$	0.1426	+ 1.57†	1
hp2003 PB2FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.6633	+ 0.67†	3	$\mathcal{E}_{\forall(b),std(dir)}$	0.6720	+ 1.99†	1
0.6589 PB2FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.6599	+ 0.15†	2	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
hp2004 PB2FU	$\mathcal{E}_{\exists(b),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dir)}$	–	–	–
0.5677 PB2FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.5835	+ 2.78†	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.5956	+ 4.91†	1
np2003 PB2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.6575	- 0.89	1	$\mathcal{E}_{\forall(b),std(dir)}$	0.6621	- 0.20	2
0.6634 PB2FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.6636	+ 0.03	1	$\mathcal{E}_{\forall(at),std(dir)}$	0.6658	+ 0.36†	3
np2004 PB2FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.7385	+ 1.99†	4	$\mathcal{E}_{\forall(b),std(dir)}$	0.7173	- 0.94	1
0.7241 PB2FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.7287	+ 0.64	1	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.1323	+ 3.12†	1	$\mathcal{E}_{\forall(b),std(dir)}$	0.1220	- 4.91†	2
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dir)}$	0.1327	+ 3.43†	1
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dir)}$	0.1319	+ 0.92†	3
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dir)}$	0.1319	+ 0.92†	1
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.7302	- 0.56	1	$\mathcal{E}_{\forall(b),std(dir)}$	–	–	–
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.7303	- 0.54	3	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dir)}$	0.6822	+ 2.86†	2
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),std(dir)}$	0.6883	+ 3.78	1
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.6997	+ 0.82†	1	$\mathcal{E}_{\forall(b),std(dir)}$	0.6904	- 0.52	2
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.6929	- 0.16	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.6951	+ 0.16	1
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.6941	+ 1.43†	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.6894	+ 0.75	1
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.6957	+ 1.67	1	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
td2003 DLHF	$\mathcal{E}_{\exists(b),std(dir)}$	0.15	+ 3.09	4	$\mathcal{E}_{\forall(b),std(dir)}$	0.1515	+ 4.12†	1
0.1455 DLHFP	$\mathcal{E}_{\exists(at),std(dir)}$	0.1418	- 2.54	3	$\mathcal{E}_{\forall(at),std(dir)}$	0.1483	+ 1.92	2
td2004 DLHF	$\mathcal{E}_{\exists(b),std(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),std(dir)}$	–	–	–
0.1371 DLHFP	$\mathcal{E}_{\exists(at),std(dir)}$	0.1364	- 0.51†	1	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),std(dir)}$	0.6727	+ 0.25	4	$\mathcal{E}_{\forall(b),std(dir)}$	0.6785	+ 1.12	2
0.6710 DLHFP	$\mathcal{E}_{\exists(at),std(dir)}$	0.6758	+ 0.72	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.6769	+ 0.88	1
hp2004 DLHFU	$\mathcal{E}_{\exists(b),std(dir)}$	0.6155	- 1.96	1	$\mathcal{E}_{\forall(b),std(dir)}$	0.6274	- 0.06	3
0.6278 DLHFP	$\mathcal{E}_{\exists(at),std(dir)}$	0.6290	+ 0.19	3	$\mathcal{E}_{\forall(at),std(dir)}$	0.6169	- 1.74	2
np2003 DLHFP	$\mathcal{E}_{\exists(b),std(dir)}$	0.5241	0.00	1	$\mathcal{E}_{\forall(b),std(dir)}$	0.5333	+ 1.76†	1
0.5241 DLHFA	$\mathcal{E}_{\exists(at),std(dir)}$	0.5348	+ 2.04†	1	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
np2004 DLHFU	$\mathcal{E}_{\exists(b),std(dir)}$	0.4986	+ 0.16†	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.5041	+ 1.27†	2
0.4978 DLHFP	$\mathcal{E}_{\exists(at),std(dir)}$	0.5017	+ 0.78	1	$\mathcal{E}_{\forall(at),std(dir)}$	0.5004	+ 0.52	4

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 BM25FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.1790	- 3.61†	4	$\mathcal{E}_{\forall(b),std(dir)}$	0.1612	-13.19	2
0.1857 BM25FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.1538	-17.18	3	$\mathcal{E}_{\forall(at),std(dir)}$	0.1822	- 1.88	1
td2004 BM25F	$\mathcal{E}_{\exists(b),std(dir)}$	0.1196	+ 2.31	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.1146	- 1.97	1
0.1169 BM25FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.1134	- 2.99	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.1175	+ 0.51	2
hp2003 BM25FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.7645	+ 1.72	3	$\mathcal{E}_{\forall(b),std(dir)}$	0.7432	- 1.12	2
0.7516 BM25FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.7516	0.00	3	$\mathcal{E}_{\forall(at),std(dir)}$	0.7624	+ 1.44†	2
hp2004 BM25FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.6606	+ 1.96	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.6565	+ 1.33	2
0.6479 BM25FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.6605	+ 1.94	1	$\mathcal{E}_{\forall(at),std(dir)}$	0.6690	+ 3.26†	6
np2003 BM25F	$\mathcal{E}_{\exists(b),std(dir)}$	0.7189	+ 1.14†	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.7119	+ 0.15	3
0.7108 BM25FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.7117	+ 0.13	2	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
np2004 BM25F	$\mathcal{E}_{\exists(b),std(dir)}$	0.6626	- 1.21	3	$\mathcal{E}_{\forall(b),std(dir)}$	0.6554	- 2.28	2
0.6707 BM25FU	$\mathcal{E}_{\exists(at),std(dir)}$	0.6792	+ 1.27†	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.6460	- 3.68	1
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.1422	- 2.27	1	$\mathcal{E}_{\forall(b),std(dir)}$	0.1565	+ 7.56	2
0.1455 DLHFP	$\mathcal{E}_{\exists(at),std(dir)}$	0.1284	-12.00	1	$\mathcal{E}_{\forall(at),std(dir)}$	0.1511	+ 3.85	2
td2004 PL2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.1318	+ 0.84	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.1359	+ 3.98	3
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),std(dir)}$	0.1321	+ 1.07	1	$\mathcal{E}_{\forall(at),std(dir)}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),std(dir)}$	0.6699	+ 0.59†	3	$\mathcal{E}_{\forall(b),std(dir)}$	0.6710	+ 0.75	1
0.6660 BM25FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.6849	+ 2.84	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.6936	+ 4.14*	2
hp2004 PB2FU	$\mathcal{E}_{\exists(b),std(dir)}$	0.6040	+ 8.73	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.5517	- 0.68	1
0.5555 DLHFA	$\mathcal{E}_{\exists(at),std(dir)}$	0.5898	+ 6.17	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.5940	+ 6.93†	1
np2003 PL2FP	$\mathcal{E}_{\exists(b),std(dir)}$	0.6934	+ 1.29	2	$\mathcal{E}_{\forall(b),std(dir)}$	0.7070	+ 3.27	1
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.7111	+ 3.87†	2	$\mathcal{E}_{\forall(at),std(dir)}$	0.7128	+ 4.12	1
np2004 PB2F	$\mathcal{E}_{\exists(b),std(dir)}$	0.7104	+ 2.30	4	$\mathcal{E}_{\forall(b),std(dir)}$	0.7261	+ 4.57	1
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),std(dir)}$	0.7040	+ 1.38	1	$\mathcal{E}_{\forall(at),std(dir)}$	0.6827	- 1.68	1

Table B.6: Evaluation of score-independent directory aggregate-level experiments $\mathcal{E}_{\exists(f),std(dir)}$ and $\mathcal{E}_{\forall(f),std(dir)}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1529	- 4.79†	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.1606 PL2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1520	- 5.35†	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.1619	+ 0.81†	1
td2004 PL2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1324	+ 1.92	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.1327	+ 2.16†	3
0.1299 PL2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1370	+ 5.47†	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
hp2003 PL2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7489	+ 0.73†	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.7460	+ 0.34†	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6672	- 0.03	2
0.6674 PL2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.7062	+ 5.81†	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2003 PL2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6685	- 0.42	1
0.6713 PL2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6703	- 0.15	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2004 PL2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7276	+ 1.49	3	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.7169	0.00	2
0.7169 PL2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.7163	- 0.08	2	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
td2003 PB2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1411	- 0.42	4	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.1386	- 2.19	1
0.1417 PB2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1402	- 1.06	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.1381	- 2.54	1
td2004 PB2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1450	+ 3.28	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.1409	+ 0.36	2
0.1404 PB2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.1431	+ 1.92	2

