Limsopatham, Nut (2014) *A framework for enhancing the query and medical record representations for patient search.* PhD thesis.

A Framework for Enhancing the Query and Medical Record Representations for Patient Search

Nut Limsopatham

School of Computing Science

University of Glasgow

A thesis submitted for the degree of

Doctor of Philosophy

September 2014

Abstract

Electronic medical records (EMRs) are digital documents stored by medical institutions that detail the observed symptoms, the conducted diagnostic tests, the identified diagnoses and the prescribed treatments. These EMRs are being increasingly used worldwide to improve healthcare services. For example, when a doctor compiles the possible treatments for a patient showing some particular symptoms, it is advantageous to consult the information about patients who were previously treated for those same symptoms. However, finding patients with particular medical conditions is challenging, due to the implicit knowledge inherent within the patients' medical records and queries - such knowledge may be known by medical practitioners, but may be hidden from an information retrieval (IR) system. For instance, the mention of a treatment such as a drug may indicate to a practitioner that a particular diagnosis has been made for the patient, but this diagnosis may not be explicitly mentioned in the patient's medical records. Moreover, the use of negated language (e.g. 'without', 'no') to describe a medical condition of a patient (e.g. the patient has no fever) may cause a search system to erroneously retrieve that patient for a query when searching for patients with that medical condition (e.g. find patients with fever).

This thesis focuses on enhancing the search of EMRs, with the aim of identifying patients with medical histories relevant to the medical conditions stated in a text query. During retrieval, a healthcare practitioner indicates a number of inclusion criteria describing the medical conditions of the patients of interest. To attain effective retrieval performance, we hypothesise that, in a patient search system, both the information needs and patients' histories should be represented based upon *the medical decision process*. In particular, this thesis argues that since the medical decision process typically encompasses four aspects (symptom, diagnostic test, diagnosis and treatment), a patient search system should take into account these aspects and apply inferences to recover the possible implicit knowledge. We postulate that considering these aspects and their derived implicit knowledge at three different levels of the retrieval process (namely, sentence, medical record and inter-record levels) enhances the retrieval performance. Indeed, we propose a novel framework that can gain insights from EMRs and queries, by modelling and reasoning upon information during retrieval in terms of the four aforementioned aspects at the three levels of the retrieval process, and can use these insights to enhance patient search.

Firstly, at the sentence level, we extract the medical conditions in the medical records and queries. In particular, we propose to represent only the medical conditions related to the four medical aspects in order to improve the accuracy of our search system. In addition, we identify the context (negative/positive) of terms, which leads to an accurate representation of the medical conditions both in the EMRs and queries. In particular, we aim to prevent patients whose EMRs state the medical conditions in the contexts different from the query from being ranked highly. For example, preventing patients whose EMRs state “no history of dementia” from being retrieved for a query searching for patients with dementia.

Secondly, at the medical record level, using external knowledge-based resources (e.g. ontologies and health-related websites), we leverage the relationships between medical terms to infer the wider medical history of the patient in terms of the four medical aspects. In particular, we estimate the relevance of a patient to the query by exploiting association rules that we extract from the semantic relationships between medical terms using the four aspects of the medical process. For example, patients with a medical history involving a *CABG surgery* (treatment) can be inferred as relevant to a query searching for a patient suffering from *heart disease* (diagnosis), since a CABG surgery is a treatment of heart disease.

Thirdly, at the inter-record level, we enhance the retrieval of patients in two different manners. First, we exploit knowledge about how the four medical aspects are handled by different hospital departments to gain a better understanding about the appropriateness of EMRs created by different departments for a given query. We propose to aggregate EMRs at the department level (i.e. inter-record level) to extract implicit knowledge (i.e. the expertise of each department) and model this department’s expertise, while ranking patients. For instance, patients having EMRs from the cardiology department are likely to be relevant to a query searching for patients who suffered from a heart attack. Second, as a medical query typically contains several medical conditions that the relevant patients should satisfy, we propose to explicitly model the relevance towards multiple query medical conditions in the EMRs related to a particular patient during retrieval. In particular, we rank highly those patients that match all the stated medical conditions in the query by adapting coverage-based diversification approaches originally proposed for the web search domain.

Finally, we examine the combination of our aforementioned approaches that exploit the implicit knowledge at the three levels of the retrieval process to further improve the retrieval performance by adapting techniques from the fields of data fusion and machine learning. In particular, data fusion techniques, such as CombSUM and CombMNZ, are used to combine the relevance scores computed by the different approaches of the proposed framework. On the other hand, we deploy state-of-the-art learning to rank approaches (e.g. LambdaMART and AdaRank) to learn from a set of training data an effective combination of the relevance scores computed by the approaches of the framework. In addition, we introduce a novel selective ranking approach that uses a classifier to effectively apply one of the approaches of the framework on a per-query basis.

This thesis draws insights from a thorough evaluation and analysis of the proposed framework using a standard test collection provided by the TREC Medical Records track. The experimental results show the effectiveness of the framework. In particular, the results demonstrate the importance of dealing with the implicit knowledge in patient search by focusing on the medical decision criteria aspects at the three levels of the retrieval process.

Acknowledgements

This thesis has been one of the most interesting journeys of my life. As well as excitement and joyfulness, I have learned tremendously from this journey. I owe my deepest gratitude to several people for their immense support during the course of my PhD.

First and foremost, I am sincerely and heartily grateful to my supervisor, Iadh Ounis, and co-supervisor Craig Macdonald. Without their insightful advice and extensive support, this thesis would not have been possible.

I dedicate this thesis to my family, especially my parents and my aunt: Kitti Limsopatham and Suree Raksuriya, and Vipa Limsopatham. Their endless support and belief in me made it possible for me to complete this PhD.

I am also thankful to Jeremy Singer, my second supervisor, for his encouragement and support over the course of my PhD, to Djoerd Hiemstra and Yanjun Qi for mentoring during the SIGIR Doctoral Consortium, and to Gareth Jones and Simon Rogers for their thoughtful feedback during my PhD viva.

I would like to also thank my colleagues at the TerrierTeam research group for collaboration over the past four years: Rodrygo Santos, Richard McCreadie, Dyaa Albakour, Eugene Kharitonov and Graham McDonald. Special thanks to Richard for reading parts of this thesis.

Contents

Abstract i

Acknowledgements iv

1 Introduction 1

1.1 Introduction 1

1.2 Motivation 2

1.3 Thesis Statement 4

1.4 Contributions 5

1.5 Origins of Material 7

1.6 Thesis Outline 7

2 Information Retrieval 10

2.1 Introduction 10

2.2 Text Operations 11

2.2.1 Tokenisation 11

2.2.2 Stopword Removal 11

2.2.3 Stemming 12

2.2.4 Thesauri and Controlled Vocabulary 12

2.3 Indexing 13

2.4 Document Retrieval 15

2.4.1 The Best Match Weighting Model 15

2.4.2 Divergence From Randomness 16

2.5 Query Expansion 17

2.5.1 Pseudo-Relevance Feedback 18

2.5.2 Query Expansion using External Resources 19

2.6 Machine Learning for IR 20

2.6.1 Classification 20

2.6.2 Regression 21

2.6.3 Learning to Rank 21

2.7 Aggregate Ranking 22

2.8 Evaluation 24

2.8.1 Test Collections 24

2.8.2 Evaluation Measures 25

2.9 Conclusions 26

3 Searching in the Medical Domain 27

3.1 Introduction 27

3.2 Medical Information 28

3.3 Searching Knowledge-based Information 29

3.3.1 Dealing with Medical Terminology using a Conceptual Representation 32

3.3.2 Query Expansion for Dealing Medical Terminology 33

3.4 Searching Patient-Specific Information 35

3.4.1 Medical Records Track at TREC 36

3.4.2 Existing Medical Records Search Approaches 38

3.5 Conclusions 45

4 A Framework for Patient Search 47

4.1 Introduction 47

4.2 Our Patient Search System 48

4.2.1 The Query and Medical Record Representation Unit 49

4.2.2 The Patient Ranking Unit 51

4.2.3 Our Framework for Enhancing the Representations of Queries and Medical Records 52

4.3 Baseline Patient Search System 59

4.3.1 Experimental Setup 60

4.3.2 Research Questions 60

4.3.3 Experiments with the Adaptation of Aggregate Ranking Approaches for Ranking Patients 61

4.3.4 Experiments with the Task-Specific Representation Approach 62

4.3.5 Analysis and Discussion 65
4.4 Conclusions 67

5 Negation Handling 70
5.1 Introduction 70
5.2 Non-Supervised Approaches to Handle Negation 71
5.2.1 The NegFlag Approach 72
5.2.2 Term Dependence for Negation 72
5.3 A Learned Approach to Handle Negation 73
5.3.1 Context Identification 74
5.3.2 Context-based Penalisation 74
5.3.3 Penalising Weight Estimation 75
5.3.4 Learning Procedure 76
5.4 Experiments 79
5.4.1 Experimental Setup 79
5.4.2 Research Questions 80
5.4.3 Experiments with the Non-Supervised Approaches 80
5.4.4 Experiments with the Learned Approach 83
5.4.5 Analysis and Discussion 89
5.5 Conclusions 98

6 Reasoning using Medical Concept Relationships 100
6.1 Introduction 100
6.2 Extracting and Representing Medical Concept Relationships from External Resources as Association Rules 101
6.3 A Bayesian-based Approach for Inferring Conceptual Relationships 102
6.4 A Stochastic Analysis for Inferring Conceptual Relationships 104
6.5 Modelling Conceptual Relationships when Ranking Medical Records 105
6.6 Experiments 108
6.6.1 Experimental Setup 108
6.6.2 Research Questions 109
6.6.3 Experiment with our Bayesian-based Approach for Inferring Conceptual Relationships 110
6.6.4 Experiments with a Stochastic Analysis for Inferring Conceptual Relationships 112

6.6.5 Analysis and Discussion 113

6.7 Conclusions 117

7 Leveraging Knowledge about Hospital Departments' Expertise 120

7.1 Introduction 120

7.2 Leveraging Department-Level Evidence 121

7.2.1 The Extended Voting Techniques for Department-Level Evidence 121

7.2.2 Leveraging Department-Level Evidence when Ranking Medical Records . . . 123

7.3 Obtaining Department-Level Evidence 123

7.3.1 A Voting Paradigm for Modelling Department-Level Evidence 124

7.3.2 A Federated Search Technique for Modelling Department-Level Evidence . . . 125

7.4 Experiments 126

7.4.1 Experimental Setup 127

7.4.2 Research Questions 128

7.4.3 Experiments with Our Aggregate Scoring Approach 128

7.4.4 Experiments with Our Record Scoring Approach 134

7.4.5 Learning to Leverage Department-Level Evidence 138

7.4.6 Per-Query Performance Analysis 141

7.5 Conclusions 141

8 Modelling Relevance towards Multiple Inclusion Criteria 144

8.1 Introduction 144

8.2 Motivation & Problem Definition 145

8.3 An Approach for Modelling Relevance towards Multiple Inclusion Criteria 147

8.3.1 The Extended Patient Model 148

8.3.2 The Extended Two-Stage Model 148

8.3.3 Inclusion Criteria Extraction 150

8.4 Experiments 151

8.4.1 Experimental Setup 151

8.4.2 Research Questions 154

8.4.3 Experiments when Setting λ Uniformly 154

8.4.4 Experiments with the Learned λ Values 159

8.4.5 Analysis and Discussion 162

8.5 Conclusions 166

9 Combination of Approaches in Patient Search 168

9.1 Introduction 168

9.2 Data Fusion Approaches for Combining Relevance Scores 169

9.3 Learning to Rank for Combining the Components of Our Framework 171

9.4 Selective Query Dependent Retrieval 171

9.4.1 Learning Features 172

9.4.2 The Learning Procedure 172

9.5 Experiments 173

9.5.1 Experimental Setup 173

9.5.2 Research Questions 174

9.5.3 Experiments with Our Approaches for Combining Relevance Scores 174

9.5.4 Analysis and Discussion 177

9.6 Combination Approaches for the Two Units of Our Search System 180

9.6.1 Learning to Combine Representations 180

9.6.2 Learning to Selectively Rank the Patients' Medical History 184

9.7 Conclusions 188

10 Conclusions and Future Work 191

10.1 Contributions and Conclusions 191

10.1.1 Contributions 191

10.1.2 Conclusions 194

10.2 Directions for Future Work 196

10.3 Closing Remarks 198

References 211

List of Figures

2.1 Medical concepts identified by the MetaMap tool from a phrase ‘He presents to the ER today with hematuria.’ 14

3.1 A sample snapshot of a document from TREC 2004 and 2005 Genomics track ad hoc retrieval task. 30

3.2 An example of a transesophageal echocardiography medical record, from the cardiology department. 37

3.3 A mock example of narrative medical records. 38

4.1 Our patient search system. 49

4.2 Medical concepts extracted by the MetaMap tool from the query ‘Patients with diabetes mellitus who also have thrombocytosis’ 51

4.3 Illustrative examples when ranking patients using the patient and the two-stage models. 53

4.4 Components of the proposed framework. 55

4.5 An example of using association rules to infer the medical conditions of patient with ‘heart disease’. 57

4.6 An example of a chest X-ray medical record, radiology department. 58

4.7 The difference between the performances of our task-specific representation approach and the term-based representation approach on the TREC 2011 and 2012 Medical Records track, on a per-query basis. 66

4.8 The retrieval performances achieved by the different voting techniques tested, while varying the number of voting documents. 68

5.1 The differences between the retrieval performances of our NegFlag approach and the common baseline, when applied with a term-based representation, evaluated using the TREC 2011 Medical Records Track. 89

5.2 The differences between the retrieval performances of our NegFlag approach and the common baseline, when applied with a term-based representation, evaluated using the TREC 2012 Medical Records Track. 90

5.3 The differences between the retrieval performances of our NegFlag approach and the common baseline, when applied with our task-specific representation, evaluated using the TREC 2011 Medical Records Track. 90

5.4 The differences between the retrieval performances of our NegFlag approach and the common baseline, when applied with our task-specific representation, evaluated using the TREC 2012 Medical Records Track. 91

5.5 bpref performances of our term dependence for negation approach when applied with a term-based representation, while varying $window\_size$. 92

5.6 infNDCG performances of our term dependence for negation approach when applied with a term-based representation, while varying $window\_size$. 92

5.7 bpref performances of our term dependence for negation approach when applied with our task-specific representation, while varying $window\_size$. 93

5.8 infNDCG performances of our term dependence for negation approach when applied with our task-specific representation, while varying $window\_size$. 93

5.9 The difference between the performances of our learned approach and the common baseline, when applied with the Bo1 QE, evaluated using the TREC 2011 and 2012 Medical Records Track. 95

5.10 Ablation study of feature importance. 96

6.1 An example of identifying candidate concept expansions using our approach on query *"patients with osteoporosis"*. 103

6.2 Transforming (a) the collection *D* into (b) a bipartite graph *G*. 104

6.3 The retrieval performances of our Bayesian-based approach for inferring the conceptual relationships and the task-specific representation baseline on a per-query basis, evaluated using the TREC 2011 and 2012 Medical Records Track. 115

6.4 The retrieval performances of our stochastic approach for inferring the conceptual relationships and the task-specific representation baseline on a per-query basis, evaluated using the TREC 2011 and 2012 Medical Records Track. 116

6.5 The retrieval performances of our Bayesian-based approach for inferring the conceptual relationships and the Bo1 query baseline on a per-query basis, evaluated using the TREC 2011 and 2012 Medical Records Track. 117

6.6 The retrieval performances of our stochastic approach for inferring the conceptual relationships and the Bo1 baseline on a per-query basis, evaluated using the TREC 2011 and 2012 Medical Records Track. 118

7.1 An example of a transesophageal echocardiography (*TEE*) medical record, from the cardiology department.[1] 121

7.2 Examples of medical records from hospital departments. 124

7.3 The bpref performances of the aggregate scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with a term-based representation. 129

7.4 The infNDCG performances of the aggregate scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with a term-based representation. 130

7.5 The bpref performances of the aggregate scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with the task-specific representation. 131

7.6 The infNDCG performances of the aggregate scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with the task-specific representation. 132

7.7 The bpref performances of the record scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with a term-based representation. 134

7.8 The infNDCG performances of the record scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with a term-based representation. 135

7.9 The bpref performances of the record scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with the task-specific representation. 136

7.10 The infNDCG performances of the record scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with the task-specific representation. 137

7.11 Comparison of the retrieval performances of the aggregate scoring approach and the baseline, when applied with BM25 and the expCombSUM voting technique 142

8.1 Illustrative examples of ranking patients using the two main existing approaches. Note that the relevance scores computed using CombSUM are normalised to maintain probability estimates. 146

8.2 Medical concepts extracted by the MetaMap tool from query 102: 'Patients with diabetes mellitus who also have thrombocytosis', as the query inclusion criteria. 152

8.3 The retrieval performances of our proposed approach when applied within the patient model (IC-Cover-P), within the patient ranking stage (IC-Cover-2P) and within the medical record ranking stage (IC-Cover-2R) of the two-stage model, in terms of bpref for TREC 2011, as we vary the mixture parameter λ between 0 and 1. 164

9.1 The comparison of the performances of our four components on the TREC 2011 and 2012 Medical Records track, on a per-query basis. 178

9.2 The standard deviations of the retrieval performances achieved by each component of our framework on a per-query basis, on the TREC 2011 and 2012 Medical Records track. 179

9.3 The per-query performance difference between of the data fusion approaches that consider and do not consider *the agreement between the components*. 180

9.4 The difference between the retrieval performance obtained using the patient model and the two-stage model on each query, in terms of bpref, infNDCG and infAP, respectively. 188

List of Tables

2.1 An example of an inverted index 15

3.1 Examples of interchangeable pairs of replacements. 34

3.2 Name variants for *TGF-beta1*. 35

3.3 Examples of sentences in EMRs where the presence of the query term ‘cancer’ does not always indicate that the patient has ‘cancer’. 43

3.4 Comparing approaches to handle negated language in medical records search. 44

4.1 List of 16 of the MetaMap’s 133 semantic types that we consider for our proposed approach, based on the four aspects of the medical decision process. 50

4.2 An example of medical concepts obtained from the query *“patients with diabetes mellitus who also have thrombocytosis”* using our task-specific representation approach. . 52

4.3 The negation representation process for a medical record - italicised terms occur in a negative context. 56

4.4 Voting techniques used in our experiments. 61

4.5 Comparing retrieval performances of different patient ranking approaches on TREC 2011 and 2012 Medical Records track’s queries. Statistical significance (paired t-test) at $p < 0.05$ over the patient model baselines (DPH and BM25) are denoted $^\blacktriangle$ and $^\triangle$, respectively. 62

4.6 Our examined four variants of selecting medical concepts identified by the MetaMap tool to represent the medical records and queries 63

4.7 The comparison of the retrieval performances of different query and medical record representation approaches on the TREC 2011 and 2012 Medical Records track’s test topics. Statistical significance (paired t-test) at $p < 0.05$ over the term-based representation baseline and the corresponding conceptual representation baseline are denoted $^\blacktriangle$ and $^\triangle$, respectively. 64

4.8 Analysis of the task-specific representation approach w.r.t. the aspects of medical concepts found in the queries. The numbers between the parentheses indicate the number of queries impacted (benefited/harmed) compared to the total number of queries. . . . 67

5.1 Examples of sentences in EMRs where the presence of the query term ‘cancer’ does not always indicate the relevance. 71

5.2 The NegFlag process for a medical record - italicised terms occur in a negative context. 72

5.3 An example of how our learned approach deals with negation, where w_n is the weight of a term. 75

5.4 List of 13 features used to predict the penalising weight of the opposite context terms. . 78

5.5 Retrieval performances of our non-supervised approaches to handle negation in comparison to existing approaches, when applied with a term-based representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the traditional approach baselines are denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively. 81

5.6 Retrieval performances of our non-supervised approaches to handle negation in comparison to existing approaches, when applied with the task-specific representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the traditional approach baselines are denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively. 82

5.7 Retrieval performances of our learned approach to handle negation in comparison to existing approaches, when applied with a term-based representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the common baseline is denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively. 84

5.8 Retrieval performances of learned approach to handle negation in comparison to existing approaches, when applied with the task-specific representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the common baseline is denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively. 85

5.9 Retrieval performances of our learned approach to handle negation in comparison to existing approaches, when applied with a term-based representation and a query expansion approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the traditional approach baselines are denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively. 88

5.10 Analysis of our approach w.r.t. the aspects of medical concepts found in the queries. The numbers between the parentheses indicate the number of queries impacted compared to the total number of queries. 95

5.11 Retrieval performances of applying our negation handling approach to deal with other types of contexts in comparison to existing approaches, when applied with a term-based representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the common baseline is denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively. 98

6.1 Examples of medical concepts associated to ‘osteoporosis’ defined in our associated rules database. 102

6.2 List of the used features. 107

6.3 List of resources used for extracting the conceptual relationships related to the four aspects of the medical decision process. 109

6.4 Number of association rules extracted from each domain-specific resource. 109

6.5 The types of the extracted association rules in terms of the four medical aspects. 110

6.6 Retrieval performances of our two approaches for inferring relationships between medical concepts in comparison with the three baselines on TREC 2011 and 2012 Medical Records track’s queries. Statistical significance (paired t-test) at $p < 0.05$ over the task-specific representation and the Bo1 & Semantic QE baselines are denoted ^{t} and ^{s}, respectively. 112

6.7 Retrieval performances of our two approaches for inferring relationships between medical concepts in comparison with the three baselines on TREC 2011 and 2012 Medical Records track’s queries. Statistical significance (paired t-test) at $p < 0.05$ over the task-specific representation and the Bo1 & Semantic QE baselines are denoted ^{t} and ^{s}, respectively. 112

6.8 Analysis of our Bayesian-based approach w.r.t. the aspects of medical concepts found the queries. The numbers between the parentheses indicate the number of queries impacted (benefited/harmed) compared to the total number of queries. 114

6.9 Analysis of our stochastic approach w.r.t. the aspects of medical concepts found the queries. The numbers between the parentheses indicate the number of queries impacted (benefited/harmed) compared to the total number of queries. 114

7.1 Statistics of hospital departments in the collection. 127

7.2 Comparison of the retrieval performances of the aggregate scoring approach with the expCombSUM and expCombMNZ baselines on TREC 2011 and 2012 Medical Records track's queries. Statistical significance (paired t-test) at $p < 0.05$ over expCombSUM and expCombMNZ are denoted $^{\oplus}$ and $^{\ominus}$, respectively. 140

8.1 Statistics of the inclusion criteria extracted from the TREC queries 152

8.2 Comparison of the retrieval performances of the IC-Cover approach (λ = 0.5) against the common baseline (λ = 0) on the TREC Medical Records Track 2011 and 2012. Statistical significance (paired t-test, $p < 0.05$) over the corresponding baseline is denoted $^{\blacktriangle}$. The column denoted $\triangle$ (resp. $\bigtriangledown$) shows the number of queries improved (resp. harmed) in relation to the corresponding baseline. 156

8.3 Comparison of the retrieval performances of the IC-Cover approach (λ = 0.5) against the common baseline (λ = 0) on the TREC Medical Records Track 2011 and 2012, when applied with the NegFlag approach. Statistical significance (paired t-test, $p < 0.05$) over the corresponding baseline is denoted $^{\blacktriangle}$. The column denoted $\triangle$ (resp. $\bigtriangledown$) shows the number of queries improved (resp. harmed) in relation to the corresponding baseline. 158

8.4 List of the used features. 160

8.5 Comparison of the retrieval performances using various λ values within the proposed IC-Cover approach. Statistical significances (paired t-test, $p < 0.05$) over the settings when $\lambda = 0$, $\lambda = 1$, $\lambda = 0.5$, and when using a learned technique (5-fold) are denoted $^{\oplus}$, $^{\ominus}$, $^{\odot}$ and $^{\otimes}$, respectively. The column denoted $\triangle$ (resp. $\bigtriangledown$) shows the number of queries improved (resp. harmed) in relation to the baseline where $\lambda = 0$. 161

8.6 Comparison of the retrieval performances using various λ values within the proposed IC-Cover approach, when applied with the NegFlag approach. Statistical significances (paired t-test, $p < 0.05$) over the settings when $\lambda = 0$, $\lambda = 1$, $\lambda = 0.5$, and when using a learned technique (5-fold) are denoted $^{\oplus}$, $^{\ominus}$, $^{\odot}$ and $^{\otimes}$, respectively. The column denoted $\triangle$ (resp. $\bigtriangledown$) shows the number of queries improved (resp. harmed) in relation to the baseline where $\lambda = 0$. 161

8.7 Analysis of IC-Cover-P(SUM) when applied with BM25 and the NegFlag approach w.r.t. the number of inclusion criteria extracted from the queries. The numbers between the parentheses indicate the percentage compared to the total number of queries. 166

9.1 The components of our framework and their retrieval performances. 173

9.2 The comparison of the retrieval performances of the proposed approaches for combining the components of our framework on TREC 2011 and 2012 Medical Records track's queries. 175

9.3 List of the learning features used to predict the importance of the relevance scores from the bag-of-words (BoW) and bag-of-concepts (BoC) representations. 182

9.4 The retrieval performances of different representation approaches on the TREC 2011 and 2012 Medical Records track test collection. Statistical significance (paired t-test) at $p < 0.05$, at $p < 0.01$, and at $p < 0.001$ over a baseline are denoted ^{a}, ^{aa} and ^{aaa}, respectively. ^{a} is ^{1}, ^{2}, ^{3} or ^{4} to represent the bag-of-words representation (BoW), the task-specific representation (BoC), the score combination ($\delta = 2$) baselines, or our learned approach (5-fold) approach, respectively. 183

9.5 List of the query features used by our classifier to decide on when to apply either the patient model or the two-stage model. 185

9.6 The retrieval performances of different patient ranking approaches on the TREC Medical Records track's collection. Statistical significance (paired t-test) at $p < 0.05$, at $p < 0.01$, and at $p < 0.001$ over an alternative approach are denoted ^{x}, ^{xx} and ^{xxx}, respectively. Where ^{x} is $^{\oplus}$, $^{\ominus}$, $^{\otimes}$, $^{\oslash}$ or $^{\odot}$ and refers to the patient model, the two-stage model, CombMAX, CombSUM, and our selective approach (5-fold), respectively. . . . 186

Chapter 1

Introduction

1.1 Introduction

Government-led initiatives worldwide have digitised the storage of the medical records of patients within healthcare service providers (e.g. hospitals), resulting in the emergence of electronic medical records (EMRs) (Kotsiopoulos *et al.*, 2003; Tambouris & Makropoulos, 1999). These EMRs have been developed and extensively used worldwide to enhance healthcare services. Indeed, the EMR initiatives have generated vast quantities of records that could aid medical practitioners in identifying effective treatments for patients showing particular symptoms (Edinger *et al.*, 2012; Hersh, 2008*a*; Voorhees & Hersh, 2012; Voorhees & Tong, 2011). For example, when a doctor compiles the possible effective treatments for patients with skin cancer, it could be advantageous to be able to search for patients that have previously been admitted to the hospital with that disease. Importantly, when conducting comparative effectiveness research within the medical domain (Edinger *et al.*, 2012; Voorhees & Hersh, 2012; Voorhees & Tong, 2011), which aims to compare the effectiveness of different medical procedures, it is crucial to identify patients (known as cohorts) with particular medical conditions. To search for such cohorts, healthcare practitioners develop a set of inclusion criteria that describe the medical conditions of the patients of interest. These criteria typically include medical conditions (e.g. symptoms, diseases), which are formulated into a textual query used to search for patients that have medical records matching the specified criteria.

Existing information retrieval (IR) techniques have been deployed to retrieve patients based on the relevance of their medical records towards a given query (we refer to this task as *patient search*). For example, Leveling *et al.* (2012) deployed BM25 (Robertson *et al.*, 1994) to rank the concatenations of medical records associated to each patient. However, Edinger *et al.* (2012) showed that existing IR approaches may fail in this task in several scenarios. For example, the common use of negated language and medical jargon in the medical records makes patient search a challenging task. In particular, negated language is extensively used in medical records to indicate an absence of a medical condition (Koopman *et al.*, 2010). Moreover, medical terms have several associated terms, such as acronyms, synonyms and hyponyms, which are used inconsistently by practitioners (Aronson & Rindflesch, 1997; Hersh, Hickam, Haynes & McKibbon, 1994; Srinivasan, 1996; Trieschnigg *et al.*, 2010). When reading medical records, healthcare practitioners commonly use what we refer to as *their implicit knowledge* to deal with these challenges (e.g. the negated language and the medical jargon). This implicit knowledge is the information that is commonly known by healthcare practitioners but may be hidden from the IR system. For instance, when searching for a patient with a particular disease, healthcare practitioners will also look for patients with medical records that contain any of the medical jargon related to that disease.

Existing works (e.g. Demner-Fushman *et al.* (2011); King *et al.* (2011); Zhu & Carterette (2013)) have studied approaches for dealing with the implicit knowledge that resulted in an effective retrieval performance. For example, Demner-Fushman *et al.* (2011) used real medical experts to interactively improve the queries issued to their system, while King *et al.* (2011) applied query expansion techniques by using the UMLS metathesaurus[1] to improve the representation of a query. In this thesis, we investigate a novel framework that exploits implicit knowledge to enhance a patient search system at the three different levels of the retrieval process (i.e. sentence, record and inter-record levels).

In the remainder of this chapter, we firstly discuss the motivation for the work in this thesis in Section 1.2. Next, we present the statement and contributions of this thesis in Sections 1.3 and 1.4, respectively. The origins of the material are then described in Section 1.5. Finally, we provide an overview of the structure for the remainder of the thesis in Section 1.6.

1.2 Motivation

Existing studies attempted to deal with the patient search task (i.e. retrieving patients based on the relevance of their medical records towards a given query) by using traditional IR techniques. However, according to Edinger *et al.* (2012), in several cases, using existing IR techniques to retrieve patients with a given medical condition is not effective, since patient search task has unique characteristics, such as the use of complex terminology and negated language in medical records and queries. Examples of medical queries are:

Q137: Patients with inflammatory disorders receiving TNF-inhibitor treatments

Q179: Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression

[1] `http://www.nlm.nih.gov/research/umls/`

Recently, several approaches have been proposed to uncover implicit knowledge in patient search. Demner-Fushman *et al.* (2011) and King *et al.* (2011) proposed to handle the negated language in medical records by simply disregarding terms with a negated context during indexing. However, disregarding terms with a negated context could not prevent patients whose medical records state that they do not have the medical condition in the query from being retrieved. For example, the patient with a record such as "he is admitted with acute diabetes ... he has no evidence of hypertension ..." may still be retrieved for a query aiming to find patients with diabetes and hypertension, since the term *diabetes* is contained both in the query and the patient's record. In addition, this approach is not effective when searching for patients who do not have a particular medical condition (e.g. "find patients with diabetes who have no evidence of hypertension")

Moreover, the medical terminology is known to be rich and difficult to deal with (Hersh, 2008*a*; Hersh, Hickam, Haynes & McKibbon, 1994; Srinivasan, 1996). The emergence of medical ontologies provides an opportunity for a retrieval system to leverage their semantic relationships to gain more understanding of the medical records and queries. Indeed, an ontology provides a vocabulary describing a domain of interest, with the definitions of classes, relations, functions, and their associated named entities (Gruber, 1993). For instance, Demner-Fushman *et al.* (2011) proposed to expand each query with related terms derived from RxNorm[1] (a drug dictionary) and the UMLS metathesaurus to improve the representation of a query to aid relevance estimation. Meanwhile, Zuccon *et al.* (2012) derived the medical conditions related to the query based on the hierarchy of the SNOMED CT ontology[2] and estimated the emphasis to be placed on each condition using term and inverse document frequencies. However, while these works leverage individual resources, it is unclear how to simultaneously use multiple resources to derive related terms, and how much emphasis to place on them within a single system.

Furthermore, medical records describe the medical conditions of a patient throughout the medical process (Hersh, 2008*a*; Silfen, 2006). These medical conditions include four main aspects (namely, symptoms, diagnostic tests, diagnoses, and treatments), which are related to the *medical decision process* (Limsopatham, Macdonald & Ounis, 2013*a*; Silfen, 2006). Medical practitioners typically take into account the information related to the medical decision process (Silfen, 2006) when consulting with patients. For example, knowing that a patient is visiting a hospital with a *'chest pain'* (symptom), a healthcare practitioner may suspect that the patient has *'heart disease'* (diagnosis). Given the symptom, the practitioner compiles a set of diagnostic procedures, such as *'chest X-ray'* (diagnostic test), for the patient. Once the practitioner is confident that the patient suffers from *'heart disease'* (diagnosis),

[1] `http://www.nlm.nih.gov/research/umls/rxnorm/`

[2] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

the practitioner may prescribe a treatment, such as *'coronary artery bypass surgery'*, for the patient. Therefore, an effective patient search system should model and infer in terms of these four aspects of the medical decision process.

Existing works have shown that the structure of documents could be leveraged within a retrieval process to enhance the retrieval performance (e.g. Broder *et al.* (2010); Robertson *et al.* (2004); Salton & Buckley (1991); Salton *et al.* (1993); Wilkinson (1994)). For instance, Salton *et al.* (1993) separated individual documents into sentences, and combined the relevance scores of both the documents and their sentences when ranking those documents. Meanwhile, Broder *et al.* (2010) aggregated web documents to the site level (i.e. the website containing those web documents) to extract new evidence for ranking the web documents. They argued that the information extracted from the site level was useful for improving retrieval performance, as the site-level information of a website encompassed the overall information of the web documents within that website. Inspired by these existing works, we propose to leverage information at three levels of the retrieval process, including sentence, record and inter-record levels to enhance a patient search system.

1.3 Thesis Statement

This thesis argues that since the medical decision process typically encompasses four aspects (symptom, diagnostic test, diagnosis and treatment), a patient search system should take into account these aspects and apply inferences to uncover possible implicit knowledge that is missing from the medical records and queries. We postulate that considering these aspects and their derived implicit knowledge at different levels of the retrieval process (namely, sentence, record and inter-record levels) will enhance the effectiveness of a patient search system. Indeed, we propose to build a framework that can gain insights from the medical records and queries, by modelling and reasoning during retrieval in terms of the four aforementioned aspects at the three levels of the retrieval process:

- At the *sentence level*, using a medical concept extraction tool, we will detect the medical conditions related to the aforementioned four aspects in the medical records and queries, and we will use them to better match relevant EMRs. In addition, by using negation detection tools, we identify the context (negative/positive) of terms, which we can use to generate a more accurate representation of the targeted medical conditions.

- At the *record level*, we will extract the relationships between medical terms using knowledge-based resources (e.g. ontologies, websites) such that we can infer the wider medical history of the patients in terms of the four medical aspects.

- at the *inter-record level*, we will use knowledge about how the four aspects are handled by different hospital departments to gain understanding about the appropriateness of medical records from different departments for a given query. Meanwhile, we will use the relevance towards different medical conditions related to the four aspects stated in a query to ensure the retrieval of patients whose medical records satisfy most of the query medical conditions.

Therefore, by considering the four aspects of the medical decision process at the three aforementioned levels of the framework, we argue that the effectiveness of the patient search system will be improved.

1.4 Contributions

We contribute a framework for enhancing the representations of queries and medical records in patient search. This proposed framework allows the effective retrieval of patients based on the relevance of their medical records towards the queries. The framework contains four components, each of which leverages the implicit knowledge at one of the three different levels of the retrieval process (namely, sentence, record and inter-record levels).

In the course of the thesis, several research questions regarding the proposed framework for enhancing the representations of queries and medical records are addressed. We investigate several approaches that instantiate the various components of the proposed framework and uncover implicit knowledge from the medical records and queries, in order to enhance the retrieval performance of a patient search system. We thoroughly evaluate these approaches and investigate how to effectively combine them. The four components of our framework are as follows:

1. **Negation Handling:** The first component of the framework aims to handle negated language in patient search by uncovering the context (i.e. positive or negative) of the medical conditions stated in each sentence in the medical records and queries. To instantiate this component, we propose novel approaches that consider the contexts of terms when matching the medical records with a query. In addition, our approaches demote patients whose records have medical conditions occurring in the opposite context of the query.

   Traditional patient search approaches in the literature simply ignore the negated language in the queries and medical records, or disregard any medical conditions with a negated context during indexing and/or retrieval. In this thesis, we show that the accurate representations of the contexts of the medical conditions in the queries and medical records enable a search system to more effectively retrieve patients whose medical records contain the query medical conditions with the right context. Furthermore, we propose a supervised approach that learns how to handle negated language in medical records and queries.

2. **Conceptual Reasoning:** The second component uncovers relationships between medical conditions. At the record level of the retrieval process, these uncovered relationships are leveraged to infer the relevance of patients towards a given query. To instantiate this component, we introduce novel approaches for leveraging association rules of relationships between medical conditions. These rules, extracted from several existing medical resources, permit the discovery of related medical conditions and thereby enhance the patient retrieval process.

   Existing works show inconsistent results when using medical resources to infer the medical conditions related to the query in patient search (Zuccon *et al.*, 2012). In addition, there are limited studies on deploying medical resources to infer associated medical conditions and their appropriate weighting. We propose approaches that weight the association rules representing the relationships between medical conditions using either Bayesian probabilities or random walks on the relationship graph, to allow effective medical condition inference from medical resources.

3. **Department Expertise:** The third component extracts knowledge gained from sets of medical records issued from individual hospital departments and promotes patients whose medical records are issued from departments that have expertise in the medical conditions stated in the query. For instance, patients having records from a hospital's cardiology department are likely to be relevant to a query such as "find patients suffering from heart attack". To instantiate this component, we investigate an adaptation of aggregate ranking approaches to measure the relevance of hospital departments towards a given query. Then, when scoring records issued by a given department, we take into account the relevance of that department.

4. **Inclusion Criteria Coverage:** The fourth component infers the relevance of a patient by taking into account the relevance of his/her associated medical records with respect to each of the inclusion criteria (i.e. medical conditions) stated in the query. We introduce a novel approach to rank highly patients whose medical records are relevant to all of the medical conditions stated in the query. For example, when searching for patients with several medical conditions (e.g. "find patients with diabetes and hypertension"), intuitively the patients who have all of the medical conditions should be ranked higher. In this thesis, we investigate the adaptation of a coverage-based search result diversification technique to promote patients who are relevant to all of the medical conditions stated in the query.