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
hp2003 PB2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6621	+ 0.49	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6651	+ 0.94†	1
0.6589 PB2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6636	+ 0.71†	2	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.6648	+ 0.90†	1
hp2004 PB2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.5833	+ 2.75	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.5677 PB2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2003 PB2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6657	+ 0.35	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6604	- 0.45	2
0.6634 PB2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6579	- 0.83	2	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2004 PB2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7188	- 0.73	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.7241 PB2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.7092	- 2.06	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.1259	- 1.87†	1
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7343	0.00	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.7398	+ 0.75	1
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.7422	+ 1.08$^{\dagger *}$	3	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.7359	+ 0.22	1
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6712	+ 1.21†	2	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.6632	0.00	2
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7019	+ 1.14	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6934	- 0.09	2
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.7070	+ 1.87	3	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6993	+ 2.19†	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
td2003 DLHF	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1454	- 0.07	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.1420	- 2.41	1
0.1455 DLHFP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1457	+ 0.14	3	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.1453	- 0.14	1
td2004 DLHF	$\mathcal{E}_{\exists(b),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.1371 DLHFP	$\mathcal{E}_{\exists(at),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6729	+ 0.28	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6708	- 0.03	1
0.6710 DLHFP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6796	+ 1.28†	2	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.6713	+ 0.05	1
hp2004 DLHFU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6222	- 0.89	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6086	- 3.06	3
0.6278 DLHFP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6289	+ 0.18	2	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.6228	- 0.80†	3
np2003 DLHFP	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.5241	0.00	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.5286	+ 0.86†	1
0.5241 DLHFA	$\mathcal{E}_{\exists(at),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2004 DLHFU	$\mathcal{E}_{\exists(b),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.4978 DLHFP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.5060	+ 1.65†	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
td2003 BM25FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1810	- 2.53	3	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.2000	+ 7.70	1
0.1857 BM25FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1865	+ 0.43	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.1796	- 3.28	3
td2004 BM25F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1170	+ 0.09	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.1130	- 3.34	2
0.1169 BM25FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1162	- 0.60†	3	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.1164	- 0.43	1
hp2003 BM25FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7459	- 0.76	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.7468	- 0.64	2
0.7516 BM25FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.7701	+ 2.46$^{\dagger *}$	2	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.7574	+ 0.77†	1
hp2004 BM25FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6496	+ 0.26	3	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6561	+ 1.27	1
0.6479 BM25FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6532	+ 0.82	4	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.6655	+ 2.72	1
np2003 BM25F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7197	+ 1.25	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.7124	+ 0.23	1
0.7108 BM25FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.7009	- 1.39	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2004 BM25F	$\mathcal{E}_{\exists(b),lrg(dir)}$	–	–	–	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6393	- 4.68	1
0.6707 BM25FU	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6727	+ 0.30†	4	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1547	+ 6.32†	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.1527	+ 4.95	3
0.1455 DLHFP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.1469	+ 0.96	3	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.1552	+ 6.67$^{\dagger *}$	1

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2004 PL2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.1295	- 0.92	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.1384	+ 5.89	1
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.13	- 0.54	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.1372	+ 4.97	1
hp2003 DLHFU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6712	+ 0.78	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.6994	+ 5.02$^{\dagger *}$	3
0.6660 BM25FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6768	+ 1.62	6	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.6872	+ 3.18	1
hp2004 PB2FU	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6042	+ 8.77	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.5714	+ 2.86	3
0.5555 DLHFA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.5550	- 0.09	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2003 PL2FP	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.6872	+ 0.38	1	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6859	+ 0.19	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–
np2004 PB2F	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.7132	+ 2.71	2	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.6910	- 0.49	1	$\mathcal{E}_{\forall(at),lrg(dir)}$	–	–	–

Table B.7: Evaluation of score-independent directory aggregate-level experiments $\mathcal{E}_{\exists(f),lrg(dir)}$ and $\mathcal{E}_{\forall(f),lrg(dir)}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1512	- 5.85†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1601	- 0.31†	1
0.1606 PL2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1522	- 5.23†	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1607	+ 0.06†	1
td2004 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1370	+ 5.47†	4	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1314	+ 1.15†	2
0.1299 PL2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1352	+ 4.08	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1323	+ 1.85†	2
hp2003 PL2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6787	+ 1.69†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	–	–	–
0.6674 PL2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6680	+ 0.09	3	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6674	0.00	2
np2003 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6727	+ 0.21†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6779	+ 0.98	3
0.6713 PL2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6719	+ 0.09	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6687	- 0.39	1
np2004 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.7440	+ 3.78†	1	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7347	+ 2.48	2
0.7169 PL2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7348	+ 2.50†	3	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7391	+ 3.10	1
td2003 PB2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1402	- 1.06†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1403	- 0.99	2
0.1417 PB2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1430	+ 0.92	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1459	+ 2.96$^{\dagger *}$	4
td2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1444	+ 2.85$^{\dagger *}$	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1431	+ 1.92†	2
0.1404 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1443	+ 2.78†	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1414	+ 0.71†	1
hp2003 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6648	+ 0.90	1	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6644	+ 0.83†	4
0.6589 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6705	+ 1.76†	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6671	+ 1.24†	2
hp2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.5699	+ 0.39†	1	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.5694	+ 0.30†	2
0.5677 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	–	–	–
np2003 PB2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6570	- 0.96	1	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6683	+ 0.74	2
0.6634 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6570	- 0.96	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6675	+ 0.62	1
np2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.7221	- 0.28†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7466	+ 3.11†	5
0.7241 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7298	+ 0.79	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7060	- 2.50	1
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1383	+ 7.79†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	–	–	–
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1348	+ 5.07†	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1335	+ 4.05†	3
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1342	+ 2.68†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1302	- 0.38†	1
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1318	+ 0.84†	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1322	+ 1.15†	1
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.7328	- 0.20	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7457	+ 1.55†	3
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7391	+ 0.65	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7397	+ 0.74	2

continued on next page

continued from previous page								
Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6880	+ 3.74†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	–	–	–
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6897	+ 4.00†	4
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.7158	+ 3.14$^{\dagger *}$	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7060	+ 1.73	2
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6960	+ 0.29†	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6990	+ 0.72	1
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6925	+ 1.20	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7181	+ 4.94$^{\dagger *}$	2
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7223	+ 5.55$^{\dagger *}$	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7236	+ 5.74$^{\dagger *}$	1
td2003 DLHF	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1482	+ 1.86	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1522	+ 4.60†	3
0.1455 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1462	+ 0.48	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1479	+ 1.65	2
td2004 DLHF	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1351	- 1.46†	1
0.1371 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1355	- 1.17†	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6763	+ 0.79	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6795	+ 1.27†	2
0.6710 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6649	- 0.91	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6702	- 0.12	2
hp2004 DLHFU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6555	+ 4.41†	3	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6294	+ 0.25	3
0.6278 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6228	- 0.80	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6317	+ 0.62†	3
np2003 DLHFP	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.5279	+ 0.73†	4	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	–	–	–
0.5241 DLHFA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.5249	+ 0.15†	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.5369	+ 2.44†	1
np2004 DLHFU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.5178	+ 4.02†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.5137	+ 3.19†	4
0.4978 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.5082	+ 2.09	3	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.4956	- 0.44	1
td2003 BM25FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1964	+ 5.76†	3	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1830	- 1.45	2
0.1857 BM25FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1944	+ 4.68	3	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1692	- 8.89†	2
td2004 BM25F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.12	+ 2.65†	4	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.12	+ 2.65	2
0.1169 BM25FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1136	- 2.82	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1173	+ 0.34	2
hp2003 BM25FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.7544	+ 0.37	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7628	+ 1.49†	5
0.7516 BM25FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7624	+ 1.44†	3	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7547	+ 0.41†	2
hp2004 BM25FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6844	+ 5.63*	5	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6612	+ 2.05	2
0.6479 BM25FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6552	+ 1.13	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6898	+ 6.47$^{\dagger *}$	2
np2003 BM25F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.7224	+ 1.63†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7148	+ 0.56†	4
0.7108 BM25FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7126	+ 0.25	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7021	- 1.22	1
np2004 BM25F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6790	+ 1.24†	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6844	+ 2.04†	1
0.6707 BM25FU	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6807	+ 1.49†	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6810	+ 1.54†	2
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1432	- 1.58	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1607	+10.45	3
0.1455 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1484	+ 1.99†	3	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1571	+ 7.97†	4
td2004 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.1433	+ 9.64	3	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.1287	- 1.53	1
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.1337	+ 2.30	3	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.1322	+ 1.15	5
hp2003 DLHFU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6670	+ 0.15	3	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6943	+ 4.25†	5
0.6660 BM25FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6796	+ 2.04	5	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.6589	- 1.07	3
hp2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.5612	+ 1.03	3	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.6132	+10.39	3
0.5555 DLHFA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.5877	+ 5.80	1	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.5707	+ 2.74	1
np2003 PL2FP	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.6899	+ 0.77	1	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7213	+ 5.36†	4
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.6946	+ 1.46	3	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7013	+ 2.44	1
np2004 PB2F	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.7312	+ 5.30	2	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.7373	+ 6.18*	3
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.7382	+ 6.31	2	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.7460	+ 7.43†	1

Table B.8: Evaluation of score-dependent experiments $\mathcal{E}_{\exists(f),L(SU)_{pl}}$ and $\mathcal{E}_{\forall(f),L(SU)_{pl}}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1510	- 5.98†	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1611	+ 0.31†	1
0.1606 PL2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1617	+ 0.68†	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	–	–	–
td2004 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1339	+ 3.08	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1310	+ 0.85†	2
0.1299 PL2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1317	+ 1.39	4	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1315	+ 1.23†	2
hp2003 PL2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6782	+ 1.62†	3	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6796	+ 1.83†	2
0.6674 PL2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6807	+ 1.99†	1
np2003 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6663	- 0.74	3	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6664	- 0.73	5
0.6713 PL2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.67	- 0.19	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6687	- 0.39	1
np2004 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7407	+ 3.32†	1	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7364	+ 2.72	4
0.7169 PL2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7242	+ 1.02	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7376	+ 2.89	1
td2003 PB2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1406	- 0.78	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1412	- 0.35†	1
0.1417 PB2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1426	+ 0.64	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1408	- 0.64	5
td2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1450	+ 3.28$^{\dagger *}$	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1423	+ 1.35†	2
0.1404 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1476	+ 5.13†	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1406	+ 0.14	1
hp2003 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6660	+ 1.08†	1	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	–	–	–
0.6589 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6604	+ 0.23†	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6693	+ 1.58†	2
hp2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.5699	+ 0.39†	1	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.5735	+ 1.02†	1
0.5677 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.5862	+ 3.26†	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	–	–	–
np2003 PB2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6570	- 0.96	1	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6670	+ 0.54†	2
0.6634 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6570	- 0.96	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6574	- 0.90	2
np2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7425	+ 2.54	4	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7461	+ 3.04†	3
0.7241 PB2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7295	+ 0.75	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7226	- 0.21	1
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1427	+11.22†	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	–	–	–
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1353	+ 5.46†	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	–	–	–
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1302	- 0.38†	1
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1318	+ 0.84†	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1322	+ 1.15†	1
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7340	- 0.04	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.74	+ 0.78	1
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7353	+ 0.14	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7358	+ 0.20	2
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6706	+ 1.12	3	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	–	–	–
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6654	+ 0.33	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6986	+ 5.34†	4
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7088	+ 2.13	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7011	+ 1.02	2
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6962	+ 0.32†	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7023	+ 1.20	1
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7123	+ 4.09†	4	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7159	+ 4.62$^{\dagger *}$	2
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7237	+ 5.76$^{\dagger *}$	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7236	+ 5.74$^{\dagger *}$	1
td2003 DLHF	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1451	- 0.27	4	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1529	+ 5.09†	1
0.1455 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1465	+ 0.69	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1471	+ 1.10	1
td2004 DLHF	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1351	- 1.46†	1
0.1371 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6743	+ 0.49	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6752	+ 0.63	3
0.6710 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6689	- 0.31	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6689	- 0.31	1
hp2004 DLHFU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6481	+ 3.23†	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6563	+ 4.54†	4
0.6278 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6254	- 0.38	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6209	- 1.10†	3
np2003 DLHFP	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.5319	+ 1.49†	5	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	–	–	–
0.5241 DLHFA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.5364	+ 2.35†	1
np2004 DLHFU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.5104	+ 2.53†	2
0.4978 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.5086	+ 2.17†	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	–	–	–