1.5 Origins of Material

The material in this thesis is based on a number of conference publications:

- Chapter 4: The overall structure of the proposed framework was published in the doctoral consortium of SIGIR 2013 (Limsopatham, 2013), while our proposed approach that retrieves patients based upon the medical conditions related to the four medical aspects was introduced in ECIR 2013 (Limsopatham, Macdonald & Ounis, 2013*a*) and OAIR 2013 (Limsopatham, Macdonald & Ounis, 2013*b*). In addition, we initially proposed to adapt voting techniques to rank patients based on the relevance of their medical records in TREC 2011 (Limsopatham *et al.*, 2011).
- Chapter 5: The proposed approaches to handle negated language when ranking patients are based on work published in TREC 2011 (Limsopatham *et al.*, 2011), SIGIR 2012 (Limsopatham *et al.*, 2012) and CIKM 2013 (Limsopatham, Macdonald & Ounis, 2013*c*).
- Chapter 6: The approach to leverage association rules extracted from existing medical resources to infer the relevance of a patient was initially published in OAIR 2013 (Limsopatham, Macdonald & Ounis, 2013*b*).
- Chapter 7: The approach to take into account the relevance of hospital departments when ranking patients was published in ECIR 2013 (Limsopatham, Macdonald & Ounis, 2013*e*).
- Chapter 8: Our approach to model the relevance towards multiple medical conditions in a given query was published in CIKM 2014 (Limsopatham *et al.*, 2014).
- Chapter 9: The learned approaches based on a classification and a regression technique to combine relevance scores generated using different techniques were initially published in CIKM 2013 (Limsopatham, Macdonald & Ounis, 2013*d*) and SIGIR 2013 (Limsopatham, Macdonald, McCreadie & Ounis, 2013), respectively.

1.6 Thesis Outline

The remainder of this thesis is organised as follows:

- Chapter 2 provides a background on the important IR concepts used in this thesis. Specifically, we first introduce the architecture of a basic IR system, which includes indexing, retrieval, and text operations where documents and queries are parsed and processed before indexing and retrieval. Then, we discuss query expansion approaches that aim to improve the representation of the query to represent the searchers' information needs more effectively. In addition, we discuss more advanced research in IR that is used or extended in the later chapters of this thesis, including machine learning for IR and aggregate ranking. We then discuss the evaluation of an IR system in the context of the Text REtrieval Conference (TREC).

- Chapter 3 discusses existing works on IR in the medical domain. In particular, we describe the characteristics of medical documents and the existing approaches to rank medical documents. In addition, we identify the knowledge gap that this thesis addresses, which is how to uncover implicit knowledge and improve the representations of queries and medical records within a patient search system. We then introduce the TREC Medical Records track, which provides the testbed used for evaluating the patient search approaches that are proposed in this thesis.

- Chapter 4 describes the proposed framework, which defines four components for dealing with the implicit knowledge problem in patient search. Using the test collection provided by the TREC Medical Records track (discussed in Chapter 3), we investigate and identify a baseline patient search system that our framework builds upon. We empirically investigate each of the four components of the framework in Chapters 5, 6, 7 and 8.

- Chapter 5 instantiates the Negation Handling component by introducing supervised and non-supervised approaches that demote patients whose records have medical conditions occurring in the opposite context of the query. We evaluate the effectiveness of the proposed approaches and analyse the queries for which this component is particularly effective.

- Chapter 6 investigates the second component of the framework, namely Conceptual Reasoning. In particular, we propose and thoroughly evaluate two approaches that leverage the association rules of relationships between the medical conditions obtained from existing medical resources, to infer the relevance of patients.

- In Chapter 7, we tackle Department Expertise component of the proposed framework. We introduce approaches that promote patients whose medical records are issued from the hospital departments that have expertise in the medical conditions stated in the query, by aggregating the medical records associated to each hospital department. This inter-record evidence is evaluated using approaches based on a federated search and expert search.

- In Chapter 8, we examine the last component of our framework (i.e. the Inclusion Criteria Coverage component). In particular, we introduce an approach for promoting patients whose medical records are relevant to multiple medical conditions stated in a query. We propose a probabilistic approach that adapts a technique from coverage-based search result diversification (Agrawal *et al.*, 2009; Carbonell & Goldstein, 1998; Santos *et al.*, 2010*a*) to measure the relevance towards each of the query's medical conditions, and rank highly the patients whose medical records are likely to be relevant to all of those medical conditions. We empirically evaluate the proposed approach and analyse the queries for which our approach is likely to be effective.

- In Chapter 9, we investigate methodologies for combining the approaches examined within the framework, which uncover the implicit knowledge at the three levels of the retrieval process. In particular, we employ data fusion (Shaw & Fox, 1994) (e.g. CombSUM) and machine learning techniques (Liu, 2009) (e.g. classification, regression and learning to rank) to combine the rankings produced by those approaches.

- Chapter 10 closes this thesis by highlighting the contributions and the conclusions drawn from each of the individual chapters. Finally, we discuss possible research directions for future work.

Chapter 2

Information Retrieval

2.1 Introduction

This chapter provides an overview of information retrieval techniques that this thesis relies on. Indeed, we describe existing works related to our framework that improves the retrieval effectiveness of a patient search system, and position our work in the literature. The remainder of this chapter is organised as follows:

- Section 2.2 provides the background of the text operation processes for documents and queries before building an index and retrieving documents.
- Section 2.3 describes the indexing process, which enables quick access to the information items in a collection. Specifically, we discuss the indexing strategies and the index structure that are used in our IR system.
- In Section 2.4, we explain in detail the retrieval process used to match documents to a query. In particular, we discuss approaches for matching and ranking documents to a given query.
- Section 2.5 describes techniques to reformulate the queries to better describe the information needs of searchers. We describe query expansion techniques, which leverage either the statistics obtained from the index or external resources.
- Section 2.6 discusses the adaptation of techniques from the field of machine learning (e.g. classification and regression) to improve effectiveness of an IR system.
- In Section 2.7, we discuss related work on aggregate search, which aims to rank groups of documents representing particular entities. For instance, in enterprise search, an expert search system is used to rank experts in a particular area, based on the documents related to them. Specifically,

in Chapter 4, we examine the adaptation of several voting techniques (Macdonald, 2009) to rank patients based on the relevance of their medical records towards a query.

- Section 2.8 discusses the evaluation methodologies of an IR system. Importantly, we discuss the Text REtrievel Conference (TREC) (Voorhees & Harman, 2005) that provides testbed for facilitating IR research. We use a test collection provided by TREC to evaluate the framework proposed in this thesis.

2.2 Text Operations

Text operations transform documents and queries before indexing and retrieval, so that they can be processed efficiently and effectively within an IR system (Baeza-Yates & Ribeiro-Neto, 1999). In particular, performing text operations on both documents and queries can be viewed as controlling the vocabulary used during indexing and retrieval, as documents and queries are pre-processed and represented using the extracted terms, which may be stemmed, while some non-informative terms may also be removed. The remainder of this section discusses text operation techniques that are commonly used in an IR system. We use the following sentences from a medical record from the medical transcription samples website[1] as an example document:

```
"He presents to the ER today with hematuria that began while he was
sleeping last night. He denies any pain, nausea, vomiting or diarrhea."
```

2.2.1 Tokenisation

Tokenisation is a process of extracting *words* from texts in documents and queries, based on boundary conditions, such as a white-space character and punctuations. In addition, these formed words are normally converted to lower-case. After the tokenisation operation, the example document can be represented as follows (we use '|' to represent a separation between words):

```
"he|presents|to|the|er|today|with|hematuria|that|began|while|he|was|
sleeping|last|night|he|denies|any|pain|nausea|vomiting|or|diarrhea"
```

2.2.2 Stopword Removal

Words that are used in many documents in the collection are less effective for identifying relevant documents (Baeza-Yates & Ribeiro-Neto, 1999; Croft *et al.*, 2009; Fox, 1992). Indeed, common words, such

[1] http://www.medicaltranscriptionsamples.com/

as, a and the, appear in nearly all documents; hence, they cannot be used to distinguish between relevant and non-relevant documents. Such words, which are frequently referred to as *stopwords*, are not useful for an IR system and are therefore filtered out when indexing or retrieving documents (Fox, 1992; Lo *et al.*, 2005). A stopword list can be constructed by using top *k* most frequent words in the collection. However, more often the stopword list is pre-defined manually. Stopwords may contain articles, prepositions and conjunctions. In addition, stopwords removal can result in a marked reduction of the index size (Baeza-Yates & Ribeiro-Neto, 1999), which leads to efficient retrieval. After a stopword removal process, the example document is reduced as follows:

```
"present|er|today|hematuria|began|sleeping|last|night|denies|pain|
nausea|vomiting|diarrhea"
```

2.2.3 Stemming

Words in human language have syntactical variations, which may prevent an effective match between the words in a query and a document (Baeza-Yates & Ribeiro-Neto, 1999). For example, alleviated, alleviation, alleviates and alleviating are variants of alleviate. A document containing the word “alleviating” is not exactly matched with the query “alleviation”. This problem can be lessened by the stemming process, which represents a word in both documents and queries using its normalised form (i.e. stem). Indeed, a stemming process reduces the different forms of words, which occur because of syntactical variations (e.g. plurals, gerunds, tenses), by using a set of transformation rules. Among the existing stemming algorithms, Porter’s stemmer (Porter, 1997) is the most popular (Croft *et al.*, 2009). Porter’s stemmer uses a set of rules for removing suffixes of words to transform words to their stems. After stemming using the Porter’s stemmer, the example document is then represented as follows:

```
"present|er|today|hematuria|began|sleep|last|night|deni|pain|nausea|
vomit|diarrhea"
```

Once the documents and the queries are stemmed, words in the documents are used to build an index, which will be discussed in Section 2.3, while words in queries are used to retrieve documents using a process discussed in Section 2.4.

2.2.4 Thesauri and Controlled Vocabulary

Alternatively, apart from the terms in the documents and queries, the semantic meanings of terms can be used as words when indexing and retrieving documents (Baeza-Yates & Ribeiro-Neto, 1999). Such techniques are commonly used in search systems for specific domains, e.g. the medical domain. In particular, when representing words using the semantic meanings of terms, a thesaurus, which provides lists of words grouped together according to the similarity of the meanings, is used to provide a controlled-vocabulary for indexing and retrieving documents. Using a controlled vocabulary, phrases and synonyms of terms are represented as a single word (i.e. concept). For example, 'heart disease', 'cardiovascular disease' and 'CVD', which have the same semantic meanings, can be represented using the same concept within a search system. This leads to the normalisation of indexing concepts, where a clear semantic meaning of concepts is used when indexing. Meanwhile, the retrieval is based on concepts instead of terms, which could improve the performance of a search system (Baeza-Yates & Ribeiro-Neto, 1999). In the medical domain, the MetaMap tool (Aronson & Lang, 2010) is widely used to identify medical concepts based on the UMLS Metathesaurus[1] in documents and queries (Aronson, 1994; Aronson & Lang, 2010; Demner-Fushman *et al.*, 2011; King *et al.*, 2011; Qi & Laquerre, 2012). In Figure 2.1, we show the result of using the MetaMap tool to identify medical concepts from the phrase "He presents to the ER today with hematuria". Note that for readability, we show only parts of the example document processed using the MetaMap tool.

In this thesis, we investigate how a thesauri can be leveraged for the accurate representations of medical conditions in queries and medical records in Chapters 4, 6 and 8.

2.3 Indexing

Once the documents are pre-processed with text operations, they are used to create a suitable data structure (e.g. an inverted index) that enables fast access to the representations of those documents, where each document has a unique identifier (i.e. document-ID or docid). Typically, an integer is used as docid in an inverted index, since it requires less storage than a string. An *inverted index* is normally used to enable an efficient term-based access to the representations of documents in an IR system through the docids (Rijsbergen, 1979). Specifically, the inverted index contains term-posting lists, each of which can link to the representations of documents in which a term occurs. Table 2.1 shows an example of an inverted index, which contains three representations of documents. In this example, the term 'diarrhea' occurs once in document 1 and twice in document 3. Importantly, the following data structures may also be created during the indexing of documents to provide a quick access to additional information required during the retrieval process.

- *Lexicon* : a lexicon stores information about each term in the collection, which may include the number of occurrences of the term in the collection, the number of documents in the collection

[1] `http://www.nlm.nih.gov/research/umls/`

```
Input: "He presents to the ER today with hematuria."

Phrase: "He"

Phrase: "presents"

Meta Candidates (Total=2; Excluded=0; Pruned=0; Remaining=2)
   966   Present [Quantitative Concept]
   966   Present (Presentation) [Idea or Concept]
Meta Mapping (966):
   966   Present [Quantitative Concept]
Meta Mapping (966):
   966   Present (Presentation) [Idea or Concept]

Phrase: "to the ER today"
Meta Candidates (Total=2; Excluded=0; Pruned=0; Remaining=2)
   861   Today (ToDay brand) [Antibiotic,Organic Chemical]
   861   Today (Today (temporal qualifier)) [Temporal Concept]
Meta Mapping (861):
   861   Today (ToDay brand) [Antibiotic,Organic Chemical]
Meta Mapping (861):
   861   Today (Today (temporal qualifier)) [Temporal Concept]

Phrase: "with hematuria"
Meta Candidates (Total=1; Excluded=0; Pruned=0; Remaining=1)
  1000   Haematuria (Hematuria) [Finding]
Meta Mapping (1000):
  1000   Haematuria (Hematuria) [Finding]
```

Figure 2.1: Medical concepts identified by the MetaMap tool from a phrase ‘He presents to the ER today with hematuria.’

the term occurs in, and the identifier of the term (i.e. term-id). In addition, the lexicon provides a direct access to the posting list in the inverted index of each term.

- *Document Index* : a document index allows access to information regarding each document, including the document length and the frequencies of terms occurring in each document.

- *Direct/Forward Index*: a direct index stores term-related information, such as the occurrence statistics of terms within each document in the collection (Ounis *et al.*, 2006). The direct index enables a quick access to term frequencies within a document, which are required in some IR approaches, e.g. when performing query expansion.

term	posting-list (List(docid,term frequency))
present	(1,3),(2,4),(3,3)
er	(1,1),(3,1)
today	(1,1),(3,2)
hematuria	(1,3)
began	(1,1),(2,2)(3,2)
sleep	(1,3),(2,1)
last	(1,1),(2,1)
night	(1,2),(3,2)
deni	(1,2),(2,2),(3,4)
nausea	(1,2),(2,1),(3,1)
vomit	(1,1),(2,2),(3,1)
diarrhea	(1,1),(3,2)

Table 2.1: An example of an inverted index

2.4 Document Retrieval

An IR system produces a ranked list of documents according to their relevance towards a query. Ideally, an IR system ranks a document that is more likely to be relevant to a given query higher than other documents, so that searchers would find relevant documents at the top of the list. The ranked list of documents is produced using a *term weighting model*, which computes the relevance score of each document based on the occurrence of query terms. However, the relevance of a document is subjective and vary among searchers; hence, there is no perfect term weighting model (Voorhees & Harman, 2005). Term weighting models based on the probability ranking principle (PRP) (Robertson *et al.*, 1981) have been shown to be the most effective while having sound theoretical foundation. Indeed, term weighting models within the probability ranking principle aim to rank documents in the collection in the order of decreasing probability of relevance towards the searcher's query. The remainder of this section discusses two families of probabilistic models for estimating the relevance of a document.

2.4.1 The Best Match Weighting Model

From the probability ranking principle, Robertson *et al.* (1981) introduced the best match (BM) model, which calculates the weight of a term t, based on the number of documents in the collection (N), the number of documents containing t (N_t), the number of relevant documents that t occurs in (r), and the number of relevant documents in the collection (R), as follows:

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

However, the relevance information may not always be available. Croft & Harper (1988) simplified Equation (2.1) when the relevance information from the collection cannot be obtained, as follows:

$$w_s = \log \frac{N - N_t + 0.5}{N_t + 0.5} \tag{2.2}$$

Note that Equation (2.2) is similar to the inverse document frequency (IDF) component in the classical TF-IDF weighting model (Salton, 1971). To overcome the lack of relevance information problem, Robertson *et al.* (1981) combined the probability ranking principle with the 2-Poisson model (Harter, 1975). Indeed, according to Harter (1975)'s 2-Poisson distribution, terms occur randomly across the documents in the collection; however, they occur more densely in an elite set of documents, which are relevant to queries containing those terms. The well-known BM25 model, which deployed the approximation of the 2-Poisson model of term frequencies, can be calculated as follows:

$$score(d, q) = \sum_{t \in q} w_s \cdot \frac{(k_1 + 1) \cdot tfn}{k_1 + tfn} \cdot \frac{(k_3 + 1) \cdot qtf}{k_3 + qtf} \tag{2.3}$$

where k_1 and k_3 are parameters that control the saturation of the normalised term frequency tfn of term t in the document d, and the saturation of qtf, the term frequency of t in the query (i.e. the number of occurrence of t in the query), respectively. Importantly, tfn is computed as follows:

$$tfn = \frac{tf}{1 + b + b \cdot \frac{l}{avg_l}} \tag{2.4}$$

where tf is the number of occurrences of t in the document d, and avg_l is the average length of all documents in the collection. b is a parameter for the normalisation of document length l. Robertson *et al.* (1994) suggested to set b, k_1 and k_3 to 0.75, 1.2 and 1000, respectively. In Chapters 4, 5, 6, 7 and 8, we use BM25 as a baseline approach for ranking medical records based on their relevance towards a given query, before using an aggregate ranking approach (see Section 2.7) to combine the relevance scores to rank patients.

2.4.2 Divergence From Randomness

Amati (2003) introduced the Divergence from Randomness (DFR) framework for term weighting, which is also based on the previously discussed 2-Poisson model. The term weighting models in the DFR family are based on the idea that a term t is informative for a document d, if the term's distribution in d is different from the random distribution (i.e. the distribution of t in the entire collection). Within the DFR models, the relevance score of a document d is calculated as follows:

$$score(d, q) = \sum_{t \in q} Inf_1 \cdot Inf_2 \tag{2.5}$$

where Inf_1 and Inf_2 are the functions that measures the informative content of the term t related to the document collection and to the elite set of the term (i.e. the set of documents in which a term t occurs), respectively. Inf_1 and Inf_2 can be calculated as:

$$Inf_1 = -\log_2 Prob_1(tf|Collection) \tag{2.6}$$

$$Inf_2 = 1 - Prob_2(tf|E_t) \tag{2.7}$$

where Inf_1 uses $Prob_1$ to measure the probability that the term frequency of term t is tf by chance. On the other hand, Inf_2 measures the information gain based on the elite set of document E_t by considering the occurrence of the term t in the document d. The term t is informative if it occurs rarely in the collection but, in contrast, appears frequently in a particular set of document E_t.

In this thesis, we use the DPH weighting model from DFR to rank the patients' medical records towards a given query. DPH is a parameter-free model, where all the parameters are derived directly from the collection. DPH calculates the relevance score of a document d for a query q as follows:

$$score_{DPH}(d, q) = \sum_{t \in q} \left(\frac{(1-F)^2}{tf+1} \cdot (tf \cdot \log_2(tf \cdot \frac{avg_l}{l} \cdot \frac{N}{TF})) + 0.5 \cdot \log_2(2 \cdot \pi \cdot tf \cdot (1-f)) \right) \tag{2.8}$$

where $F = \frac{tf}{l}$, tf is the term frequency in the document d, l is the length of d, avg_l is the average length of documents in the collection, N is the number of documents in the collection, and TF is the term frequency across all documents in the collection.

2.5 Query Expansion

As discussed in Section 2.4, an IR system produces the ranked list of documents based on their relevance towards a given searcher query. However, an IR system may not retrieve documents that satisfy the information need of the searcher. In this case, the searcher may reformulate the query and resubmit it to the IR system. This process can be continued until the searcher finds a relevant document or the searcher abandons the IR system. To alleviate the searchers' dissatisfaction, query expansion approaches (e.g. Amati (2003); Bendersky & Croft (2008); Voorhees (1994)) aim to automatically modify or rewrite the original query to better match the vocabulary of relevant documents. Importantly, query expansion approaches aim to alleviate vocabulary mismatch between the relevant documents and a query (Amati, 2003; Bendersky & Croft, 2008; Diaz & Metzler, 2006; Voorhees, 1994). Next, in Sections 2.5.1

and 2.5.2, we discuss query expansion approaches using pseudo-relevance feedback and information from external resources, respectively.

2.5.1 Pseudo-Relevance Feedback

Pseudo-relevance feedback (PRF) leverages the local statistics of the top ranked documents initially retrieved using the original query to expand the query. Indeed, PRF assumes that the top ranked documents (i.e. pseudo-relevant documents) are relevant; hence, they are useful for expanding the query to retrieve more relevant documents. Importantly, informative terms are extracted from the pseudo-relevant documents based on the statistics of their occurrence in the pseudo-relevant documents and/or the statistics of their occurrence in the entire collection.

In this thesis, we deploy the DFR Bo1 weighting model (Amati, 2003) to rank the informativeness of terms in the pseudo-relevant documents. DFR Bo1, which is based on the Bose-Einstein statistics, measures the informativeness of a term based on the divergence of the term's distribution in the pseudo-relevant feedback documents from its distribution in the entire collection. The more the divergence is measured, the more likely that the term is informative and related to the query. Specifically, Bo1 from the Divergence from Randomness framework calculates the weight $w(t)$ of a term t in the pseudo-relevant documents, as follows:

$$w(t) = tf_x \cdot log_2 \frac{1 + P_n}{P_n} + log_2(1 + P_n) \tag{2.9}$$

where tf_x is the frequency of term t in the set of pseudo-relevant documents x. $P_n = \frac{TF}{N}$, where TF is the term frequency of t in all documents in the collection, and N is the number of documents in the collection. In practice, the number of pseudo-relevant documents ranges from 3 to 10, while the number of expanded terms should be more than the number of the pseudo-relevant documents (Amati, 2003). In this thesis, we follow Amati (2003) and set the number of pseudo-relevant documents and expanded terms to 3 and 10, respectively.

Once we identify the expanded terms, a parameter-free function is used to calculate the query term weight $qtw(t)$ of each term t in the expanded query, as follows (Amati, 2003):

$$\begin{aligned} qtw(t) &= \frac{qtf}{qtf_{max}} + \frac{w(t)}{\lim_{TF \to tfw} w(t)} \\ &= TF_{max} \log_2 \frac{1 + P_{n,max}}{P_{n,max}} + \log_2(1 + P_{n,max}) \end{aligned} \tag{2.10}$$

where qtf is the frequency of a term t in the query, and qtf_{max} is the maximum of qtf. $\lim_{TF \to tfw} w(t)$ is the upper bound of $w(t)$, TF_{max} is the frequency TF of the term with the maximum $w(t)$ in the pseudo-relevant documents. $P_{n,max} = \frac{TF_{max}}{N}$. Note that if an original query term t does not appear in the most informative term list extracted from the pseudo-relevant documents, its query term weight $qtw(t)$ remains equal to the original one.

2.5.2 Query Expansion using External Resources

External resources (e.g. WordNet (Miller *et al.*, 1990), UMLS Metathesaurus, MeSH[1]) have been used to improve the representation of the query in searching medical documents. However, existing works report inconsistent results of deploying external resources in query expansion. For example, Hersh *et al.* (2000) proposed to expand a query with terms derived from UMLS Metathesaurus and weight those terms equal to the original query terms. Indeed, they investigated different approaches to derive terms using the hierarchical structure of the UMLS Metathesaurus (e.g. synonyms and parent-children relationships). However, their experiments with the OHSUMED test collection (Hersh, Buckley, Leone & Hickam, 1994) showed the degradation in the retrieval performance in most cases. Nevertheless, they showed that query expansion using external resources is useful in some specific cases. Therefore, when to effectively apply the query expansion is an important question to be tackled. Meanwhile, Zhou *et al.* (2008) effectively expanded a medical query by using PubMed's Automatic Term Mapping tool[2] to identify medical terms in the query. They expanded the medical terms with synonyms, hypernyms (more generic terms), hyponyms (more specific terms), and variants of the terms (e.g. abbreviations), which are derived from a number of medical resources (e.g. Entrez Gene[3], UMLS Metathesaurus and FDA[4]. Indeed, the suitability of external resources used for query expansion for the document collection highly impacts the retrieval performance (Stokes *et al.*, 2009).

In patient search, several participating groups at the TREC Medical Records track applied query expansion using external resources. King *et al.* (2011) improved the retrieval performance by deriving related terms from the UMLS Metathesaurus by focusing on particular types of related terms, such as synonyms and more specific terms. In contrast, Qi & Laquerre (2012) reported that query expansion with synonyms and hyponyms from the UMLS Metathesaurus decreased the retrieval performance. In Chapter 6, we investigate effective approaches that derive expanded terms and their term weight from association rules extracted from external resources to improve the representation of the queries.

[1] https://www.nlm.nih.gov/mesh/

[2] http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_040.html

[3] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

[4] http://www.fda.gov/cder/drugsatfda/datafiles/default.html

2.6 Machine Learning for IR

Machine learning research aims to automatically detect patterns in data, and use these patterns to describe the data or predict unseen data (Murphy, 2012). Machine learning is usually divided into two types: *supervised* and *non-supervised* learning. Supervised (or predictive) learning approaches learn a mapping from an input x to an output y, given a training set of data that contains examples of mapping between inputs and outputs. For example, using examples of spam and non-spam emails to predict whether a new email is a spam. On the other hand, non-supervised (i.e. descriptive) learning approaches aim to discover patterns in a set of data. For example, to discover a set of terms that searchers normally use for retrieving particular web-pages from a query log. These patterns are learned using *features* (i.e. attributes or covariates), which are some properties about the input data. Machine learning approaches have been used in many tasks related to information retrieval, such as document categorisation (Salles *et al.*, 2010), spam filtering (Cormack *et al.*, 2011) and learning to rank (Liu, 2009).

In the remainder of this section, we provide background of supervised machine learning approaches that our work is developed upon.

2.6.1 Classification

The goal of classification is to learn a mapping from an input x to an output y, where y is in a set Y. $|Y|$ is the number of classes. For example, a spam classifier predicts whether a web-page (i.e. an instance) is a spam or not (i.e. $|Y| = 2$). Indeed, classification is a supervised machine learning that uses a set of labelled training data to learn a model to predict the class of an instance of an unseen (i.e. test) data. Different classification approach such as, decision trees, instance-based learning and support vector machines (SVM) have been proposed. Decision trees use a tree-like structure, where each node in the decision trees represents a feature in an instance to be classified. Instances are classified starting at the root node through the leaf nodes, which are the resultant classes. An example of algorithms to build decision trees is C4.5 (Quinlan, 1993). Instance-based learning approaches, such as the nearest neighbour algorithm (Aha, 1997) avoid training the classification model beforehand, but instead perform the induction process to classify an instance by comparing a new instance to the training instances to find the closest match.

In this thesis, we use a decision trees-based technique called *Gradient Boosted Regression Trees (GBRT)* (Tyree *et al.*, 2011) (as implemented in the jforests package Ganjisaffar *et al.* (2011)[1]), as it has been shown to be effective for several IR tasks (e.g. Ganjisaffar *et al.* (2011); Tyree *et al.* (2011)).

[1] `http://code.google.com/p/jforests`

GBRT is a machine learning technique that is based on tree averaging (i.e. averaging the regression scores computed from each tree). To train the model, for each iteration, a new tree that could lessen the remaining errors of the current model is added to the model. We deploy GBRT to selectively chose when to apply a particular ranking function on a per-query basis in Chapter 9.

2.6.2 Regression

Regression is a special type of classification, where the outputs in Y that a learned model aims to map to is a continuous number. An example of a regression task is to predict an appropriate term weight for a query term using a set of training query (Lease *et al.*, 2009). Indeed, any classification approaches can be adapted for a regression task.

We also deploy the GBRT for regression tasks in our approaches proposed in this thesis. Specifically, in Chapter 5, GBRT is used to predict a penalising weight to demote non-relevant patients. We use the GBRT to learn the importance of the expanded terms when applying a query expansion technique in Chapter 6. We learn the level of emphasis on the relevance of departments when ranking patients whose medical records are issued from those departments in Chapter 7. In Chapter 8, we learn how to effectively promote the relevance towards medical conditions stated in a query. Finally, in Chapter 9, we use GBRT to learn an effective combination of rankings produced by different ranking approaches.

2.6.3 Learning to Rank

Learning to rank (LTR or LETOR) is a machine learning technique tailored for information retrieval. Many factors (e.g. the reliability and the completeness of information items) impact the perception of searchers about the relevance of a document. As a result, these factors may be taken into account when designing an algorithm to effectively rank information items. LTR enables an IR system to take into account these factors as features when learning a ranking model. Typically, LTR re-ranks an initial ranking (the so-called *sample*) retrieved using an IR approach to promote documents with the desired factors to the top ranks.

LTR can be categorised into three types based on its evaluation objective (Liu, 2009). First, *point-wise* approaches learn to judge the relevance of each document individually. For example, Fuhr (1989) built a ranking function using least squares polynomial (LSP) to estimate the relevance of each document. Second, *pair-wise* approaches learn a model that decides which one of a pair documents should be ranked higher. For instance, the RankNet (Burges *et al.*, 2005) algorithm uses the cross entropy (Boer *et al.*, 2002) as a loss function when ranking a ranking of a pair of documents. Third, *list-wise* approaches optimise a retrieval measure by considering the whole ranking (see Section 2.8.2). For example, AdaRank (Wu *et al.*, 2010) applies a boosting method and an exponential loss function when optimising a targeted retrieval measure of a ranking. Liu (2009) and Macdonald *et al.* (2013) suggested that LTR based on the list-wise paradigm is more effective than the others. In addition, the size of the sampled documents and the used measure of training the learned model greatly impact the retrieval performance (Macdonald *et al.*, 2013). In this work, we deploy a list-wise approach to learn a ranking model that considers relevance based on different components of our framework. Specifically, in Chapter 9, we investigate the adaptation of LTR approaches, including AFS (Automatic Feature Selection) (Metzler, 2007), AdaRank (Xu & Li, 2007) and LambdaMART (Wu *et al.*, 2008) to combine rankings produced by different approaches of our framework.

2.7 Aggregate Ranking

Beyond document retrieval, the IR community has investigated the retrieval of entities based on the relevance of their associated documents (i.e. aggregate ranking). Existing approaches for ranking entities use a set of documents related to each entity to represent that entity. For example, in expert search, an IR system aims to find persons who are experts in the topic expressed as a query, given a collection of documents related to a set of persons. A person is not represented as a single document, but instead a set of documents related to a particular person are used to infer the person's expertise. In the remainder of this section, we discuss aggregate ranking approaches for expert search, which we adapt for the patient search task that aims to rank patients based on the relevance of their medical records towards a given query.

Following Balog *et al.* (2006), we categorise aggregate ranking approaches into two groups: *Model 1* and *Model 2*. First, Model 1-based ranking approaches rank persons (i.e. candidates) based on the relevance of their corresponding *virtual document*. A virtual document is a concatenation of documents associated to a particular candidate. A virtual document can be represented as a language model of terms within documents associated to each candidate, before using a language model for IR (Hiemstra, 2001) to rank the virtual documents. On the other hand, ranking approaches within the Model 2 paradigm calculate the relevance score of a candidate by combining the relevance scores of documents associated to that candidate, by using an aggregate function. For example, Balog *et al.* (2006) summed up the relevance scores, computed using a language model (Hiemstra, 2001), of documents associated to a candidate, as the relevance score of the candidate. Macdonald & Ounis (2006) proposed the Voting Model, where documents vote for the relevance of their associated candidates. Different voting techniques, which are inspired by data-fusion techniques, can be used to aggregate the relevance scores of candidates. For example, the CombSUM voting technique calculates the relevance score of a candidate c for query q, as follows:

$$score\_candidate_{CombSUM}(c,q) = \sum_{d \in R(q) \cap profile(c)} score(d,q) \tag{2.11}$$

where $score(d,q)$ is the relevance score of document d, $R(q) \cap profile(c)$ is the set of documents associated to the candidate c that are also in the ranking $R(q)$, which is initially retrieved using a retrieval model, such as, BM25 and DFR DPH. Note that the Model 2 proposed by Balog *et al.* (2006) is equivalent to using the CombSUM voting technique to aggregate the relevance scores of documents ranked using a language model (Macdonald & Ounis, 2011).

In this work, we investigate how to effectively adapt aggregate ranking techniques from expert search in the patient search task. Beside the virtual document technique in Model 1, and the CombSUM voting technique in the Model 2 paradigm, in Chapter 4 we also examine other voting techniques that have been shown to be effective (Macdonald, 2009; Macdonald & Ounis, 2006), including CombMAX, expCombSUM and expCombMNZ. The later three voting techniques calculate the relevance score of a candidate c, as follows (Macdonald, 2009):

$$score\_candidate_{CombMAX}(c,q) = \max_{d \in R(q) \cap profile(c)} score(d,q) \tag{2.12}$$

$$score\_candidate_{expCombSUM}(c,q) = \sum_{d \in R(q) \cap profile(c)} e^{score(d,q)} \tag{2.13}$$

$$score\_candidate_{expCombMNZ}(c,q) = |R(q) \cap profile(c)| \cdot \sum_{d \in R(q) \cap profile(c)} e^{score(d,q)} \tag{2.14}$$

While the approaches based on Model 1 and the CombSUM voting technique equally weight the importance of the relevance of each document, CombMAX, expCombSUM and expCombMNZ put more emphasis on highly ranked documents, when estimating the relevance of candidates. Specifically, CombMAX uses the highest relevance score of documents associated to a particular candidate, as the relevance score of that candidate, while expCombSUM and expCombMNZ give more emphasis on the documents with high relevance scores. Importantly, the expCombMNZ voting model also takes into account the number of voting documents for each candidate in $R(q)$.

Later, in Chapter 4, we investigate which aggregate ranking approaches are effective for the patient search task, while in Chapter 7, we propose to extend existing voting techniques to take into account the hospital department that issues each medical record when ranking patients.

2.8 Evaluation

We have discussed the fundamental components of an IR system. Next, we describe the evaluation of the performance of an IR system based on the relevance of the retrieved documents for a given query. In general, IR research is evaluated using a test collection, which consists of (1) a set of documents, (2) a set of queries expressing the information needs of the searchers, and (3) a set of relevance judgements indicating which documents in the collection are relevant to each of the queries.

2.8.1 Test Collections

Classical IR evaluation is based on the Cranfield experiment paradigm (Cleverdon, 1962), where the relevance judgements are based on manually assessing a document to mark whether it is relevant or not to a given query. Note that the Cranfield experiments used full relevance judgement on a small collection, where every document was judged for each query. Recently, the number of documents used in a test collection has markedly increased (e.g. the medical records test collection used in this thesis consists of 101,711 medical records (Voorhees & Hersh, 2012; Voorhees & Tong, 2011)) to reflect information access in reality; as a result, having complete relevance judgement is not practical. To avoid judging all documents in the collection, a *pooling* technique is used to create a small subset of documents representing samples of relevant documents to be assessed (Sparck Jones & van Rijsbergen, 1975). Specifically, a sample pool is created from a union of documents retrieved using different IR systems and (maybe) some randomly sampled documents. However, the number of documents in the sample pool may still exceed the availability of resources for making relevance judgements, in which case, some techniques (e.g. selecting only documents within particular rank cut-off) are deployed when selecting documents from the pool for assessing.

The Text REtrievel Conference (TREC) (Voorhees & Harman, 2005) is a series of IR workshops for advancing information retrieval technologies organised by the National Institute of Standards and Technology (NIST) and the Disruptive Technology Office of the U.S. Department of Defence. TREC has provided test collections for supporting IR research on different topics of interests (e.g. web search, genomics search and patient search). The test collections provided by TREC consist of a set of documents, a set of queries, and a set of relevance judgements. The relevance judgements from TREC are created using the aforementioned pooling method, where all of the participating groups contribute to the pool a ranking of the documents retrieved from the provided set of documents. The ranking of documents is generated by a system (so-called *a run*) of a participating group. Note that a particular group can submit several runs. In general, only the top N (e.g. 100) documents retrieved in each run are judged by the human assessors, which are later used as the gold-standard relevance judgements. Ideally, the runs should be generated from multiple IR systems, so that the documents in the pool are diverse enough for producing complete relevance judgements.

2.8.2 Evaluation Measures

Next, we discuss the evaluation measures usually used to evaluate an IR system, given a test collection. For a particular task, suitable evaluation measures are closely correlated with the searchers' satisfaction. Indeed, most measures are derived from *precision* and *recall*. Precision is the proportion of the relevant documents retrieved and the numbers of the retrieved documents, while recall is the proportion of relevant documents retrieved compared with the number of relevant documents in the collection.

In this thesis, we use precision at 10 (P10), bpref (Buckley & Voorhees, 2004), infNDCG (Yilmaz *et al.*, 2008) and infAP (Yilmaz *et al.*, 2008) to evaluate our proposed approaches, as they focus on precision. Indeed, precision is important for a patient search system, as incorrectly recruiting patients whose medical conditions are not relevant to the query can be harmful for them. In particular, these measures are officially used for the TREC Medical Records track (Voorhees & Hersh, 2012; Voorhees & Tong, 2011). P10 is the precision of the top 10 retrieved documents, while the others are the measures that are designed to cope with incomplete judgements. Specifically, even though the pooling technique is used to create a reasonable sample for relevance judgements, the size of the pool may still be too big for the available resource for making the relevance judgements. Importantly, the relevance judgement for the used patient search collection is incomplete (Voorhees & Hersh, 2012; Voorhees & Tong, 2011), as the medical experts assessing the relevance are limited. Indeed, the bpref measure deals with the incompleteness of relevance judgement by penalising an IR system that ranks non-relevant documents higher than relevant documents. For a query with R relevant documents, bpref is calculated as follows (Buckley & Voorhees, 2004):

$$bpref = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{R} \tag{2.15}$$

where r is a relevant document in the ranked list, $|n$ ranked higher than $r|$ is a number of judged non-relevant documents that are ranked higher than r.

On the other hand, infNDCG and infAP use a sample technique to infer the MAP (Mean Average Precision) and nDCG (normalised Discounted Cumulative Gain) measures (Yilmaz *et al.*, 2008). Indeed, MAP is measured by averaging all of the precision values calculated after each relevant document is retrieved. nDCG uses graded judgement, where a document can be judged using different levels of relevance e.g. highly relevant, relevant and non-relevant, which rewards an IR system that ranks highly relevant documents at the top of the ranked list.

2.9 Conclusions

We have discussed a comprehensive background of fundamental information retrieval research that this work builds upon in order to investigate how existing IR techniques can be extended and applied for the patient search task. Indeed, we started by discussing the key components of an IR system, including text operations, indexing and document retrieval in Sections 2.2, 2.3 and 2.4. Specifically, we discussed text operation techniques, such as tokenisation, stopword removal, and stemming, which are typically performed on both documents and queries as a preparation process before indexing and retrieval. Then, in Section 2.3, we discussed the indexing data structures, such as inverted index, that allow efficient access to representatives of documents within an IR system. Section 2.4 provided an overview of probabilistic retrieval models, where the retrieval approaches from the best match (BM) and the divergence from randomness (DFR) families were introduced. Next, Sections 2.5 and 2.6 described advanced approaches to rank documents. Specifically, query reformulation approaches, including pseudo relevance feedback and semantic-based query expansion, that enhance the representation of a given query were explained in Section 2.5, while machine learning based approaches, including classification, regression and learning to rank, which are typically used in IR research, were discussed in Section 2.6. Going beyond document ranking, we introduced aggregate ranking approaches, which aim to rank aggregates of documents in Section 2.7. For example, to rank patients based on the relevance of their medical records towards a given query, the relevance of a patient is inferred by the relevance of the aggregate of the patient's medical records towards the query.

In the next chapter, we discuss how existing IR approaches have been deployed to search documents in the medical domain and identify the knowledge gaps that are addressed in this thesis.

Chapter 3

Searching in the Medical Domain

3.1 Introduction

In the previous chapter, we introduced the fundamental concepts of an information retrieval (IR) system, which includes components such as indexing and retrieval. In addition, we discussed advanced topics in IR, including query expansion, machine learning for IR and aggregate ranking. These provide background knowledge for the further development of an IR system specialised for the medical domain, which will be discussed in this chapter.

In this chapter, we present background knowledge for medical IR. According to Hersh (2008*a*), medical information can be categorised into two groups (namely, *knowledge-based information* and *patient-specific information*). The knowledge-based information is the medical knowledge that can be directly applied to several patients (e.g. treatment procedures). On the other hand, the patient-specific information details the health and medical conditions of an individual patient.

This thesis is primarily scoped within the boundaries of the searching of patient-specific information. We aim to search for patients having a medical history relevant to a given query. In this chapter, we discuss the existing problems in the literature and how existing algorithms and techniques deal with them. In addition, we identify knowledge gaps and position our work with respect to the literature. Note that we also discuss existing works in the searching of knowledge-based information, as they face many of the same problems. Importantly, research on the searching of knowledge-based information is more mature and, as will be shown later, several existing works in the searching of patient-specific information were adopted from approaches developed for the searching of knowledge-based information.

The remainder of this chapter is organised as follows:

- In Section 3.2, we define the medical information in the context of this thesis. Indeed, we further describe the knowledge-based information and the patient-specific information.

- Section 3.3 discusses current IR approaches for searching knowledge-based information.
- In Section 3.4, we introduce existing IR techniques for searching patient-specific information.
- Finally, in Section 3.5, we provide conclusions for this chapter.

3.2 Medical Information

Advances in information technology (IT) and medicine, and their intersection, has resulted in the rapid growth of medical information in the form of electronic documents (Hersh, 2008*a*). The development of electronic medical documents is driven by several factors, including increased emphasis on ensuring patient safety and preventing medical errors, as well as, advances in biotechnologies such as gene chips (Hersh, 2004, 2008*a*). Consequently, a massive amount of data has been generated that has also led to an explosion of new scientific knowledge (Hersh, 2008*a*). Moreover, the scientific literature is also another cause driving the growth of medical research data. Experiments identify new genes, diseases, and other biological processes that require further investigation. Furthermore, the literature itself becomes a source of experiments as they are knowledge that drives new hypotheses by researchers. On the other hand, hospitals started to migrate traditional medical records to *Electronic Medical Records (EMRs)* in order to reduce cost and improve the quality of medical care service. In addition, knowledge from EMRs can also be consulted when researching for diseases and treatments. As a consequence, there is an urgent need in IT to facilitate access to medical information.

As previously discussed in Section 3.1, medical information can be categorised into two groups, which are *knowledge-based information* and *patient-specific information* (Hersh, 2008*a*). Knowledge-based information is derived from observational or experimental research. It provides knowledge, which is ready to be applied to patients. For example, knowledge-based information from clinical research provides clinicians, administrators, and researchers with knowledge from experiments and observations, which can be directly applied to individual patients. The information is commonly provided in the form of books and journal papers.

On the other hand, patient-specific information, often in the form of medical records, provides information about the health and medical conditions of a particular patient, which could be helpful in clinical decision support to improve the healthcare safety and quality for patients. Medical records can be either in a format of structured, as in a laboratory value, or narrative text. For example, the history and physical reports contain both laboratory values and narrative text explaining the patient's condition. In particular, the medical records detail the clinical decision process related to a patient, covering symptoms, diagnostic tests, diagnoses and treatments. To enhance the quality of healthcare services, EMRs have been increasingly used worldwide, which has resulted in the creation of a tremendous numbers of EMRs (Hersh *et al.*, 2013). The emergence of EMRs is enabling healthcare practitioners to consult with existing medical practices to tailor a medical procedure to a specific patient (Hersh, 2008*a*).

3.3 Searching Knowledge-based Information

In order to deal with the large increase in the amount of medical information discussed in the previous section, advanced IT facilities have to be developed. One of the major needs is a technology to search for relevant pieces of information from a large set of documents. In this section, we describe the particular information needs underlying the searching of knowledge-based information, and how they have been handled in the context of a prominent information retrieval evaluation forum. Specifically, we discuss how existing works deal with the search of knowledge-based information, and how they are related to the work proposed in this thesis.

Along with the growth of the Internet and the Web, a phenomenal expansion of web-based medical document collections have been witnessed in the recent years. Knowledge-based information collections, such as PubMed[1], provide medical literature from journals and books. Searchers of these collections aim to retrieve documents pertaining to a specific medical scenario (Liu & Chu, 2007). For example, in diagnosing a potential lung cancer patient, a physician may use a query "lung cancer diagnosis" to search for the latest diagnosis techniques. In addition, the advances in information technology have increased the number of publicly available websites that provide information about healthcare and treatment (e.g. `http://www.patientslikeme.com/`, `http://www.mymediconnect.net/`). In particular, the number of users using search engines to search for information related to personal health has been growing. Hersh (2008*b*) reported that 80% of search engine users have searched for websites or documents related to their health condition. Moreover, about 98% of US physicians use the Internet to find documents related to healthcare (Hersh, 2008*b*).

To facilitate research in the searching of medical knowledge-based documents. TREC (see Section 2.8.1) introduced the TREC Genomics track (Hersh & Voorhees, 2009). The track pertains to the evaluation of techniques for the searching of knowledge-based information. In particular, four different tasks were introduced during the five years of the Genomics track, including *ad hoc retrieval*, *summarisation*, *text categorisation* and *question-answering*. However, we will discuss only the ad hoc retrieval and question-answering tasks, as they are related to the ranking of documents.

[1] `http://www.ncbi.nlm.nih.gov/pubmed`

```
  <MedlineCitation Owner="NLM" Status="Completed">
     <PMID>11830988</PMID>
     <ArticleTitle>[Amalgam. XI. Glass-ionomer as a possible substitute of
       amalgam: longevity]</ArticleTitle>
     <Abstract>
     <AbstractText>The clinical use of glass-ionomer increases, also for
       restorative goals. The longevity of glass ionomer restorations is
       among others determined by premature contact with saliva and by
       acid erosion. The adherence to the dental hard tissues may be increased
       by acid pretreatment. The longevity data presented here indicate
       that glass ionomer restorations do not last as long as amalgam restorations.
       In the deciduous teeth the material seems to be a more acceptable substitute
       for amalgam. The same holds true for restorations which are not submitted
       to stress, such as class V.
     </AbstractText>
     </Abstract>
     <Affiliation>Vakgroep Cariologie en Endodontologie, Academisch Centrum
     Tandheelkunde Amsterdam (ACTA), Louwesweg 1, 1066 EA Amsterdam.</Affiliation>
  <MeshHeadingList>
  <MeshHeading>
  <DescriptorName MajorTopicYN="N">Acrylic Resins</DescriptorName>
  <QualifierName MajorTopicYN="Y">standards</QualifierName>
  </MeshHeading>
  <MeshHeading>
  <DescriptorName MajorTopicYN="N">Dental Amalgam</DescriptorName>
  <QualifierName MajorTopicYN="Y">adverse effects</QualifierName>
  </MeshHeading>
  <MeshHeading>
  <DescriptorName MajorTopicYN="N">Dental Bonding</DescriptorName>
  </MeshHeading>
  <MeshHeading>
  <DescriptorName MajorTopicYN="N">Dental Cements</DescriptorName>
  </MeshHeading>
  </MeshHeadingList>
</MedlineCitation>
```

Figure 3.1: A sample snapshot of a document from TREC 2004 and 2005 Genomics track ad hoc retrieval task.

1. **Ad Hoc Retrieval:** The *ad hoc* retrieval task (Hersh *et al.*, 2004, 2005) represents the situation of a researcher tackling a new area and using an IR system to find relevant medical literature. The aim of the task is to find medical documents that are relevant to the search user's information needs described in the form of text. The medical documents of the *ad hoc* retrieval task consist of the collection of MEDLINE abstracts in ASCII text format from 1994 to 2003 (4,591,008 records) (Hersh *et al.*, 2004, 2005). An example of a document from the corpus is shown in Figure 3.1. Examples of queries are "Provide information about the role of the gene Interferon-beta in the disease Multiple Sclerosis" and "Generating transgenic mice".

2. **Question-Answering:** For the question-answering task, the information need is to find short, specific answers to questions provided in the query topics, instead of finding relevant documents as in the *ad hoc* task. In particular, the task focuses on the retrieval of short passages that specifically

address an information need from a corpus of full-text medical documents from journals (Hersh *et al.*, 2007). The topics were expressed as questions and the systems were evaluated on how well they retrieved relevant information. The queries contain one or more biological objects, processes, and some explicit relationships between them. An example of the queries is “What is the role of PrnP in mad cow disease?”.

From the examples of medical queries and a medical document (see Figure 3.1) previously shown, we can see the presence of medical terms e.g. ‘Interferon-beta’, ‘Multiple Sclerosis’, ‘transgenic mice’ and ‘glass-ionomer’. Indeed, medical terminology used in medical documents and queries is rich. Medical practitioners and researchers selectively use different medical terms to refer to a particular medical concept. For example, ‘multiple sclerosis’ can be referred to as ‘MS’ or ‘insular sclerosis’.

On the other hand, the ShARe/CLEF eHealth evaluation lab (Suominen *et al.*, 2013) was initiated in 2013. The evaluation lab consists of three tasks: (1) identification of disorders from clinical reports (2) mapping abbreviations in clinical reports to UMLS codes, and (3) information retrieval from a collection of certified health web pages. We discuss only the third task because it is related to the searching of knowledge-based information. In particular, the information retrieval task simulates the situation where a non-expert wants to find information about a particular medical condition on the Internet (Suominen *et al.*, 2013). For instance, after getting a discharge summary report, a patient wants to know more about his/her health conditions. The medical documents and queries provided by the ShARe/CLEF eHealth evaluation lab also contain many medical terms (Suominen *et al.*, 2013).

Traditional approaches in IR suffer from the query-document mismatch problem (Büttcher *et al.*, 2004), which is one scenario of the implicit knowledge issue discussed in Chapter 1. For example, if we are searching for treatments of lung cancer using the query “lung cancer treatment”, those systems may not be able to retrieve some relevant documents because in the medical documents the treatments of lung cancer are commonly referred to using specialised terms, such as “lung excision” or “chemotherapy”. In contrast, healthcare practitioners have knowledge about this kind of information and can effectively find the relevant documents. Importantly, approaches for both the *ad hoc* retrieval and the question-answering tasks are confronted with this problem. Meanwhile, for the information retrieval task of the ShARe/CLEF eHealth evaluation lab, the search users may not have the background knowledge of the medical domain, while the medical documents in a certified medical website are typically written by medical experts. Hence, it is crucial for a search system to deal with the different levels of the medical terms used in the medical documents and the queries. In the remainder of this section, we describe two groups of approaches that have been used to deal with the complexity of medical terminology.

3.3.1 Dealing with Medical Terminology using a Conceptual Representation

Medical terminology, which can be complex, inconsistent, and ambiguous, poses an important challenge when searching in the medical domain (Srinivasan, 1996; Trieschnigg *et al.*, 2010). For example, 'heart disease' can be referred to as 'coronary artery disease', 'coronary heart disease', or 'CHD'. This means that traditional search systems may not be able to retrieve medical documents relevant to a query, if those documents contain only synonyms of the query terms.

Zhou *et al.* (2008) proposed a conceptual retrieval approach and used domain-specific knowledge for the question-answering task. In particular, they used the query translation functionality of PubMed to extract gene names, biological objects, and biological processes (i.e. MeSH terms) in a query and defined them as concepts. Then, during document ranking, the similarity between a query and a document is determined by two components (namely, similarity of concepts and similarity of words). Document *D1* will be ranked higher than document *D2* only if the similarity of the concepts between the query and *D1* is more than that of the query and *D2*, or if the similarity of the concepts are the same for both *D1* and *D2*, the similarity of the words between *D1* and the query is more than that of *D2* and the query. In addition, they expanded the concept query (i.e. the identified medical terms) using external domain-specific resources, including Entrez Gene[1], MeSH, and the ADAM database of abbreviations. They reported that, either applying their proposed conceptual IR model alone, or with expansions using domain-specific knowledge, the performance of the system was improved significantly.

Aronson (1994) and Hersh, Hickam, Haynes & McKibbon (1994) proposed *bag-of-concepts* (BoC) approaches to represent the medical documents and queries using concepts from medical resources, such as MeSH and UMLS Metathesaurus[2]. Under these approaches, 'heart disease', 'coronary artery disease', 'coronary heart disease', and 'CHD', which share the same meaning are represented with the same concept. For instance, Aronson (1994) deployed MetaMap (Aronson & Lang, 2010) to identify medical concepts in medical documents and queries and represented them in the forms of the UMLS Concept Unique Identifiers (CUIs). Intuitively, these approaches should alleviate the terminology mismatch problem. However, empirical studies (Srinivasan, 1996; Trieschnigg *et al.*, 2010) have shown that the performance of BoC representations can be inconsistent, and sometimes outperformed by a traditional bag-of-words representation (BoW), since not all documents and queries could be effectively represented using medical concepts. For example, medical concepts may not be found in some queries. To cope with this problem, other works (e.g. Srinivasan (1996); Trieschnigg *et al.* (2010)) combined the

[1] https://www.ncbi.nlm.nih.gov/gene/

[2] http://www.nlm.nih.gov/research/umls/

relevance scores of both BoW and BoC, when inferring the relevance of a document. In particular, Srinivasan (1996) proposed the so-called *score combination* approach that linearly combines the relevance scores from both BoW and BoC, when inferring the relevance of a document d towards a query Q, as follows (Srinivasan, 1996):

$$score(d, Q) = \delta \cdot score_{BoW}(d, Q) + score_{BoC}(d, Q) \tag{3.1}$$

where δ is a parameter to emphasise the relevance score computed using BoW.

3.3.2 Query Expansion for Dealing Medical Terminology

Another area of works that deal with the complexity of medical terminology in the searching of knowledge-based information is how to effectively represent queries and documents using medical concepts, instead of terms.

Existing works introduced approaches to address the complexity of medical terminology by using domain-specific knowledge (Büttcher *et al.*, 2004; Fujita, 2004; Hersh *et al.*, 2004; Huang *et al.*, 2005). Indeed, most of the existing works made attempts to use domain-specific knowledge to cope with three major issues in medical IR, namely the frequent use of (possibly non-standardised) acronyms, the presence of homonyms (the same words referring to many meanings), and the possibility of synonyms (two or more words referring to one meaning) (Büttcher *et al.*, 2004). Within the context of this thesis, we argue that these three issues can be linked to the issue of implicit knowledge, since the information about acronyms, hyponyms and synonyms are commonly known by medical researchers, but they are not necessarily available to traditional IR systems. Most effective approaches aimed to expand medical queries with related terms from controlled vocabularies and ontologies (Büttcher *et al.*, 2004; Fujita, 2004; Hersh *et al.*, 2004; Huang *et al.*, 2005; Stokes *et al.*, 2007; Zhou *et al.*, 2008). For example, Fujita (2004), whose system achieved the best performance for the ad hoc task at TREC 2004 Genomics track, expanded query using domain-specific databases (i.e. LocusLink[1] and MeSH[2]), and applied pseudo-relevance feedback before using BM25 to rank documents. Also on the ad hoc retrieval task, Büttcher *et al.* (2004) tackled the issues of synonyms and acronyms. Since an acronym can be seen as a synonym (i.e. the abbreviated form of an acronym is used to refer to its corresponding long-form), the acronym and synonym issues were dealt with at the same time. In particular, Büttcher *et al.* (2004) expanded the acronyms, gene names/symbols, and protein names/symbols in a query using external referenced data from AcroMed[3], euGenes[4] and LocusLink. For each query, they selected the best 10 expansions, based

[1] http://www.ncbi.nlm.nih.gov/LocusLink

[2] http://www.ncbi.nlm.nih.gov/mesh

[3] http://medstract.med.tufts.edu/acro1.1/index.htm

[4] http://http://iubio.bio.indiana.edu/

alpha	a
beta	b
gamma	g
epsilon	e
I	1
II	2
III	3
receptor	r
gene	genetic
p	protein
mutation	mutant

Table 3.1: Examples of interchangeable pairs of replacements.

on the number of occurrences in the corpus, from those resources. The weight of the terms that were expanded were changed to 1.4 and the weight of the expansions were 0.95. The original terms with no expansion remained with weight 1.0. The expanded queries were used to rank medical documents based on the scores produced from BM25.

Lu *et al.* (2009) also reported that query expansion using external resources was helpful in their experiments with the TREC 2006 and 2007 Genomics track. They suggested to use the Automatic Term Mapping (ATM) provided by PubMed, which assigns different search tags to terms in a query. The PubMed's ATM functionality translates the query into the PubMed syntax. Then, they expanded the terms tagged with *[Text Words]*, *[MeSH Terms]*, and *[All Field]* using MeSH.

Zhong & Huang (2006) advised that the traditional keyword-based retrieval approaches failed in the medical domain because using one or a few keywords to represent a particular medical condition (i.e. medical concept) is not enough for retrieving relevant documents in the domain, due to synonym ambiguity and name variant problems. They proposed to firstly extract medical concepts from a query topic using the BioNLP abbreviation extraction function (Chang *et al.*, 2002). Secondly, for each topic, they expanded the identified medical concepts by adding lexicon variants for each concept using the same technique as Huang *et al.* (2005) (the variant is determined by a break-point—a position where a string can be broken down into parts separated by a space—, and replacement—replacing a substring with a different string that does not change the original meaning, for example, ii with 2), and retrieving documents by treating all variants as the same concept. Table 3.1 shows examples of pairs of replacements and Table 3.2 depicts name variants for the term *TGF-beta1*. This approach archived 8.91% better performance than the baseline where the approach was not applied, and 7.96% higher than the best automatic run in TREC 2005 Genomics Track (Hersh *et al.*, 2005) in terms of *MAP (Mean Average Precision)*.

TGFbeta1	TGFbeta 1	TGF beta1
TGF beta 1	TGFbetaI	TGFbeta I
TGF betaI	TGF beta I	TGFb1
TGFb 1	TGF b1	TGF b 1
TGFbI	TGFb I	TGF b I

Table 3.2: Name variants for *TGF-beta1*.

Huang *et al.* (2005) introduced a novel query expansion algorithm for dealing with variants of medical terms. In particular, each of the original query terms is split by hyphens, the adjacent of two letters with different cases (except for the first and the second letter), and the adjacent of a letter and a digit. For example, the term *185delAG* is expanded with *185del AG* and *185 del AG*. In addition, Huang *et al.* (2005)'s algorithm also created variants of terms based on a substring that can be replaced by a different string without changing of the original meaning. For example, the number *2* in *COP2* can be replaced by *ii*, and *alpha* can be replaced by *a*. The experiments with the ad hoc task of TREC 2005 Genomics track showed that this approach was effective.

To summarise, query expansion techniques have been proposed to deal with the searching of medical knowledge-based documents. Most of these approaches derived variants for query terms using domain-specific resources e.g. MeSH and AcroMed. These variants were added into a query to better describe the information underlying the query. Importantly, several useful resources were identified as sources for query expansions. Later, in Section 3.4.2.3, we discuss how query expansion approaches were adapted to the task of the searching of medical records, which are patient-specific information.

We have discussed existing approaches that deal with problems in the searching of knowledge-based information. These problems also prevail in the searching of the patient-specific information. In the next section, we describe existing works on the searching of patient-specific information, discuss how approaches used in the searching of knowledge-based information were adapted in the searching of patient-specific information, and position the work in this thesis in the context of the literature.

3.4 Searching Patient-Specific Information

This section discusses search of patient-specific information, which this thesis focuses on. As mentioned in Section 3.2, patient-specific information is generally in the form of medical records, which detail the clinical decision process relating to a patient, covering symptoms, diagnostic tests, diagnoses and treatments. Medical records have been converted into an electronic format to improve the quality of healthcare services and the patients safety (Hersh, 2008*a*). These electronic medical records can be used to find patients who have a medical history relevant to a so-called inclusion criteria, in order to possibly recruit those patients for a clinical trial (Voorhees & Hersh, 2012; Voorhees & Tong, 2011). The main searchers are healthcare providers, administrators, and researchers who may want to analyse the results of a particular medical procedure (e.g. an effective approach to combating a particular disease). Indeed, their information needs focus on finding patients whose medical records show that the corresponding patient has the same medical conditions (Averbuch *et al.*, 2004) (e.g. disease, treatment) as the ones stated in a query. For example, in a comparative effectiveness trial (see Chapter 1, Section 1.1), a healthcare practitioner may develop inclusion criteria to describe the patients of interest. The criteria include medical and health conditions (e.g. symptoms, diseases), which are used as a query to search for patients that have medical records matching the criteria (Voorhees & Hersh, 2012; Voorhees & Tong, 2011). In order to facilitate this process, an effective information retrieval system is used for identifying the patients whose medical records are relevant to these inclusion criteria.

In this thesis, we focus on the task of ranking patients based on the relevance of their medical records towards a query. However, the relevance of medical records is still considered during ranking patients since existing approaches for ranking aggregates (previously discussed in Section 2.7) that we adapt to rank patients can also take into account the relevance of individual documents.

In particular, our work is evaluated using the TREC Medical Records track's test collection, which is a standard test collection widely used for evaluating patient search systems. In the next section, we discuss the TREC Medical Records track.

3.4.1 Medical Records Track at TREC

To foster research on the searching of patients from their medical records, the Text REtrieval Conference (TREC) initiated the Medical Records track (Voorhees & Tong, 2011) in 2011. The Medical Records track ran between 2011 and 2012. In particular, the TREC Medical Records track uses the NLP Repository corpus of medical records provided by the University of Pittsburgh[1]. This corpus provides anonymised medical histories of patients throughout their visits to a hospital, including their detailed EMRs from various hospital departments. As illustrated in Figure 3.2, the EMRs are semi-structured documents containing hospital and medical information about a patient issued during the patient's visits to the hospital, such as the issuing department information (*type* and *subtype* tags), codes identifying admission diagnosis – in the form of International Classification of Diseases codes (*admit_diagnosis* tag), and a textual description of the patient made by the clinician (*report_text* tag). An example of narrative text in a mock medical record is shown in Figure 3.3.

[1] http://www.dbmi.pitt.edu/nlpfront

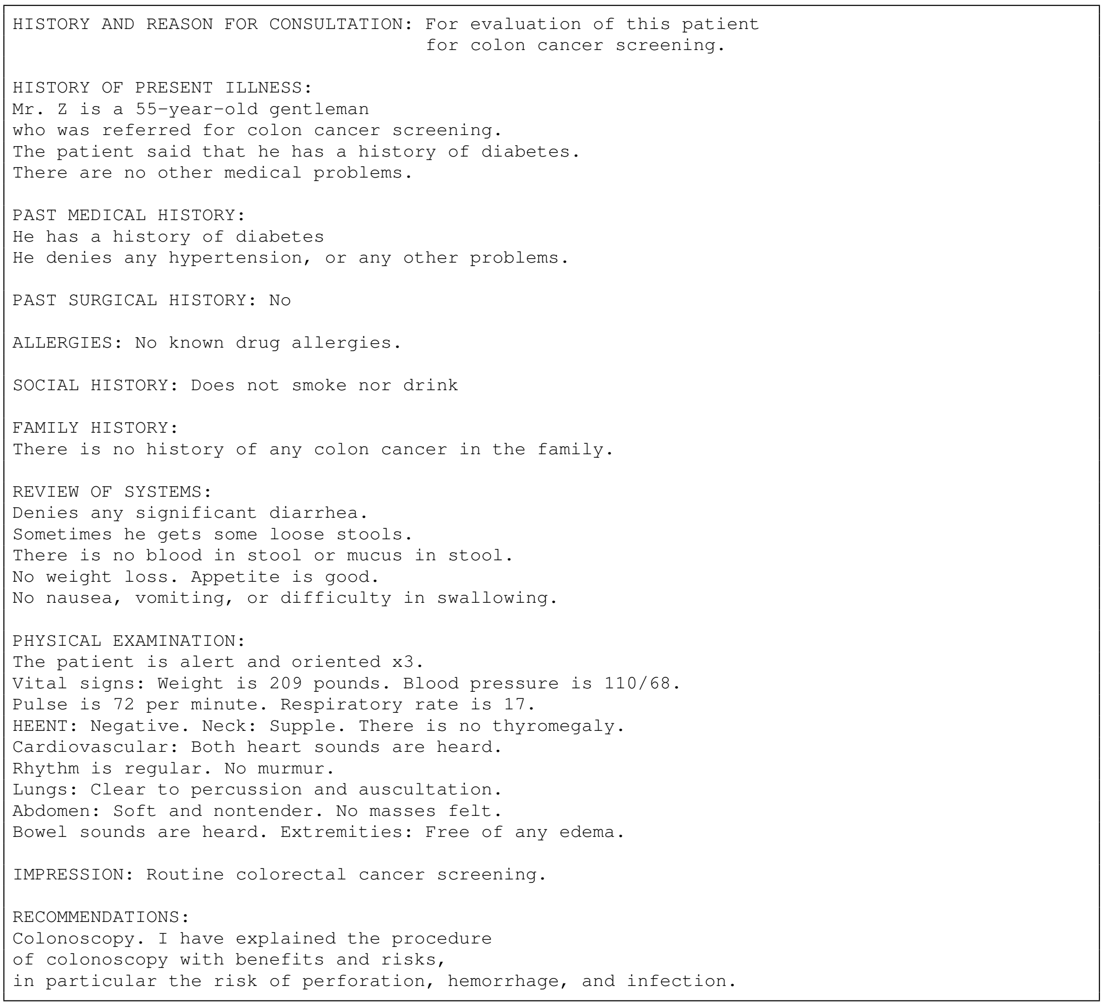
```
<report>
<type>ECHO</type>
<subtype>TEE</subtype>
<admit_diagnosis> 414.12</admit_diagnosis>
...
<report_text>
... (report text here) ...
</report_text>
</report>
```

Figure 3.2: An example of a transesophageal echocardiography medical record, from the cardiology department.

The patient search task in the context of the TREC Medical Records track (Voorhees & Hersh, 2012; Voorhees & Tong, 2011) aims to find patients having a medical history relevant to the query, based upon these patients' medical records. In particular, a patient search system ranks patients with respect to the relevance of their medical records towards a given query. The TREC medical records collection consists of 100,711 medical records, which can be mapped to 17,265 patient visits. A *patient visit* is a set of medical records associated to a visit to the hospital of a patient. A patient visit is used to represent a *patient* as a unit of retrieval, since relating multiple visits to a particular patient is made impossible by a de-identification process when building the medical records repository, to address privacy concerns (Voorhees & Hersh, 2012; Voorhees & Tong, 2011). The TREC 2011 and 2012 Medical Records track provides 34 and 47 queries, respectively, where each query describes medical conditions of the targeted patients (i.e. inclusion criteria). Examples of queries include:

Q101: Patients with hearing loss
Q102: Patients with complicated GERD who receive endoscopy
Q137: Patients with inflammatory disorders receiving TNF-inhibitor treatments
Q179: Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression

From the example of medical record in Figure 3.3 and the examples of queries, we can see that, similar to documents and queries for the searching of knowledge-based information (see Section 3.3), medical records and queries contain several medical jargons e.g. 'colon cancer', 'colonoscopy', 'hearing loss' and 'bipolar depression'. In contrast, we also observe the frequent use of negated language to describe medical conditions of a patient. For example, "He *denies* any hypertension" and "*No* nausea, vomiting, or difficulty in swallowing" in the example medical record, and the query "Patients taking atypical antipsychotics *without* a diagnosis schizophrenia or bipolar depression".



```
HISTORY AND REASON FOR CONSULTATION: For evaluation of this patient
                                     for colon cancer screening.

HISTORY OF PRESENT ILLNESS:
Mr. Z is a 55-year-old gentleman
who was referred for colon cancer screening.
The patient said that he has a history of diabetes.
There are no other medical problems.

PAST MEDICAL HISTORY:
He has a history of diabetes
He denies any hypertension, or any other problems.

PAST SURGICAL HISTORY: No

ALLERGIES: No known drug allergies.

SOCIAL HISTORY: Does not smoke nor drink

FAMILY HISTORY:
There is no history of any colon cancer in the family.

REVIEW OF SYSTEMS:
Denies any significant diarrhea.
Sometimes he gets some loose stools.
There is no blood in stool or mucus in stool.
No weight loss. Appetite is good.
No nausea, vomiting, or difficulty in swallowing.

PHYSICAL EXAMINATION:
The patient is alert and oriented x3.
Vital signs: Weight is 209 pounds. Blood pressure is 110/68.
Pulse is 72 per minute. Respiratory rate is 17.
HEENT: Negative. Neck: Supple. There is no thyromegaly.
Cardiovascular: Both heart sounds are heard.
Rhythm is regular. No murmur.
Lungs: Clear to percussion and auscultation.
Abdomen: Soft and nontender. No masses felt.
Bowel sounds are heard. Extremities: Free of any edema.

IMPRESSION: Routine colorectal cancer screening.

RECOMMENDATIONS:
Colonoscopy. I have explained the procedure
of colonoscopy with benefits and risks,
in particular the risk of perforation, hemorrhage, and infection.
```

Figure 3.3: A mock example of narrative medical records.

The track's official primary measures are bpref for TREC 2011 and infNDCG for TREC 2012, respectively.

In this next section, we discuss how existing approaches address the patient search task.

3.4.2 Existing Medical Records Search Approaches

As previously discussed in Section 1.2, traditional IR approaches are not effective for the patient search task. Edinger *et al.* (2012) showed that existing IR approaches (e.g. Demner-Fushman *et al.* (2011); King *et al.* (2011); Leveling *et al.* (2012)) failed in retrieving the patients whose medical records are relevant to a given query. In particular, patient search has unique characteristics, such as the uses of medical terminology and of negated language (Voorhees & Hersh, 2012; Voorhees & Tong, 2011); hence, there is a need for specific techniques to deal with such characteristics. In particular, we categorise the main problems of patient search and discuss how they are dealt within the literature.

3.4.2.1 Patient Ranking Models

Patient search in the context of the search task provided by the TREC Medical Records track aims to find patients having a medical history relevant to a query based on their medical records. In particular, a patient search system predicts and ranks patients with respect to the relevance of their medical records towards a query. As previously discussed in Section 2.7, this task can be handled using well-established approaches from expert search task.

Recall from Section 2.7 that aggregate ranking techniques can be categorised into two approaches. First, Model 1 ranks patients based on the relevance of the concatenation of their medical records (Balog *et al.*, 2006). Second, Model 2 uses ranked documents to rank expert persons (e.g. Voting Model (Macdonald & Ounis, 2006) and Model 2 (Balog *et al.*, 2006)).

In the context of patient search, both Model 1 and Model 2 can be adapted to rank patients based on the relevance of their medical records. Approaches sharing the same paradigm as Model 1, are referred to as *patient model* approaches (e.g. Demner-Fushman *et al.* (2011); King *et al.* (2011)). They represent a patient by combining the associated medical records into the form of a single *patient document*, and use the latter as a unit of retrieval. For example, Demner-Fushman *et al.* (2011) and King *et al.* (2011) effectively deployed the patient model by using a term weighting model (e.g. BM25 or a language model) to rank the patient documents. On the other hand, approaches based on Model 2 are referred to as *two-stage model* approaches (e.g. Limsopatham *et al.* (2011); Zhu & Carterette (2012)). They initially ranked the medical records based on their relevance towards the query, and then they estimated the relevance of patients by aggregating the relevance scores of their associated medical records that have been retrieved for the query. For example, Limsopatham *et al.* (2011) used the expCombSUM voting technique from the Voting Model (Macdonald & Ounis, 2006) to effectively aggregate the relevance scores for a patient.

However, it is not clear in the literature under which conditions a particular ranking approach should be deployed to rank patients. Demner-Fushman *et al.* (2011) and King *et al.* (2011) showed that the patient model was effective for the patient search task, while Zhu & Carterette (2012) reported that the two-stage model was also effective. In addition, Zhu & Carterette (2012) proposed to use a data fusion technique (Shaw & Fox, 1994), such as CombSUM and CombMAX, to merge the relevance scores from both the *patient model* and the *two-stage model* in order to exploit the effectiveness of both patient ranking approaches.

However, there is no extensive study comparing the effectiveness of the patient and the two-stage models. Therefore, in this thesis, we identify effective ranking approaches for the patient search task. In particular, in Chapter 4, we thoroughly compare the performances of approaches based on the two models within the same settings to identify which models are more effective for the patients search task. This will provide a common baseline that the work in this thesis further builds upon.

Moreover, Edinger *et al.* (2012) found that if a query contained several inclusion criteria (i.e. medical conditions), existing patient ranking approaches (i.e. both the patient and the two-stage models) could not effectively retrieve patients who were relevant to all or most of the medical conditions. For example, for a query to find patients with ‘lung cancer’ and ‘hypertension’, an effective patient search system should rank patients who are relevant to both conditions higher that those who are relevant to only one or none of the conditions. In Chapter 8, we investigate how to deal with this problem by adapting techniques from coverage-based search result diversification in web search (Agrawal *et al.*, 2009; Carbonell & Goldstein, 1998; Santos *et al.*, 2010*a*).

3.4.2.2 Modelling Medical Information

Most of the existing approaches rank patients' medical records using traditional ranking models (e.g. a language model (Hiemstra, 2001)). For example, King *et al.* (2011) used BM25 which is based on probabilistic ranking principle (see Section 2.4.1) to rank patients, which were represented using the terms in medical records of individual patients.

Lin & Demner-Fushman (2006) and Abhyankar *et al.* (2014) suggested to model medical documents (e.g. journal papers and medical records) during retrieval, by extracting four components from a given query. The four components are called *PICO*, which consist of *Problem/Population*, *Intervention*, *Comparison*, and *Outcome*. Then, when ranking medical documents, each of the PICO components are focused on. Indeed, the approach relies on inputs from knowledge experts to decide levels of importance of each PICO component, when ranking medical documents.

On the other hand, as discussed in Section 1.2, the medical records describe the medical conditions of a patient (Hersh, 2008*a*; Silfen, 2006), which include four main aspects (namely, symptoms, diagnostic tests, diagnoses, and treatments) related to the *medical decision process* (Silfen, 2006). These kinds of information are important for medical practitioners to take into account when consulting patients (Silfen, 2006). The medical conditions related to these four aspects are commonly used to describe the conditions of a patient. For example, knowing that a patient is visiting a hospital with *‘chest* *pain'* (symptom), a healthcare practitioner may suspect that the patient has *'heart disease'* (diagnosis). Hence, given the symptom, the practitioner can compile a set of diagnostic procedures such as *'chest X-ray'* (diagnostic test) for the patient. Once the practitioner is confident that the patient suffers from *'heart disease'* (diagnosis), the practitioner may prescribe a treatment to the patient such as *'coronary artery bypass surgery'*. In this work, we propose to model the relevance and representations of information need and medical records in terms of the medical conditions related to the four medical aspects. Later, in Chapters 4 and 8, we introduce our approaches that model medical information in terms of the medical decision process within a search system.

3.4.2.3 Medical Terminology

Patient search systems also encounter the problem of medical terminology in medical records and queries. To tackle this, most existing approaches deployed the same kinds of techniques used in the prior works in the searching of medical knowledge-based documents.

Conceptual Representations

Several approaches (e.g. Koopman *et al.* (2011, 2012); Qi & Laquerre (2012)) for patient search used BoC approaches, which were developed for the searching of knowledge-based documents (see Section 3.3.1), to represent medical records and queries. For instance, Koopman *et al.* (2011) extracted medical concepts from medical records and queries using the MetaMap tool, and represented medical records and queries using the CUIs of all of the identified medical concepts. Koopman *et al.* (2011) showed that the BoC approach was more effective than the BoW approach. Qi & Laquerre (2012), whose system ranked among the top groups of participants at the TREC 2012 Medical Records track also used the same BoC approach and showed that it was effective. In addition, Koopman *et al.* (2012) proposed a graph-based approach for weighting patients based on the occurrences of medical concepts in their medical records. They built a graph model based on the co-occurrences of medical concepts within a particular window of concepts, and calculated the PageRank (Brin & Page, 1998) scores for each medical concepts. Then, during retrieval, they scored patients based on the PageRank scores of the query concepts that matched their medical records. Meanwhile, King *et al.* (2011) showed that using both BoC and BoW further improved retrieval performance. In contrast to the searching of the medical knowledge-based document (see Section 3.3.1) that in general the BoC approach is less effective than traditional BoW approach, BoC was effective for the patient search task.

In this thesis, we also deploy the BoC approach. However, using all of the medical concepts identified to represent medical records and queries may be too general. Instead, we focus only on specific types of medical concepts when representing medical records and queries (see Chapter 4).