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 BM25FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1954	+ 5.22†	3	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1730	- 6.84	1
0.1857 BM25FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1753	- 5.60†	4	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1908	+ 2.75†	4
td2004 BM25F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1193	+ 2.05	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1217	+ 4.11$^{\dagger *}$	2
0.1169 BM25FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1135	- 2.91	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1190	+ 1.80	4
hp2003 BM25FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7515	- 0.01	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7620	+ 1.38†	3
0.7516 BM25FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7636	+ 1.60	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7511	- 0.07	1
hp2004 BM25FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6509	+ 0.46	4	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6563	+ 1.30	2
0.6479 BM25FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6693	+ 3.30†	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6832	+ 5.45$^{\dagger *}$	2
np2003 BM25F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7094	- 0.20	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7192	+ 1.18†	4
0.7108 BM25FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7151	+ 0.60	1	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7184	+ 1.07†	3
np2004 BM25F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6779	+ 1.07†	3	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6844	+ 2.04†	1
0.6707 BM25FU	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6773	+ 0.98†	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6810	+ 1.54†	2
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1404	- 3.51	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1561	+ 7.29	1
0.1455 DLHFP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.15	+ 3.09†	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1445	- 0.69	2
td2004 PL2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.1405	+ 7.50	2	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.1289	- 1.38	1
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.1367	+ 4.59†	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.1350	+ 3.29	3
hp2003 DLHFU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6600	- 0.90	1	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7038	+ 5.68$^{\dagger *}$	5
0.6660 BM25FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.6772	+ 1.68	4	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6608	- 0.78	4
hp2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.5628	+ 1.31	3	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.6038	+ 8.69	4
0.5555 DLHFA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.5933	+ 6.80	2	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.5607	+ 0.94	1
np2003 PL2FP	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.6944	+ 1.43†	3	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7149	+ 4.43†	5
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7017	+ 2.48	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.6904	+ 0.85	3
np2004 PB2F	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.7432	+ 7.03*	4	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.7368	+ 6.11	3
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.7231	+ 4.13	3	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.7468	+ 7.55†	1

Table B.9: Evaluation of score-dependent experiments $\mathcal{E}_{\exists(f),L(SU)_{in}}$ and $\mathcal{E}_{\forall(f),L(SU)_{in}}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.1606 PL2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.16	- 0.37†	1	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1609	+ 0.19†	2
td2004 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.1320	+ 1.62†	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1325	+ 2.00†	2
0.1299 PL2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.1318	+ 1.46	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1319	+ 1.54†	2
hp2003 PL2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.7018	+ 5.15†	1	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6722	+ 0.72†	1
0.6674 PL2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6774	+ 1.50†	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6946	+ 4.08†	1
np2003 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6661	- 0.77	1	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6663	- 0.74	1
0.6713 PL2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6736	+ 0.34†	1	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6688	- 0.37	1
np2004 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.7518	+ 4.87†	4	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7264	+ 1.33	3
0.7169 PL2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.7353	+ 2.57†	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.7202	+ 0.46	1
td2003 PB2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.1456	+ 2.75	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1465	+ 3.39$^{\dagger *}$	3
0.1417 PB2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.1449	+ 2.26	5	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1436	+ 1.34	2
td2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.1456	+ 3.70$^{\dagger *}$	4	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1413	+ 0.64	1
0.1404 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.1475	+ 5.06†	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1385	- 1.35	2
hp2003 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6604	+ 0.23	1	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6649	+ 0.91	2

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
0.6589 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6713	+ 1.88†	1	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6716	+ 1.93†	2
hp2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.5744	+ 1.18†	1	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.5744	+ 1.18†	1
0.5677 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.5824	+ 2.59†	1	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.5895	+ 3.84†	2
np2003 PB2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6616	- 0.27	3	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.6634 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6676	+ 0.63	1
np2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.7086	- 2.14	3	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7308	+ 0.93†	4
0.7241 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.7518	+ 3.83†	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6986	- 3.52	1
td2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.1283 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.1283	0.00†	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	–	–	–
td2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	–	–	–
hp2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.7359	+ 0.22	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7365	+ 0.30	1
0.7343 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.7399	+ 0.76	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.7435	+ 1.25	1
hp2004 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6690	+ 0.87	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.67	+ 1.03†	2
0.6632 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6853	+ 3.33†	4
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.7040	+ 1.44	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7043	+ 1.48†	5
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.7014	+ 1.07†	3	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6978	+ 0.55	1
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6874	+ 0.45	3	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6883	+ 0.58	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.7067	+ 3.27†	1
td2003 DLHF	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.1439	- 1.10	3	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1436	- 1.31	2
0.1455 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.1464	+ 0.62†	3	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1504	+ 3.37	4
td2004 DLHF	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.1371 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6743	+ 0.49	1	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6769	+ 0.88	1
0.6710 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6788	+ 1.16†	4	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6683	- 0.40	2
hp2004 DLHFU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6279	+ 0.02	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.6278 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6166	- 1.78	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6297	+ 0.31†	2
np2003 DLHFP	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.5360	+ 2.27†	5	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.5241	0.00	1
0.5241 DLHFA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.5302	+ 1.16$^{\dagger *}$	1
np2004 DLHFU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.4941	- 0.74	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	–	–	–
0.4978 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.5081	+ 2.07	3	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.4966	- 0.24	1
td2003 BM25FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.1911	+ 2.91	1	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1836	- 1.13	3
0.1857 BM25FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.1925	+ 3.66	3	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1829	- 1.51	3
td2004 BM25F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.1191	+ 1.88	3	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1185	+ 1.37	1
0.1169 BM25FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.1167	- 0.17	1	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1209	+ 3.42$^{\dagger *}$	3
hp2003 BM25FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.7569	+ 0.71	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7540	+ 0.32	3
0.7516 BM25FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.7712	+ 2.61	4	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.7572	+ 0.75†	1
hp2004 BM25FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6696	+ 3.35	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6643	+ 2.53	2
0.6479 BM25FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6718	+ 3.69$^{\dagger *}$	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6831	+ 5.43$^{\dagger *}$	2
np2003 BM25F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.7160	+ 0.73	3	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7127	+ 0.27	3
0.7108 BM25FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.7152	+ 0.62	4	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.7026	- 1.15	1
np2004 BM25F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6851	+ 2.15†	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6718	+ 0.16	2
0.6707 BM25FU	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6745	+ 0.57	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	–	–	–
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1655	+13.75$^{\dagger *}$	3
0.1455 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1470	+ 1.03	1
td2004 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.1377	+ 5.36	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.1343	+ 2.75	3

continued on next page

continued from previous page

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
0.1307 I(n$_e$)C2FP	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.1316	+ 0.69	2	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.1355	+ 3.67	2
hp2003 DLHFU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6544	- 1.74	1	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6829	+ 2.54	1
0.6660 BM25FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.6775	+ 1.73	6	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6806	+ 2.19	4
hp2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.5998	+ 7.97	4	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.5939	+ 6.91*	3
0.5555 DLHFA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.5739	+ 3.31	3	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.5751	+ 3.53	1
np2003 PL2FP	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.6935	+ 1.30	2	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.6914	+ 0.99	2
0.6846 I(n$_e$)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.7266	+ 6.13$^{\dagger *}$	3	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.6925	+ 1.15	2
np2004 PB2F	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.7407	+ 6.67	4	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.7173	+ 3.30	2
0.6944 I(n$_e$)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.7280	+ 4.84	1	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.7155	+ 3.04	1

Table B.10: Evaluation of score-dependent experiments $\mathcal{E}_{\exists(f),L(SU')_{pl}}$ and $\mathcal{E}_{\forall(f),L(SU')_{pl}}$.

Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
td2003 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	–	–	–
0.1606 PL2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1599	- 0.44†	2
td2004 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.1333	+ 2.62†	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1329	+ 2.31†	2
0.1299 PL2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1348	+ 3.77†	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1290	- 0.69†	2
hp2003 PL2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	–	–	–
0.7435 PL2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	–	–	–
hp2004 PL2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6807	+ 1.99†	1	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6722	+ 0.72†	1
0.6674 PL2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6774	+ 1.50†	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	–	–	–
np2003 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6640	- 1.09	1
0.6713 PL2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6711	- 0.03	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6688	- 0.37	1
np2004 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.7472	+ 4.23	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7283	+ 1.59	2
0.7169 PL2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7477	+ 4.30†	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.7180	+ 0.15	1
td2003 PB2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.1453	+ 2.54	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1452	+ 2.47	3
0.1417 PB2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1456	+ 2.75	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1387	- 2.12	3
td2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.1436	+ 2.28†	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1406	+ 0.14	1
0.1404 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1480	+ 5.41†	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1426	+ 1.57	2
hp2003 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6606	+ 0.26	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	–	–	–
0.6589 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6604	+ 0.23†	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6718	+ 1.96†	2
hp2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.5744	+ 1.18†	1	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.5743	+ 1.16†	1
0.5677 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.5831	+ 2.71†	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.5884	+ 3.65†	2
np2003 PB2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6649	+ 0.23	3	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6576	- 0.87	2
0.6634 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6569	- 0.98	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6642	+ 0.12	1
np2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.7112	- 1.78	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7255	+ 0.19†	2
0.7241 PB2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7264	+ 0.32	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.7220	- 0.29	1
td2003 I(n$_e$)C2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	–	–	–
0.1283 I(n$_e$)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	–	–	–
td2004 I(n$_e$)C2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1313	+ 0.46†	1
0.1307 I(n$_e$)C2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1336	+ 2.22†	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	–	–	–
hp2003 I(n$_e$)C2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.7304	- 0.53	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7221	- 1.66	3
0.7343 I(n$_e$)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7371	+ 0.38	3	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.7406	+ 0.86	2
hp2004 I(n$_e$)C2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6593	- 0.59	2
0.6632 I(n$_e$)C2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6853	+ 3.33†	2

continued on next page

continued from previous page								
Setting	$\mathcal{E}$	MAP	+/-%	B	$\mathcal{E}$	MAP	+/-%	B
np2003 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.7073	+ 1.92	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6908	- 0.46	2
0.6940 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7124	+ 2.65$^{\dagger *}$	5	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6994	+ 0.78	1
np2004 I(n_e)C2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6842	- 0.02	1	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	–	–	–
0.6843 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7030	+ 2.73	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.7067	+ 3.27†	1
td2003 DLHF	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.1432	- 1.58	3	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1434	- 1.44	2
0.1455 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1455	0.00	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1537	+ 5.64*	3
td2004 DLHF	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.1373	+ 0.15†	1	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	–	–	–
0.1371 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1393	+ 1.60†	3	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	–	–	–
hp2003 DLHFU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6723	+ 0.19	3	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6807	+ 1.45	1
0.6710 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6719	+ 0.13	4	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6706	- 0.06†	2
hp2004 DLHFU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6282	+ 0.06†	5	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	–	–	–
0.6278 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6468	+ 3.03	3	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6213	- 1.04†	2
np2003 DLHFP	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.5314	+ 1.39†	3	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.5244	+ 0.06	1
0.5241 DLHFA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.5239	- 0.04	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.5301	+ 1.14$^{\dagger *}$	1
np2004 DLHFU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.5017	+ 0.78†	1	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	–	–	–
0.4978 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.5099	+ 2.43	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	–	–	–
td2003 BM25FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.1873	+ 0.86	1	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1809	- 2.58	3
0.1857 BM25FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1783	- 3.98	3	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1861	+ 0.22	3
td2004 BM25F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.1196	+ 2.31†	4	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1217	+ 4.11†	2
0.1169 BM25FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1185	+ 1.37†	4	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1223	+ 4.62$^{\dagger *}$	3
hp2003 BM25FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.7518	+ 0.03	4	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7551	+ 0.47	3
0.7516 BM25FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7758	+ 3.22†	4	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.7529	+ 0.17	2
hp2004 BM25FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6668	+ 2.92	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6577	+ 1.51	1
0.6479 BM25FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6746	+ 4.12†	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6899	+ 6.48$^{\dagger *}$	2
np2003 BM25F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.7081	- 0.38	3	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7142	+ 0.48	3
0.7108 BM25FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7216	+ 1.52†	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6989	- 1.67	1
np2004 BM25F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6897	+ 2.83†	4	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6720	+ 0.19	2
0.6707 BM25FU	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6825	+ 1.76†	2	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	–	–	–
td2003 I(n_e)C2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	–	–	–	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1570	+ 7.90†	3
0.1455 DLHFP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1437	- 1.24	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1503	+ 3.30	1
td2004 PL2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.1343	+ 2.75	3	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.1327	+ 1.53	2
0.1307 I(n_e)C2FP	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.1357	+ 3.83	3	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.1376	+ 5.28	4
hp2003 DLHFU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6590	- 1.05	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6816	+ 2.34	1
0.6660 BM25FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.6807	+ 2.21	4	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6719	+ 0.89	2
hp2004 PB2FU	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.5962	+ 7.33†	4	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.6001	+ 8.03	3
0.5555 DLHFA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.5949	+ 7.09†	5	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.5640	+ 1.53	1
np2003 PL2FP	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6943	+ 1.42	2	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7059	+ 3.11	4
0.6846 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7131	+ 4.16	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.7009	+ 2.38	2
np2004 PB2F	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.6893	- 0.73	3	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.7217	+ 3.93	2
0.6944 I(n_e)C2FA	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.7269	+ 4.68	1	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.6914	- 0.43	1

Table B.11: Evaluation of score-dependent experiments $\mathcal{E}_{\exists(f),L(SU')_{in}}$ and $\mathcal{E}_{\forall(f),L(SU')_{in}}$.

Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	B
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b)}$	0.5455	-1.41	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b)}$	0.5557	+0.43	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at)}$	0.5486	-0.85	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at)}$	0.5775	$+4.37^{\dagger *}$	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at)}$	0.4381	$+5.41^{\dagger}$	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),avg(dom)}$	0.5490	-0.78	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),avg(dom)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),avg(dom)}$	0.7590	$+4.64^{\dagger *}$	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),avg(dom)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),avg(dom)}$	0.5537	+0.07	2
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),avg(dom)}$	0.4152	-0.09	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),avg(dom)}$	0.5626	+1.68	2
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),avg(dom)}$	0.4452	$+7.12^{\dagger}$	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),std(dom)}$	0.5474	-1.07	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),std(dom)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),std(dom)}$	0.5777	$+4.41^{\dagger *}$	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),std(dom)}$	0.4212	+1.34	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),std(dom)}$	0.5554	+0.38	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),std(dom)}$	0.4265	+2.62	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),std(dom)}$	0.5622	+1.61	2
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),std(dom)}$	0.4233	+1.85	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),lrg(dom)}$	0.5561	+0.51	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),lrg(dom)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.5398	-2.44	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),lrg(dom)}$	0.4013	-3.44^{*}	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),lrg(dom)}$	0.5455	-1.41	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),lrg(dom)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),lrg(dom)}$	0.5541	+0.14	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),lrg(dom)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),avg(dir)}$	0.5512	-0.38	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),avg(dir)}$	0.4115	-0.99	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),avg(dir)}$	0.5745	$+3.84^{\dagger *}$	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),avg(dir)}$	0.4147	-0.22	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),avg(dir))}$	0.5499	-0.61	2
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),avg(dir))}$	0.4108	-1.15	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),avg(dir)}$	0.5626	$+1.68^{\dagger}$	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),avg(dir)}$	0.4395	$+5.75^{\dagger}$	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),std(dir)}$	0.5371	-2.93	4
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),std(dir)}$	0.4190	+0.82	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),std(dir)}$	0.5685	+2.75	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),std(dir)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),std(dir)}$	0.5619	+1.55	4
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),std(dir)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),std(dir)}$	0.5648	$+2.08^{\dagger *}$	2
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),std(dir)}$	0.4374	+5.25	1

continued on next page

continued from previous page

Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	B
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.5513	-0.36	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),lrg(dir)}$	0.4235	+1.96†	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),lrg(dir)}$	0.5554	+0.38	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),lrg(dir)}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.5516	-0.31	4
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),lrg(dir)}$	0.4053	-2.48	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.5613	+1.45†	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),lrg(dir)}$	0.4421	+6.37†	1

Table B.12: Evaluation of the score-independent document-level and aggregate-level experiments with limited relevance information. The table displays the evaluation results of a decision mechanism, which is trained and evaluated with different mixed tasks.

Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	B
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.5412	-2.19	4
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),L(SU)_{pl}}$	0.4144	-0.29	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.5539	+0.11	4
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),L(SU)_{pl}}$	0.4215	+1.42†	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.5462	-1.28	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),L(SU)_{pl}}$	0.4218	+1.49	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.5685	+2.75†	2
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),L(SU)_{pl}}$	0.4215	+1.42	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.5481	-0.94	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),L(SU)_{in}}$	0.4038	-2.84	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.5702	+3.05	4
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),L(SU)_{in}}$	0.4207	+1.23†	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.5331	-3.65	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),L(SU)_{in}}$	0.4243	+2.09	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.5743	+3.80$^{\dagger *}$	2
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),L(SU)_{in}}$	0.4264	+2.60	2
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.5505	-0.51	4
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),L(SU')_{pl}}$	0.4154	-0.04	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.5698	+2.98†	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),L(SU')_{pl}}$	0.4201	+1.08†	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.5471	-1.12	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),L(SU')_{pl}}$	0.4178	+0.53	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.5663	+2.35†	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),L(SU')_{pl}}$	0.4156	0.00	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.5285	-4.48	2
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(b),L(SU')_{in}}$	0.4154	-0.05	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.5742	+3.78†	3
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(b),L(SU')_{in}}$	0.4194	+0.91	1
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	0.5464	-1.25	3

continued on next page

continued from previous page							
Task	Retrieval approaches		Baseline	$\mathcal{E}$	MAP	+/- %	B
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\exists(at),L(SU')_{in}}$	–	–	–
mq2003	DLHFP	BM25F	0.5533	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.5733	$+3.61^{\dagger *}$	1
mq2004	DLHFP	PB2F	0.4156	$\mathcal{E}_{\forall(at),L(SU')_{in}}$	0.4157	+0.02	1

Table B.13: Evaluation of the score-dependent experiments with limited relevance information. The table displays the evaluation results of a decision mechanism, which is trained and evaluated with different mixed tasks.

Bibliography

Achlioptas, D., Fiat, A., Karlin, A. & McSherry, F. (2001). Web search via hub synthesis. *in* 'Proceedings of the 42nd Annual Symposium of Foundations of Computer Science'. pp. 611–618. 3.4.1

Adamic, L. (2001). 'Network Dynamics: The World Wide Web. PhD Thesis, Stanford University'. 3.2.2

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics* **74(1)**, 47–97. 3.2.2

Albert, R., Jeong, H. & Barabási, A.-L. (1999). Diameter of the World Wide Web. *Nature* **401**, 130–131. 3.2.2

Allan, J. (1996). Automatic hypertext link typing. *in* 'Proceedings of the 7th ACM conference on Hypertext'. ACM Press. pp. 42–52. 3.2.1

Amati, G. (2003). Probability Models for Information Retrieval based on Divergence from Randomness. PhD thesis. Department of Computing Science, University of Glasgow. UK. 2.3.3, 2.3.3.3, 7.4.3.1

Amati, G. (2006). Frequentist and Bayesian approach to Information Retrieval. *in* 'Proceedings of the 28th European Conference on Information Retrieval (ECIR'06). To appear'. 2.3.3.4

Amati, G. & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems* **20**(4), 357–389. 2.3.3, 2.3.3.4, 4.4.1, 7.4.2.2

Amati, G., Carpineto, C. & Romano, G. (2004). Query Difficulty, Robustness, and Selective Application of Query Expansion. *in* 'Proceedings of the 26th European Conference in Information Retrieval (ECIR'04)'. pp. 127–137. 3.6.2

Amati, G., Ounis, I. & Plachouras, V. (2003). The dynamic absorbing model for the web. Technical Report TR-2003-137. Department of Computing Science, University of Glasgow. 3.4.4.3, 4.5.2.4

Amento, B., Terveen, L. & Hill, W. (2000). Does authority mean quality? predicting expert quality ratings of web documents. *in* 'Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 296–303. 3.3.2.1

Amitay, E., Carmel, D., Darlow, A., Herscovici, M., Lempel, R., Soffer, A., Kraft, R. & Zien, J. (2003). Juru at TREC 2003 - Topic Distillation Using Query-Sensitive Tuning and Cohesiveness Filtering. *in* 'NIST Special Publication 500-255: The Twelfth Text REtrieval Conference (TREC 2003)'. pp. 255–262. 3.6.2, 5.3.1

Amitay, E., Carmel, D., Darlow, A., Lempel, R. & Soffer, A. (2002). Topic Distillation with Knowledge Agents. *in* 'NIST Special Publication 500-251: The Eleventh Text Retrieval Conference (TREC 2002)'. 3.6.2, 5.2.1

Aslam, J. A. & Montague, M. (2001). Models for metasearch. *in* 'Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 276–284. 3.4.4.1

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Addison Wesley. 2.1

Bailey, P., Craswell, N. & Hawking, D. (2003). Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing & Management* **39**(6), 853–871. 4.2

Bar-Yossef, Z., Broder, A., Kumar, R. & Tomkins, A. (2004). Sic transit gloria telae: towards an understanding of the web's decay. *in* 'Proceedings of the 13th international conference on World Wide Web (WWW13)'. ACM Press. pp. 328–337. 3.3.2

Barabási, A.-L. (2002). *Linked: The New Science of Networks.* Perseus Publishing. 3.2.2

Barabási, A. L. & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science* **286**, 509–512. 3.2.2

Baron, L. (1996). Labeled, Typed links as Cues when Reading Hypertext Documents. *Journal of the American Association for Information Science* **47**(12), 896–908. 3.2.1

Bartell, B. T., Cottrell, G. W. & Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. *in* 'Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. pp. 173–181. 3.4.4.1

Beitzel, S., Jensen, E., Chowdhury, A. & Grossman, D. (2003). Using titles and category names from editor-driven taxonomies for automatic evaluation. *in* 'Proceedings of the 12th international Conference on Information and Knowledge Management (CIKM)'. ACM Press. pp. 17–23. 3.6.1

Beitzel, S., Jensen, E., Chowdhury, A., Grossman, D. & Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. *in* 'Proceedings of the 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM Press. pp. 321–328. 7.2.1

Belew, R. (2000). *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW.* Cambridge University Press, Cambridge, UK. 2.3, 3.2.1

Berners-Lee, T., Masinter, L. & McCahill, M. (1994). RFC 1738 - Uniform Resource Locators (URL). RFC 1738. IETF. 4.5.1

Bernstein, Y. & Zobel, J. (2004). A scalable system for identifying co-derivative documents. *in* 'The Eleventh Symposium on String Processing and Information Retrieval (SPIRE'04)'. 3.2.3

Bharat, K. & Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. *in* 'Proceedings of the 7th international conference on World Wide Web (WWW7)'. Elsevier Science Publishers B. V.. pp. 379–388. 3.2.2

Bharat, K. & Broder, A. (1999). Mirror, mirror on the Web: A study of Host Pairs with Replicated Content. *in* 'Proceedings of the 8th international conference on World Wide Web (WWW8)'. 3.2.3

Bharat, K. & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. *in* 'Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 104–111. 3.3.2.2, 3.4.1, 5.1

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press. 5.5.3

Blair, D. (2001). Some thoughts on the reported results of trec. *Information Processing & Management* **38(3)**, 445–451. 2.4

Bomhoff, M., Huibers, T. & Van der Vet, P. (2005). User Intentions in Information Retrieval. *in* 'Proceedings of the 5th Dutch Information Retrieval Workshop (DIR'5)'. pp. 47–54. 3.6.1, 7.2.1

Borodin, A., Roberts, G., Rosenthal, J. & Tsaparas, P. (2001). Finding authorities and hubs from link structures on the world wide web. *in* 'Proceedings of the 10th international conference on World Wide Web (WWW10)'. ACM Press. pp. 415–429. 3.3.2.2

Botafogo, R. A. & Shneiderman, B. (1991). Identifying aggregates in hypertext structures. *in* 'Proceedings of the third annual ACM conference on Hypertext'. ACM Press. pp. 63–74. 3.3.1

Botafogo, R., Rivlin, E. & Shneiderman, B. (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems* **10**(2), 142–180. 3.3.2

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* **30**(1–7), 107–117. 3.2.2, 3.3.1, 3.3.2.1, 3.4.2, 4.5.3, 4.8, 6.2.2.1

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum* **36**(2), 3–10. 1.2, 3.2.4, 3.5.1, 3.6.1

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph Structure in the Web. *in* 'Proceedings of the 9th international conference on World Wide Web (WWW9)'. 3.2.2

Bruza, P. (1990). Hyperindices: A novel aid for searching in hypermedia. *in* 'Proceedings of the ACM European Conference on Hypertext '90 (ECHT '90)'. pp. 109–122. 3.2.1

Bush, V. (1945). As we may think. *The Atlantic Monthly* **176(1)**, 101–108. 3.2.1

Callan, J. & Connell, M. (2001). Query-based sampling of text databases. *ACM Transactions on Information Systems* **19**(2), 97–130. 7.4.2.1

Can, F., Nuray, R. & Sevdik, A. B. (2004). Automatic performance evaluation of web search engines. *Information Processing & Management* **40**(3), 495–514. 3.5.2

Carmel, D., Cohen, D., Fagin, R., Farchi, E., Herscovici, M., Maarek, Y. & Soffer, A. (2001). Static index pruning for information retrieval systems. *in* 'Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM Press. pp. 43–50. 4.2

Carriere, J. & Kazman, R. (1997). WebQuery: Searching and visualizing the Web through connectivity. *in* 'Proceedings of the 6th internaltional conference on World Wide Web (WWW6)'. 3.3.2, 3.3.2.2

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P. & Rajagopalan, S. (1998). Automatic resource list compilation by analyzing hyperlink structure and associated text. *in* 'Proceedings of the 7th international conference on World Wide Web (WWW7)'. 3.4.1

Chakrabarti, S., Joshi, M. & Tawde, V. (2001). Enhanced topic distillation using text, markup tags, and hyperlinks. *in* 'Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 208–216. 3.4.1

Chambers, J., Cleveland, W. & Tukey, P. (1983). *Graphical methods for data analysis.* Duxbury Press. 5.5.3

Chowdhury, A. & Soboroff, I. (2002). Automatic evaluation of world wide web search services. *in* 'Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 421–422. 3.5.2

Clarke, C., Craswell, N. & Soboroff, I. (2004). Overview of the TREC 2004 Terabyte track. *in* 'NIST Special Publication 500-261: The 13th Text REtrieval Conference (TREC 2004)'. pp. 80–88. 2.3.3.4, 3.5.1

Clarke, C., Scholer, F. & Soboroff, I. (2005). The TREC 2005 Terabyte Track. *in* 'Proceedings of the 14th Text REtrieval Conference (TREC 2005)'. 3.5.1

Cleverdon, C. (1997). The cranfield tests on index language devices. *in* 'Readings in information retrieval'. Morgan Kaufmann Publishers Inc.. pp. 47–59. 2.4

Cohn, D. & Chang, H. (2000). Learning to probabilistically identify authoritative documents. *in* 'Proceedings of the 17th International Conference on Machine Learning'. ACM Press. 3.3.2.2