Query Expansion

Beside using medical concepts to represent medical records and queries, existing works also leveraged medical resources for query expansion to deal with the medical terminology. King *et al.* (2011) effectively deployed UMLS Metathesaurus and their collection of medical reference encyclopedias in their query expansion technique to improve retrieval performance. Importantly, they added expansions into the query and set a particular weight equally for all of the expansions. Similarly, Demner-Fushman *et al.* (2011) exploited medical resources, such as UMLS, RxNorm, and knowledge from medical experts, within their retrieval framework to achieve the highest retrieval performance among the participating systems reported in the TREC 2011 Medical Records track. However, the approach proposed by Demner-Fushman *et al.* (2011) needs inputs from medical experts. In contrast, Qi & Laquerre (2012) showed that query expansion by using synonyms and more specific concepts derived from the UMLS Metathesaurus harmed the retrieval performance.

Several participants at the TREC Medical Records track showed that pseudo-relevance feedback was effective for the patient search task (e.g. King *et al.* (2011); Qi & Laquerre (2012); Voorhees & Hersh (2012); Zhu & Carterette (2012); Zhu *et al.* (2014)). For example, Qi & Laquerre (2012) showed that using pseudo-relevance feedback and a vector space model within a patient search system that used the BoC representation approach was effective. Meanwhile, within a BoW representation, Zhu *et al.* (2014) followed Diaz & Metzler (2006) and expanded the queries by deriving related terms and their weights from different collections of documents, such as the TREC Medical Records track's collection itself, the TREC 2007 Genomics track's collection (Hersh *et al.*, 2007) and the TREC 2009 ClueWeb09 Category B dataset[1] (excluding Wikipedia pages). Zhu *et al.* (2014) showed that this approach significantly improved retrieval performance.

While pseudo-relevance feedback is likely to be effective for the patient search task, the retrieval performance achieved when using query expansion with knowledge-based resources (e.g. UMLS Metathesaurus, SNOMED CT[2]) depended mostly on the used resources. Importantly, it is not clear in the literature how we can effectively use several resources to derive related terms and their weight within a single approach. In particular, existing works (e.g. King *et al.* (2011); Qi & Laquerre (2012)) simply

[1] `http://lemurproject.org/clueweb09.php`

[2] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

Patient admitted with **cancer**
Diagnosed and found no evidence of **cancer**
Negative result on **cancer** screening test

Table 3.3: Examples of sentences in EMRs where the presence of the query term ‘cancer’ does not always indicate that the patient has ‘cancer’.

added expansions into the queries. However, we hypothesise that expansions are not equally important. Moreover, existing works derived expansions from different resources individually. In this thesis, we investigate novel approaches that combine several resources to enable the inferences of relationships between medical concepts over different resources (Chapter 6). In addition, in Chapter 7, we investigate approaches to deal with the medical terminology by using knowledge extracted from aggregates of medical records, instead of using external resources or individual medical records.

3.4.2.4 Negated Language

One of the major challenges of searching patients from their medical records is the use of negated language (Koopman *et al.*, 2010). Negation is commonly used in medical records to indicate that the patient does not suffer from a particular medical condition (Koopman *et al.*, 2010). As a result, the presence of a query term does not always imply that the record is relevant to the query (Averbuch *et al.*, 2004). In particular, the relevance also depends on the context of the query terms occurring in the medical records. For example, while all the three sentences, shown in Table 3.3, contain the query term ‘cancer’, only the first sentence indicates that the disease is pertaining to the patient, while the other two sentences are non-relevant towards a query searching for patients suffering from ‘cancer’.

However, prior work dealing with negation in documents is limited. Table 3.4 summarises and compares the main existing work on negated language in the medical domain. Specifically, classical/traditional IR approaches (e.g. BM25) do not explicitly deal with negation. They simply ignore the presence of negated language in the medical records and query (Koopman *et al.*, 2010). On the other hand, most of the previous works that deal with negated language (e.g. boolean retrieval model, post-retrieval filtering (Salton & McGill, 1986), vector negation (Widdows, 2003)) only tackled it within queries. For example, the boolean retrieval model focuses on retrieving documents by firstly forming a list of documents containing the query terms and then removing the documents with the occurrence of the negated query terms from the retrieved list. Nevertheless, these approaches do not take into account negated language in documents. Indeed, most phrases indicating negation (e.g. no, not) are commonly seen as stopwords and are usually discarded during indexing (Chapman *et al.*, 2001).

There have been recent attempts to deal with negation in patient search (e.g. Demner-Fushman *et al.* (2011); King *et al.* (2011); Zhu & Carterette (2012)). In particular, these approaches commonly

Approach	Improve query representation	Discarding negated language	Improve document representation	Learning to penalise non-relevant documents
Traditional approach (e.g. Koopman *et al.* (2010))	✖	✖	✖	✖
Post-retrieval filtering Salton & McGill (1986)	✔	✖	✖	✖
Vector negation (Widdows, 2003)	✔	✖	✖	✖
Boolean retrieval model	✔	✖	✖	✖
Discarding negated sentences (e.g. Demner-Fushman *et al.* (2011); King *et al.* (2011); Koopman & Zuccon (2014); Zhu & Carterette (2012))	✖	✔	✔	✖

Table 3.4: Comparing approaches to handle negated language in medical records search.

deploy a negation detection tool (e.g. NegEx (Chapman *et al.*, 2001) or NegFinder (Mutalik *et al.*, 2001)) to deal with negated language in medical records. For example, several top performing search systems (e.g. Demner-Fushman *et al.* (2011); King *et al.* (2011); Zhu & Carterette (2012)) at the TREC 2011 and 2012 Medical Records track showed that dealing with negation in medical records improved retrieval performance. King *et al.* (2011), Zhu & Carterette (2012) and Demner-Fushman *et al.* (2011) proposed to disregard terms with the negated language during indexing and retrieval. Indeed, they removed from a search system parts of the sentences that contain the negated context identified using the NegEx algorithm. King *et al.* (2011) reported that by removing the negated sentences from a search system, the retrieval performance could be improved by 5%. Koopman & Zuccon (2014) proposed a language model-based approach that enabled a search system to disregard terms with a negated context on a per-query basis. In particular, when dealing with negation, they considered the mixture of the language models of terms with a positive context and terms with a negated context, where the mixture parameter enabled the system to choose whether to disregard or negatively score the terms with a negated context. In this thesis, we propose a novel approach that aims to prevent non-relevant patients from being retrieved by demoting the relevance scores of patients whose medical records do not have the medical conditions stated in the query (Chapter 5).

3.5 Conclusions

We have described two types of medical information in the context of this thesis. Firstly, the knowledge-based information is knowledge (e.g. in the forms of books or journals) that is ready to be applied when conducting research or when consulting patients. Secondly, the patient-specific information is about the description of the medical and health conditions of a particular patient (e.g. in the form of medical records). We have highlighted different types of information needs underlying the use of search systems for both the knowledge-based information and the patient-specific information. In addition, we discussed the main problems when searching medical information, most of which are related to implicit knowledge, and reviewed existing techniques in the literature that dealt with such problems.

In particular, we found the following knowledge gaps in the literature.

- *How to effectively rank patients?*

  As discussed in Section 3.4.2.1, existing works adapted either *the patient model* or *the two-stage model* for the patient search task. However, very little work investigated which models are more effective for the task. Hence, in Section 4.3.3 of Chapter 4, we will thoroughly evaluate the two models for ranking patients in order to identify an effective baseline system that our proposed framework will build upon.

- *How to model medical information?*

  We have shown in Section 3.4.2.2 that most of the existing works used traditional approaches to model the medical information. However, we argue that the information needs in patient search are specific. In particular, when practitioners search for patients, they are likely to specify the medical conditions of the patients of interest, as a query. Therefore, we will propose an approach that focuses a search system on particular types of medical conditions in Section 4.2.1. In addition, in Chapter 8, we will propose a novel ranking approach that models the relevance towards each of the medical conditions stated in the query.

- *How to effectively deal with the medical terminology?*

  In Section 3.4.2.3, we have shown that existing works dealt with the complexity of the medical terminology using either conceptual representation approaches or query expansion approaches. Later in Chapter 4, we will introduce a conceptual representation that is different from those in the literature, in that the proposed approach uses only the medical concepts related to some specific types of medical conditions. This enables the search system to focus on the medical conditions crucial to healthcare practitioners. Meanwhile, in Chapter 6 we will propose novel query expansion approaches that take into account the relationships between medical concepts from several resources at the same time. In addition, our query expansion approaches also calculate the importance of each derived medical concept. Furthermore, in Chapter 7, we will introduce novel approaches that deal with the complexity of the medical terminology in a different manner. In particular, our approaches aggregate the medical records at the department level to deal with the medical terminology complexity.

- *How to deal with negated language in the medical records and queries?*

  In Chapter 5, we will introduce approaches for dealing with negated language in patient search, which are different from the existing works discussed in Section 3.4.2.4 in that the proposed approaches could prevent non-relevant patients from being retrieved by demoting the relevance scores of patients whose medical records contain the query terms in an opposite context.

In the next chapter, we will introduce a novel framework for patient search to uncover the implicit knowledge when searching for patient-specific information, which is the main focus of this thesis.

Chapter 4

A Framework for Patient Search

4.1 Introduction

In the previous chapter, we discussed existing works related to the searching of medical documents. We showed that the problems caused by the presence of implicit knowledge within patient search have not been thoroughly examined in the literature. Indeed, as highlighted by Section 3.4.2, for the patient search task, the relevance of a patient depends on both the presence of the query terms within the patient's medical records, and the implicit knowledge encoded within those records. For example, knowledge about the relationships between medical terms can be used to infer the relevance of a patient at each individual record. In particular, we could infer that a patient has a particular disease, if his/her record shows evidence of having a particular treatment for that disease.

To deal with the problems of this implicit knowledge, in this chapter, we propose a framework for enhancing the query and medical record representations, by modelling and reasoning during retrieval in terms of the four aspects of the medical decision process (namely, symptoms, diagnostic tests, diagnoses and treatments), which were described in Section 3.4.2.2. In particular, as previously discussed in Section 3.4.2.2, medical conditions related to these aspects are crucial information that healthcare practitioners consider when consulting a patient. Our framework acts at three different levels of the patient retrieval process (namely, at sentence level, record level and inter-record level) to uncover the implicit knowledge in patient search. The remainder of this chapter is structured as follows:

- In Section 4.2, we present our patient search system, which consists of (1) *the query and medical record representation unit*, (2) *the patient ranking unit* and (3) *our framework for patient search.* The proposed framework for patient search uncovers implicit knowledge within the medical records and queries by focusing on and reasoning about the medical conditions related to the aforementioned four aspects of the medical decision process at the three different levels of the patient retrieval process.

- In Section 4.3, we identify an effective patient search baseline that our framework will be built upon in the next chapters. Then, we provide further analysis and discussion of the effectiveness of the identified baseline.
- In Section 4.4, we provide a summary of this chapter.

4.2 Our Patient Search System

Our work focuses on searching for patients based on the relevance of their medical records with respect to a query. Recall that the aim of our patient search system is to find, from their medical records, the patients who have similar medical conditions (e.g. disease, treatment) as stated in a query (as described in Section 3.4.1). Figure 4.1 shows the architecture of our patient search system. The queries and medical records are first processed by the query and medical record representation unit using a term-based or a conceptual representation approach. In particular, as previously discussed in Section 3.4.2.3, a term-based representation approach uses terms to represent the queries and medical records. Meanwhile, a conceptual representation approach represents the queries and medical records using concepts from medical resources (e.g. UMLS Metathesaurus[1]). Then, our proposed framework enhances the applied representation by leveraging the implicit knowledge uncovered from the three levels of the retrieval process. Finally, the patient ranking unit deploys either the patient model or the two-stage model described in Section 3.4.2.1 to rank patients based on the relevance of their medical records towards the queries. Note that existing patient search approaches in the literature deploy the same kind of system in Figure 4.1, but without the proposed framework. For example, Koopman *et al.* (2012) used a conceptual representation approach in the query and medical record representation unit, and applied a patient model approach (i.e. using a graph-based model to rank the patient documents) in the patient ranking unit (see Section 3.4.2.3). Even though some existing works have tackled a particular implicit knowledge problem (reviewed in Section 3.4.2), we deal with several different implicit knowledge problems within a single framework.

In this section, we discuss each unit of the patient search system and introduce our framework for enhancing the representations of the queries and medical records by uncovering and leveraging implicit knowledge.

[1] `http://www.nlm.nih.gov/research/umls/`

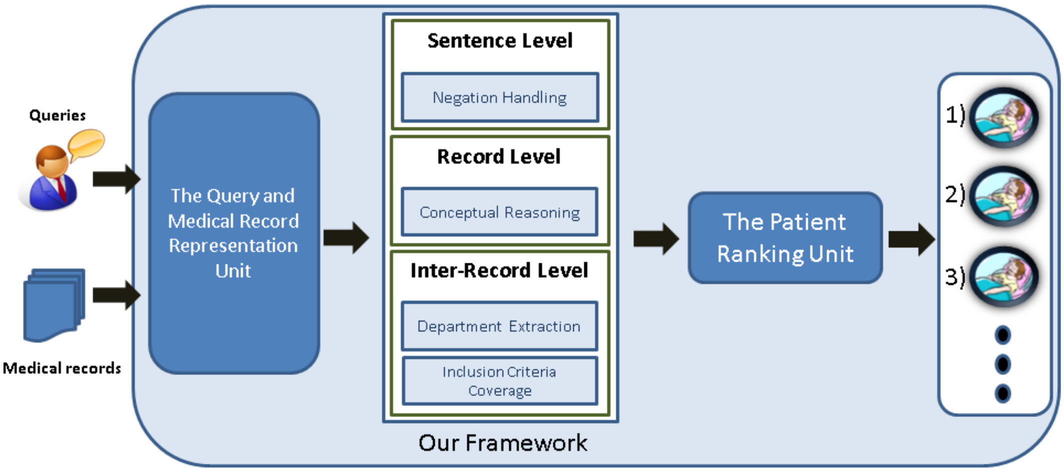


Figure 4.1: Our patient search system.

4.2.1 The Query and Medical Record Representation Unit

As previously mentioned in Section 3.4.2.3, existing patient search systems showed that both the term-based and the conceptual representation approaches were effective. Hence, within the query and medical record representation unit, we can choose to apply either approaches. However, as argued in Section 3.4.2.3, in the existing literature, the comparison between the two representation approaches for the patient search task has not been extensively studied.

In addition, as previously discussed in Section 3.4.2.2, the users of a patient search system firstly choose a set of inclusion criteria that describe the medical conditions of the targeted patients and then use these as a query. In particular, these inclusion criteria are mainly related to medical concepts concerning the four aspects of the medical decision process (namely, symptom, diagnostic test, diagnosis and treatment), previously described in Section 3.4.2.2. We argue that when modelling each query and medical record, we need to consider which concepts are related to the user's information need. To achieve this, we introduce our *task-specific representation* approach, which is a conceptual representation approach and aims to represent the patients' medical records and queries using medical concepts extracted from each sentence within the medical records and the queries.

Our task-specific representation approach focuses on the concepts related to each of the four aspects of the medical decision process. In particular, this approach represents the medical records and queries using informative medical concepts extracted from them. For example, the medical conditions 'cancer', 'carcinoma', and 'malignant tumour' share a similar meaning; hence, they are represented with the same

MetaMap's Semantic Type	Aspects of the Medical Decision Process			
	Symptom	Diagnostic test	Diagnosis	Treatment
Body Location or Region	✔	✔	✔	✔
Body Part, Organ, or Organ Component	✔	✔	✔	✔
Clinical Drug	–	–	–	✔
Diagnostic Procedure	–	✔	–	–
Disease or Syndrome	–	–	✔	–
Finding	✔	–	–	–
Health Care Activity	–	✔	–	✔
Injury or Poisoning	✔	–	–	–
Intellectual Product	–	✔	–	✔
Medical Device	–	✔	–	✔
Mental or Behavioral Dysfunction	✔	–	✔	–
Neoplastic Process	✔	✔	✔	✔
Pathologic Function	✔	–	–	–
Pharmacologic Substance	–	–	–	✔
Sign or Symptom	✔	–	–	–
Therapeutic or Preventive Procedure	–	–	–	✔

Table 4.1: List of 16 of the MetaMap's 133 semantic types that we consider for our proposed approach, based on the four aspects of the medical decision process.

medical concept. Differing from existing works in conceptual representation, which were reviewed in Section 3.4.2.3, the task-specific representation approach uses only medical concepts related to the four medical aspects. We argue that by using concepts in this manner, we can generate a more accurate representation of each medical record and query.

Our proposed approach is defined as follows. To represent medical conditions within a search system, we deploy MetaMap[1] – a medical concept recognition tool based on the UMLS Metathesaurus that is widely used in the literature (see Section 3.4.2.3) – to identify medical concepts from sentences in the medical records and queries, and represent them in the form of a UMLS Concept Unique Identifier (CUI). Importantly, we use only the medical concepts related to the four aforementioned aspects, which we identify based on their MetaMap's semantic types.[2] In Table 4.1, we list the defined 16 MetaMap semantic types that are associated to the four medical aspects of interest. For each of MetaMap's semantic type, we list the aspects of the medical decision process that are associated with the concepts related to that type. For example, medical concepts that have the semantic type of *Disease or Syndrome* are associated with the *diagnosis* aspect.

Figure 4.2 shows an example of using the MetaMap tool to extract medical concepts from the query "Patients with diabetes mellitus who also have thrombocytosis", where 'C0011849' and 'C0836924' are

[1] In this thesis, we use the MetaMap version 2011v2.

[2] `http://metamap.nlm.nih.gov/SemanticTypeMappings_2011AA.txt`

```
Input: "Patients with diabetes mellitus who also
        have thrombocytosis"

Phrase: "Patients"

Phrase: "with diabetes mellitus"
Meta Candidates (4):
  1000 C0011849:Diabetes Mellitus [Disease or Syndrome]
          Diabetes
   861 C0011847:Diabetes [Disease or Syndrome]
   789 C0241863:DIABETIC [Finding]
Meta Mapping (1000):
  1000 C0011849:Diabetes Mellitus [Disease or Syndrome]

Phrase: "who also"

Phrase: "have"

Phrase: "thrombocytosis."
Meta Candidates (1):
  1000 C0836924:Thrombocytosis [Disease or Syndrome]
Meta Mapping (1000):
  1000 C0836924:Thrombocytosis [Disease or Syndrome]
```

Figure 4.2: Medical concepts extracted by the MetaMap tool from the query ‘Patients with diabetes mellitus who also have thrombocytosis’

the CUIs extracted from the query with the highest confidence (identified as *Meta Mapping* concepts by MetaMap in Figure 4.2), and shown in Table 4.2. Note that, we could use the medical concepts identified with any level of confidence. For example, instead of using only medical concepts with the highest confidence (as shown in Table 4.2), we can use all of the extracted medical concepts regardless of their confidence levels. With this variant, we can extract four different CUIs from the example query, including ‘C0011849’, ‘C0011847’, ‘C0241863’ and ‘C0836924’.

Later, in Section 4.3.4, we will investigate different methods for choosing the medical concepts with different levels of confidence in both the task-specific representation and an existing conceptual representation approach. We will also examine which representation approach should be used within a patient search system.

4.2.2 The Patient Ranking Unit

Next, we discuss the patient ranking unit of our patient search system shown in Figure 4.1. As discussed in Section 3.4.2.1, existing patient search approaches have typically adapted aggregate ranking

Concept (CUI)	MetaMap's Definition	Related Aspects
C0011849	Diabetes Mellitus [Disease or Syndrome]	Diagnosis
C0836924	Thrombocytosis [Disease or Syndrome]	Diagnosis

Table 4.2: An example of medical concepts obtained from the query *"patients with diabetes mellitus who also have thrombocytosis"* using our task-specific representation approach.

approaches to rank patients based on the relevance of their medical records towards a given query. Recall that approaches for ranking aggregates can be categorised into two groups, namely *the patient* and *the two-stage models* (see Section 3.4.2.1). We illustrate the retrieval behaviour of these two groups in Figures 4.3(a) and 4.3(b), respectively. For instance, in Figure 4.3(a), the patient model ranks patients based upon the relevance of the concatenations of all medical records associated to those patients, with respect to the query. For a query q, the patient model calculates the relevance score of a patient p as follows:

$$score(p, q) = score(D_p, q) \tag{4.1}$$

where the patient document D_p is created by concatenating the medical records associated to the patient p. $score(D_p|q)$, the relevance score of D_p towards the query q, can be calculated using any retrieval model, such as BM25 (as defined Section 2.4).

On the other hand, the second group of approaches combines the relevance scores of the medical records associated to a particular patient to estimate the relevance of that patient by using an aggregate function, such as a voting technique from the Voting Model (see Section 2.7). For example, the expCombSUM voting technique calculates the relevance score of a patient p to a query q as follow:

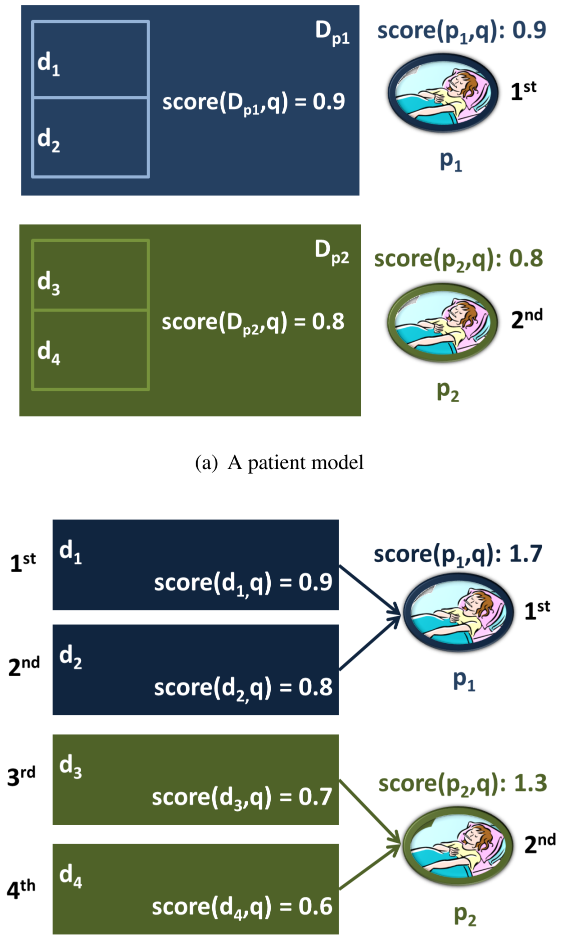
$$score(p, q) = \sum_{d_i \in R(q) \cap profile(p)} score(d_i, q) \tag{4.2}$$

where $R(q) \cap profile(p)$ is the set of medical records associated with the patient p that are also in the ranking $R(q)$, which is initially obtained using a retrieval model, such as, BM25 or DFR DPH. $score(d_i, q)$ is the relevance score of medical record d_i for the query q.

Later, in Section 4.3.4, we will investigate the retrieval effectiveness of different variants of the patient and the two-stage models when ranking patients based on the relevance of their medical records towards the queries.

4.2.3 Our Framework for Enhancing the Representations of Queries and Medical Records

In this section, we introduce a framework that uncovers the implicit knowledge from the medical records and queries. We argue that the existing approaches for ranking documents may not be effective for



(a) A patient model

(b) A two-stage model (CombSUM)

Figure 4.3: Illustrative examples when ranking patients using the patient and the two-stage models.

ranking patients based on the relevance of their medical records, because of the specificities of the medical records and queries (as previously discussed in Section 3.4.2). For example, the common use of a negative qualifier in the medical records to indicate that the patient does not possess a particular medical condition (e.g. “the patient denies having a fever”). Without considering the context of medical terms, a patient whose medical records state that the patient does not have a medical condition is likely to be retrieved for a query searching for patients with that medical condition, since the mention of that condition is matched. Hence, as such context terms are typically not considered by a patient search system, they can be viewed as implicit knowledge, which an effective search system should aim to uncover. In particular, we argue that implicit knowledge, which is known by healthcare practitioners, but may be hidden from traditional search systems, should be discovered and leveraged to improve the retrieval effectiveness. This implicit knowledge includes:

- The context (either positive or negative) of medical terms in the medical records and queries.
- The relationships between medical terms that could be inferred using medical knowledge sources.
- The latent knowledge that could be discovered from the aggregates of medical records (e.g. the expertise of a particular hospital department can be discovered from the aggregates of the medical records issued by that department).
- The knowledge that healthcare practitioners apply when identifying the important medical conditions in the query and ensuring that the retrieved patients have medical histories that cover all (or most) of those medical conditions.

To achieve this, our proposed framework works at three different levels of the retrieval process to uncover and leverage implicit knowledge in order to improve retrieval performance, namely sentence, record and inter-record levels. This will allow our framework to deal with different types of implicit knowledge. For example, we can deal with negated language at the sentence level. In this thesis, we propose four components, each of which works at different levels of the retrieval process. In Figure 4.4, we show the four components of our framework, two of which work at the inter-record level, while the other two work at the sentence and the record levels, respectively. Note that there may be other types of implicit knowledge that the current framework does not deal with (e.g. the knowledge about the development of disease over time). However, a new dedicated component can be developed and added to the framework.

For the remainder of this section, we introduce each component of the proposed framework.

4.2.3.1 Negation Handling

The first component of the proposed framework, namely *the Negation Handling component*, uncovers implicit knowledge at the sentence level. This component aims to recognise and represent the context of medical terms within sentences in each medical record, as well as in the user query. As previously discussed (see Section 3.4.2.4), negative qualifiers, e.g. ‘no’ or ‘without’, are commonly used in medical records, to indicate the absence of a medical condition of a patient, typically a symptom. Moreover, by demonstrating that a medical practitioner has tested and found that a symptom is not occurring, the practitioner is also implicitly ruling out other possible diagnoses. As a result, it is critical that the context of each term or concept (whether it occurs in a negative or a positive context) is considered.

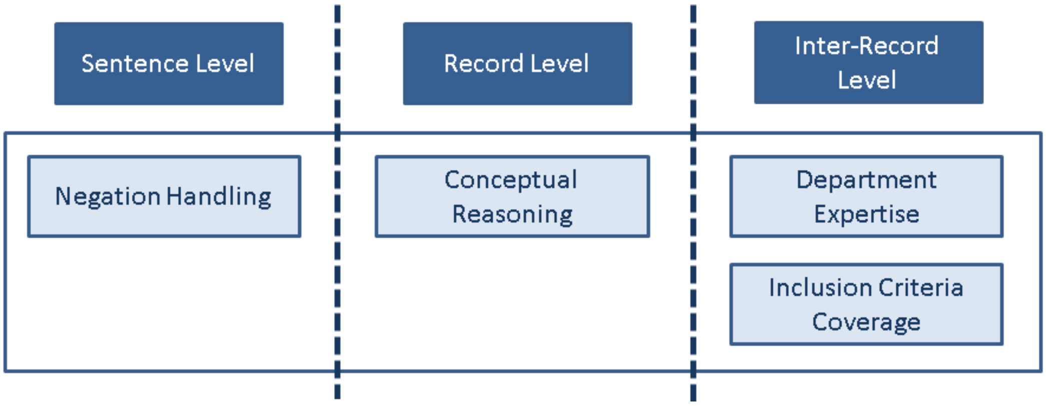


Figure 4.4: Components of the proposed framework.

The presence of a query term alone in a patient's medical records is insufficient for an IR system to determine that the patient is relevant to the query. For example, a patient whose medical record states that "the patient denies experiencing fever" may be ranked highly for the query "find a patient with fever", even though the patient does not have fever. Hence, it is of paramount importance that negation be handled effectively in patient search. The key question that we aim to address is how can we handle negated language in both the medical records and queries.

The Negation Handling component focuses on enabling a search system to differentiate between terms with different contexts (i.e. positive or negative). Indeed, we argue that negated language should be explicitly represented to disambiguate the context of terms, such that during retrieval only terms with the same context as the query are matched. In particular, we first use a negation detection tool to disambiguate the context of terms, and then represent those in the negated context differently during indexing and retrieval. This is based on the intuition that the context of terms in the medical records and queries should be understood and disambiguated, so that a query searching for the presence of medical conditions would rank highly patients whose medical records contain those conditions within positive contexts. On the other hand, a query searching for an absence of medical conditions would return only patients associated to medical records with the query medical conditions appearing in a negative context in the top ranks. Within the Negation Handling component, we use a natural language processing (NLP) tool to detect negated language. In particular, we deploy the NegEx algorithm (Chapman *et al.*, 2001) to recognise the (negative or positive) context of terms in records and queries. The NegEx algorithm uses a set of predefined regular expression to detect negated phrases in a sentence. Examples of phrases that indicate the use of negated language are "no signs of", "ruled out", "no" and "negative for". Then, we

Original record	Patient does not have fever
Negation recognition	Patient does *not have fever*
Removing stopwords	Patient *fever*
Negation representation	Patient *n$fever*

Table 4.3: The negation representation process for a medical record - italicised terms occur in a negative context.

represent terms with different contexts differently, so that terms in the records and queries are matched only when they are in the same context. For example, *fever* is represented as *fever* or *n$fever*, if it is in a positive or negative context, respectively, as shown in Table 4.3. We also penalise patients whose medical records contain the query terms with the opposite context in order to prevent the patients whose medical records have an occurrence of any query terms with the opposite of the intended context from being ranked highly. For example, for a query 'hypertension', patients whose medical records contain the term 'n$hypertension' should not be highly ranked. Chapter 5 discusses approaches that instantiate our first component and presents our experimental results.

4.2.3.2 Reasoning using Medical Concept Relationships

The second component of the proposed framework (i.e. *the Conceptual Reasoning component*) uncovers the implicit knowledge at the record level. In particular, we aim to exploit medical resources (e.g. ontologies, and health-related websites), which have been extensively developed as discussed in Section 3.4.2.3, to improve the representation of the user queries to aid relevance estimation when ranking. In particular, the idea is to leverage the semantic relationships (e.g. simple relationships such as synonyms, hyponyms, or the more complex relationships between the treatment and diagnostic test, and between the diagnostic test and diagnosis) extracted from medical resources to permit reasoning about the matches between the requested information need and the medical histories of the patients.

Recall that in Section 3.4.2.3, we showed that, traditionally, search systems have resorted to semantic resources (e.g. MeSH[1], MedDRA[2], DOID[3]) for semantic query expansion to improve the representation of queries. However, existing query expansion approaches are not tailored to the relationships between the medical decision process. We argue that the medical queries and patient histories can be characterised in terms of the four medical aspects of the medical decision process (as described in Section 3.4.2.2) and that this can be used to enhance retrieval. For example, in the query "find patients with complicated GERD who receive endoscopy", *'GERD'* is a diagnosis and *'endoscopy'* is a diagnostic

[1] http://www.nlm.nih.gov/mesh

[2] http://www.meddramsso.com/

[3] http://purl.bioontology.org/ontology/DOID

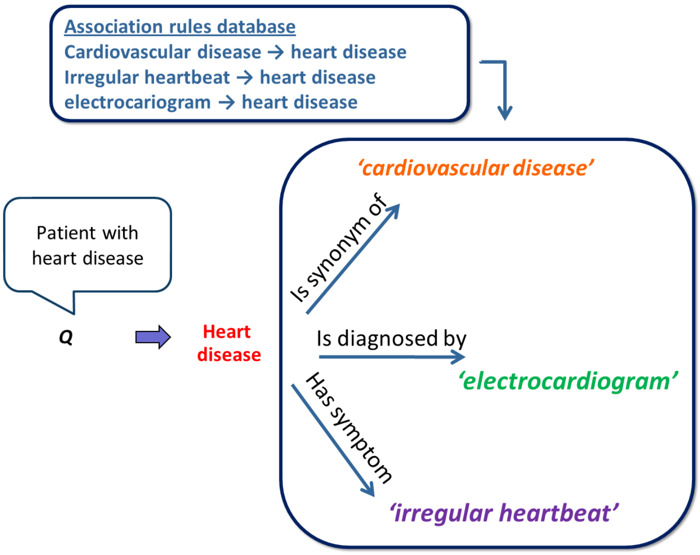

Association rules database
Cardiovascular disease → heart disease
Irregular heartbeat → heart disease
electrocariogram → heart disease
Patient with heart disease
Q
Heart disease
'cardiovascular disease'
Is synonym of
Is diagnosed by
'electrocardiogram'
Has symptom
'irregular heartbeat'

Figure 4.5: An example of using association rules to infer the medical conditions of patient with 'heart disease'.

test. By using the semantic relationships from the medical resources and ontologies, we can improve the representation of queries, thereby permitting improved relevance estimation for a patient history with respect to the four medical aspects. For instance, we could infer that patients treated with *amiodarone* (treatment) are suffering from *heart disease* (diagnosis), since amiodarone is a medicine for patients with heart disease (the semantic relationship between the treatment and the diagnosis). Thus, enriching queries that include heart disease with the concept amiodarone should result in an enhanced retrieval performance, as it provides new evidence from which relevance can be inferred.

In particular, the second framework component is different from existing works in that while most of the existing techniques exploit the hierarchical information of knowledge-based resources such as ontologies to aid retrieval in the form of semantic query expansion, we leverage these resources through inference rules tailored to the relationships within the four aforementioned aspects (e.g. treatment-diagnosis) to infer associations between records and queries. In Figure 4.5, we show an example of medical concepts that are inferred from the query "patient with heart disease" using the extracted association rules. Indeed, 'cardiovascular disease', 'electrocardiogram', and 'irregular heartbeat' are the medical conditions that can be inferred from 'heart disease', since they are synonym, diagnostic test, and symptom of 'heart disease', respectively. Chapter 6 further discusses and evaluates our component to reasoning on conceptual relationships.

```
<report>
<dep_info> CHEST-RAD </dep_info>
<admit_diagnosis>786.50</admit_diagnosis>
...
<report_text>
... (report text here) ...
</report_text>
</report>
```

Figure 4.6: An example of a chest X-ray medical record, radiology department.

4.2.3.3 Leveraging Knowledge about Hospital Departments' Expertise

Our third component (namely, *the Department Expertise component*) uncovers the implicit knowledge from the aggregates of medical records issued from individual hospital departments. We aim to extract additional information about the medical records based upon where they were created. Each medical record is created by a particular hospital department, e.g. radiology. This information is recorded within each medical record.

In contrast to existing patient search approaches that have improved the representations of the medical records and queries by using information from external resources or individual medical records (as reviewed in Section 3.4.2.3), we argue that the medical records of a patient with a given disease are likely to be predominantly from the particular hospital department that specialises in that disease. For example, patients with heart disease are likely to have records from the hospital's cardiology department. Therefore, given a query about a particular medical condition, it is likely that medical records from a department specialising in that medical condition will be relevant to the query. Hence, inspired by recent works (Broder *et al.*, 2010; Metzler *et al.*, 2009) on aggregating the Web documents to domain- or site-level to extract new evidence for ranking, we aim to aggregate the medical records at the department level (i.e. inter-record level) and thereby extract implicit medical knowledge about the hospital departments to enhance retrieval performance. To illustrate, Figure 4.6 shows an example of a chest X-ray medical record from the radiology department (as defined by *CHEST-RAD* in the dep_info field). We leverage background knowledge about hospital departments to better rank patients having medical records from those departments.

In Chapter 7, we propose approaches to instantiate this component and evaluate their retrieval effectiveness.

4.2.3.4 Modelling Relevance towards Multiple Inclusion Criteria

The final component of the proposed framework (namely, *the Inclusion Criteria Coverage component*) also uncovers the implicit knowledge at the inter-record level. We aim to model the relevance with respect to different query medical conditions (i.e. inclusion criteria) when ranking patients based on the relevance of their medical records. As discussed in Section 3.4, queries for the patient search task often contain several medical conditions. For example, a query "Find patients with diabetes and heart disease" requires patients who are suffering from both 'diabetes' and 'heart disease'. We argue that it is important to identify the medical conditions that the query focuses on (i.e. 'diabetes' and 'heart disease') and rank patients who have all of these medical conditions higher than those who have only one condition.

However, as discussed in Section 3.4.2.1, existing approaches for ranking patients do not model the relevance towards multiple medical conditions stated in the query. This component aims to address this knowledge gap and attain effective retrieval performance, by considering the relevance of a patient towards each of the inclusion criteria stated in the query separately, such that it is known if all criteria are matched by the patient. Our component models relevance towards each of the inclusion criteria by adapting techniques from search result diversification to promote patients whose medical records are relevant to several different inclusion criteria.

In Chapter 8, we introduce an approach to instantiate this component and evaluate our proposed approach, which aims to highly rank patients that are relevant to multiple inclusion criteria in the query.

4.3 Baseline Patient Search System

In this section, we describe how we evaluate the effectiveness of patient search. This forms the basis under which we examine the performance of the task-specific representation approach later in this section and the other components of the framework in subsequent chapters. In Section 4.3.1, we describe the experimental setup, while Section 4.3.2 discusses the research questions investigated in this chapter. In Section 4.3.3, we investigate the effectiveness of applying aggregate ranking approaches (as discussed in Section 4.2.2) for the patient search task. In Section 4.3.4, we examine the effectiveness of the representation approaches discussed in Section 4.2.1. Finally, we further analyse the performances of our representation and patient ranking approaches in Section 4.3.5.

4.3.1 Experimental Setup

To evaluate the representation approaches (Section 4.2.1) and the patient ranking approaches (Section 4.2.2) discussed in this chapter, as well as approaches proposed in this thesis, we use the test collection provided by the TREC 2011 and 2012 Medical Records track, previously introduced in Section 3.4.1. The aim is to retrieve patient *visits* based on the relevance of their associated medical records towards a query. To avoid privacy issues, a visit, which contains a set of medical records associated with a patient during a visit to the hospital, is used to represent a patient. Recall that the test collection consists of 101,711 medical records, which are associated to 17,265 patient visits, and includes 34 and 47 queries from TREC 2011 and 2012, respectively. A query describes the medical conditions of the targeted patients.

We use the Terrier retrieval platform[1] (Ounis *et al.*, 2006) to index the medical records corpus, applying Porter's English stemmer and removing stopwords. For a patient model-based approach (i.e. using a patient model in the patient ranking unit in Figure 4.1), we estimate the relevance towards a given query of the patient documents (Equation (4.1)) using two effective weighting models, namely, DPH from the Divergence from Randomness (DFR) framework and BM25, discussed in Section 2.4. DFR DPH is a parameter-free model, where all parameter values are automatically derived from the collection statistics. On the other hand, we use the default parameters for BM25, where $k_1 = 1.2, k_3 = 1000, b = 0.75$. For the two-stage model, we use either BM25 or DFR DPH to initially rank the medical records for a given query, before deploying different voting techniques from the Voting Model, including CombSUM, CombMAX, expCombSUM and expCombMNZ, to aggregate the relevance scores of medical records to estimate the relevance of their associated patients, as they have shown to be effective for other aggregate ranking tasks (Macdonald, 2009; Macdonald & Ounis, 2006). Table 4.4 describes how each of the four voting techniques used calculates the relevance of a patient. In addition, we measure statistically significant differences between the retrieval performance achieved by our approaches and the baselines using the paired t-test[2] at $p < 0.05$.

4.3.2 Research Questions

Our aim is to determine an effective ranking patient baseline. We investigate the effectiveness of existing term-based and conceptual representation approaches, in comparison with the novel task-specific representation approach (see Section 4.2.1). In addition, we examine how to effectively adapt existing

[1] http://terrier.org

[2] Note that the paired t-test assumes that the difference in performance between the two approaches being compared is normally distributed. Visual inspection leaves us confident that this assumption is reasonable.

Voting technique	Description
CombSUM	sum of the relevance scores of retrieved records associated to a given patient.
CombMAX	the maximum of the relevance scores of retrieved records associated to a given patient.
expCombSUM	sum of the exponential of the relevance scores of the retrieved records associated to a given patient.
expCombMNZ	the product of expCombSUM and the number of retrieved records associated to a given patient.

Table 4.4: Voting techniques used in our experiments.

aggregate ranking approaches for the patient ranking task. We investigate the following four research questions within Sections 4.3.3, 4.3.4 and 4.3.5:

RQ 1. Which aggregate ranking approach is the most effective for ranking patients based on the relevance of their medical records?

RQ 2. What is the effective way to select medical concepts extracted using the MetaMap tool with different levels of confidence to represent the medical records and queries?

RQ 3. Is the task-specific representation approach that uses only medical concepts related to the four medical aspects effective for the patient search task?

RQ 4. What is the effective number of voting medical records for ranking patients when using a voting technique (i.e. a two-stage model)?

4.3.3 Experiments with the Adaptation of Aggregate Ranking Approaches for Ranking Patients

In this section, we evaluate the approaches based on the patient and the two-stage models for the patient search task. We initially limit the number of voting medical records for the two-stage model ($|R(Q)|$ in Equation (4.2)) to 5,000, as we will show later in Section 4.3.5.2 that it is an effective setting.

Table 4.5 compares the retrieval performances in terms of bpref for TREC 2011 and infNDCG for TREC 2012 between the patient model and the two-stage model, when applied with a term-based representation. From Table 4.5, we observe that CombMAX, expCombMNZ and expCombSUM are more effective than the patient model (DPH and BM25). In particular, we observe that using expCombSUM to aggregate the relevance scores of the medical records initially ranked using BM25 is the most effective approach under both bpref and infNDCG for TREC 2011 and 2012, respectively. The attained retrieval performances significantly outperform the DPH instance of the patient model (paired t-test,

Approach	bpref (TREC 2011)	infNDCG (TREC 2012)
The Patient Models		
DPH	**0.4542**	0.3779
BM25	0.4539	**0.3944**
The Two-Stage Models		
CombSUM (+DPH)	0.3656	0.3367
CombSUM (+BM25)	0.3663	0.3338
CombMAX (+DPH)	0.4827	0.4119
CombMAX (+BM25)	0.4976▲,△	0.4193▲
expCombMNZ (+DPH)	0.4774	0.4254
expCombMNZ (+BM25)	0.4725	0.4274▲
expCombSUM (+DPH)	0.4871	0.4167
expCombSUM (+BM25)	**0.5018**▲,△	**0.4343**▲

Table 4.5: Comparing retrieval performances of different patient ranking approaches on TREC 2011 and 2012 Medical Records track's queries. Statistical significance (paired t-test) at $p < 0.05$ over the patient model baselines (DPH and BM25) are denoted ▲ and △, respectively.

$p < 0.05$) for both TREC 2011 and 2012. Indeed, the performance improvements over the DPH baseline are 10.48% and 14.92% for TREC 2011 and 2012, respectively. These results are in line with the previous works in expert search, which indicate that the two-stage model is often more effective than the patient model (Balog *et al.*, 2006). In particular, we suggest that the expCombSUM and expCombMNZ voting techniques should be used to rank patients based on the relevance of their medical records. Importantly, answering the first research question, the two-stage model is more effective than the patient model. In particular, using the expCombSUM voting technique to aggregate the relevance scores of medical records ranked using the BM25 weighting model is significantly (paired t-test, $p < 0.05$) better than using either DPH or BM25 to rank patient documents directly. As we have shown that the two-stage model is more effective than the patient model, the remainder of the thesis will focus only on the approaches based on the two-stage model.

4.3.4 Experiments with the Task-Specific Representation Approach

Next, we evaluate the retrieval performance of the task-specific representation approach (introduced in Section 4.2.1), when applied with the CombSUM, CombMAX, expCombMNZ and expCombSUM voting techniques. As before, we limit the number of voting documents to 5,000. We compare the achieved retrieval performances with an existing conceptual representation approach that uses the medical concepts identified with any semantic types to represent the medical records and queries (i.e. the approach used by Qi & Laquerre (2012), see Section 3.4.2.3), as well as the term-based representation approach used in the previous experiment. As discussed in Section 4.2.1, the MetaMap tool used identifies the

Concepts Identified in Medical Records	Concepts Identified in Queries	
	Any Level of Confidence	Highest Level of Confidence
Any Level of Confidence	V1	V2
Highest Level of Confidence	V3	V4

Table 4.6: Our examined four variants of selecting medical concepts identified by the MetaMap tool to represent the medical records and queries

medical concepts in the medical records and queries with different levels of confidence. We evaluate both the task-specific representation approach and the existing conceptual representation approach, when using different variants of choosing medical concepts identified with the different confidence levels. In particular, we examine four variants, as shown in Table 4.6.

The first variant (V1) uses all medical concepts identified in the medical records and queries with any level of confidence. Second, V2 uses all medical concepts identified in the medical records with any level of confidence, while using only medical concepts identified in the queries with the highest confidence[1]. Third, V3 uses only medical concepts identified in the medical records with the highest confidence, but using medical concepts identified in the queries with any level of confidence. Fourth, V4 uses only the medical concepts with the highest confidence identified in the medical records and queries.

Table 4.7 compares the retrieval effectiveness of the task-specific representation approach with the existing conceptual representation approach, when applied with the four variants previously described. In addition, the TREC Median baseline and the most effective term-based representation approach in Table 4.5 are also reported (i.e. the term-based representation, when using expCombSUM to aggregate the relevance scores of medical records initially ranked using BM25). From Table 4.7, we firstly observe that using the medical concepts identified in the medical records and queries at any confidence levels to represent the medical records and queries (i.e. the variant V1) results in the most effective retrieval performances for both the task-specific representation approach and the conceptual representation baseline. To answer our second research question, medical concepts identified using the MetaMap tool with any level of confidence should be used to represent the medical records and queries. This is due to the fact that sometimes the MetaMap tool may erroneously identify medical concepts from a medical record or a query. Hence, using all possible identified concepts can avoid missing important medical concepts. Next, we find that the task-specific representation approach markedly outperforms the corresponding conceptual representation baseline for both TREC 2011 and 2012 for up to 13.2% and 12.3%, respectively. Importantly, the task-specific representation approach significantly (paired t-test, $p < 0.05$) outperforms the corresponding conceptual representation approach, for most of the voting techniques

[1] Identified as “Meta Mapping”, as shown in Figure 4.2

Approach	bpref (TREC 2011)		infNDCG (TREC 2012)	
TREC Median	0.4120		0.4244	
Term-based Representation				
expCombSUM (+BM25)	0.5018		0.4343	
V1 (Using concepts with any confidence level)				
	Task-specific	Conceptual	Task-specific	Conceptual
CombSUM (+DPH)	$0.3918^{\triangle}$	0.3619	0.3703	0.3421
CombSUM (+BM25)	$0.3914^{\triangle}$	0.3662	0.3899	0.3633
CombMAX (+DPH)	$0.5087^{\triangle}$	0.4429	$0.4686^{\triangle}$	0.4224
CombMAX (+BM25)	$0.5180^{\triangle}$	0.4625	$0.4785^{\triangle}$	0.4444
expCombMNZ (+DPH)	$0.5048^{\triangle}$	0.4464	$0.4737^{\triangle}$	0.4294
expCombMNZ (+BM25)	$0.5048^{\triangle}$	0.4648	**0.4908**$^{\blacktriangle}$	**0.4640**
expCombSUM (+DPH)	$0.5183^{\triangle}$	0.4452	$0.4704^{\triangle}$	0.4221
expCombSUM (+BM25)	**0.5243**$^{\triangle}$	**0.4630**	$0.4880^{\blacktriangle,\triangle}$	0.4535
V2 (Using concepts with any confidence level for records, and with only the highest confidence for queries)				
CombSUM (+DPH)	$0.3932^{\triangle}$	0.3588	$0.3779^{\triangle}$	0.3366
CombSUM (+BM25)	$0.3979^{\triangle}$	0.3600	$0.3864^{\triangle}$	0.3520
CombMAX (+DPH)	$0.4888^{\triangle}$	0.4306	$0.4309^{\triangle}$	0.3881
CombMAX (+BM25)	$0.5002^{\triangle}$	0.4529	0.4333	0.4012
expCombMNZ (+DPH)	$0.4843^{\triangle}$	0.4305	$0.4416^{\triangle}$	0.4016
expCombMNZ (+BM25)	0.4832	0.4484	**0.4538**$^{\triangle}$	**0.4185**
expCombSUM (+DPH)	$0.4929^{\triangle}$	0.4333	$0.4218^{\triangle}$	0.3922
expCombSUM (+BM25)	**0.5074**$^{\triangle}$	**0.4544**	$0.4510^{\triangle}$	0.4165
V3 (Using concepts with the highest confidence for records, and with any confidence level for queries)				
CombSUM (+DPH)	0.3846	0.3527	$0.3214^{\triangle}$	0.2525
CombSUM (+BM25)	0.3868	0.3535	$0.3286^{\triangle}$	0.2629
CombMAX (+DPH)	$0.4425^{\triangle}$	0.3837	$0.3602^{\triangle}$	0.2528
CombMAX (+BM25)	$0.4601^{\triangle}$	0.3910	$0.3659^{\triangle}$	0.2552
expCombMNZ (+DPH)	$0.4462^{\triangle}$	0.3881	$0.3737^{\triangle}$	0.2696
expCombMNZ (+BM25)	$0.4545^{\triangle}$	**0.3955**	**0.3810**$^{\triangle}$	**0.2730**
expCombSUM (+DPH)	$0.4513^{\triangle}$	0.3842	$0.3657^{\triangle}$	0.2557
expCombSUM (+BM25)	**0.4633**$^{\triangle}$	0.3920	$0.3721^{\triangle}$	0.2553
V4 (Using concepts only with the highest confidence)				
CombSUM (+DPH)	$0.3805^{\triangle}$	0.3438	$0.3669^{\triangle}$	0.3182
CombSUM (+BM25)	$0.3889^{\triangle}$	0.3444	$0.3708^{\triangle}$	0.3265
CombMAX (+DPH)	$0.4451^{\triangle}$	0.3729	$0.4055^{\triangle}$	0.3499
CombMAX (+BM25)	$0.4624^{\triangle}$	**0.3938**	$0.4105^{\triangle}$	0.3467
expCombMNZ (+DPH)	$0.4374^{\triangle}$	0.3763	$0.4163^{\triangle}$	**0.3665**
expCombMNZ (+BM25)	$0.4491^{\triangle}$	0.3896	$0.4195^{\triangle}$	0.3622
expCombSUM (+DPH)	$0.4479^{\triangle}$	0.3750	$0.4147^{\triangle}$	0.3574
expCombSUM (+BM25)	**0.4628**$^{\triangle}$	0.3919	**0.4207**$^{\triangle}$	0.3538

Table 4.7: The comparison of the retrieval performances of different query and medical record representation approaches on the TREC 2011 and 2012 Medical Records track's test topics. Statistical significance (paired t-test) at $p < 0.05$ over the term-based representation baseline and the corresponding conceptual representation baseline are denoted $^{\blacktriangle}$ and $^{\triangle}$, respectively.

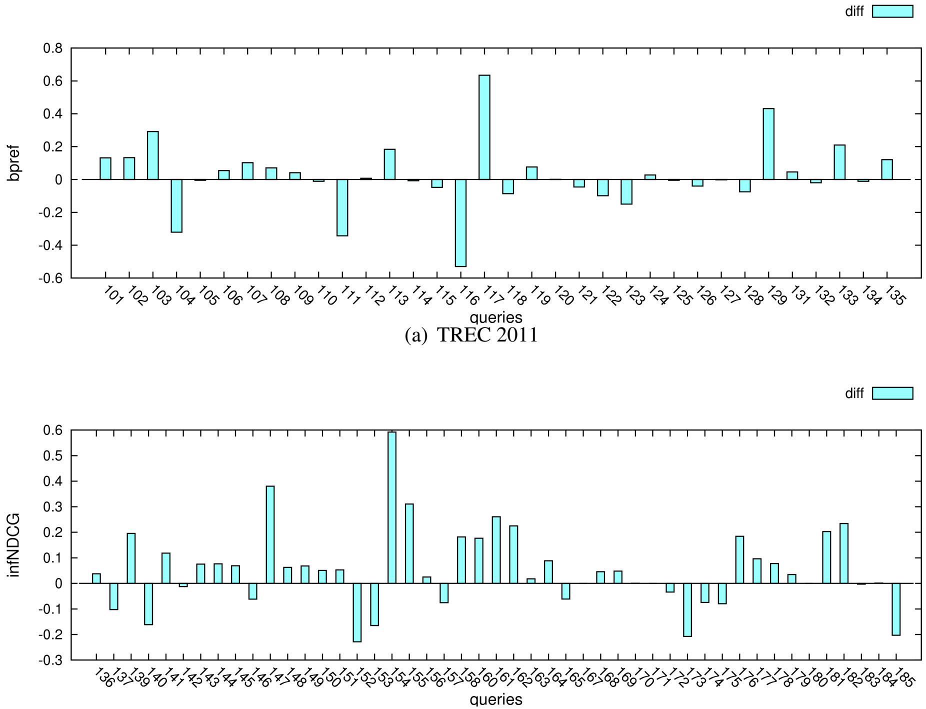
tested. On the other hand, comparing with the most effective term-based representation approach, the task-specific representation approach when applied with expCombSUM and BM25 also markedly outperforms the term-based representation baseline for both TREC 2011 (bpref 0.5243 vs 0.5018) and 2012 (infNDCG 0.4880 vs 0.4343). Indeed, for TREC 2012, the task-specific representation approach performs significantly better than the term-based representation baseline (paired t-test, $p < 0.05$). Furthermore, we find that for the queries from TREC 2011, the task-specific representation approach attains the most effective retrieval performance when applied with expCombSUM and BM25, while for TREC 2012 queries, the approach is most effective when applied with expCombMNZ and BM25. Importantly, based on the effective retrieval performance achieved by the task-specific representation approach, we find that the medical records and queries should be represented within a patient search system by using only the medical concepts related to the four medical aspects. This answers our third research question. Moreover, when comparing with the TREC Median baseline, we observe that when using with CombMAX, expCombMNZ or expCombSUM, the task-specific representation approach (with V1 variant) markedly outperforms the TREC Median baseline. For instance, using BM25 and expCombSUM, the task-specific representation performs +27.3% better than the TREC Median baseline. Due to its effectiveness, for the remaining of the thesis, we will use the V1 variant of the task-specific representation approach (i.e. using medical concepts identified in the medical records and queries with any level of confidence).

4.3.5 Analysis and Discussion

This section provides further analysis and discussion with respect to the performances of the task-specific representation approach that represents the medical records and queries by using medical concepts related to the four aspects of the medical decision process. In particular, Section 4.3.5.1 discusses the types of queries that are likely to benefit from the proposed approach. In Section 4.3.5.2, we discuss the impact of the number of medical records used to vote for the relevance of patients.

4.3.5.1 Failure Analysis

In this section, we conduct a failure analysis for the task-specific representation approach. In particular, in Figure 4.7, we compare the retrieval performances achieved by the task-specific representation approach and the term-based representation approach, on a per-query basis. BM25 and the expCombSUM voting technique are used to rank medical records and patients, respectively, since it is effective for both the term-based representation and the task-specific representation approaches, as shown in Tables 4.5 and 4.7. The limit on the number of voting documents is 5,000 as used in Sections 4.3.3 and 4.3.4. From



(a) TREC 2011

(b) TREC 2012

Figure 4.7: The difference between the performances of our task-specific representation approach and the term-based representation approach on the TREC 2011 and 2012 Medical Records track, on a per-query basis.

Figure 4.7, we first observe that the proposed approach performs better than the term-based representation approach for 17 out of 34 queries and 31 out of 47 queries for TREC 2011 and 2012, respectively. Meanwhile, the proposed approach decreases the retrieval performances for 17 out of 34 queries from TREC 2011 and 14 out of 47 queries for TREC 2012. These results show that the task-specific representation approach is effective for many queries, especially for TREC 2012 queries.

Next, in Table 4.8, we compare the numbers of queries benefited or harmed by the task-specific representation approach, by grouping the types of medical concepts that can be extracted from the queries. We find that the proposed approach is likely to be effective when a query contains medical concepts related to all of the four aspects of medical decision process (67.7%), and when a medical concept related to a treatment or a symptom is identified in a query (63.5% and 62.9% respectively). However, the approach is less likely to be beneficial if a medical concept related to a diagnostic test is identified in the query (51.9%).

Aspects of medical concepts in the query	Benefited	Harmed
Symptom	62.9% (44/70)	37.1%(26/70)
Diagnostic test	51.9% (27/52)	46.2%(24/52)
Diagnosis	60.3% (41/68)	36.8% (25/68)
Treatment	63.5% (33/52)	34.6% (18/52)
All 4	67.7% (21/32)	34.4%(11/32)

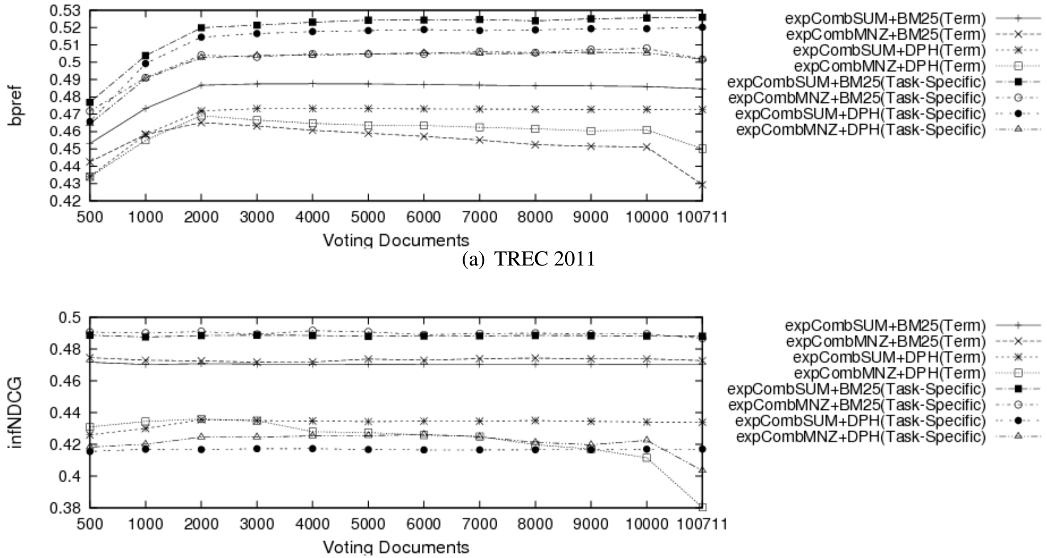
Table 4.8: Analysis of the task-specific representation approach w.r.t. the aspects of medical concepts found in the queries. The numbers between the parentheses indicate the number of queries impacted (benefited/harmed) compared to the total number of queries.

4.3.5.2 The Impact of the Numbers of Voting Documents

Next, we analyse the impact of the number of voting documents used to rank patients using the expCombSUM and expCombMNZ voting techniques, since they are effective for this patient ranking task, as shown in Section 4.3.4. In Figure 4.8, we show the achieved retrieval performance, as we vary the number of voting documents from 500 to 100,711 (i.e. all medical records in the collection). From Figure 4.8(a), we observe that for TREC 2011 queries, the retrieval performances are improved when increasing the number of voting documents from 500 to around 4,000, and then become stable. Meanwhile, for TREC 2012 queries (shown in Figure 4.8(b)), the retrieval performances are not changed markedly as we vary the number of voting documents. However, we find that the retrieval performance of the expCombMNZ voting technique decreases when the number of voting documents is very high (e.g. using all of the medical records in the collection). This is because those patients who have many medical records are promoted in the ranking by expCombMNZ, even though their associated medical records may not be relevance to the query. Overall, we observe that 5,000 is in general an effective number of voting documents, answering the fourth research question. As before, we observe that the task-specific representation approach outperforms the term-based representation approach for all variants of the examined numbers of voting documents.

4.4 Conclusions

In this chapter, we described our patient search system that consists of the query and medical record representation unit, the patient ranking unit and the novel framework for enhancing the representations of queries and medical records (Section 4.2). The query and medical record representation unit defines how queries and medical records are represented within the system (e.g. using terms or concepts) (see Section 4.2.1). We introduced the task-specific representation approach, which is a conceptual representation approach that uses only medical concept related to the medical decision process (symp-



(a) TREC 2011

(b) TREC 2012

Figure 4.8: The retrieval performances achieved by the different voting techniques tested, while varying the number of voting documents.

toms, diagnostic tests, diagnoses, and treatments). Next, the patient ranking unit used either the patient model or the two-stage model to rank patients based on the relevance of their medical records (see Section 4.2.2). In Section 4.2.3, we introduced the novel framework that uncovered the implicit knowledge from the medical records and queries at the three levels of the retrieval process (including, sentence, record and inter-record levels). In particular, the framework consists of four components, two of which uncover the implicit knowledge at the inter-record level, while the other two components performs at the sentence and the record levels of the retrieval process, respectively (see Figure 4.4).

We evaluated the effectiveness of baseline patient search systems in addition to the task-specific representation approach in Section 4.3. For the Query and Medical Record Representation Unit (Section 4.2.1), we observed that the task-specific representation approach significantly outperformed both the existing term-based and conceptual representation approaches by up to 45% (see Table 4.7). In particular, we found that using the medical concepts identified using the MetaMap tool with any level of confidence (i.e. V1 in Table 4.6) resulted in an effective retrieval performance. In addition, from Table 4.5, we found that using the expCombSUM voting technique to aggregate the relevance scores of medical records computed using BM25 (i.e. using a two-stage model) was more effective than using any patient model (bpref 0.5018 and infNDCG 0.4343). Hence, for the Patient Ranking Unit (Section 4.2.2), the two-stage model that uses BM25 to rank medical records before using the expCombSUM voting technique to aggregate relevance scores of patients is the most effective instantiation. Indeed, the effective number of voting documents for this instantiation is 5,000. Therefore, we will use this approach as a baseline, when evaluating our works in the next chapters.

Next, in Chapters 5, 6, 7 and 8, we describe in more detail and evaluate the components of the proposed framework that leverage the implicit knowledge at different levels of the retrieval process. Meanwhile, in Chapter 9, we examine the combination of these components to enhance overall performance. In the next chapter, we discuss and evaluate the first component of the framework, which deals with negated language in patient search.

Chapter 5

Negation Handling

5.1 Introduction

This chapter introduces our approaches to instantiate the first component of our framework (i.e. the Negation Handling component), which handles negated language in patient search by uncovering the implicit knowledge at the sentence level. As discussed in Sections 1.1 and 4.2.3.1, medical practitioners commonly use negated language to indicate that a patient does not have a given medical condition. For example, the sentence “patient has no fever” in a medical record indicates that the patient has been tested and was confirmed not to have fever. However, traditional information retrieval systems may not distinguish between the positive and negative contexts of terms when indexing and retrieving medical documents (see Section 3.4.2.4). Without considering whether the term occurs in a negative or positive context, the sole presence of a query term in a medical record is insufficient to imply that the patient who is associated to the record is relevant to the query. For example, when searching for patients with angina, a retrieval system might wrongly consider a patient who has a medical record stating “no evidence of angina” to be relevant. An effective patient search system should therefore highly rank only the patients whose medical records contain terms with the same contexts as the corresponding query terms. In this chapter, we propose novel approaches, which at the sentence level recognise the context of the medical conditions in medical records and queries and rank highly patients whose medical records have occurrences of the query medical conditions with the appropriate contexts. The remainder of this chapter is organised as follows:

- Section 5.2 first describes two novel non-supervised negation handling approaches. The first approach aims to represent terms with positive and negative contexts differently to enable a retrieval system to distinguish between terms with either of the two contexts. The second approach exploits

Patient admitted with **cancer**
Diagnosed and found no evidence of **cancer**
Negative result on **cancer** screening test

Table 5.1: Examples of sentences in EMRs where the presence of the query term 'cancer' does not always indicate the relevance.

the dependence between terms within medical records to demote the medical records containing query terms with a non-relevant context.

- In Section 5.3, we propose our learned approach that learns how to appropriately weight the occurrences of the opposite context of any query term, thus preventing patients whose medical records may not be relevant from being retrieved.
- In Section 5.4, we thoroughly evaluate both our non-supervised and learned approaches to deal with the negated language in patient search using the TREC Medical Records track's test collection, in comparison with the effective, common baseline identified in Chapter 4, Section 4.3.
- Section 5.5 provides concluding remarks for this chapter.

5.2 Non-Supervised Approaches to Handle Negation

As discussed in Section 3.4.2.4, the presence of a query term does not always imply that the medical record is relevant to the query. In particular, the relevance also depends on the context of query terms occurring in the medical records. For example, while all the three sentences shown in Table 5.1 contain the query term 'cancer', only the first sentence indicates that the patient has 'cancer'. The other two sentences are non-relevant towards the query searching for patients suffering from 'cancer'. In particular, several top performing search systems (e.g. Demner-Fushman *et al.* (2011); King *et al.* (2011); Zhu & Carterette (2012)) at the TREC 2011 and 2012 Medical Records track showed that dealing with the negative contexts of terms in medical records led to an effective retrieval performance (see Section 3.4.2.4).

To cope with negated language in patient search, in this section we propose two non-supervised approaches. First, term representation approach, called *NegFlag*, facilitates the handling of negative context in medical records and queries. Second, a novel term dependence approach demotes the medical records containing query terms in the opposite required context.

Original record	Patient reports palpitations but does not have fever
Negation detection	Patient reports palpitations but does *not have fever*
Removing stopwords	Patient reports palpitations *fever*
NegFlag representation	Patient reports palpitations *n$fever*

Table 5.2: The NegFlag process for a medical record - italicised terms occur in a negative context.

5.2.1 The NegFlag Approach

Our negated term representation approach, *NegFlag*, modifies the indexing and retrieval processes to distinguish between positive and negative context terms in the sentences of the medical records and queries, which are identified using the NegEx algorithm. The identified negated terms are replaced with special negated versions of those terms. Table 5.2 shows how an example sentence is processed using NegFlag, such that the term 'fever' is replaced with its negated version, 'n$fever'.

Given the query "find patients with fever", without NegFlag, a ranking model might erroneously rank the original record in Table 5.2 highly, because it contains all of the query terms. However, after NegFlag processing, as 'fever' becomes 'n$fever', the record would not score as highly.

5.2.2 Term Dependence for Negation

Even though the NegFlag helps to retrieve records containing query terms with required contexts, it does not prevent records with the occurrences of query terms with the opposite contexts, which we refer to as the *opposite context terms*, from being retrieved. For example, the NegFlag processed example in Table 5.2 might still be retrieved for the query "find patient with fever and palpitations", when the patient is known not to have fever. To alleviate this, we propose the use of term dependence models to demote medical records containing the opposite context terms.

Term dependence models (e.g. Markov Random Fields (Metzler & Croft, 2005)) have been used to improve effectiveness by scoring higher documents containing many occurrences of pairs of query terms in close proximity. In contrast, we propose to use term dependence to *demote* records containing the *opposite context* form of neighbouring terms occurring in the queries. For example, given a query "chest pain", medical records containing the pair of terms 'n$chest' and 'n$pain' should be demoted. To this effect, we score medical record d for a query q, taking negation into account, as follows:

$$score\,(d, q) = \sum_{t \in q} score(d, t) - \sum_{\langle t_1, t_2 \rangle \in q'} score(d, \langle t_1, t_2 \rangle) \tag{5.1}$$

There are two components in Equation (5.1), namely the positive scoring of query terms, and the negative dependence score for the opposite context terms. $score(d, t)$ is the score assigned to a query term t in medical record d using any term weighting model, q' is the set of the opposite context terms in q, and $\langle t_1, t_2 \rangle$ is a pair of the opposite context terms in q'. Two types of term dependence are possible (Metzler & Croft, 2005; Peng *et al.*, 2007): for full dependence (FD), $\langle t_1, t_2 \rangle$ is the set that contains unordered pairs of neighbouring terms; for sequential dependence (SD), $\langle t_1, t_2 \rangle$ is the set that contains ordered pairs of neighbouring terms. For $score(d, \langle t_1, t_2 \rangle)$, we use the binomial randomness model pBiL (Peng *et al.*, 2007) from the Divergence from Randomness (DFR) framework to score the occurrences of a pair of terms within $window\_size$ tokens in a medical record d.

5.3 A Learned Approach to Handle Negation

In this section, we propose a novel learned approach for preventing medical records clearly stating that their associated patients do not have the medical conditions stated in the query, or stating that the patients have medical conditions that the query aims to exclude, from being ranked highly. In particular, our approach consists of three components. Firstly, as we intend to promote medical records having query terms with their intended context and to demote those containing the query terms with the opposite context, we deploy the NegFlag approach (previously discussed in Section 5.2.1) to represent terms in both medical records and queries by taking their context into account. Secondly, we penalise the medical records containing the query terms with the opposite context (i.e. the opposite context terms), in order to prevent the medical records having the occurrences of the query terms with the opposite intended context from being ranked highly. For example, for a query 'hypertension', the opposite context term is 'n$hypertension'. Finally, we set an effective penalising weight for each of the opposite context terms, to reduce the relevance scores of the medical records containing these terms. Specifically, we deploy a regression technique to identify the penalising weight of an opposite context term using features (e.g. term frequency, and co-occurrence information), obtained from the query and the medical records. Specifically, our learned approach for handling negation consists of three components:

1. *Context identification*, to identify and represent the context of the terms in both medical records and queries;

2. *Context-based penalisation*, to model the penalisation of medical records containing the opposite context terms when ranking medical records;

3. *Penalising weight estimation*, to accurately weight the opposite context terms to prevent medical records that are likely to be non-relevant to the query from being ranked highly.

Next, we discuss these three components in detail.

5.3.1 Context Identification

The context identification phase is an important component of our learned approach, as it helps a search system to distinguish between a term with different contexts (e.g. diabetes and no diabetes). In particular, this component pre-processes medical records and queries by using a negation detection tool to identify negated terms. We deploy the NegFlag approach (previously introduced in Section 5.2.1), which uses the NegEx algorithm to differentiate between terms having positive and negative contexts in each sentence in both the medical records and queries. Then, terms with the negative context are replaced with their negative version before processing in an IR system.

However, as mentioned in Section 5.2.2, even though the context identification component can improve the representation of medical records and queries, it could not prevent non-relevant medical records that contain some of the query terms with the correct context from being retrieved. We introduce the second component to deal with this problem in the next section.

5.3.2 Context-based Penalisation

To decrease the likelihood that non-relevant medical records (indicated by the occurrence of the opposite context terms) are retrieved, the second component of our learned approach penalises these medical records based on the occurrences of the opposite context terms. In particular, if a query searches for a particular context (e.g. positive) of a term but a medical record contains the query term with the opposite context (e.g. negative), the medical record is likely to be non-relevant. For example, for a query “find patient with diabetes and lung cancer”, a medical record stating “patient with lung cancer who does not have diabetes” is non-relevant, since the medical record clearly states that the associated patient does not have a medical condition that the query is searching for (i.e. diabetes). However, as shown in Table 5.3, with only the context identification component, the record is represented as “lung cancer n$diabetes”. As a result, this medical record may still be retrieved since it contains two of the three query terms (i.e. ‘lung’ and ‘cancer’). The context-based penalisation component copes with this issue by reducing the relevance score of medical records, if they contain a term t' with the opposite context to its corresponding query term t (e.g. ‘n$diabetes’ is the opposite context term corresponding to query term ‘diabetes’). This component models the relevance score of a medical record based on both the occurrence of the query terms and the opposite context terms, so that the relevance score of the medical records containing the opposite context terms will be penalised, while the relevance score will be increased if the query terms with the correct context occur in the medical records. To do so, the terms having the opposite context to their corresponding query terms are added to the query with a particular weight to penalise the relevance score of medical records containing these opposite context

Context identified EMR	lung cancer n$diabetes
Context identified query	diabetes lung cancer
Context-based penalisation	diabetes lung cancer *n$diabetic*$\cdot w_1$ *n$lung*$\cdot w_2$ *n$cancer*$\cdot w_3$

Table 5.3: An example of how our learned approach deals with negation, where w_n is the weight of a term.

terms. For example, as shown in Table 5.3, ‘n$diabetes’, ‘n$lung’ and ‘n$cancer’, which are the opposite context terms of the query terms ‘diabetes’ ‘lung’ ‘cancer’, respectively, are added to the query with the penalising weight w_n. Equation (5.2) shows how the second component of the framework calculates the relevance score of a medical record d towards a query q.

$$\begin{aligned} score_{context\_penalise}(d,q) = & \sum_{t \in q} score(d,t) \\ & + \sum_{t' \in opposites(q)} w(t') \cdot score(d,t') \end{aligned} \tag{5.2}$$

where $opposites(q)$ returns a set of the opposite context terms (e.g. ‘n$diabetes’, ‘n$lung’ and ‘n$cancer’ are the opposite context terms of the query illustrated in Table 5.3), $w(t')$ is the weight of an opposite context term t' (e.g. w_1, w_2 and w_3 in Table 5.3) – typically $w(t') < 0$, to penalise occurrences of t'. $score(\cdot)$ can be calculated using any term weighting model, such as BM25. Indeed, the first part of the equation is the classical document scoring approach that estimates the relevance of a medical record based on the appearance of a query term t. On the other hand, the second part of Equation (5.2) aims to penalise the medical records that contain an opposite context term t' (e.g. ‘n$diabetes’).

From Equation (5.2), we draw attention to $w(t')$, which is a crucial parameter for effectively penalising a medical record. Indeed, there are different alternatives to estimate $w(t')$, such as giving a fixed weight to all the opposite context terms. However, to effectively estimate the weight of the opposite context terms, in the next section, we introduce the last component of our learned approach, which deploys a regression technique to learn the effective weight of the opposite context terms using several statistical features.

5.3.3 Penalising Weight Estimation

The penalising weight estimation component focuses on finding an appropriate weight to penalise the medical records containing an opposite context term (i.e. $w(t')$ in Equation (5.2)), to prevent these medical records from being ranked highly, hence leading to an effective retrieval performance. We argue that not all opposite context terms are equally important for penalising the relevance of medical records. For example, consider the query “find a patient with diabetes and lung cancer” in Table 5.3, which is represented as “diabetes lung cancer”. For this query, medical records with the term ‘n$diabetes’ should be penalised more than those containing ‘n$cancer’, as in the former, it is likely that the patient does not suffer from diabetes, while a medical record containing ‘n$cancer’ may discuss a patient who does not have another type of cancer (e.g. the patient does not have kidney cancer). In this way, the discriminative power of the opposite context terms should be considered when assigning penalising weights.

We view this problem of estimating the penalising weight of different opposite context terms as a supervised learning problem, where the objective is to predict an estimated effective penalising weight for each opposite context term, based on the retrieval performance on a training set. By doing so, we benefit from the fact that several features (e.g. term frequency and co-occurrence statistics) of the opposite context terms are taken into account to estimate the penalising weights. Indeed, a regression function $f(\cdot)$ calculates the penalising weight of an opposite context term t' using a set of features $\Phi^{t'}$, which are associated with the term t', as follows:

$$w(t') = f(\Phi^{t'}) \tag{5.3}$$

where $\Phi^{t'}$ is a set of features φ for term t'. Indeed, the regression function $f(\Phi^{t'})$ aims to approximate the weight $w(t')$ in Equation (5.2) using a particular loss function. We use the root-mean-square error (RMSE) as the loss function, calculated as:

$$RMSE(w(t'), O(t')) = \arg\min_{\varphi \in \Phi^{t'}} \sqrt{\frac{\sum^{|\Phi^{t'}|}(w(t') - O(t'))^2}{|\Phi^{t'}|}} \tag{5.4}$$

where $O(t')$ is the oracle weight for the term t' using a training dataset. The procedure to obtain $O(t')$ is discussed in Section 5.3.4.

5.3.4 Learning Procedure

This section describes the procedure to derive the weight of each opposite context term to penalise the relevance score of medical records containing the opposite context terms as discussed in Section 5.3.3. Indeed, Section 5.3.4.1 details the set of features $\Phi^{t'}$ that are used to estimate the weight $w(t')$ for the unseen queries. Then, we explain how the estimated effective penalising weight $O(t')$ for an opposite context term t' is obtained in Section 5.3.4.2. Finally, Section 5.3.4.3 describes the regression technique (learner) and the objective function (i.e. loss function) that we use to learn the penalising weights.

5.3.4.1 Learning Features

We firstly identify a set of features $\Phi^{t'}$ of an opposite context term t' to be used to train a learner to identify the penalising weight of the opposite term. These features should correlate with the weight

$O(t')$ that could bring about the optimal performance, and are generalised across terms. Table 5.4 lists the 13 features used. We focus on features that can be obtained directly from the corpus, which makes our experiments reproducible; however, there may be other features that can be explored in future work (e.g. occurrences of the term in external corpora).