Cohn, D. & Hofmann, T. (2001). The missing link - a probabilistic model of document content and hypertext connectivity. *in* T. K. Leen, T. G. Dietterich & V. Tresp, eds, 'Advances in Neural Information Processing Systems 13'. MIT Press. pp. 430–436. 3.4.1

Craswell, N. & Hawking, D. (2002). Overview of the TREC-2002 Web Track. *in* 'NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002)'. pp. 86–93. 3.5.1, 4.2, 4.3.1, 4.5.3

Craswell, N. & Hawking, D. (2004). Overview of the TREC 2004 Web Track. *in* 'NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference (TREC 2004)'. 1.1, 2.4, 3.5.1, 3.6.1, 4.2, 4.3.1, 4.4.4, 4.5.1, 7.2.1, 1, 7.4, 7.4.3.1, 8.1.2

Craswell, N., de Vries, A. & Soboroff, I. (2005). Overview of the TREC 2005 Enterprise Track. *in* 'Proceedings of the 14th Text REtrieval Conference (TREC 2005)'. 3.5.1

Craswell, N., Hawking, D. & Robertson, S. (2001). Effective site finding using link anchor information. *in* 'Proceedings of the 24th annual international ACM SIGIR

conference on Research and Development in Information Retrieval'. ACM Press. pp. 250–257. 3.4.2, 4.3.1

Craswell, N., Hawking, D., Wilkinson, R. & Wu, M. (2003). Overview of the TREC-2003 Web Track. *in* 'NIST Special Publication 500-255: The Twelfth Text REtrieval Conference (TREC 2003)'. pp. 220–236. 1.1, 3.5.1, 4.3.1, 7.4, 7.4.3.1, 8.1.2

Craswell, N., Robertson, S., Zaragoza, H. & Taylor, M. (2005). Relevance weighting for query independent evidence. *in* 'Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 416–423. 3.4.4.1, 3.4.4.3, 4.5.2.6

Croft, W. B. (2000). Combining approaches to information retrieval. *in* W. B. Croft, ed., 'Advances in Information Retrieval from the Center for Intelligent Information Retrieval'. Kluwer Academic. pp. 1–36. 1.2, 3.4

Croft, W. B. & Harper, D. (1988). Using probabilistic models of information retrieval without relevance information. *Journal of Documentation* **35**, 285–295. 2.3.1

Cronen-Townsend, S., Zhou, Y. & Croft, W. B. (2002). Predicting query performance. *in* 'Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 299–306. 3.6.2, 5.4, 7.4.2.1

Diligenti, M., Gori, M. & Maggini, M. (2002). Web page scoring systems for horizontal and vertical search. *in* 'Proceedings of the 11th international conference on World Wide Web (WWW11)'. ACM Press. pp. 508–516. 3.3.2.1

Duda, R. & Hart, P. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, USA. 5.5.1

Eiron, N. & McCurley, K. (2003*a*). Untangling compound documents on the web. *in* 'Proceedings of the fourteenth ACM conference on Hypertext and Hypermedia'. ACM Press. pp. 85–94. 3.3.1

Eiron, N. & McCurley, K. S. (2003*b*). Analysis of anchor text for web search. *in* 'Proceedings of the 26th annual international ACM SIGIR conference on Research and

Development in Information Retrieval'. ACM Press. pp. 459–460. 3.4.2, 4.3.1, 7.4.2.3, 7.4.3.3, 8.1.2

Erdös, P. & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae* **6**, 290–297. 3.2.2

Evans, D. A., Shanahan, J. G. & Sheftel, V. (2002). Topic structure modeling. *in* 'Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 417–418. 3.6.2

Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A. & Williamson, D. P. (2003). Searching the workplace web. *in* 'Proceedings of the 12th international conference on World Wide Web (WWW12)'. ACM Press. pp. 366–375. 3.4.4.1, 4.4.1

Faloutsos, M., Faloutsos, P. & Faloutsos, C. (n.d.). On Power-law Relationships of the Internet Topology. *Computer Communications Review* **29**, 251–262. 3.2.2

Fang, H., Tao, T. & Zhai, C. (2004). A formal study of information retrieval heuristics. *in* 'Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 49–56. 2.3.1

Feller, W. (1957). *An Introduction to Probability Theory and its Applications, volume 1, 2nd edition.* John Wiley and Sons. 4.5.2, 4.5.2.1, 4.5.2.2

Fisher, M. & Everson, R. (2003). When Are Links Useful? Experiments in Text Classification. *in* 'Proceedings of the 25th European Conference on Information Retrieval (ECIR'03)'. Springer-Verlag. pp. 41–56. 3.6.2

Frakes, W. & Baeza-Yates, R. (1992). *Information Retrieval Data Structures & Algorithms.* Prentice Hall, Englewood Cliffs, New Jersey. 2.2

Frei, H.-P. & Stieger, S. (1995). The Use of Semantic Links in Hypertext Information Retrieval. *Information Processing & Management* **31**, 1–13. 3.4.3

Gao, J., Cao, G., He, H., Zhang, M., Nie, J.-Y., Walker, S. & Robertson, S. (2001). TREC-10 Web Track Experiments at MSRA. *in* 'NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001)'. pp. 384–392. 4.4.1

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science* **178**, 471–479. 1.2, 3.3.2

Geller, N. (1977). On the Citation Influence Methodology of Pinski and Narin. *Information Processing & Management* **14**, 93–95. 3.3.2, 3.3.2.1

Gordon, M. & Pathak, P. (1999). Finding information on the world wide web: the retrieval effectiveness of search engines. *Information Processing & Management* **35**(2), 141–180. 3.5.2

Guinan, C. & Smeaton, A. F. (1992). Information retrieval from hypertext using dynamically planned guided tours. *in* 'Proceedings of the ACM conference on Hypertext (ECHT'92)'. ACM Press. pp. 122–130. 3.2.1

Gulli, A. & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. *in* 'Special interest tracks and posters of the 14th international conference on World Wide Web (WWW14)'. ACM Press. pp. 902–903. 3.2.2

Gurrin, C. & Smeaton, A. F. (2003). Improving the Evaluation of Web Search Systems. *in* 'Proceedings of the 25th European Conference on Information Retrieval (ECIR'03)'. Springer-Verlag. pp. 25–40. 3.6.2

Halasz, F. G. (1987). Reflections on notecards: seven issues for the next generation of hypermedia systems. *in* 'HYPERTEXT '87: Proceeding of the ACM conference on Hypertext'. ACM Press. pp. 345–365. 3.2.1

Harman, D. (1993). Overview of the First Text REtrieval Conference (TREC-1). *in* 'NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1)'. pp. 1–20. 2.4, 2.5

Harter, S. P. (1975). An algorithm for probabilistic indexing. *Journal of the American Society for Information Science* **26(4)**, 280–289. 2.3

Hauff, C. & Azzopardi, L. (2005). Age Dependent Document Priors in Link Structure Analysis. *in* 'Proceedings of the 27th European Conference on Information Retrieval (ECIR'05)'. pp. 552–554. 3.4.4.3

Haveliwala, T. H. (2002). Topic-sensitive pagerank. *in* 'Proceedings of the 11th international conference on World Wide Web (WWW11)'. ACM Press. pp. 517–526. 3.4.1

Hawking, D. (2000). Overview of the TREC-9 Web Track. *in* 'NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC 9)'. pp. 87–102. 3.5.1, 4.2

Hawking, D. & Craswell, N. (2001). Overview of the TREC-2001 Web Track. *in* 'NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001)'. pp. 61–67. 3.5.1, 3.6.1, 4.2, 4.3.1

Hawking, D. & Craswell, N. (2005). The Very Large Collection and Web Tracks. *in* 'TREC: Experiment and Evaluation in Information Retrieval, editors E.M. Voorhees and D.K. Harman'. MIT Press. pp. 199–232. 2.4, 3.5.1, 4.2

Hawking, D. & Thistlewaite, P. (1997). Overview of TREC-6 Very Large Collection Track. *in* 'NIST Special Publication 500-240: The 6th Text REtrieval Conference (TREC6)'. pp. 93–106. 3.5.1

Hawking, D., Craswell, N. & Thistlewaite, P. (1998*a*). ACSys TREC-7 Experiments. *in* 'NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7)'. pp. 299–313. 4.2

Hawking, D., Craswell, N. & Thistlewaite, P. (1998*b*). Overview of trec-7 very large collection track. *in* 'NIST Special Publication 500-242: The 7th Text REtrieval Conference (TREC7)'. pp. 91–104. 3.5.1

Hawking, D., Craswell, N., Bailey, P. & Griffihs, K. (2001). Measuring search engine quality. *Informantion Retrieval* **4**(1), 33–59. 3.5.2

Hawking, D., Craswell, N., Crimmins, F. & Upstill, T. (2004). How valuable is external link evidence when searching enterprise webs?. *in* 'Proceedings of the 15th Australasian Database Conference (ADC'04)'. 3.4.2

Hawking, D., Upstill, T. & Craswell, N. (2004). Toward better weighting of anchors. *in* 'Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 512–513. 3.4.2, 4.3.2, 4.4.1

Hawking, D., Voorhees, E., Craswell, N. & Bailey, P. (1999). Overview of the TREC-8 Web Track. *in* 'NIST Special Publication 500-246: The 8th Text REtrieval Conference (TREC8)'. pp. 131–150. 3.5.1, 4.2

He, B. & Ounis, I. (2003). A study of parameter tuning for term frequency normalization. *in* 'Proceedings of the 12th international Conference on Information and Knowledge Management (CIKM)'. ACM Press. pp. 10–16. 2.3.3.3

He, B. & Ounis, I. (2004). Inferring Query Performance Using Pre-retrieval Predictors. *in* 'The Eleventh Symposium on String Processing and Information Retrieval (SPIRE'04)'. 1.2, 3.6.2, 5.2