We focus on four types of features, namely term frequency, document frequency, frequency of co-occurrence, and query length. Indeed, the first two types include the classical term and document frequency statistics and their variants (Features 1-8), which model the ubiquity and specificity of a particular term (Lease *et al.*, 2009). Specifically, these features consist of the term occurrence statistics of both an opposite context term t' and its corresponding query term t. The higher value of these features, the more discriminative the term is. Indeed, we use the term and the document frequencies of both t and t', since the importance of the opposite context term t' may depend on the discriminative power of both the opposite context term t' and its corresponding query term t. The next set of features (Features 9-12) are related to the co-occurrence frequency. It has been shown that a term that frequently co-occurs with the query terms often relates to the query (Bai *et al.*, 2007). Therefore, it is intuitive that the medical records containing the opposite context terms that frequently co-occur with the query terms should not be highly penalised. In particular, Features 9-12 measure the co-occurrence of the opposite context term t' with the query terms using different co-occurrence variants, such as the raw number of documents where the term t' and its corresponding query term t co-occur and the EMIM (Expected Mutual Information Measure) (van Rijsbergen, 1977) of terms t' and t. Finally, since a long query tends to be more complex and hence more difficult, Feature 13 counts the number of query terms ($|Q|$). Indeed, a long query provides more evidence to infer the relevance of medical records, and hence it is possible to derive more opposite context terms to penalise the non-relevant medical records.

5.3.4.2 Estimating an Effective Penalising Weight

To identify the effective penalising weight of the opposite context terms, we follow Cao *et al.* (2008) and assume the independence between the opposite context terms, when estimating the penalising weight of each opposite context term one at a time. In particular, on the training set, when estimating the effective penalising weight of each opposite context term ($w(t')$ in Equation (5.2)), we add the opposite context term to the query, and identify the oracle weight $O(t')$ of the opposite context term t', which is the weight that provides the highest retrieval effectiveness, in terms of a particular retrieval measure (e.g. MAP or precision at 10), when ranking medical records using a particular ranking model (e.g. BM25). In particular, we sweep the penalising weight between -1 and 1 to find the best penalising weight for each opposite context term t'. We allow the penalising weight to be between -1 and 1, since it is also

Parameter	Description
$Q = t_1, t_2 ... t_n$	query Q of length n contains query terms $t_1, t_2 ... t_n$
T	# terms in the collection
N	# documents in the collection
t	a term t occurring in the query Q
t'	the term t' having a context opposite to the corresponding query term t (i.e. an opposite context term)
$\mathrm{P}(t_1, t_2)$	the maximum likelihood estimation function of the joint probability of any terms t_1 and t_2, estimated as the fraction of documents where they co-occur
$\mathrm{P}(t_1)$	the maximum likelihood estimation function of the term t_1, estimated as the fraction of documents where the term t_1 occurs

Feature types	ID	Definition
term frequency	1	$tf(t)$: raw frequency of term t in the collection
	2	$tf(t')$: raw frequency of term t' in the collection
	3	$log\frac{tf(t)}{N}$: a variant of the term frequency of t
	4	$log\frac{tf(t')}{N}$: a variant of the term frequency of t'
document frequency	5	$df(t)$: # of documents in the collection that contain term t
	6	$df(t')$: # of documents in the collection that contain term t'
	7	$log\frac{T}{df(t)+1}$: a variant of the invert document frequency of t
	8	$log\frac{T}{df(t')+1}$: a variant of the invert document frequency of t'
co-occurrence frequency	9	$\#co\text{-}occur(t', t)$: # of documents containing both terms t and t'
	10	$co\text{-}occur(t', t) = \log \mathrm{P}(t', t)$: a variant of the co-occurrence between terms t and t'
	11	$co\text{-}occur(t', Q) = \sum_{t_i \in Q} \log \mathrm{P}(t', t_i)$: a variant of the co-occurrence between the term t' and other terms in the query Q
	12	$EMIM(t', t) = \log \frac{\mathrm{P}(t', t)}{\mathrm{P}(t') \cdot \mathrm{P}(t)}$: a variant of the co-occurrence between terms t and t'
query length	13	# of the terms in the query (n)

Table 5.4: List of 13 features used to predict the penalising weight of the opposite context terms.

possible that the occurrences of an opposite context term in a medical record may infer the relevance of a medical record. For example, for a query to find patients with "hearing loss", the medical record stating "patient has difficulty in hearing" (i.e. represented as "n$hearing") is likely to be relevant. Therefore, we allow the penalising weight in Equation (5.2) to be either negative or positive, so that the learner will decide based on the term's features what is the effective penalising weight.

5.3.4.3 Learning the Penalising Weight

From a training dataset, we have examples of penalising weights for opposite context terms and their corresponding features. We then learn to predict the penalising weight of each unseen opposite context term based on its features. In particular, we deploy the Gradient Boosted Regression Trees (GBRT) as a learner, since it has been shown to be effective in several search and regression tasks (see Section 2.6).

We use the root-mean-square error (RMSE) as the loss function (Equation (5.4)) when learning the penalising weight of an opposite context term. Our proposed framework leverages term frequency, document frequency, and the co-occurrence statistics of the terms in the corpus, introduced in Section 5.3.4.1, as learning features for the GBRT learner.

5.4 Experiments

This section evaluates the effectiveness of both our non-supervised and supervised approaches for handling negated language in patient search. In particular, Section 5.4.1 describes the experimental setup to evaluate the proposed approaches. Section 5.4.2 discusses the research questions investigated in this chapter. Sections 5.4.3 and 5.4.4 discuss the experimental results with our non-supervised and learned approaches to handle negated language in patient search, respectively. Finally, Section 5.4.5 further analyses the retrieval performance of the proposed approaches.

5.4.1 Experimental Setup

We evaluate the proposed approaches for handling negated language in patient search using the same settings as in the previous chapter (see Section 4.3.1). We use the TREC Medical Records track's test collection. We compare the retrieval performance of our approaches to handle negation with the common, effective baselines identified previously in Section 4.3. In particular, we firstly use BM25 or DFR DPH to rank medical records based on their relevance towards the queries. Then, we use expCombSUM and expCombMNZ to aggregate the relevance scores of medical records to estimate the relevance of their associated patients, as they were shown to be effective in Chapter 4. In addition, we evaluate the proposed approaches when applied with both a traditional term-based and the task-specific representation baselines, respectively.

As our proposed learned approach (Section 5.3) requires training data, we apply two different training regimes. Firstly, we apply a five-fold cross-validation (5-fold) within each set of queries, since the target measure of TREC 2011 and 2012 queries are different. Secondly, we use the TREC 2012 queries as training data, when testing with the TREC 2011 queries, and vice versa. We refer to this setting as *cross-collection validation* (x-collection).[1] To estimate the oracle weight $O(t')$ as described in Section 5.3.4.2, we target the TREC primary measure (i.e. bpref and infNDCG measures for TREC 2011 and 2012, respectively).

[1] Note that as the number of queries provided by TREC is small (81 queries in total), there is a possibility of overfitting when training using our two regimes.

5.4.2 Research Questions

Our aim is to investigate the effectiveness of the proposed approaches (introduced in Sections 5.2 and 5.3) to deal with negated language in medical records and queries within a search system for retrieving patients based on the relevance of their medical records. In particular, in the remainder of this section, we investigate the following research questions:

RQ 1. Can we enhance the retrieval performance of a patient search system by representing terms along with their context?

RQ 2. Can we effectively demote patients whose medical records contain the opposite context terms to prevent non-relevant patients (as indicated by the occurrence of the opposite context terms) from being ranked highly, by using a term dependence model?

RQ 3. Can we effectively learn to demote patients whose medical records contain the opposite context terms?

RQ 4. What are the effective features for learning the penalising weights for preventing patients whose medical records contain the opposite context terms from being retrieved?

5.4.3 Experiments with the Non-Supervised Approaches

In this section, we evaluate the effectiveness of the proposed approaches for handling negated language using the TREC 2011 and 2012 Medical Records track's test collection. Specifically, we compare the effectiveness of the proposed approaches with that of the common, effective patient ranking baselines (identified in Section 4.3), where negation is not explicitly handled. In addition, we also discuss the retrieval performance achieved by the post-retrieval filtering approach (Widdows, 2003) (i.e. filtering out medical records with opposite context query terms from the initial set of medical records retrieved using a ranking model). We report the retrieval performance in terms of bpref and precision at 10 (P10) for TREC 2011, and infNDCG, infAP and P10 for TREC 2012, respectively.

In Table 5.5, we compare the retrieval performances of our non-supervised approaches with the baselines, when applied with a term-based representation. Specifically, we set $window\_size = 3$ for our term dependence approach for negation. As will be shown later in Section 5.4.5.2, $window\_size = 3$ is an effective setting. We observe that the NegFlag approach (see Section 5.2.1), applied with any used term weighting models and voting techniques, outperforms or at least performs comparably to both the common and the post-retrieval filtering baselines for all the reported measures. Specifically, the best performance that the NegFlag achieved is when BM25 is used to rank medical records and expCombSUM

Approach	TREC 2011		TREC 2012		
	bpref	P10	infNDCG	infAP	P10
Term-based representation					
BM25+expCombSUM					
>> The common baseline	0.5018	**0.5735**	0.4343	0.1828	0.4681
>> Post-retrieval filtering	0.2600$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.2265$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.1672$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.0403$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.1383$^{\blacktriangledown\blacktriangledown\blacktriangledown}$
>> NegFlag	**0.5158**$^{\blacktriangle}$	**0.5735**	**0.4567**$^{\blacktriangle\blacktriangle}$	**0.1984**$^{\blacktriangle}$	**0.5000**$^{\blacktriangle}$
>> NegFlag with SD ($window\_size$ = 3)	0.5136	0.5706	0.4537$^{\blacktriangle}$	0.1974$^{\blacktriangle}$	0.4979
>> NegFlag with FD ($window\_size$ = 3)	0.5129	**0.5735**	0.4538$^{\blacktriangle}$	0.1976$^{\blacktriangle}$	0.4979
BM25+expCombMNZ					
>> The common baseline	0.4725	0.5559	0.4274	0.1730	0.4681
>> Post-retrieval filtering	0.2554$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.2382$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.1514$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.0338$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.1000$^{\blacktriangledown\blacktriangledown\blacktriangledown}$
>> NegFlag	**0.4889**	**0.5676**	0.4435$^{\blacktriangle}$	0.1821	**0.4872**
>> NegFlag with SD ($window\_size$ = 3)	0.4886	0.5647	0.4429$^{\blacktriangle}$	0.1816	0.4830
>> NegFlag with FD ($window\_size$ = 3)	0.4881	0.5647	**0.4444**$^{\blacktriangle}$	**0.1823**	0.4830
DPH+expCombSUM					
>> The common baseline	0.4871	0.5765	0.4167	0.1703	0.4638
>> Post-retrieval filtering	0.4477	0.5235	0.3841	0.1577	0.4340
>> NegFlag	0.5055$^{\blacktriangle}$	**0.5794**	**0.4355**$^{\blacktriangle\blacktriangle}$	0.1833$^{\blacktriangle\blacktriangle}$	**0.4894**
>> NegFlag with SD ($window\_size$ = 3)	**0.5067**$^{\blacktriangle}$	0.5765	0.4346$^{\blacktriangle\blacktriangle}$	0.1835$^{\blacktriangle}$	0.4872
>> NegFlag with FD ($window\_size$ = 3)	0.5061$^{\blacktriangle}$	0.5735	0.4352$^{\blacktriangle\blacktriangle}$	**0.1837**$^{\blacktriangle\blacktriangle}$	**0.4894**
DPH+expCombMNZ					
>> The common baseline	0.4774	0.5676	0.4254	0.1698	0.4596
>> Post-retrieval filtering	0.4373$^{\blacktriangledown}$	0.5176	0.3872	0.1548	0.4170
>> NegFlag	0.4890	**0.5765**	0.4463$^{\blacktriangle\blacktriangle\blacktriangle}$	0.1818$^{\blacktriangle\blacktriangle}$	**0.4936**$^{\blacktriangle}$
>> NegFlag with SD ($window\_size$ = 3)	**0.4892**	0.5735	0.4474$^{\blacktriangle\blacktriangle\blacktriangle}$	0.1823$^{\blacktriangle\blacktriangle}$	0.4894
>> NegFlag with FD ($window\_size$ = 3)	0.4886	0.5647	**0.4488**$^{\blacktriangle\blacktriangle\blacktriangle}$	**0.1830**$^{\blacktriangle\blacktriangle}$	0.4872

Table 5.5: Retrieval performances of our non-supervised approaches to handle negation in comparison to existing approaches, when applied with a term-based representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the traditional approach baselines are denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively.

Approach	TREC 2011		TREC 2012		
	bpref	P10	infNDCG	infAP	P10
Task-specific representation					
BM25+expCombSUM					
>> The common baseline	0.5243	**0.5882**	0.4880	0.2244	0.5128
>> Post-retrieval filtering	0.3933$^{\blacktriangledown\blacktriangledown}$	0.3853$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3518$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.1301$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3404$^{\blacktriangledown\blacktriangledown\blacktriangledown}$
>> NegFlag	0.5310	0.5853	**0.4974**	**0.2288**	**0.5468**$^{\blacktriangle}$
>> NegFlag with SD ($\mathit{window\_size}$ = 3)	0.5307	0.5853	0.4972	0.2281	**0.5468**$^{\blacktriangle}$
>> NegFlag with FD ($\mathit{window\_size}$ = 3)	**0.5311**	0.5853	0.4964	0.2271	**0.5468**$^{\blacktriangle}$
BM25+expCombMNZ					
>> The common baseline	0.5048	0.5824	0.4908	0.2226	0.5149
>> Post-retrieval filtering	0.3711$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3853$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3531$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.1279$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3191$^{\blacktriangledown\blacktriangledown\blacktriangledown}$
>> NegFlag	0.5098	0.5824	**0.4981**	**0.2262**	**0.5255**
>> NegFlag with SD ($\mathit{window\_size}$ = 3)	0.5099	0.5824	0.4975	0.2255	0.5234
>> NegFlag with FD ($\mathit{window\_size}$ = 3)	**0.5100**	**0.5853**	0.4961	0.2245	0.5234
DPH+expCombSUM					
>> The common baseline	0.5183	0.5559	0.4704	0.2078	0.4957
>> Post-retrieval filtering	0.5162	0.5588	0.4727	0.2011	0.5234
>> NegFlag	0.5213	**0.5647**	0.4833$^{\blacktriangle\blacktriangle}$	**0.2105**	**0.5277**$^{\blacktriangle}$
>> NegFlag with SD ($\mathit{window\_size}$ = 3)	**0.5216**	**0.5647**	**0.4834**$^{\blacktriangle\blacktriangle}$	0.2099	0.5255$^{\blacktriangle}$
>> NegFlag with FD ($\mathit{window\_size}$ = 3)	**0.5216**	**0.5647**	0.4819$^{\blacktriangle}$	0.2087	**0.5277**$^{\blacktriangle}$
DPH+expCombMNZ					
>> The common baseline	0.5048	**0.5647**	0.4737	0.2091	0.5106
>> Post-retrieval filtering	0.5026	0.5618	0.4732	0.2036	0.5255
>> NegFlag	0.5101	0.5618	**0.4840**$^{\blacktriangle}$	**0.2107**	**0.5340**$^{\blacktriangle}$
>> NegFlag with SD ($\mathit{window\_size}$ = 3)	**0.5102**	0.5618	0.4837$^{\blacktriangle}$	0.2103	**0.5340**$^{\blacktriangle}$
>> NegFlag with FD ($\mathit{window\_size}$ = 3)	0.5101	0.5618	0.4836$^{\blacktriangle}$	0.2098	**0.5340**$^{\blacktriangle}$

Table 5.6: Retrieval performances of our non-supervised approaches to handle negation in comparison to existing approaches, when applied with the task-specific representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the traditional approach baselines are denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively.

is used to aggregate the relevance scores of medical records to rank patients. The obtained retrieval performances are significantly better than the common baseline, in terms of bpref ($p < 0.05$) for TREC 2011, infNDCG ($p < 0.01$), infAP ($p < 0.05$), and P10 ($p < 0.05$) for TREC 2012. In addition, we find that our term dependence approach can marginally improve the retrieval performance over the NegFlag approach. However, the post-retrieval filtering approach is not effective. This is likely to be because this approach simply discards all medical records containing query terms with the opposite context, while it may be possible that some of these medical records are relevant. For example, the patients with 'anemia' are likely to have 'nausea' from time to time. Correspondingly, their medical records are likely to contain 'nausea' with both positive and negative contexts. Completely ignoring these patients when searching with a query 'nausea' is likely to harm the recall of a search system.

Next, Table 5.6 reports the retrieval performances when our non-supervised approaches are applied with the task-specific representation (discussed in Chapter 4). We observe the same patterns of retrieval performances as when applied with the term-based representation, but with a smaller magnitude. The NegFlag approach outperforms the common approach baseline for many variants of the term weighting models used and voting techniques, and the term dependence approach can further improve the retrieval performances marginally. For example, when applied with BM25 and expCombSUM, the NegFlag approach and our term dependence approach (with full dependence) achieve bpref 0.5310 and 0.5311, respectively, while the retrieval performance of the common baseline is bpref 0.5243. To answer the first research question, representing terms along with their context (i.e. using the NegFlag approach) significantly improves the retrieval performance. Meanwhile, our term dependence approach further improves the retrieval performance, answering the second research question.

Next, comparing between Tables 5.5 and 5.6, we observe our non-supervised approaches are more effective when applied with the term-based representation approach than when applied with the task-specific representation approach. For example, when using BM25 and expCombSUM, the NegFlag approach could improve the retrieval performance in terms of bpref from 0.5018 to 0.5158 (+2.78%) when applied with a term-based representation approach, while the bpref performance is improved from 0.5243 to 0.5310 (+1.28%).

5.4.4 Experiments with the Learned Approach

Tables 5.7 and 5.8 compare the retrieval effectiveness of our learned approach with the common baseline, when applied with a term-based and the task-specific representation approach, respectively. In addition, the retrieval performance of the NegFlag approach and the post-retrieval filtering approach are also reported. However, as our term dependence approach, which also builds upon the NegFlag

Approach	TREC 2011		TREC 2012		
	bpref	P10	infNDCG	infAP	P10
Term-based representation					
BM25+expCombSUM					
>> The common baseline	0.5018	**0.5735**	0.4343	0.1828	0.4681
>> Post-retrieval filtering	0.2600▼▼▼	0.2265▼▼▼	0.1672▼▼▼	0.0403▼▼▼	0.1383▼▼▼
>> NegFlag	**0.5158**▲	**0.5735**	0.4567▲▲	**0.1984**▲	**0.5000**▲
>> Our learned approach (5-fold)	0.5018	0.5706	0.4516▲	0.1960▲	0.4979
>> Our learned approach (x-collection)	0.5092	0.5706	**0.4581**▲▲	0.1977▲	0.4851
>> Our learned approach (oracle)	0.5473▲▲▲	0.6000	0.4821▲▲▲	0.2120▲▲▲	0.5170▲▲
BM25+expCombMNZ					
>> The common baseline	0.4725	0.5559	0.4274	0.1730	0.4681
>> Post-retrieval filtering	0.2554▼▼▼	0.2382▼▼▼	0.1514▼▼▼	0.0338▼▼▼	0.1000▼▼▼
>> NegFlag	**0.4889**	**0.5676**	**0.4435**▲	**0.1821**	**0.4872**
>> Our learned approach (5-fold)	0.4838	0.5618	0.4422	0.1799	0.4809
>> Our learned approach (x-collection)	0.4860	0.5588	0.4420	0.1793	0.4702
>> Our learned approach (oracle)	0.5209▲▲▲	0.5912	0.4661▲▲▲	0.1920▲▲▲	0.5021
DPH+expCombSUM					
>> The common baseline	0.4871	0.5765	0.4167	0.1703	0.4638
>> Post-retrieval filtering	0.4477	0.5235	0.3841	0.1577	0.4340
>> NegFlag	**0.5055**▲	**0.5794**	0.4355▲▲	**0.1833**▲▲	**0.4894**
>> Our learned approach (5-fold)	0.4975	0.5676	0.4331▲	0.1826▲	0.4872
>> Our learned approach (x-collection)	0.5005	0.5647	**0.4357**▲	0.1830▲	0.4830
>> Our learned approach (oracle)	0.5335▲▲▲	0.5853	0.4688▲▲▲	0.1935▲▲▲	0.5149▲▲
DPH+expCombMNZ					
>> The common baseline	0.4774	0.5676	0.4254	0.1698	0.4596
>> Post-retrieval filtering	0.4373▼	0.5176	0.3872	0.1548	0.4170
>> NegFlag	**0.4890**	**0.5765**	0.4463▲▲▲	0.1818▲▲	0.4936▲
>> Our learned approach (5-fold)	0.4806	0.5500	**0.4504**▲▲▲	**0.1829**▲▲	**0.5021**▲
>> Our learned approach (x-collection)	0.4855	**0.5765**	0.4460▲▲	0.1821▲▲	0.4894
>> Our learned approach (oracle)	0.5176▲▲▲	0.6000	0.4680▲▲▲	0.1921▲▲▲	0.5149▲▲

Table 5.7: Retrieval performances of our learned approach to handle negation in comparison to existing approaches, when applied with a term-based representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the common baseline is denoted ▲, ▲▲ and ▲▲▲ (resp. ▼, ▼▼ and ▼▼▼), respectively.

Approach	TREC 2011		TREC 2012		
	bpref	P10	infNDCG	infAP	P10
Task-specific representation					
BM25+expCombSUM					
>> The common baseline	0.5243	**0.5882**	0.4880	0.2244	0.5128
>> Post-retrieval filtering	0.3933$^{\blacktriangledown\blacktriangledown}$	0.3853$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3518$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.1301$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3404$^{\blacktriangledown\blacktriangledown\blacktriangledown}$
>> NegFlag	0.5310	0.5853	**0.4974**	**0.2288**	**0.5468**$^{\blacktriangle}$
>> Our learned approach (5-fold)	0.5296	0.5822	0.4934	0.2254	**0.5468**$^{\blacktriangle}$
>> Our learned approach (x-collection)	**0.5337**$^{\blacktriangle}$	0.5853	0.4957	0.2282	0.5340
>> Our learned approach (oracle)	0.5494$^{\blacktriangle\blacktriangle\blacktriangle}$	0.5588	0.5067$^{\blacktriangle\blacktriangle}$	0.2351$^{\blacktriangle}$	0.5447$^{\blacktriangle}$
BM25+expCombMNZ					
>> The common baseline	0.5048	0.5824	0.4908	0.2226	0.5149
>> Post-retrieval filtering	0.3711$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3853$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3531$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.1279$^{\blacktriangledown\blacktriangledown\blacktriangledown}$	0.3191$^{\blacktriangledown\blacktriangledown\blacktriangledown}$
>> NegFlag	0.5098	0.5824	**0.4981**	**0.2262**	**0.5255**
>> Our learned approach (5-fold)	0.5086	0.5765	0.4977	0.2251	**0.5255**
>> Our learned approach (x-collection)	**0.5125**	**0.5853**	0.4927	0.2234	0.5170
>> Our learned approach (oracle)	0.5244$^{\blacktriangle\blacktriangle}$	0.5647	0.4977	0.2280	0.5511$^{\blacktriangle\blacktriangle}$
DPH+expCombSUM					
>> The common baseline	0.5183	0.5559	0.4704	0.2078	0.4957
>> Post-retrieval filtering	0.5162	0.5588	0.4727	0.2011	0.5234
>> NegFlag	0.5213	**0.5647**	**0.4833**$^{\blacktriangle\blacktriangle}$	**0.2105**	**0.5277**$^{\blacktriangle}$
>> Our learned approach (5-fold)	0.5207	**0.5647**	0.4809$^{\blacktriangle}$	0.2094	0.5234$^{\blacktriangle}$
>> Our learned approach (x-collection)	**0.5253**	**0.5647**	0.4814$^{\blacktriangle}$	0.2089	0.5234$^{\blacktriangle}$
>> Our learned approach (oracle)	0.5363$^{\blacktriangle\blacktriangle}$	0.5382	0.4910$^{\blacktriangle\blacktriangle\blacktriangle}$	0.2148	0.5426$^{\blacktriangle\blacktriangle\blacktriangle}$
DPH+expCombMNZ					
>> The common baseline	0.5048	**0.5647**	0.4737	0.2091	0.5106
>> Post-retrieval filtering	0.5026	0.5618	0.4732	0.2036	0.5255
>> NegFlag	0.5101	0.5618	**0.4840**$^{\blacktriangle}$	0.2107	**0.5340**$^{\blacktriangle}$
>> Our learned approach (5-fold)	0.5088	**0.5647**	0.4826	**0.2110**	0.5298
>> Our learned approach (x-collection)	**0.5105**	0.5618	0.4789	0.2070	0.5234
>> Our learned approach (oracle)	0.5218$^{\blacktriangle\blacktriangle}$	0.5559	0.4919$^{\blacktriangle\blacktriangle\blacktriangle}$	0.2158	0.5489$^{\blacktriangle\blacktriangle}$

Table 5.8: Retrieval performances of learned approach to handle negation in comparison to existing approaches, when applied with the task-specific representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the common baseline is denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively.

approach, could not markedly improve the retrieval performance (as shown in Section 5.4.3), we do not include it in this experiment.

From Tables 5.7 and 5.8, we observe the learned approach for handling negation, with both five-fold cross-validation (5-fold) and cross-collection validation (x-collection) settings, outperforms the common baseline in terms of the TREC primary measures (i.e. bpref and infNDCG for TREC 2011 and 2012, respectively). However, we find that the learned approach could not markedly improve over the NegFlag approach (akin to the context identification component of the learned approach). This means that on this setting the penalising weight estimation component of the learned approach (introduced in Section 5.3.3), could not effectively penalise non-relevant medical records. This is because the learned approach aims to demote medical records containing query terms with the opposite context from the retrieved ranking list; however, the relevance of the retrieved medical records depends only on the occurrence of a small number of query terms. As the evidence (i.e. query terms) used to retrieve medical records is limited, the proposed learned approach could not effectively demote potentially non-relevant medical records while retaining the relevant ones at the top ranks.

When comparing the retrieval performance achieved by the learned approach when applied with the term-based and the task-specific representation approaches (Table 5.7 vs Table 5.8), we observe that in general the learned approach has a potential to markedly enhance the retrieval performance when applied with the term-based representation approach. For example, when using BM25 and expCombSUM, the learned approach with a term-based representation approach and the task-specific representation approach improves the infNDCG retrieval performance by +5.48% (infNDCG 0.4343 vs 0.4581) and +1.58% (infNDCG 0.4880 vs 0.4957), respectively.

Next, in Section 5.4.4.1, we examine whether having more evidence (i.e. query terms) to infer the relevance of medical records, the learned approach could further improve the retrieval performance. In particular, we only investigate the learned approach when applied with the term-based representation approach, as it has more potential to further enhance retrieval performance (see Tables 5.7 and 5.8).

5.4.4.1 Applying Query Expansion

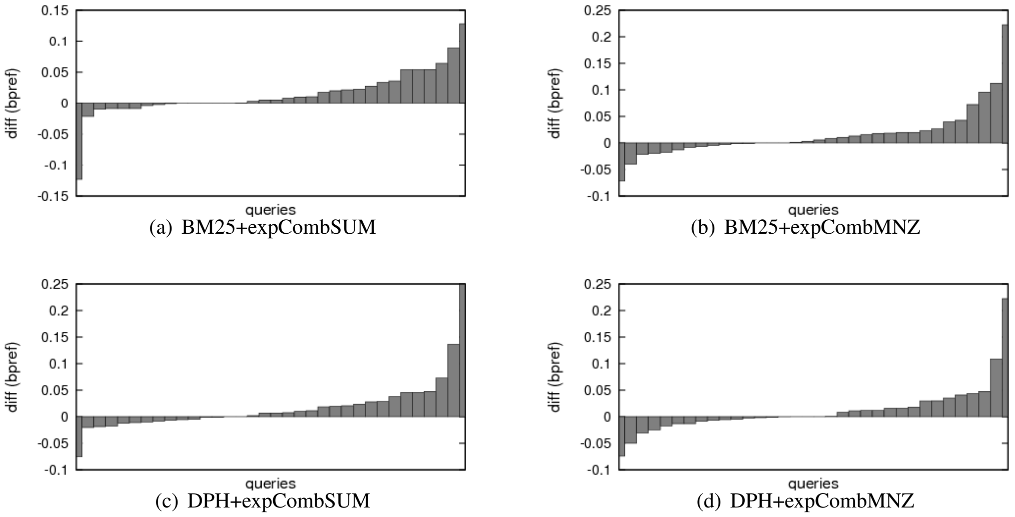
As discussed in Section 3.4.2.3, local-statistic and external corpus query expansion (QE) approaches have been shown to be effective for the patient search task. In this section, we apply such approaches to improve the query representation by adding more evidence (i.e. query terms) to the queries. In particular, we expect that if QE expands the query with more evidence (i.e. informative terms) to infer the relevance of medical records, the learned approach would effectively demote the non-relevant medical records in the ranking list, and hence improve retrieval performance. Therefore, we improve the representation of the queries by using information from both internal and external corpora. Indeed, we apply the DFR Bo1 model to expand the queries with the top 10 informative terms from the top 3 ranked documents (as suggested by Amati (2003) in Section 2.5.1) retrieved from the medical records collection of the TREC Medical Records and the MEDLINE abstract collection of the TREC 2005 Genomics (Hersh *et al.*, 2005) tracks.

Table 5.9 compares the retrieval performance, after applying the aforementioned QE technique on both the learned approach and all of the baselines (i.e. applying Bo1 in conjunction with all of the approaches in this experiment). In particular, we report the same retrieval effectiveness measures (i.e. bpref and P10 for TREC 2011, and infNDCG, infAP, and P10 for TREC 2012) as in Section 5.4.4. In addition, the highest retrieval performance that the learned approach could achieve is also discussed (i.e. when using the oracle $w(t') = O(t')$).

From Table 5.9, we firstly observe that after applying QE, the retrieval performance of the learned approach (namely, *Our learned approach*) and all of the baselines (namely, the common baseline and NegFlag) increase markedly. This shows that, overall, the QE technique could expand the queries with informative terms. Next, as expected, we find that after applying QE, the proposed learned approach further improves the retrieval performance. Indeed, with a cross-collection validation setting, the learned approach when applied with DPH and expCombSUM, outperforms the baselines for all of the reported measures. For TREC 2011, the learned approach, *Our learned approach (x-collection)*, performs significantly ($p < 0.05$) better than the common baseline where the negation is not explicitly handled, in terms of bpref (0.5786 versus 0.5569), while the performance, in terms of precision at 10, improves from 0.6527 to 0.6647. For the TREC 2012 topic set, *Our learned approach (x-collection)* significantly ($p < 0.05$) outperforms the common baseline, for all the reported retrieval measures. In particular, the learned approach outperforms the common baseline by 6.3%, 8.8%, and 14%, in terms of infNDCG, infAP, and P10, respectively. Moreover, when analysing the penalising weights, we find that our learned model could learn proper weights for penalising the relevance scores of the medical records. For example, for the query #101 “patient with hearing loss”, the learned model set the weight of the negative term of ‘hearing’ (i.e. ‘n$hearing’) to 0.2437. This could help our system to retrieve patients whose medical records stating “patient could not hear” or “patient exhibited no signs of hearing”. This confirms that the third component of the learned approach, namely the penalising weight estimation component, can effectively penalise non-relevant medical records, when several informative terms are used in a query, answering the third research question. In addition, we find that with the cross-collection validation setting, *Our learned approach (x-collection)* could perform comparably to the best possible setting (i.e.

Approach	TREC 2011		TREC 2012		
	bpref	P10	infNDCG	infAP	P10
Term-based representation					
BM25+expCombSUM+QE					
>> The common baseline	0.5491	0.6118	0.4661	0.2135	0.4660
>> NegFlag	0.5606	0.6235	0.5054$^{\blacktriangle\blacktriangle\blacktriangle}$	0.2362$^{\blacktriangle}$	0.4915
>> Our learned approach (5-fold)	0.5638	**0.6265**	0.5108$^{\blacktriangle\blacktriangle\blacktriangle}$	0.2378$^{\blacktriangle\blacktriangle}$	0.4979
>> Our learned approach (x-collection)	**0.5673**	0.6235	**0.5126**$^{\blacktriangle\blacktriangle\blacktriangle}$	**0.2397**$^{\blacktriangle\blacktriangle}$	**0.5106**$^{\blacktriangle}$
>> Our learned approach (oracle)	0.5891$^{\blacktriangle\blacktriangle}$	0.6265	0.5114$^{\blacktriangle\blacktriangle\blacktriangle}$	0.2416$^{\blacktriangle\blacktriangle}$	0.5085
BM25+expCombMNZ+QE					
>> The common baseline	0.5217	0.5912	0.4608	0.2013	0.4596
>> NegFlag	0.5325	**0.5941**	0.4881$^{\blacktriangle\blacktriangle}$	0.2172	0.4787
>> Our learned approach (5-fold)	0.5349	0.5912	0.4940$^{\blacktriangle\blacktriangle}$	**0.2198**	0.4872
>> Our learned approach (x-collection)	**0.5440**$^{\blacktriangle\blacktriangle\blacktriangle}$	0.5912	**0.4953**$^{\blacktriangle\blacktriangle\blacktriangle}$	0.2187$^{\blacktriangle}$	**0.4894**
>> Our learned approach (oracle)	0.5594$^{\blacktriangle\blacktriangle}$	0.5912	0.5017$^{\blacktriangle\blacktriangle}$	0.2276$^{\blacktriangle\blacktriangle}$	0.5170$^{\blacktriangle}$
DPH+expCombSUM+QE					
>> The common baseline	0.5569	0.6529	0.4619	0.1982	0.4702
>> NegFlag	0.5733	0.6559	0.4838$^{\blacktriangle}$	0.2127$^{\blacktriangle}$	0.5149$^{\blacktriangle}$
>> Our learned approach (5-fold)	0.5707	0.6500	0.4815	0.2152$^{\blacktriangle\blacktriangle}$	0.5213$^{\blacktriangle}$
>> Our learned approach (x-collection)	**0.5786**$^{\blacktriangle}$	**0.6647**	**0.4911**$^{\blacktriangle}$	**0.2157**$^{\blacktriangle}$	**0.5362**$^{\blacktriangle}$
>> Our learned approach (oracle)	0.5891$^{\blacktriangle\blacktriangle}$	0.6647	0.5004$^{\blacktriangle\blacktriangle}$	0.2229$^{\blacktriangle\blacktriangle}$	0.5447$^{\blacktriangle\blacktriangle}$
DPH+expCombMNZ+QE					
>> The common baseline	0.5533	0.6500	0.4615	0.1917	0.4745
>> NegFlag	**0.5634**	0.6500	0.4819$^{\blacktriangle}$	0.2078$^{\blacktriangle\blacktriangle}$	0.**5298**
>> Our learned approach (5-fold)	0.5570	0.6500	**0.4929**$^{\blacktriangle\blacktriangle}$	0.2108$^{\blacktriangle\blacktriangle}$	0.5277$^{\blacktriangle}$
>> Our learned approach (x-collection)	0.5633	**0.6588**	0.4926$^{\blacktriangle\blacktriangle\blacktriangle}$	**0.2133**$^{\blacktriangle\blacktriangle}$	0.5255$^{\blacktriangle}$
>> Our learned approach (oracle)	0.5659	0.6412	0.5055$^{\blacktriangle\blacktriangle\blacktriangle}$	0.2186$^{\blacktriangle\blacktriangle}$	0.5404$^{\blacktriangle\blacktriangle}$

Table 5.9: Retrieval performances of our learned approach to handle negation in comparison to existing approaches, when applied with a term-based representation and a query expansion approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the traditional approach baselines are denoted $^{\blacktriangle}$, $^{\blacktriangle\blacktriangle}$ and $^{\blacktriangle\blacktriangle\blacktriangle}$ (resp. $^{\blacktriangledown}$, $^{\blacktriangledown\blacktriangledown}$ and $^{\blacktriangledown\blacktriangledown\blacktriangledown}$), respectively.



(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 5.1: The differences between the retrieval performances of our NegFlag approach and the common baseline, when applied with a term-based representation, evaluated using the TREC 2011 Medical Records Track.

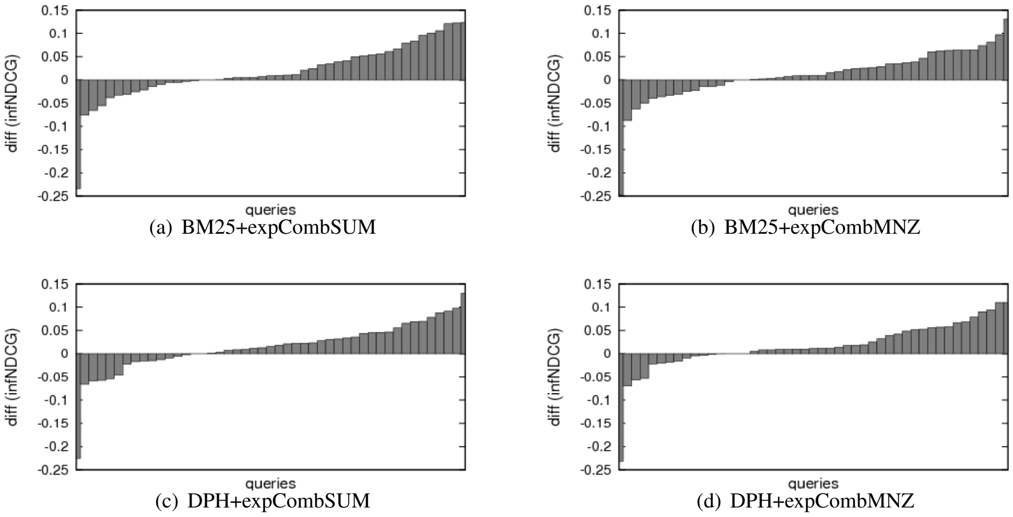
when $w(t') = O(t')$), *Our learned approach (oracle)*. This shows that the learned approach is robust and could be generalised between the training and test topic sets. However, we find that the performance improvement over the NegFlag by the learned approach is not statistically significant.

5.4.5 Analysis and Discussion

This section analyses the retrieval performances of our proposed approaches to deal with negated language in patient search. Specifically, Section 5.4.5.1 discusses the performances of the NegFlag approach on a per-query basis. Section 5.4.5.2 analyses the retrieval effectiveness of the term dependence approach for handling negation, as we vary the $window\_size$ that considers the dependence between terms in medical records. Section 5.4.5.3 discusses which types of queries are likely to be benefited from the learned approach. In Section 5.4.5.4, we analyse the impact of different features used by the learned approach. Finally, Section 5.4.5.5 discusses how to apply our approaches to deal with other types of contexts (e.g. assertive or predictive).

5.4.5.1 NegFlag

Figures 5.1, 5.2, 5.3, and 5.4 report the differences between the retrieval performances achieved by the NegFlag approach and the common baseline, on a per-query basis. Specifically, Figures 5.1 and 5.2 show the retrieval performances when applied with a term-based representation, while Figures 5.3



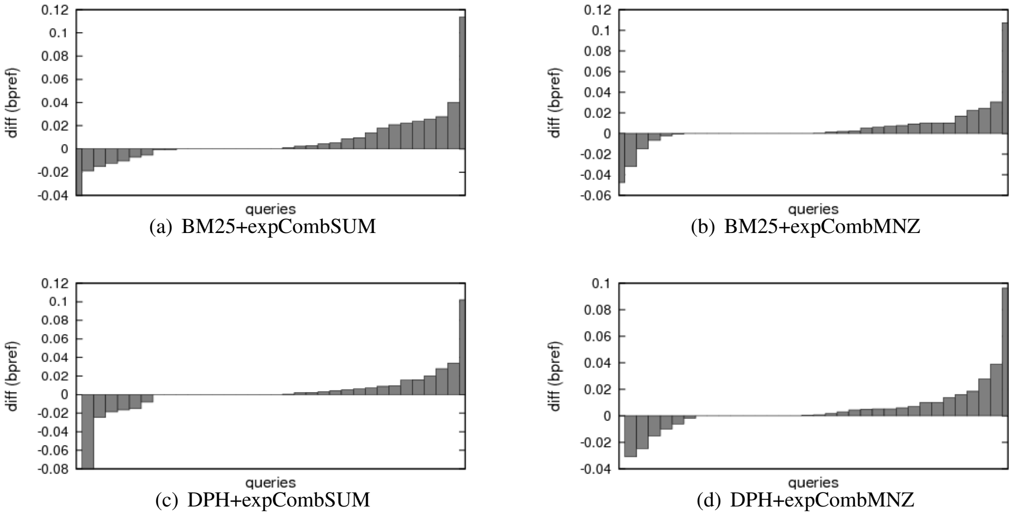
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 5.2: The differences between the retrieval performances of our NegFlag approach and the common baseline, when applied with a term-based representation, evaluated using the TREC 2012 Medical Records Track.



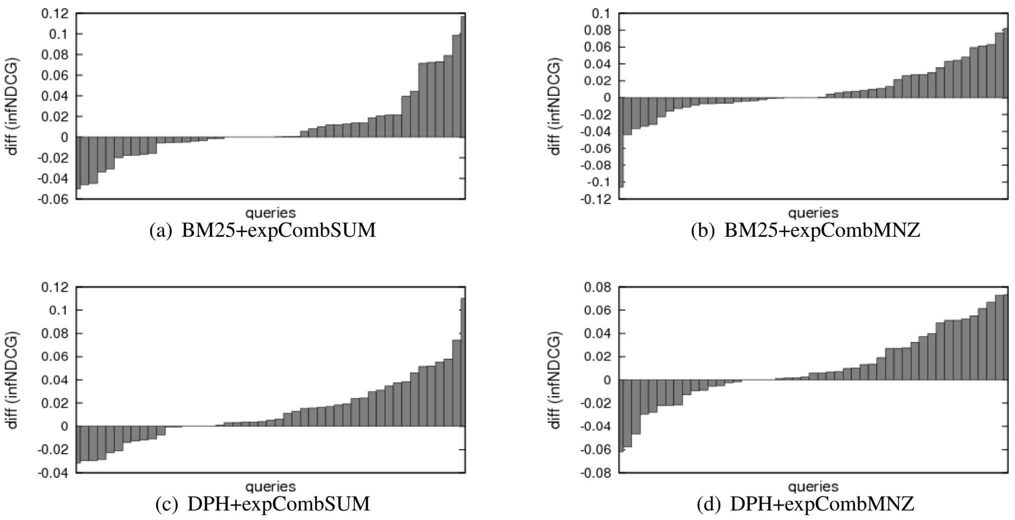
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 5.3: The differences between the retrieval performances of our NegFlag approach and the common baseline, when applied with our task-specific representation, evaluated using the TREC 2011 Medical Records Track.



(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

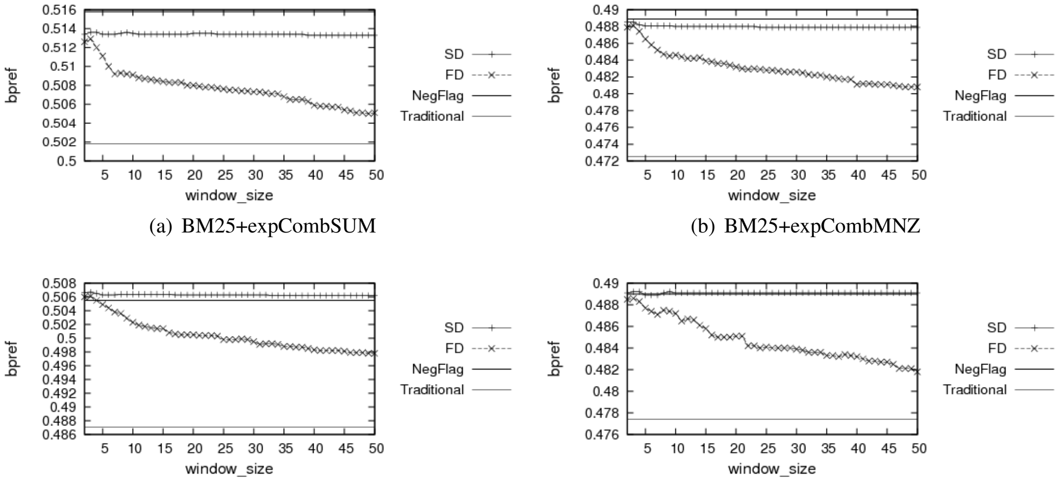
(d) DPH+expCombMNZ

Figure 5.4: The differences between the retrieval performances of our NegFlag approach and the common baseline, when applied with our task-specific representation, evaluated using the TREC 2012 Medical Records Track.

and 5.4 are for the task-specific representation. These figures show that the NegFlag approach is effective, as it improves the retrieval performances for a majority of the queries regardless of the representation approach, ranking model and voting technique used. In particular, comparing between Figures 5.1 and 5.2, we observe that the NegFlag approach benefits more queries of TREC 2012 than TREC 2011. However, the magnitude of improvements could not be compared because they are evaluated on different measures (i.e. bpref for TREC 2011 and infNDCG for TREC 2012). Meanwhile, we also observe the same pattern when applied with the task-specific representation (Figure 5.3 vs Figure 5.4). Next, when comparing the effectiveness of the NegFlag approach when using the term-based and the task-specific representation (i.e. Figure 5.1 vs Figure 5.3 and Figure 5.2 vs Figure 5.4), we observe that the magnitude of the improvement when applied with the term-based representation approach is higher.

5.4.5.2 Term Dependence for Negation

This section evaluates the effectiveness of the term dependence approach, as we vary the size of the windows when considering the dependence between terms. From Figures 5.5, 5.6, 5.7 and 5.8, we observe that different variants of the term dependence approach benefit differently from the sizes of the windows. For example, as shown in Figure 5.7, with the task-specific representation, our full dependence (FD) term dependence approach is the most effective, when the $window\_size$ is 22 (resp. 3), when using the BM25 (resp. DPH) ranking models, respectively, to rank medical records before aggregating the



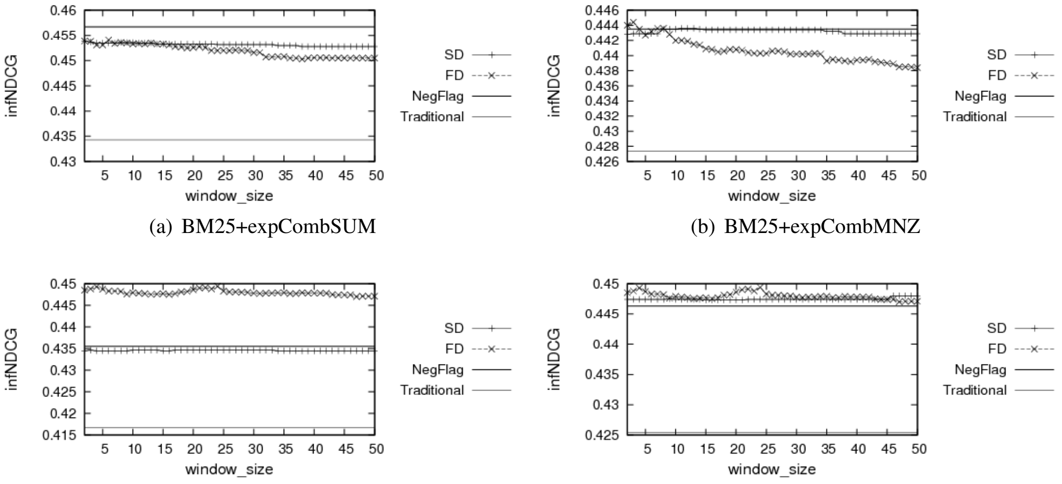
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 5.5: bpref performances of our term dependence for negation approach when applied with a term-based representation, while varying $window\_size$.



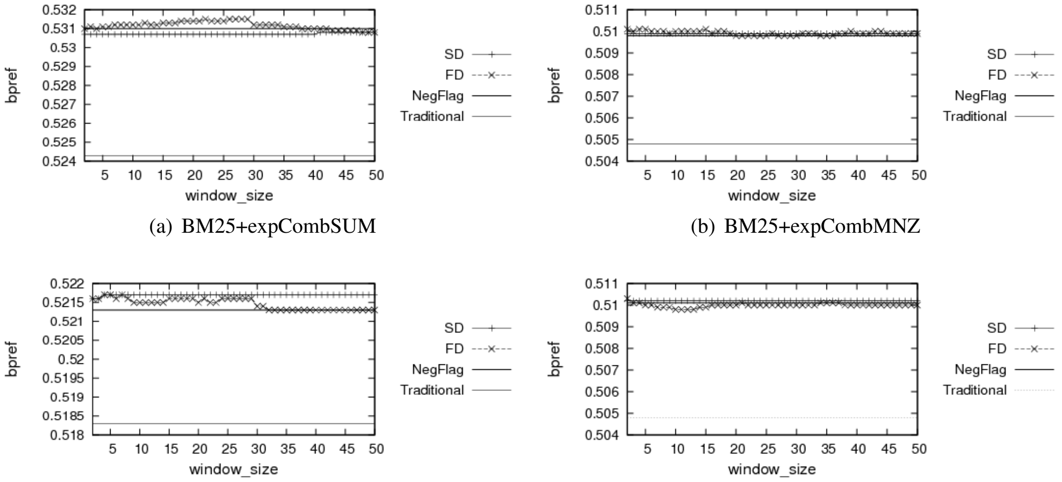
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 5.6: infNDCG performances of our term dependence for negation approach when applied with a term-based representation, while varying $window\_size$.



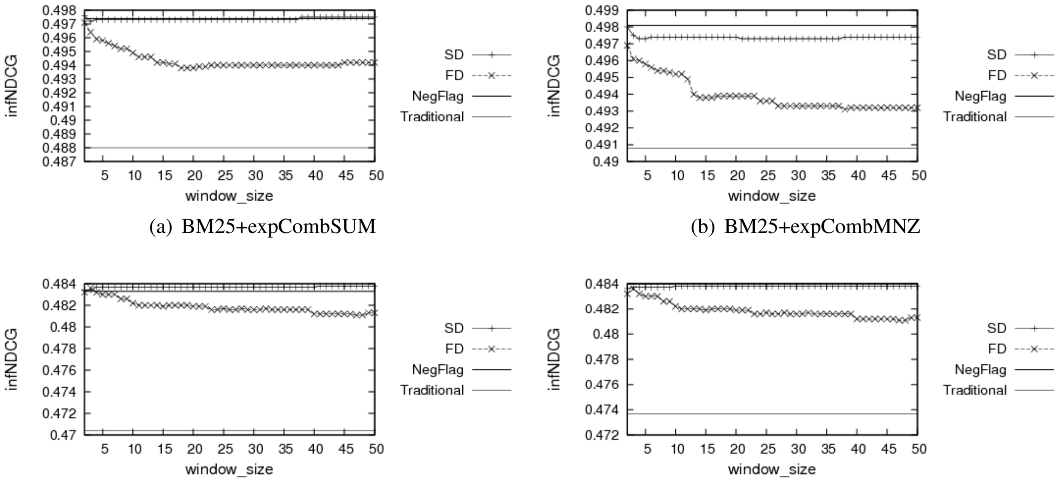
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 5.7: bpref performances of our term dependence for negation approach when applied with our task-specific representation, while varying $window\_size$.



(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 5.8: infNDCG performances of our term dependence for negation approach when applied with our task-specific representation, while varying $window\_size$.

relevance scores using expCombSUM to rank patients. However, we observe that the term dependence approach is likely to be effective when the $window\_size = 3$.

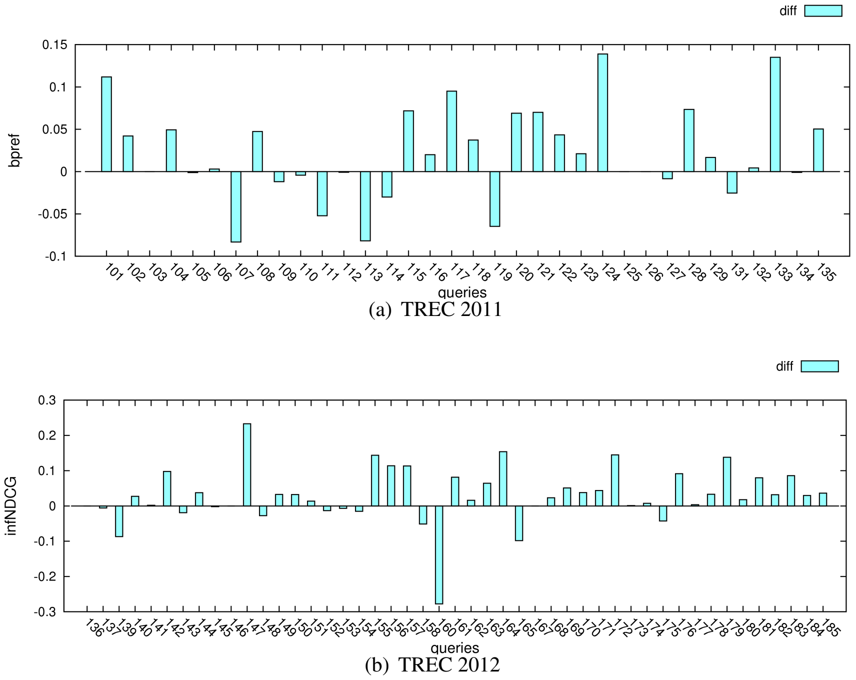
5.4.5.3 Failure Analysis

In this section, we conduct a failure analysis to investigate when the learned approach to handle negation is effective. In particular, we choose to investigate the retrieval performances of the learned approach with the cross-collection validation regime, where DPH is used to rank medical records, expCombSUM is used to estimate the relevance of a patient, and DFR Bo1 is used to expand the queries, since it is one of the most effective approaches as previously shown in Table 5.9 (i.e. it achieved bpref 0.5786 and infNDCG 0.4911). From Figure 5.9, we firstly observe that comparing with the common baseline where negation is not explicitly handled, the learned approach improves the retrieval performances for the majority of the queries. Specifically, for TREC 2011 and 2012, the learned approach improves the retrieval performances for 19 out of 34 queries and 32 out of 47 queries, respectively. On the other hand, the learned approach decreases the retrieval performances for 12 out of 34 queries from TREC 2011 and 12 out of 47 queries for TREC 2012. These results show that the learned approach is effective for many queries, especially for the queries from TREC 2012. Importantly, for the query 179: “Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression”, where negated language is used in the query, the approach markedly improves the retrieval performance from infNDCG 0.0255 to 0.1635.

Table 5.10 shows the numbers of queries benefited or harmed by the learned approach, which are grouped based on the types of medical concepts that can be extracted from the queries. Specifically, we define the types of medical concepts based on the four aspects of the medical decision process (namely, symptom, diagnostic test, diagnosis and treatment), discussed in Chapter 4. We observe that the learned approach is most likely to be effective when medical concepts related to all of the four aspects can be extracted from the query, followed by the queries that contains medical concepts related to treatment. In particular, the learned approach improves the retrieval performance for 75% of the queries that contain medical concepts related to all of the four aspects.

5.4.5.4 Our Learned Approach: Feature Importance

Next, in order to examine the importance of each proposed feature in Table 5.4, we conduct an ablation study using the same setting as in Section 5.4.5.3. Indeed, we examine removing each of the features from the feature space to examine the impact of each feature on the retrieval performance. For example, when we evaluate the importance of Feature 1, we remove Feature 1 from the feature space, while



(a) TREC 2011

(b) TREC 2012

Figure 5.9: The difference between the performances of our learned approach and the common baseline, when applied with the Bo1 QE, evaluated using the TREC 2011 and 2012 Medical Records Track.

keeping the other twelve features. The increase or reduction in the achieved retrieval effectiveness is an indicator of the importance of the feature. Specifically, the retrieval performance decreases when removing an effective feature, while the performance remains the same or increases when removing a non-effective feature. We compare the importance of each feature based on its impact on the retrieval performance, in terms of bpref and infNDCG for TREC 2011 and 2012, respectively, as they are the TREC primary measures (see Section 3.4.1). The percentage improvement or reduction, in terms of each retrieval measure, between when a particular feature is removed and when all the features are considered, are summed up to measure the overall impact of that feature. Then, we normalise this measure for each

Aspects of medical concepts in the query	Benefited	Harmed
Symptom	62.9% (44/70)	31.4%(22/70)
Diagnostic test	67.3% (35/52)	30.8%(16/52)
Diagnosis	63.2% (43/68)	27.9% (19/68)
Treatment	69.2% (36/52)	25% (13/52)
All 4	75% (24/32)	25%(8/32)

Table 5.10: Analysis of our approach w.r.t. the aspects of medical concepts found in the queries. The numbers between the parentheses indicate the number of queries impacted compared to the total number of queries.

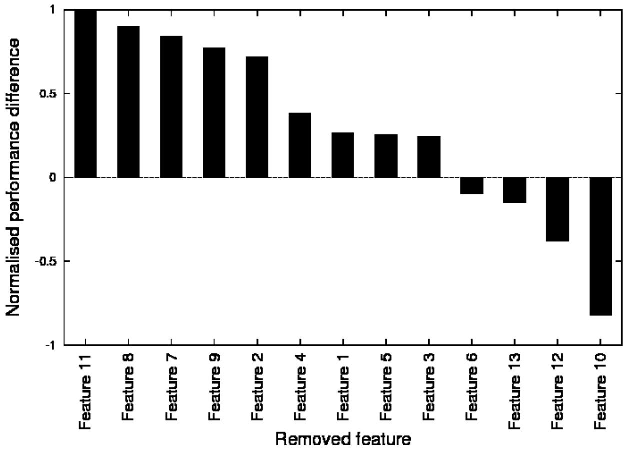


Figure 5.10: Ablation study of feature importance.

feature by dividing it with that of the most important feature (i.e. the feature with the most negative impact on retrieval effectiveness), in order to easily rank the importance of different features. We refer to the normalised measure as *feature importance*. Hence, the feature whose removal *degrades effectiveness most* has the *highest feature importance*. If the feature importance of a particular feature is zero, the feature has no impact on the retrieval performance, while a negative feature importance indicates that the feature is not useful and can be removed from the feature set.

Figure 5.10 compares the feature importance of each feature in our feature space. We observe that Features 11, 8 and 7, which are the co-occurrence between the opposite context term t' and the query Q, an IDF variant of the opposite context term t', and an IDF variant of the query term t corresponding to the opposite context term t', respectively, are the most important features, answering the fourth research question. This is intuitive as the co-occurrence of the opposite context term t' and the terms in the query Q could be used to measure the relatedness between the opposite context term and the query, while the IDF variant of the opposite context term t' and its corresponding query term t could measure the informativeness of both associated terms. In contrast, Features 10, 12 and 13, namely the two variants of the co-occurrence frequency between the opposite context term t' and its corresponding query term t, and the query length, respectively, are the least importance features. Indeed, adding these features to the feature set has a negative impact on retrieval effectiveness. Overall, we find that 9 out of the proposed 13 features are beneficial to obtaining effective estimation of the penalising weights for opposite context terms.

5.4.5.5 Handling Other Types of Contexts

Besides negation, other types of contexts (e.g. the description of medical conditions of patient's family members) are also commonly used in the medical domain. For example, the medical condition 'diabetes' in the sentence "the patient is likely to have diabetes, if the blood glucose level does not decrease in the next two months" has a predictive context, since the healthcare practitioner does not confirm that the patient currently has diabetes. Without handling these kinds of contexts, a patient search system may consider patients having medical records containing the query terms with these contexts to be relevant. However, unlike the negative context, other contexts may not be used as a strong evidence to demote the medical records. For instance, for the aforementioned sentence, the patient is not confirmed to have diabetes. Hence, our approaches proposed so far consider only the negative context of terms in medical records and queries (see Sections 5.2 and 5.3).

In this section, we examine the impact of other types of contexts on the retrieval of patients. In particular, to deal with different types of contexts, we propose *a context representation approach* (i.e. *ConTextFlag*) that extends NegFlag (see Section 5.2.1) by deploying the ConText algorithm (Chapman *et al.*, 2007) to identify contexts of terms in medical records and queries. The ConText algorithm uses the same techniques as the NegEx algorithm used in the NegFlag approach, but it can also detect other contexts, apart from negation. Once the contexts of terms are identified, we prefix a term with 'n$', 'p$', and 'o$' before being processed by a search system, if the term is in a negative context, predictive context, or is describing a medical condition of the patient's family member, respectively. This allows a search system to distinguish between terms with different contexts.

Table 5.11 compares the retrieval performances of ConTextFlag with the NegFlag and the common baseline, when applied with a term-based representation. We examine the performance of our ConTextFlag approach using only the term-based representation, because besides the negative context, other types of contexts cannot be obtained from the MetaMap tool that is used to extract medical concepts for the task-specific representation. From Table 5.11, we observe that the ConTextFlag significantly outperforms the common baseline, which does not take into account contexts of terms, in terms of the TREC primary measures (i.e. bpref and infNDCG) for different variants of the ranking models and voting techniques used. However, the ConTextFlag approach performs comparably to the NegFlag approach in terms of both bpref and infNDCG. This shows that other types of contexts do not as markedly impact the retrieval performance of a patient search system as the negative context. Recall that the ConTextFlag approach also improves the representation of the negated terms.

Approach	TREC 2011		TREC 2012		
	bpref	P10	infNDCG	infAP	P10
Term-based representation					
BM25+expCombSUM					
>> The common baseline	0.5018	0.5735	0.4343	0.1828	0.4681
>> NegFlag	**0.5158**▲	0.5735	**0.4567**▲▲	**0.1984**▲	0.5000▲
>> Context representation	0.5151	**0.5853**	0.4538▲	0.1980▲	**0.5021**▲
BM25+expCombMNZ					
>> The common baseline	0.4725	0.5559	0.4274	0.1730	0.4681
>> NegFlag	**0.4889**	0.5676	**0.4435**▲	0.1821	**0.4872**
>> Context representation	0.4897	**0.5853**▲	0.4425	**0.1827**	0.4787
DPH+expCombSUM					
>> The common baseline	0.4871	0.5765	0.4167	0.1703	0.4638
>> NegFlag	**0.5055**▲	0.5794	0.4355▲▲	0.1833▲▲	**0.4894**
>> Context representation	0.5045	**0.5824**	**0.4357**▲	**0.1851**▲	0.4872
DPH+expCombMNZ					
>> The common baseline	0.4774	0.5676	0.4254	0.1698	0.4596
>> NegFlag	**0.4890**	0.5765	**0.4463**▲▲▲	**0.1818**▲▲	**0.4936**▲
>> Context representation	0.4886	**0.5794**	0.4420▲	0.1817▲	0.4809

Table 5.11: Retrieval performances of applying our negation handling approach to deal with other types of contexts in comparison to existing approaches, when applied with a term-based representation approach. Statistically significant improvement (resp. decrease) (paired t-test) at $p < 0.05$, $p < 0.01$ and $p < 0.001$ compared to the common baseline is denoted ▲, ▲▲ and ▲▲▲ (resp. ▼, ▼▼ and ▼▼▼), respectively.

5.5 Conclusions

We have motivated the need for a patient search system to handle negated language, which is commonly used in the medical domain. We have discussed the first component of our thesis framework (see Section 4.2.3), that uncovers the implicit knowledge at the sentence level of the retrieval process (Figures 4.1 and 4.4), and handles negated language in patient search.

To tackle the first component of the framework, we proposed novel approaches that enable a search system to retrieve patients whose medical records contain query terms with a right context. In particular, we introduced two non-supervised (Section 5.2) and one supervised approach (Section 5.3). Specifically, in Section 5.2.1, we introduced the NegFlag approach to distinguish between the contexts of terms before being processed within a search system. Section 5.2.2 introduced a use of term dependence to demote medical records containing the query terms in the opposite context. Then, also building up on the NegFlag, in Section 5.3, we proposed our learned approach that demoted medical records containing the query terms with the opposite context of the query's intent. Specifically, our learned approach prevented non-relevant medical records from being ranked highly, by demoting those containing occurrences of opposite context terms (i.e. a term having the opposite context to its corresponding query term). We deployed a regression technique to effectively estimate the penalising weight for the opposite context query terms. In particular, we used the Gradient Boosted Regression Trees (GBRT) to learn the weight of an opposite context term, using features including co-occurrence frequencies.

In Section 5.4, we evaluated the retrieval performances of our proposed approaches using the same test collection used in Sections 4.3.3 and 4.3.4 of Chapter 4. We compare the effectiveness of the proposed approaches with the common baseline, previously identified in Section 4.3.4. Our experimental results showed that the proposed approaches significantly outperformed several strong baselines, including the common baseline. In particular, in Section 5.4.3, we have shown that the NegFlag approach significantly improved the retrieval performance of the common baseline (see in Tables 5.5 and 5.6). Indeed, the NegFlag approach outperformed this common baseline for the majority of the queries (see Figures 5.1, 5.2, 5.3 and 5.4). Meanwhile, the term dependence approach to prevent non-relevant patients from being retrieved further improved the retrieval performance, as shown in Tables 5.5 and 5.6). The term dependence approach was effective when the $window\_size$ was set to 3 (see Section 5.4.5.2). On the other hand, we observed that our learned approach was also effective. As shown in Table 5.9, when applied with a query expansion technique, our learned approach further improved the retrieval performance over the NegFlag approach. In Section 5.4.5.4, we found that the co-occurrence between the opposite context term (see Section 5.3) and the query terms and an IDF variant of the opposite context term were the most effective features for learning the penalising weight.

Next, in Chapter 6, we discuss the second component of the framework (i.e. the Conceptual Reasoning component), which aims to infer the relationships between medical conditions, by uncovering the implicit knowledge at the record level of the retrieval process. We introduce approaches to instantiate this component and evaluate the proposed approaches using the test collection and the baseline identified in Chapter 4.

Chapter 6

Reasoning using Medical Concept Relationships

6.1 Introduction

This chapter discusses our approaches that instantiate the Conceptual Reasoning component of the proposed framework (previously defined in Section 4.2.3.2). Recall that the Conceptual Reasoning component aims to uncover the implicit knowledge at the record level by inferring relationships between medical conditions, with respect to the four aspects of the medical decision process (i.e. symptom, diagnostic test, diagnosis and treatment) (see Section 3.4.2.2). As previously discussed in Section 3.4.2.3, medical practitioners use different terms to describe a particular medical condition of a patient (e.g. using 'carcinoma' or 'malignant tumour' to refer to 'cancer' as they share the same meaning). However, such information may be hidden from existing search systems (we referred to this as implicit knowledge in Section 1.2).

To deal with this implicit knowledge, in this chapter, we propose to improve the representations of medical queries by inferring the relationships of medical conditions that are extracted from the queries. In particular, we focus on medical conditions that are related to the four aspects of the medical decision process, which are the information that healthcare practitioners take into account when dealing with patients (see Section 3.4.2.2). To do so, we firstly deploy the task-specific representation approach (introduced in Section 4.2.1) to represent both the medical records and queries using only concepts related to the four aforementioned aspects. Then, using these conceptual representations, we propose two novel query expansion (QE) approaches to further improve the query representation by exploiting two types of resource. The first type consists of external resources about the possible relationships between the medical conditions (i.e. medical concepts). For example, if a patient is prescribed a medicine that treats a particular disease, then we can infer that the patient suffers from that disease. To infer relationships between medical concepts, we propose two approaches, which are based on Bayes' theorem and a stochastic analysis. As a second type of resources, we use the collection of medical records itself. In particular, we extract the most informative concepts from the top-ranked documents retrieved using the original query. The concepts inferred using these two types of resources are used to expand the original queries to improve their representations.

The remainder of this chapter is organised as follows.

- Section 6.2 discusses how we extract possible relationships between medical concepts from external resources and represent them in the forms of association rules.
- In Section 6.3, we propose a novel approach that uses Bayes' theorem to infer relationships between medical concepts using the association rules presented in Section 6.2.
- Section 6.4 introduces a second novel approach that uses a stochastic analysis to infer the relationships between medical concepts using the association rules.
- Section 6.5 discusses how we can combine the medical concepts derived from the external resources (i.e. using either the Bayesian-based approach (Section 6.3) or the stochastic analysis-based approach (Section 6.4)) with the medical concepts derived from the medical record collection, by using a pseudo-relevance feedback approach, to improve the representation of the query.
- In Section 6.6, we evaluate the proposed approaches to exploit the relationships between the medical concepts to improve the retrieval performance of a patient search system.
- Section 6.7 provides concluding remarks for this chapter.

6.2 Extracting and Representing Medical Concept Relationships from External Resources as Association Rules

Driven by the medical decision process (as described in Section 4.2.1), we extract directed association rules representing the relationships between concepts from two different types of medical resources[1], which are ontology-based and free-text-based resources, respectively. We use different strategies for extracting conceptual relationships from each type of resource. For the ontology-based resources (e.g. MedDRA[2] and DOID[3]), we use the semantic relationships of concepts within each ontology to represent

[1] We provide the list of all of the used resources in Section 6.6.1.

[2] http://www.meddramsso.com

[3] http://purl.bioontology.org/ontology/DOID

Medical Concept	Related Concept
Dowager's hump	Osteoporosis
DEXA	Osteoporosis
Prolia	Osteoporosis
Boniva	Osteoporosis

Table 6.1: Examples of medical concepts associated to 'osteoporosis' defined in our associated rules database.

the relationships between concepts. For instance, 'Coronary heart disease' is a particular type of 'heart disease'.

For the free-text-based resources (e.g. `http://www.rxlist.com`), we use MetaMap to identify concepts related to the four aforementioned medical aspects from the free-text (we described how medical concepts could be identified from sentences in Section 4.2.1), and then assume the existence of relationships between the identified concepts. For example, from a drug indication in the `http://www.rxlist.com` website, which states that *"Boniva (ibandronate sodium) is indicated for the treatment and prevention of osteoporosis in postmenopausal women"*, MetaMap can identify the concepts *'Boniva'* (treatment) and *'osteoporosis'* (diagnosis). Assuming the relationships between medical concepts found in the drug description, we surmise that there is an association between the two concepts. Next, the extracted association rules are stored in a database. For instance, as shown in Table 6.1, the rules associated with the concept *'osteoporosis'* are *'Dowager's hump'*→*'osteoporosis'*, *'DEXA'*→*'osteoporosis'*, *'Prolia'*→*'osteoporosis'*, and *'Boniva'*→*'osteoporosis'*. These association rules provide new evidence to infer the possible relevance of medical records, which are then aggregated to estimate the relevance of their associated patients (see Section 4.2.2). For instance, we can infer that patients taking the *'Boniva'* medicine suffer from *'osteoporosis'*, since *'Boniva'* is a treatment for *'osteoporosis'*.

6.3 A Bayesian-based Approach for Inferring Conceptual Relationships

We argue that the relationships between medical concepts related to the medical decision process could be leveraged to deal with the complexity of medical terminology. For example, if we have evidence that a patient is taking the *'olmesartan'* medicine (treatment), we can infer that the patient suffers from *'hypertension'* (diagnosis), since *'olmesartan'* is a treatment for *'hypertension'*. Therefore, using external domain-specific resources, we propose to reformulate the queries by using the association rules extracted from the medical resources (see Section 6.2).

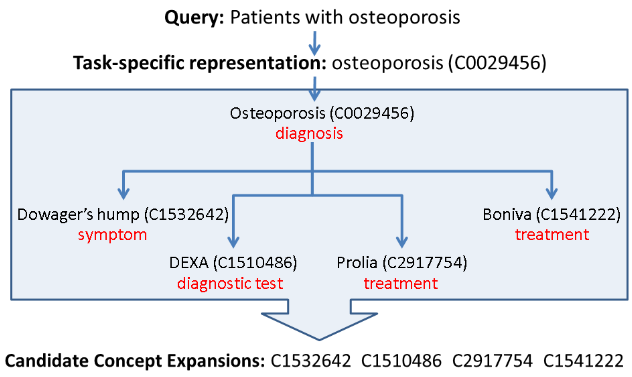

Query: Patients with osteoporosis
Task-specific representation: osteoporosis (C0029456)
Osteoporosis (C0029456)
diagnosis
Dowager's hump (C1532642)
symptom
DEXA (C1510486)
diagnostic test
Prolia (C2917754)
treatment
Boniva (C1541222)
treatment
Candidate Concept Expansions: C1532642 C1510486 C2917754 C1541222

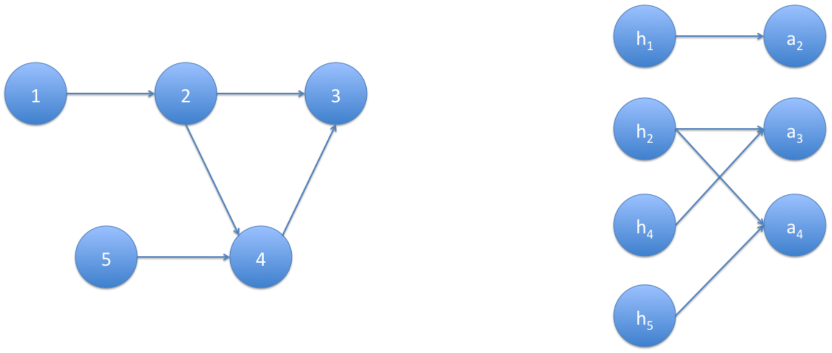
Figure 6.1: An example of identifying candidate concept expansions using our approach on query *"patients with osteoporosis"*.

Specifically, we first retrieve a set of candidate concept expansions (denoted $inferred(q)$) corresponding to the query concepts from the extracted association rules. Then, to prevent excessively general candidate concepts being added to the query, we estimate the association of a query concept and each candidate concept expansion using a Bayesian probabilistic score computed based on the occurrences of concepts in the association rules (both derived from ontologies and free-text resources). The higher the probability, the stronger the relationship between the two concepts. Indeed, the Bayesian probabilistic score of the association between query concept t and its corresponding concept t' is estimated as follows:

$$w_a(t,t') = p(t'|t) = \frac{p(t' \cap t)}{p(t)} \tag{6.1}$$

where $p(t' \cap t)$ is the maximum likelihood that the concept t' co-occurs with the query concept t within all the extracted association rules, and $p(t)$ is the maximum likelihood that the concept t is contained in an association rule. This maximum likelihood is calculated based on the frequency of the occurrences of each concept in the whole association rules database.

Figure 6.1 shows an example of how our approach identifies candidate concept expansions for the query *"patients with osteoporosis"*. In particular, we first obtain the concept *'osteoporosis'* from the query using the task-specific representation approach (see Section 4.2.1). Then, the concept *'osteoporosis'* is used to retrieve related concepts from the database of association rules previously extracted from the medical resources (see Section 6.2). The retrieved candidate concept expansions include *'Dowager's hump'*, *'DEXA'*, *'Prolia'* and *'Boniva'*, which are the symptom, diagnostic test, and treatments associated with the original query concept (i.e. *'osteoporosis'*). As discussed in Section 2.5.1, it has been shown that QE is effective when the highly informative terms are added to the original query. Hence,



(a) A collection *D* of documents *1, 2, 3, 4, 5*

(b) A bipartite graph *G*

Figure 6.2: Transforming (a) the collection *D* into (b) a bipartite graph *G*.

we follow the work of Amati (2003) described in Section 2.5.1 and use only the *top 10* candidate concept expansions, which are ranked based on the score computed using Equation (6.1), to expand the original query.

6.4 A Stochastic Analysis for Inferring Conceptual Relationships

Next, we discuss our stochastic approach for identifying relationships between medical concepts in the association rule database built in Section 6.2. Indeed, we propose to measure the strength of the relationships between the medical concepts in the association rule database using a stochastic analysis of some random walk behaviour through the association rules.

A stochastic analysis of random-walk behaviour has been studied before in IR (e.g. Blanco & Lioma (2012); Brin & Page (1998); Lempel & Moran (2001)). For example, Brin & Page (1998)'s PageRank used a stochastic analysis of random walks on the hyperlink of the entire Web to determine the importance of each web page. Lempel & Moran (2001) introduced a stochastic approach for link-structure analysis (SALSA), which performed random walks on a sub-graph derived from the link-structure of an initial set of retrieved results. In particular, consider a collection *D* of documents (e.g. *1, 2, 3, 4 and 5*) and directed relationships between the documents (see Figure 6.2(a)). For example, there is a relationship from document 1 to document 2. The approach first converts the collection *D* to a bipartite graph *G*, of which the two parts are considered *hubs* and *authorities*. Hubs are documents (i.e. nodes) that link to other documents, while authorities are documents that are linked to by other documents. The SALSA algorithm performs random walks to uncover the relationships between these documents.