He, B. & Ounis, I. (2005*a*). A study of the dirichlet priors for term frequency normalisation. *in* 'Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 465–471. 2.3.3.3

He, B. & Ounis, I. (2005*b*). Term Frequency Normalisation Tunning for BM25 and DFR models. *in* 'Proceedings of the 27th European Conference on Information Retrieval (ECIR'05)'. pp. 200–214. 2.3.3.3, 7.4.2.2

Heydon, A. & Najork, M. (1999). Mercator: A scalable, extensible web crawler. *World Wide Web* **2**(4), 219–229. 3.2.2

Hiemstra, D. (1998). A linguistically motivated probabilistic model of information retrieval. *in* 'ECDL '98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries'. Springer-Verlag. London, UK. pp. 569–584. 2.3.2

Hoel, P. G. (1984). *Introduction to Mathematical Statistics, 5th Ed.*. Wiley. 6.2.1

Hofmann, T. (1999). Probabilistic latent semantic indexing. *in* 'Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 50–57. 3.3.2.2

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science* **44**, 161–174. 3.2.4

Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E. & Lukose, R. M. (1998). Strong regularities in World Wide Web surfing. *Science* **280**, 95–97. 3.3.3

Jansen, B. J. & Pooch, U. W. (2001). A review of web searching studies and a framework for future research. *Journal of the American Society of Information Science* **52**(3), 235–246. 1.2, 3.2.4, 7.4.2.1

Jardine, K. & Sibson, R. (1971). *Mathematical taxonomy.* Wiley. 5.4.1

Jiang, X.-M., Song, W.-G. & Zeng, H.-J. (2005). Applying Associative Relationship on the Clickthrough Data to Improve Web Search. *in* 'Proceedings of the 27th European Conference on Information Retrieval (ECIR'05)'. pp. 475–486. 3.3.3

Jin, R. & Dumais, S. (2001). Probabilistic combination of content and links. *in* 'Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 402–403. 3.4.3

Jin, R., Hauptmann, A. G. & Zhai, C. X. (2002). Title language model for information retrieval. *in* 'Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 42–48. 4.3.1

Joachims, T. (2002). Optimizing search engines using clickthrough data. *in* 'Proceedings of the 8th ACM SIGKDD international conference on Knowledge Discovery and Data Mining'. ACM Press. pp. 133–142. 3.3.3

Joachims, T., Granka, L., Pan, B., Hembrooke, H. & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *in* 'Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 154–161. 3.3.3, 8.2

Kamps, J., Mishne, G. & de Rijke, M. (2004*a*). Language models for searching in web corpora. *in* 'NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference (TREC 2004)'. 3.3.1

Kamps, J., Mishne, G. & de Rijke, M. (2004*b*). Language Models for Searching in Web Corpora. *in* 'NIST Special Publication 500-261: The Thirteenth Text Retrieval conference (TREC 2004)'. 4.5.1

Kamps, J., Monz, C., de Rijke, M. & Sigurbjornsson, B. (2003). Approaches to Robust and Web Retrieval. *in* 'NIST Special Publication 500-255: The Twelfth Text Retrieval Conference (TREC 2003)'. pp. 594–599. 4.4.1

Kang, I.-H. & Kim, G. (2003). Query type classification for web document retrieval. *in* 'Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 64–71. 3.6.1, 5.2.1

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. *in* 'Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms'. ACM Press. pp. 668–677. 3.3.2.2, 3.3.2.2, 5.1, 5.2

Kraaij, W., Westerveld, T. & Hiemstra, D. (2002). The importance of prior probabilities for entry page search. *in* 'Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 27–34. 3.3.1, 3.4.4.2, 3.4.4.3, 4.5.1

Kullback, S. (1959). *Information Theory and Statistics*. Jown Wiley & Sons, New York, USA. 5.4.1, 5.4.1

Kwok, K. L., Deng, P., Dinstl, N. & Chan, M. (2002). TREC2002 Web, Novelty and Filtering Track Experiments using PIRCS. *in* 'NIST Special Publication 500-251: The Eleventh Text Retrieval Conference (TREC)'. pp. 520–528. 3.3.1

Lafferty, J. & Zhai, C. (2001). 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 Press. pp. 111–119. 2.3.2

Lafferty, J. & Zhai, C. (2003). Probabilistic relevance models based on document and query generation. *in* 'Language Modelling and Information Retrieval, Kluwer International Series on Information Retrieval, vol 13'. 2.3.2

Lavrenko, V. & Croft, W. B. (2001). Relevance based language models. *in* 'Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 120–127. 2.3.2

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the Web. *Nature* **400**, 107–109. 3.2.2

Lebanon, G. & Lafferty, J. (2002). Cranking: Combining rankings using conditional probability models on permutations. *in* 'Proceedings of the 19th International Conference on Machine Learning'. 3.4.4.1

Lee, J. H. (1997). Analyses of multiple evidence combination. *in* 'Proceedings of the 20th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 267–276. 3.4.4.1

Lempel, R. & Moran, S. (2000). The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks (Amsterdam, Netherlands: 1999)* **33**(1-6), 387–401. 3.3.2.2

Levy, P. (1995). *Que est ce que le virtuel.* Editions La Decouverte. 3.2.1

Lewis, D. (1996). The TREC-4 Filtering Track. *in* 'NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)'. pp. 165–180. 2.5

Li, L., Shang, Y. & Zhang, W. (2002). Improvement of hits-based algorithms on web documents. *in* 'Proceedings of the 11th international conference on World Wide Web (WWW11)'. ACM. 3.4.1

Li, W.-S., Kolak, O., Vu, Q. & Takano, H. (2000). Defining logical domains in a web site. *in* 'Proceedings of the eleventh ACM on Hypertext and hypermedia'. ACM Press. pp. 123–132. 3.3.1

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* **37**, 145–151. 5.4.1, 5.4.1

Lindgren, B. (1971). *Elements of Decision Theory.* The Macmillan Company, New York. 5.2

Luce, R. & Raiffa, H. (1957). *Games and Decisions.* John Wiley and Sons, New York. 8.2

Luhn, H. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* **2**, 159–165. 2.2

Manmatha, R., Rath, T. & Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. *in* 'Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 267–275. 3.4.4.1, 4.5.2.6, 5.4

Manning, C. & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. 2.2

McBryan, O. A. (1994). GENVL and WWWW: Tools for taming the Web. *in* 'Proceedings of the 1st international conference on World Wide Web (WWW1)'. 1.2, 3.4.2

Meng, W., Yu, C. & Liu, K.-L. (2002). Building Efficient and Effective Metasearch Engines. *ACM Computing Surveys* **34**, 48–89. 3.4.4.1

Metzler, D. & Croft, W. B. (2005). 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'. ACM Press. pp. 472–479. 4.3.2

Ng, A. Y., Zheng, A. X. & Jordan, M. I. (2001). Link Analysis, Eigenvectors and Stability. *in* 'IJCAI'. pp. 903–910. 3.3.2.1

Nottelmann, H. & Fuhr, N. (2003). From Uncertain Inference to Probability of Relevance for Advanced IR Applications. *in* 'Proceedings of the 25th European Conference on IR Research (ECIR'03)'. pp. 235–250. 5.4.2

Ogilvie, P. & Callan, J. (2003). Combining document representations for known-item search. *in* 'Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 143–150. 3.4.4.2, 4.4.1

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. & Johnson, D. (2005). Terrier Information Retrieval Platform. *in* 'Proceedings of the 27th European Conference on IR Research (ECIR 2005)'. pp. 517–519. 4.2

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report. Stanford University, Stanford, CA. 4.5.2, 5.2

Pandurangan, G., Raghavan, P. & Upfal, E. (2002). Using PageRank to Characterize Web Structure. *in* '8th Annual International Computing and Combinatorics Conference (COCOON)'. 4.5.2.6

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J. & Giles, C. L. (2002). Winners Don't Take All: Characterizing the Competition for Links on the Web. *Proceedings of the National Academy of Sciences* **99**(8), 5207–5211. 3.2.2

Pinski, G. & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management* pp. 297–312. 3.3.2, 3.3.2.1

Pirolli, P. & Pitkow, J. E. (1999). Distributions of surfers' paths through the world wide web: Empirical characterizations. *World Wide Web* **2**(1-2), 29–45. 3.3.3, 5.4.1

Pirolli, P., Pitkow, J. & Rao, R. (1996). Silk from a sow's ear: extracting usable structures from the web. *in* 'Conference proceedings on Human factors in computing systems'. ACM Press. pp. 118–125. 3.3.2

Plachouras, V. & Ounis, I. (2004). Usefulness of hyperlink structure for query-biased topic distillation. *in* 'Proceedings of the 27th annual international ACM SIGIR conference on Research and Developement in Information Retrieval'. ACM Press. pp. 448–455. 2.3.3.4, 3.3.1, 4.3, 4.5.1, 4.5.1, 7.4.2.1

Plachouras, V. & Ounis, I. (2005). Dempster-Shafer theory for a query-biased combination of evidence on the Web. *Information Retrieval* **8**, 197–218. 3.6.2, 5.2.1

Plachouras, V., Cacheda, F., Ounis, I. & Van Rijsbergen, C. J. (2003). University of Glasgow at the Web track: Dynamic Application of Hyperlink analysis using the Query Scope. *in* 'NIST Special Publication 500-251: The Twelfth Text Retrieval Conference (TREC 2003)'. 3.3.1, 4.4.1, 4.5.1

Plachouras, V., He, B. & Ounis, I. (2004). University of Glasgow at TREC2004: Experiments in Web, Robust and Terabyte tracks with Terrier. *in* 'Proceeddings of the 13th Text REtrieval Conference (TREC 2004)'. 2.3.3.4, 3.6.2, 4.3, 4.4.1, 4.5.1

Plachouras, V., Ounis, I. & Amati, G. (2005). The Static Absorbing Model for the Web. *Journal of Web Engineering* **4**, 165–186. 4.5.2.6

Plachouras, V., Ounis, I. & Cacheda, F. (2004). Selective Combination of Evidence for Topic Distillation using Document and Aggregate-level Information. *in* 'Proceedings of RIAO 2004 (Recherche d'Information Assistee par Ordinateur - Computer assisted information retrieval)'. pp. 610–622. 5.3.1

Plachouras, V., Ounis, I., Amati, G. & Van Rijsbergen, C. J. (2002). University of Glasgow at the Web track of TREC 2002. *in* 'NIST Special Publication 500-251: The Eleventh Text Retrieval Conference (TREC)'. pp. 645–651. 4.4.1

Ponte, J. M. & Croft, W. B. (1998). A language modeling approach to information retrieval. *in* 'Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 275–281. 2.3.2

Porter, M. F. (1980). An algorithm for suffix stripping. *Program* **14(3)**, 130–137. 2.2, 4.2

Press, W., Flannery, B., Teukolsky, S. & Vetterling, W. (1992). *Numerical Recipes: The Art of Scientific Computing.* 2nd edn. Cambridge University Press. Cambridge (UK) and New York. 4.3.2

R Development Core Team (2005). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. 1

Raggett, D., Le Hors, D. & Jacobs, I. (1999). 'Html 4.01 specification'. 3.2.1, 4.3.1

Raghavan, S. & Garcia-Molina, H. (2001). Crawling the hidden web. *in* 'Proceedings of the 27th International Conference on Very Large Data Bases (VLDB'01)'. Morgan Kaufmann Publishers Inc.. pp. 129–138. 3.2.2

Ribeiro-Neto, B. & Muntz, R. R. (1996). A Belief Network Model for IR. *in* 'Proceedings of the 19th annual international ACM SIGIR conference on Research and Developement in Information Retrieval'. pp. 253–260. 3.4.3

Richardson, M. & Domingos, P. (2002). The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. *in* 'Advances in Neural Information Processing Systems 14'. 3.4.1

Robertson, S. (2002). 'On Bayesian Models and Event Spaces in Information Retrieval. Workshop on Mathematical/Formal Methods in Information Retrieval, ACM SIGIR Conference'. 2.3.2

Robertson, S. & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Informmation Science* **27**, 129–146. 2.3, 2.3.1

Robertson, S. & Walker, S. (1994). 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'. Springer-Verlag New York, Inc.. pp. 232–241. 2.3.1

Robertson, S., Van Rijsbergen, C. J. & Porter, M. F. (1981). Probabilistic models of indexing and searching. *in* 'Proceedings of the 3rd annual ACM conference on Research and Development in Information Retrieval'. Butterworth & Co.. Kent, UK, UK. pp. 35–56. 2.3, 2.3.1

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. & Gatford, M. (1994). Okapi at TREC-3. *in* 'NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC-3)'. pp. 109–126. 2.3.1, 4.3.2

Robertson, S., Zaragoza, H. & Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. *in* 'Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM'04)'. ACM Press. pp. 42–49. 3.4.4.2, 4.4.1, 4.4.2

Rose, D. E. & Levinson, D. (2004). Understanding user goals in web search. *in* 'Proceedings of the 13th international conference on World Wide Web (WWW13)'. ACM Press. pp. 13–19. 3.2.4, 3.6.1, 7.2.1

Salton, G. & McGill, M. (1986). *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc. 2.3

Salton, G., Fox, E. A. & Wu, H. (1983). Extended boolean information retrieval. *Commun. ACM* **26**(11), 1022–1036. 2.3

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, overlap. *Journal of the American Society for Information Science* **39(3)**, 197–216. 3.4.4.1

Savoy, J. (1996). An extended vector-prcessing scheme for searching information in Hypertext systems. *Information Processing & Management* **32(2)**, 155–170. 3.4.3

Savoy, J. & Picard, J. (2001). Retrieval effectiveness on the web. *Information Processing & Management* **37(4)**, 543–569. 3.4.3, 4.2

Savoy, J. & Rasolofo, Y. (2001). Report on trec-10 experiment: Distributed collections and entrypage searching. *in* 'NIST Special Publication 500-250: The Tenth Text Retrieval Conference (TREC 2001)'. 3.3.1, 4.5.1

Savoy, J., Rasolofo, Y. & Perret, L. (2003). Report on the TREC 2003 Experiment: Genomic and Web Searches. *in* 'NIST Special Publication 500-255: The Twelfth Text Retrieval Conference (TREC 2003)'. pp. 739–750. 4.4.1

Shaw, J. & Fox, E. (1994). Combination of multiple searches. *in* 'NIST Special Publication 500-226: Overview of the Third Text REtrieval Conference (TREC-3)'. pp. 105–109. 3.4.4.1

Shivakumar, N. & Garcia-Molina, H. (1998). Finding near-replicas of documents on the Web. *in* 'Proceedings of the International Workshop on the World Wide Web and Databases (WebDB'98)'. 3.2.3

Siegel, S. & Castellan, J. J. (1988). *Nonparametric statistics for the Behavioral Sciences, 2nd Ed.*. McGraw-Hill. 6.2.1

Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E. & Ziviani, N. (2000). Link-based and content-based evidential information in a belief network model. *in* 'Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 96–103. 3.4.3

Silverman, B. (1986). *Density Estimation.* Chapman & Hall, London. 5.5.3

Silverstein, C., Marais, H., Henzinger, M. & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum* **33**(1), 6–12. 1.2, 3.2.4, 7.4.2.1

Singhal, A., Buckley, C. & Mitra, M. (1996). Pivoted document length normalization. *in* 'Proceedings of the 19th annual international ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM Press. pp. 21–29. 3.4.4.3

Spark-Jones, K. & Van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of documentation* **32**, 59–75. 2.4

Spertus, E. (1997). ParaSite: Mining Structural Information on the Web. *in* 'Proceedings of the 6th international conference on World Wide Web (WWW6)'. 3.2.3

Tajima, K., Hatano, K., Matsukura, T., Sano, R. & Tanaka, K. (1999). Discovery and retrieval of logical information units in web. *in* 'Proceedings of the 1999 ACM Digital Libraries Workshop on Organizing Web Space'. 3.3.1

Tajima, K., Mizuuchi, Y., Kitagawa, M. & Tanaka, K. (1998). Cut as a querying unit for WWW, Netnews, and E-mail. *in* 'Proceedings of ACM Hypertext '98'. pp. 235–244. 3.3.1

Tomiyama, T., Karoji, K., Kondo, T., Kakuta, Y., Takagi, T., Aizawa, A. & Kanazawa, T. (2003). Meiji University Web, Novelty and Genomics Track Experiments. *in* 'NIST Special Publication 500-261: The Thirteenth Text Retrieval Conference (TREC 2004)'. 4.4.1

Tomlinson, S. (2005). European web retrieval experiments with hummingbird searchservertm at clef 2005. *in* 'Working Notes for the CLEF 2005 Workshop'. 4.5.1

Trigg, R. (1983). 'A Network-Based Approach to Text Handling for the Online Scientific Community. PhD Thesis, Dept. of Computer Science, University of Maryland'. 3.2.1

Tsikrika, T. & Lalmas, M. (2004). Combining evidence for web retrieval using the inference network model: an experimental study. *Information Processing & Management* **40**(5), 751–772. 3.4.4.2

Turtle, H. & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems* **9**(3), 187–222. 3.4.3, 3.4.4.2

Upstill, T., Craswell, N. & Hawking, D. (2003). Query-independent evidence in home page finding. *ACM Transactions on Information Systems* **21**, 286–313. 3.4.2, 3.4.4.3

Van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edition.* Butterworth-Heinemann. 2.1, 2.3, 2.4

Voorhees, E. (2003). Overview of the TREC 2003 Robust Retrieval Track. *in* 'NIST Special Publication 500-255: The Twelfth Text REtrieval Conference (TREC 2003)'. pp. 69–77. 3.6.2

Voorhees, E. (2004). Overview of the TREC 2004 Robust Retrieval Track. *in* 'Proceedings of the 13th Text REtrieval Conference (TREC 2004)'. 3.6.2

Wald, A. (1950). *Statistical Decision Functions.* Wiley. 5.2

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature* **393**, 440–442. 3.2.2

Westerveld, T., Hiemstra, D. & Kraaij, W. (2001). Retrieving Web Pages Using Content, Links, URLs and Anchors. *in* 'NIST Special Publication: 500-250 The Tenth Text REtrieval Conference (TREC 2001)'. pp. 663–673. 3.3.1, 3.4.4.2, 3.4.4.3, 4.5.1

Williams, H. & Zobel, J. (1999). Compressing Integers for Fast File Access. *Computer Journal* **42**, 193–201. 2.2

Witten, I., Moffat, A. & Bell, T. (1994). *Managing Gigabytes: Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, New York. 2.2

Wong, A. & You, M. (1985). Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on PAttern Analysis and Machine Intelligence* **PAMI-7**, 599–609. 5.4.2

Yom-Tov, E., Fine, S., Carmel, D. & Darlow, A. (2005). Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. *in* 'Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 512–519. 1.2, 3.6.2

Yuret, D. (1994). 'From Genetic Algorithms To Efficient Optimization. Master Thesis, MIT, A.I. Technical Report No. 1569.'. 4.3.2

Zaragoza, H., Craswell, N., Taylor, M., Saria, S. & Robertson, S. (2004). Microsoft Cambridge at TREC-13: Web and HARD tracks. *in* 'NIST Special Publication 500-261: The Thirteenth Text Retrieval Conference (TREC 2004)'. 3.4.4.2, 4.4.1, 4.4.1, 4.4.2, 4.5.1

Zhai, C. & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *in* 'Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval'. ACM Press. pp. 334–342. 2.3.2