Algorithm 6.1 The Adapted SALSA Algorithm

```
Initialise $a(s) \leftarrow 1$, $h(s) \leftarrow 1$ for all concepts $s \in C \cup C_a$, where $C_a$ is a set of medical concepts
  in the association rule database that link to or are linked to by a concept in $C$.
Repeat the following steps until convergence:
    Update the authority score of each concept $s$:
        $a(s) \leftarrow \sum_{\{x|x \text{ points to } s\}} h(x)$
    Update the hub score of each concept $s$:
        $h(s) \leftarrow \sum_{\{x|s \text{ points to } x\}} a(x)$
    Normalise the authority scores and the hub scores.
```

Different from existing work, we propose to apply the SALSA algorithm to uncover medical concepts that are related to a given query and identify the strength of their relationships, using the link structure within our association rule database. Later, we discuss in Section 6.5 how these medical concepts are used to enrich the query.

Specifically, we adapt the SALSA algorithm to perform the analysis of the links between the concepts in the association rules. In this work, the nodes (e.g. *1, 2, 3, 4 and 5* in Figure 6.2(a)) are the medical concepts in the association rule database, while an edge is a directed relationship between two concepts (i.e. nodes). Consider a bipartite graph G where the two parts are hubs and authorities, a directed edge between hub h and authority a is indicated by a conceptual relationship between them (e.g. the edge between concept *1* and concept *2* – $h_1 \rightarrow a_2$ in Figure 6.2(b)). A good hub points to many authorities, and a good authority is pointed by many hubs. The authoritative medical concepts related to the query q should be linked to by many medical concepts in the sub-graph induced by a set of medical concepts C in the query q. A random walk on the sub-graph induced by C will visit those authoritative concepts with high probability. Hence, our adaptation of the SALSA algorithm (Lempel & Moran, 2001) identifies medical concepts that are related to those medical concepts occurring in the query q, by using the iterative algorithm presented in Algorithm 6.1.

The medical concepts with high authority scores are the concepts highly related to the query q. To incorporate these medical concepts while ranking patients based on the relevance of their medical records, using the same strategy as in Section 6.3, we use only the *top 10* candidate concept expansions ranked based on their authoritative scores computed using Algorithm 6.1.

6.5 Modelling Conceptual Relationships when Ranking Medical Records

We leverage both the local statistics from the top-ranked medical records and the association rules extracted from the external resources (Section 6.2) to improve the representation of a given query. In particular, we apply a pseudo-relevance feedback approach to identify informative concepts from the top-ranked medical records retrieved using the original query. As discussed in Section 2.5.1, when applying a pseudo-relevance feedback approach, concepts occurring in the top-ranked medical records are firstly weighted and ranked using a term weighting model. Then, the top-ranked concepts (i.e. the most informative concepts) are used to expand the original query. For example, for the query *"patients with vascular disease"*, which is represented as *"C0042373"* (*'vascular disease'*), a pseudo-relevance feedback QE approach could identify related concepts, such as *'C0190932'* (*'femoral-popliteal artery bypass graft'*) and *'C0014098'* (*'endarterectomy'*), which correspond to treatments for that disease. However, the related diagnostic procedures and symptoms, such as *'C0202896'* (*'carotid angiogram'*), might not be added to the query, if they do not appear in the top-ranked medical records. Hence, we also expand the query with medical concepts inferred using the association rules (Section 6.2). We identify candidate expansion concepts and their weights using either the Bayesian-based approach or the stochastic approach introduced in Section 6.3 or 6.4.

In particular, we estimate the relevance of a medical record d towards the query q as follows:

$$\begin{aligned} score(d,q) = & \sum_{t'' \in q_e} qtw(t'') \cdot score(d,t'') \\ & + \lambda_r \cdot \sum_{t \in q} \sum_{t' \in inferred(q)} w_a(t,t') \cdot score(d,t') \end{aligned} \tag{6.2}$$

where t'' and $qtw(t'')$ are a medical concept and its term weight in the expanded query q_e, which is reformulated using the occurrence statistics of medical concepts in the top-ranked medical records using any pseudo-relevance feedback model (e.g. Bo1 from the Divergence from Randomness (DFR) framework (described in Section 2.5.1, Equation (2.5.1)). $score()$ can be calculated using any term weighting model such as BM25, $inferred(q)$ returns a set of expanded concepts, which are related to the concepts in the original query q. We use either of the approaches previously proposed in Section 6.3 or 6.4 to extract the expanded concepts (i.e. $t' \in inferred(q)$) and their weights (i.e. $w_a()$) from the association rule database. In particular, the Bayesian-based approach and the stochastic approach uses Equation (6.1) and the authority score ($a()$) computed using Algorithm 6.1, respectively, to calculate $w_a()$. λ_r is a parameter to weight the importance of the relevance scores computed for the medical concepts derived using either of the two proposed approaches.

To learn the λ_r value for each query, we use the Gradient Boosted Regression Trees (GBRT) regression technique (see Section 2.6), previously used in Section 5.4.4.1 of Chapter 5. To learn an effective value of parameter λ_r for an unseen query, we use 12 features, which measure the predicted difficulty of the query. An effective feature should indicate the level of emphasis on the relevance scores computed

ID	Feature
1	Clarity Score (Cronen-Townsend *et al.*, 2002)
2	SCQ (Zhao *et al.*, 2008)
3	MAXCQ (Zhao *et al.*, 2008)
4	NSCQ (Zhao *et al.*, 2008)
5	AvICTF (Carmel & Yom-Tov, 2010)
6	AvIDF (Carmel & Yom-Tov, 2010)
7	EnIDF (Carmel & Yom-Tov, 2010)
8	Query Scope (ω) (He & Ounis, 2006)
9	γ_1 (He & Ounis, 2006)
10	γ_2 (He & Ounis, 2006)
11	AvPMI (Carmel & Yom-Tov, 2010)
12	Query length (He & Ounis, 2006)

Table 6.2: List of the used features.

for the concepts inferred using the association rules. Our intuition for using these features is that if the query is difficult, then it might be beneficial to take into account the relevance estimated from the inferred concepts. In particular, if using only the information from the original query is difficult for a search system, the two proposed approaches might bring novel evidence that can improve the representation of the query and hence enhance the retrieval performance. Table 6.2 lists the 12 query performance predictors computed on the original query, which are well-known for measuring the difficulty of a query. Specifically, the first set of features (Features 1-4), including the clarity score (Cronen-Townsend *et al.*, 2002), SCQ (Zhao *et al.*, 2008), MaxSCQ (Zhao *et al.*, 2008) and NSCQ (Zhao *et al.*, 2008), consider the ambiguity of a query by measuring the coherence of the language used in each medical record. The more similar the query model is to the collection model, the better the retrieval performance would be expected. The next set of features (Features 5-8) measure the specificity of each query. Indeed, queries with explicit intents could result in a better performance than queries with general terms. The features include Average Inverse Collection Term Frequency (AvICTF) (Carmel & Yom-Tov, 2010), Average Inverse Document Frequency (AvIDF) (Carmel & Yom-Tov, 2010), EnIDF (Carmel & Yom-Tov, 2010), and the query scope (ω) (He & Ounis, 2006). Features 9-10 measure the distribution of informativeness among the query terms (i.e. γ_1 and γ_2 (He & Ounis, 2006)), as a query with informative terms could attain an effective retrieval performance. Next, Feature 11, the Average of the Pointwise Mutual Information over all query term pairs (AvPMI) (Carmel & Yom-Tov, 2010), focuses on the relationship between query terms. The more co-occurrences among the query terms, the better the chance that the relevant documents are being retrieved. Finally, Feature 12 is the number of non-stopword query terms.

6.6 Experiments

This section evaluates the retrieval effectiveness of the proposed approaches for inferring the relationships between medical concepts to enhance the representation of a given query. Specifically, Section 6.6.1 discusses the experimental setup to evaluate the proposed approaches. Section 6.6.2 depicts the research questions that are investigated in this chapter. Next, in Sections 6.6.3 and 6.6.4, we discuss the experimental results of the proposed approaches for inferring the relationships of medical concepts using either Bayes' theorem or a stochastic analysis, respectively. Finally, we further analyse the retrieval performance of these two approaches in Section 6.6.5.

6.6.1 Experimental Setup

To evaluate the two approaches for inferring the conceptual relationships between medical concepts to deal with the medical terminology (discussed in Section 6.1) we use the similar settings to those in Chapter 5. Specifically, we use the TREC Medical Records track's test collection. As it has been shown to be effective in Section 4.3.3, we initially rank medical records using the BM25 weighting model, and then use the expCombSUM voting technique to aggregate the relevance scores of patients. We represent the medical records and queries using the task-specific representation approach introduced in Section 4.2.1 of Chapter 4.

To create the association rule database (discussed in Section 6.2), we use the external resources listed in Table 6.3, which are representatives of both ontology-based and free-text-based medical resources. Table 6.4 shows the number of association rules between concepts extracted from the 7 domain-specific resources. The types of relationships that are extracted from each resource are described in Column 2 of Table 6.3. Note that there are some association rules that overlap between resources. In total, there are 11,373,014 extracted association rules in our database. Meanwhile, Table 6.5 shows the number of the extracted association rules with respect to the four aspects of the medical decision process. For example, the rules with the type 'diagnosis → symptom' are the association rules that can be used to infer the related symptoms that the medical records of the relevant patients are likely to contain if the query contains a particular diagnosis. Note that there are some association rules that are duplicated among different types of rules since, using the approach described in Section 4.2.1 (see Table 4.1), some medical concepts can be categorised into several medical aspects.

For the local-statistic query expansion technique used in Equation (6.2) of Section 6.5, we deploy a parameter-free Bose-Einstein statistics-based (Bo1) model from the DFR framework (see Section 2.5.1)

Resource	Description of the Extracted Association Rules
DOID hierarchy	Hierarchical relationships between concepts within the same aspects e.g. a general disease and a specific type of the general disease
MeSH	Hierarchical relationships between concepts within the same aspects e.g. a general disease and a specific type of the general disease
MedDRA	Hierarchical relationships between concepts within the same aspects e.g. a general disease and a specific type of the general disease
DOID	Relationships between concepts across the aspects e.g. a disease and its symptoms
`http://www.rxlist.com`	Relationships between concepts across the aspects e.g. a medicine and the diseases that it can remedy
`http://www.webmd.com`	Relationships between concepts across the aspects e.g. a diagnostic test and the diseases that it can diagnosed
UMLS	Both hierarchical and across-aspects relationships e.g. a general disease and a specific type of the general disease

Table 6.3: List of resources used for extracting the conceptual relationships related to the four aspects of the medical decision process.

Resources	Association Types	# of Rules
DOID hierarchy	Specific-general	2,046
MeSH	Specific-general	53,915
MedDRA	Specific-general	86,109
DOID	Across aspects	9,664
`http://www.rxlist.com`	Across aspects	5,433
`http://www.webmd.com`	Across aspects	3,694
UMLS	Specific-general & Across aspects	11,212,153

Table 6.4: Number of association rules extracted from each domain-specific resource.

to extract informative concepts from the top-ranked medical records, as it has been shown to be effective in Section 5.4.4.1 of Chapter 5.

As both the Bayesian-based and the stochastic approaches require the λ_r parameter in Equation (6.2) to be properly set, we apply two different training regimes (namely, *5-fold and x-collection validations*), which were described in Section 5.4.1.

6.6.2 Research Questions

We empirically evaluate the Bayesian-based approach (Section 6.3) and the stochastic approach (Section 6.4) to infer the relationships between medical concepts by using association rules. For the remainder of this chapter, we aim to examine the following research questions:

RQ 1. Can the Bayesian-based approach improve the performance of a patient search system?

Rule type	# of rules
diagnosis → diagnosis	2,900,788
diagnosis → diagnostic test	1,907,492
diagnosis → symptom	2,306,712
diagnosis → treatment	2,018,061
diagnostic test → diagnosis	1,905,521
diagnostic test → diagnostic test	2,726,341
diagnostic test → symptom	2,023,443
diagnostic test → treatment	2,695,757
symptom → diagnosis	2,303,486
symptom → diagnostic test	2,020,294
symptom → symptom	4,205,071
symptom → treatment	2,096,350
treatment → diagnosis	2,015,362
treatment → diagnostic test	2,698,453
treatment → symptom	2,097,231
treatment → treatment	6,597,028

Table 6.5: The types of the extracted association rules in terms of the four medical aspects.

RQ 2. Can the stochastic approach improve the performance of a patient search system?

RQ 3. What types of queries are likely to benefit from the Bayesian-based approach?

RQ 4. What types of queries are likely to benefit from the stochastic approach?

6.6.3 Experiment with our Bayesian-based Approach for Inferring Conceptual Relationships

We first evaluate the effectiveness of the Bayesian-based approach for inferring the relationships between medical concepts. We compare the retrieval effectiveness of the Bayesian-based approach with the following baselines:

- The task-specific representation baseline: using the task-specific representation approach previously defined in Section 4.2.1 of Chapter 4.
- The Bo1 baseline: applying the Bo1 query expansion technique with the task-specific representation approach.
- The Bo1 & Semantic QE baseline: using both the Bo1 query expansion technique and a semantic query expansion technique that adds all of the related medical concepts into the queries (as used by King *et al.* (2011)). Note that this baseline uses the same set of external resources as the Bayesian-based approach.

To have a fair train/test setting, we use the *cross-collection validation* and *5-fold cross validation* as described in Section 6.6.1 when setting λ_r within the Bayesian-based approach. Furthermore, to see how the parameter setting impacts on the retrieval performance and the potential effectiveness of our approach, the performance achieved when using the best λ_r for each retrieval measure on each test collection (i.e. best possible setting) is also reported, denoted *oracle*.

Table 6.6 compares the retrieval performances of the Bayesian-based approach with the three baselines on the TREC 2011 and 2012 Medical Records track test collections. From Table 6.6, we observe that the Bo1 baseline outperforms the task-specific representation baseline (i.e. no QE) for both TREC 2011 and 2012. Specifically, the retrieval performances improve from bpref 0.5243 to 0.5403 and from infNDCG 0.4880 to 0.5129 for TREC 2011 and 2012, respectively. Meanwhile, the Bo1 & Semantic QE baseline decreases the retrieval performance.

In contrast, the Bayesian-based approach, using either a 5-fold or an x-collection training regime, outperforms the task-specific representation baseline (both TREC 2011 and 2012). Specifically, the Bayesian-based approach with the 5-fold regime (5-fold) outperforms all of the three baselines used. The Bayesian-based approach (5-fold) achieves a bpref of 0.5474 and an infNDCG of 0.5133. Importantly, our approach (5-fold) significantly (paired t-test, $p < 0.05$) outperforms the task-specific representation baseline in terms of infNDCG for TREC 2012 (0.5133 vs 0.4880).

When comparing the Bayesian-based approach with the Bo1 & Semantic QE baseline, which also uses the same set of resources as the Bayesian-based approach, we observe that the Bayesian-based approach, either using the 5-fold or x-collection training regime, significantly ($p < 0.05$) outperforms this baseline for both the TREC 2011 and 2012 queries.

Additionally, as expected, we find that with a proper setting of the parameter λ_r (i.e. the λ_r that results in the best retrieval performance for each query), the Bayesian-based approach (oracle) can achieve a better retrieval performance. In particular, the bpref retrieval performance is increased to 0.5646, while the infNDCG retrieval performance is improved to 0.5393 (+7.69% and +10.51% over the task-specific representation baseline, respectively). This shows the potential of the Bayesian-based approach if the λ_r value could be effectively set.

In answer to our first research question, the Bayesian-based approach is effective for the patient search task as it outperforms the three baselines, namely the task-specific representation baseline, the Bo1 baseline and the Bo1 & Semantic QE baseline. Indeed, the Bayesian-based approach performs significantly better than the Bo1 & Semantic QE baseline for both TREC 2011 and TREC 2012. Meanwhile, the Bayesian-based approach significantly outperforms the task-specific baseline for the TREC 2012 queries.

Approach	bpref (TREC 2011)	infNDCG (TREC 2012)
Task-specific representation	0.5243	0.4880
+ Bo1	0.5403	0.5129
+ Bo1 & Semantic QE	0.3013	0.2228
+ Our Bayesian-based approach (5-fold)	**0.5474**^{s}	**0.5133**^{t,s}
+ Our Bayesian-based approach (x-collection)	0.5464^{s}	0.5125^{t,s}
+ Our Bayesian-based approach (oracle)	0.5646^{t,s}	0.5393^{t,s}

Table 6.6: Retrieval performances of our two approaches for inferring relationships between medical concepts in comparison with the three baselines on TREC 2011 and 2012 Medical Records track's queries. Statistical significance (paired t-test) at $p < 0.05$ over the task-specific representation and the Bo1 & Semantic QE baselines are denoted ^{t} and ^{s}, respectively.

Approach	bpref (TREC 2011)	infNDCG (TREC 2012)
Task-specific representation	0.5243	0.4880
+ Bo1	**0.5403**	**0.5129**
+ Bo1 & Semantic QE	0.3013	0.2228
+ Our stochastic approach (5-fold)	0.5355^{s}	0.4846^{s}
+ Our stochastic approach (x-collection)	0.5376^{s}	0.4761^{s}
+ Our stochastic approach (oracle)	0.5586^{t,s}	0.5311^{t,s}

Table 6.7: Retrieval performances of our two approaches for inferring relationships between medical concepts in comparison with the three baselines on TREC 2011 and 2012 Medical Records track's queries. Statistical significance (paired t-test) at $p < 0.05$ over the task-specific representation and the Bo1 & Semantic QE baselines are denoted ^{t} and ^{s}, respectively.

Later, in Section 6.6.5, we further analyse the Bayesian-based approach, in comparison with the task-specific representation and the Bo1 baselines.

6.6.4 Experiments with a Stochastic Analysis for Inferring Conceptual Relationships

Next, we discuss the evaluation of the stochastic approach for inferring the conceptual relationships using a stochastic analysis to improve the representation of a given query. We compare the retrieval effectiveness of our approach with the same baselines as in Section 6.6.3. Table 6.7 shows the retrieval performances in terms of bpref and infNDCG for TREC 2011 and 2012, respectively.

From Table 6.7, we observe that the stochastic approach significantly (paired t-test, $p < 0.05$) outperforms the Bo1 & Semantic QE baselines for both TREC 2011 and 2012 queries. In addition, the stochastic approach outperforms the task-specific representation baseline for the TREC 2011 queries. Specifically, our approach with the 5-fold and the x-collection training regimes achieves bpref 0.5355 and 0.5376, respectively, in comparison with bpref 0.5243 for the task-specific representation baseline. However, for the TREC 2012 queries, our approach (either with the 5-fold or the x-collection training regime) could not improve the retrieval performance over the task-specific representation baseline.

Meanwhile, we observe that the Bo1 baseline performs better than the stochastic approach. This is partly due to the limited number of available queries for training the λ_r parameter (in Equation (6.2)). In answer to our second research question, the stochastic approach is effective for the patient search task, as it outperforms both the task-specific representation and the Bo1 & Semantic QE baselines. It significantly (paired t-test, $p < 0.05$) outperforms the latter baseline for both TREC 2011 and 2012 queries. However, we note that it could not outperform the Bo1 baseline. We further discuss the comparison of the stochastic approach and both the task-specific and the Bo1 baselines in Sections 6.6.5.2 and 6.6.5.3, respectively.

When considering the oracle setting with λ_r set to the most effective value for each query, we observe that the stochastic approach could further improve the retrieval performance significantly. This shows the potential of the stochastic approach. However, we leave for future work the study of a more effective training of λ_r.

Next, when comparing the two proposed approaches (see Tables 6.6 and 6.7), we observe that the Bayesian-based approach is more effective than the stochastic approach for both TREC 2011 and TREC 2012. For example, with the 5-fold training regime, the Bayesian-based approach achieves infNDCG 0.5133, while the infNDCG performance of the stochastic approach is 0.4846.

6.6.5 Analysis and Discussion

This section further discusses the performance of the Bayesian-based and the stochastic approaches for inferring the relationships between medical concepts. Specifically, Section 6.6.5.1 discusses which types of queries are likely to benefit from these proposed approaches, while Sections 6.6.5.2 and 6.6.5.3 compare the performance of the two proposed approaches with the task-specific representation baseline and the Bo1 baseline, respectively, on a per-query basis.

6.6.5.1 Failure Analysis

In this section, we analyse the performances of the two proposed approaches in comparison with the task-specific representation baseline to identify which types of queries that are likely to benefit from each of the two approaches. Tables 6.8 and 6.9 show the numbers of queries benefited or harmed by the Bayesian-based and the stochastic approaches, respectively, which are grouped based on the types of the medical concepts in the queries. The types of medical concepts are based on the four aspects of the medical decision process discussed in Chapter 4.

From Tables 6.8 and 6.9, we observe that both the Bayesian-based approach and the stochastic approach are more likely to be effective when the query contains medical concepts related to diagnostic

Aspects of medical concepts in the query	Benefited	Harmed
Symptom	52.9% (37/70)	42.9%(30/70)
Diagnostic test	57.7% (30/52)	38.5%(20/52)
Diagnosis	55.9% (38/68)	38.2% (26/68)
Treatment	55.8% (29/52)	40.4% (21/52)
All 4	50% (16/32)	46.9%(15/32)

Table 6.8: Analysis of our Bayesian-based approach w.r.t. the aspects of medical concepts found the queries. The numbers between the parentheses indicate the number of queries impacted (benefited/harmed) compared to the total number of queries.

Aspects of medical concepts in the query	Benefited	Harmed
Symptom	52.9% (37/70)	41.4%(29/70)
Diagnostic test	55.8% (29/52)	40.4%(21/52)
Diagnosis	51.5% (35/68)	41.2% (28/68)
Treatment	53.8% (28/52)	38.5% (20/52)
All 4	50% (16/32)	43.8%(14/32)

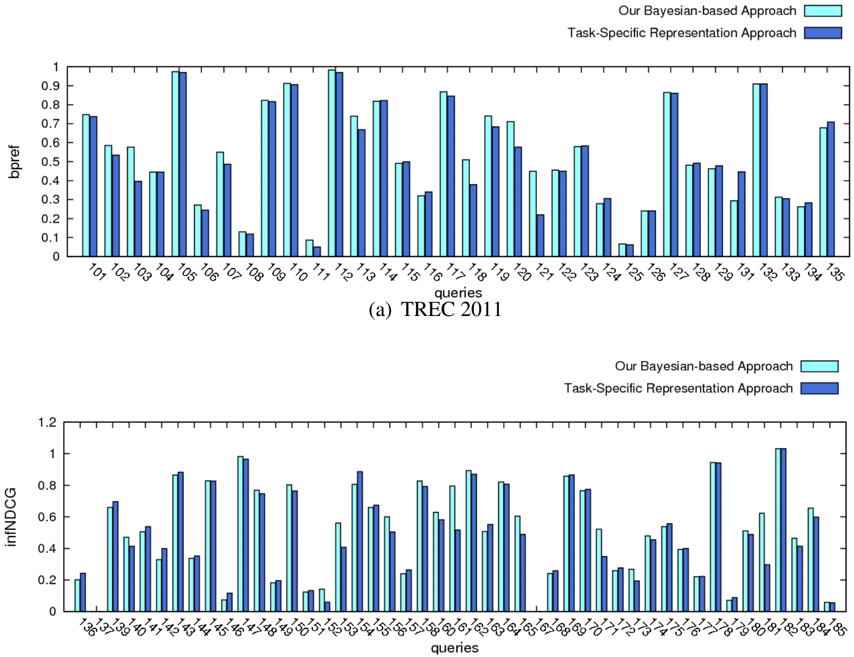
Table 6.9: Analysis of our stochastic approach w.r.t. the aspects of medical concepts found the queries. The numbers between the parentheses indicate the number of queries impacted (benefited/harmed) compared to the total number of queries.

tests. In particular, the Bayesian-based approach and the stochastic approach improve the retrieval performance for 55.9% and 55.8% of the queries containing medical concepts related to diagnostic tests, respectively. Meanwhile, we observe that the queries that contain medical concepts related to all of the four aspects are less likely to benefit from our two approaches. In answer to the third and the fourth research questions, both the Bayesian-based approach and the stochastic approach are likely to benefit the queries containing medical concepts related to diagnostic tests.

6.6.5.2 Comparison with the Task-Specific Representation Baseline

Next, we compare the retrieval performance of the two proposed approaches with the task-specific representation baseline on a per-query basis. We choose to discuss only the 5-fold setting as it is the most effective.

We first discuss the performance of the Bayesian-based approach. From Figure 6.3, we observe that in general the Bayesian-based approach performs better than the task-specific representation baseline. In particular, for the TREC 2011 queries, the Bayesian-based approach outperforms the task-specific representation baseline for 22 out of 34 queries, while it performs worse than the baseline for 10 queries. We observe that the Bayesian-based approach is likely to be more effective for difficult queries. For instance, the Bayesian-based approach could improve the retrieval performance of the queries with a bpref retrieval performance less than 0.25 for 5 out of 6 queries (e.g. queries# 108, 111,



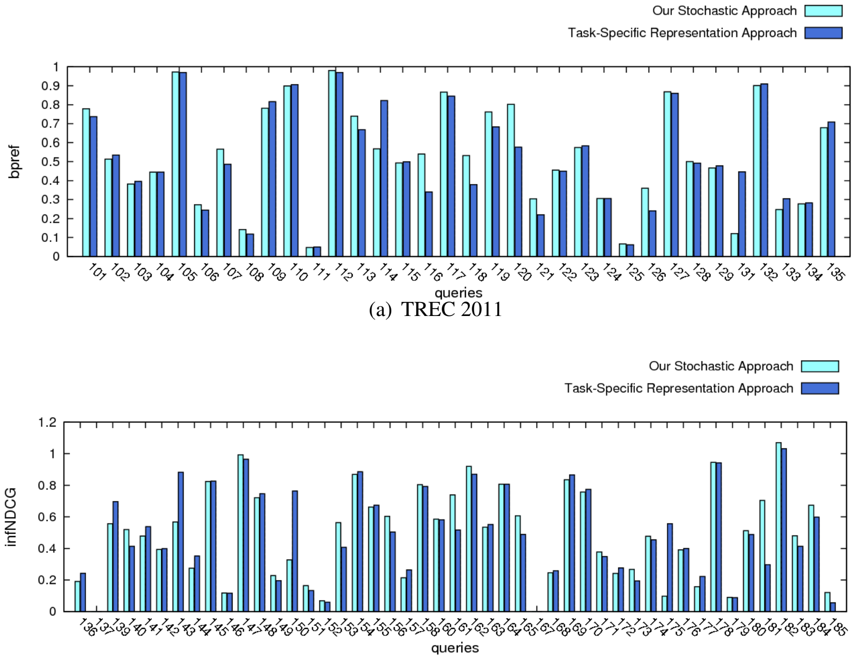
(a) TREC 2011

(b) TREC 2012

Figure 6.3: The retrieval performances of our Bayesian-based approach for inferring the conceptual relationships and the task-specific representation baseline on a per-query basis, evaluated using the TREC 2011 and 2012 Medical Records Track.

121, and 125). Meanwhile, for the TREC 2012 queries, the Bayesian-based approach outperforms the task-specific representation baseline for 23 out of 47 queries, while performing worse than the baseline for 22 queries. Moreover, we observe that the Bayesian-based approach could effectively expand queries with related concepts. For example, in query #125 “Patients co-infected with Hepatitis C and HIV”, the Bayesian-based approach could expand the query with concepts, such as C0001175 (AIDS) and C0024419 (Waldenstrom Macroglobulinemia), which are diseases related to HIV and Hepatitis C, respectively.

In Figure 6.4, we show the per-query retrieval effectiveness of the stochastic approach in comparison with the task-specific representation baseline. Firstly, we observe that the stochastic approach outperforms this baseline for the majority of the queries. Indeed, our approach is more effective for 18 out of 34 queries and for 24 out of 47 queries from TREC 2011 and 2012, respectively. Similar to the Bayesian-based approach, we also observe that the stochastic approach tends to be effective for difficult queries. Specifically, for the queries where the task-specific representation baseline obtains a retrieval performance less than 0.25 (bpref for TREC 2011 and infNDCG for TREC 2012), the stochastic approach outperforms the task-specific representation baseline for 7 out of 11 queries for TREC 2011 and for 5 out of 6 for TREC 2012.



(a) TREC 2011

(b) TREC 2012

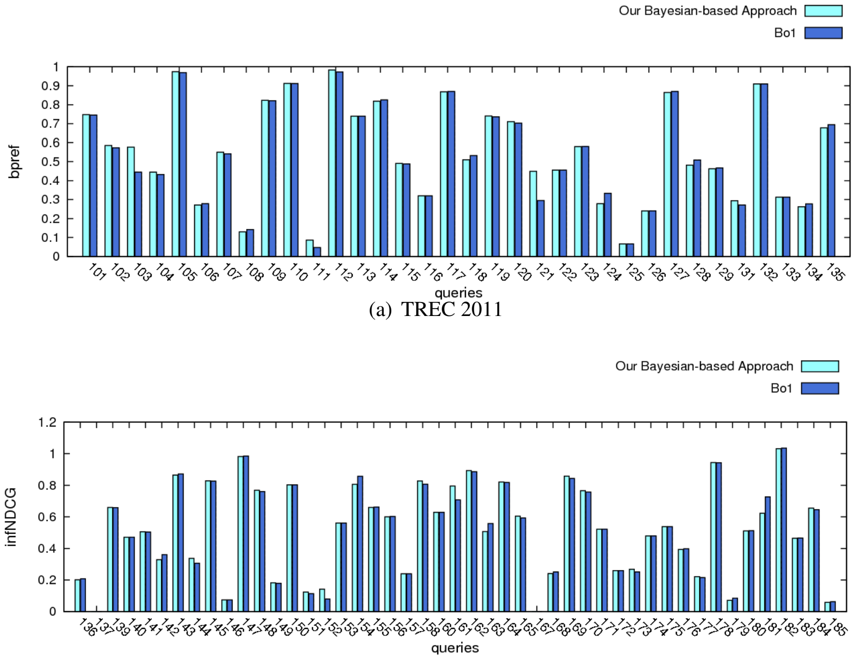
Figure 6.4: The retrieval performances of our stochastic approach for inferring the conceptual relationships and the task-specific representation baseline on a per-query basis, evaluated using the TREC 2011 and 2012 Medical Records Track.

6.6.5.3 Comparison with the Bo1 Baseline

This section discusses the retrieval performances of the two proposed approaches in comparison with the Bo1 baseline, which is a stronger baseline. We compare the retrieval performances on a per-query basis.

We first discuss the retrieval effectiveness of the Bayesian-based approach. From Figure 6.5, we observe that the Bayesian-based approach performs comparably to the Bo1 baseline. Indeed, our approach outperforms the Bo1 baseline for 16 out of 34 queries and for 21 out of 47 queries for TREC 2011 and TREC 2012, respectively. Meanwhile, the Bo1 baseline performs better than our approach for 12 and 20 queries for TREC 2011 and 2012.

Next, we discuss the retrieval performance of the proposed stochastic approach. From Figure 6.6, we observe that the Bo1 baseline benefits more queries than the stochastic approach. Specifically, the Bo1 baseline outperforms the stochastic approach for 16 (resp. 23) out of 34 (resp. 47) queries for TREC 2011 (resp. TREC 2012), while the stochastic approach performs better for 11 and 20 queries for TREC 2011 and 2012, respectively. However, this is expected, as the Bo1 baseline achieves a better retrieval performance than the stochastic approach, as shown in Table 6.7. Nevertheless, when



(a) TREC 2011

(b) TREC 2012

Figure 6.5: The retrieval performances of our Bayesian-based approach for inferring the conceptual relationships and the Bo1 query baseline on a per-query basis, evaluated using the TREC 2011 and 2012 Medical Records Track.

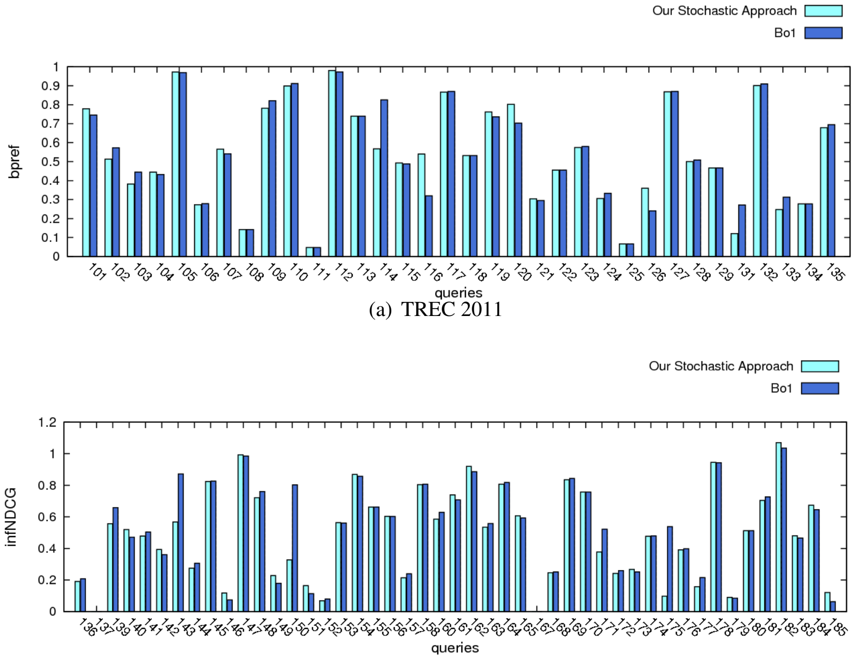
considering the performance on *the difficult queries* (i.e. queries that obtain retrieval performance $<$ 0.25 when query expansion is not applied), we observe that the stochastic approach outperforms the Bo1 baseline performs better than the Bo1 baseline for 8 out of 16 queries across the TREC 2011 and 2012, while it performs worse than the Bo1 baseline for only 3 queries.

Importantly, from this analysis, we observe that the Bayesian-based approach is in general effective, and in particular more effective than the Bo1 baseline. On the other hand, the stochastic approach is more effective for the difficult queries.

From the per-query performance comparison in Sections 6.6.5.2 and 6.6.5.3, we observe that both the Bayesian and the stochastic approaches are more likely to benefit the difficult queries (i.e. the queries that obtain retrieval performance $<$ 0.25 when not applying query expansion).

6.7 Conclusions

We have investigated approaches to instantiate the Conceptual Reasoning component of our framework (described in Section 4.2), which uncovers implicit knowledge at the record level. In particular, we discussed two approaches to infer the relationships between medical conditions, which build upon the



(a) TREC 2011

(b) TREC 2012

Figure 6.6: The retrieval performances of our stochastic approach for inferring the conceptual relationships and the Bo1 baseline on a per-query basis, evaluated using the TREC 2011 and 2012 Medical Records Track.

task-specific representation approach introduced in Section 4.2.1 of Chapter 4. We firstly showed how to extract association rules between medical concepts from external resources, including free-text documents and ontologies, in Section 6.2. Then, we investigated two approaches to take into account both medical concepts inferred using a pseudo-relevance feedback approach and the association rules of relationships between medical concepts. These two approaches infer the relationships between medical concepts in the association rule database by using a Bayes' theorem (Sections 6.3) and a stochastic analysis (Section 6.4), respectively. From the experimental results, we observed that with a fair setting the Bayesian-based approach improved the retrieval performance by up to 10.51% (see Table 6.6, Section 6.6.3). On the other hand, we observed that the stochastic approach were likely to perform effectively for difficult queries (see Table 6.7, Section 6.6.4). These results support the thesis statement detailed in Section 1.3, in that the relationships between the medical conditions (i.e. medical concepts) related to the medical decision process could be used to uncover implicit knowledge in patient search at the record level of the retrieval process.

In the next chapter, we introduce approaches to instantiate the Department Expertise component of the framework, which uncovers implicit knowledge at the inter-record level. In particular, we focus on leveraging knowledge gained from aggregates of the medical records issued by particular hospital departments to improve patient search. For instance, we aim to automatically detect from the aggregates of medical records that the cardiology department has expertise in heart diseases; hence, for a query that searches for patients with heart diseases, those patients who have medical records issued from the cardiology department should be promoted.

Chapter 7

Leveraging Knowledge about Hospital Departments' Expertise

7.1 Introduction

In this chapter, we discuss the Department Expertise component of our framework. Recall that this component aims to extract implicit knowledge at the inter-record level of the retrieval process (see Section 4.2). In particular, we propose to explicitly make available to an information retrieval (IR) system some of the implicit knowledge, by exploiting insights gained from aggregates of medical records. For a medical query, we propose to weight the importance of each hospital department for that query by considering the medical records created by that department. In particular, we leverage this evidence (i.e. *the department-level evidence*) to emphasise on the medical records that were created by the departments whose expertise is relevant to the query. As shown in Figure 7.1, a medical record contains information about its type and subtype, which are related to the hospital department that the medical record is issued from. We form the department-level evidence from medical records that share the same *type* and *subtype* tags shown in Figure 7.1. We propose to use this department-level evidence to give higher importance to the medical records from the hospital departments that specialise in the medical condition(s) stated in each medical query. For example, for a query about heart disease, a higher importance is given to medical records from the cardiology department. We argue that the modelling and use of the department-level evidence by a patient search system will lead to enhanced retrieval performance.

The remainder of this chapter is organised as follows:

- In Section 7.2, we introduce our two approaches to model department-level evidence, when estimating the relevance of patients based on their medical records.

```
<report>
<type>ECHO</type>
<subtype>TEE</subtype>
<admit_diagnosis> 414.12</admit_diagnosis>
...
<report_text>
... (report text here) ...
</report_text>
</report>
```

Figure 7.1: An example of a transesophageal echocardiography (*TEE*) medical record, from the cardiology department.[2]

- Section 7.3 discusses our two techniques for obtaining department-level evidence from aggregates of medical records, which are adapted from a voting (Section 7.3.1) and federated search (Section 7.3.2) paradigms, respectively.
- In Section 7.4, we discuss the evaluation of the proposed approaches to leverage department-level evidence to enhance the retrieval performance of a patient search system.
- Section 7.5 summarises the key findings from this chapter.

7.2 Leveraging Department-Level Evidence

In this section, we describe our two approaches for leveraging department-level evidence, when ranking patients based on the relevance of their medical records towards a medical query. Specifically, in Section 7.2.1, we introduce the first approach that enables the voting techniques (see Section 4.2.2) to take into account the department-level evidence when aggregating the relevance scores of the medical records to rank patients. On the other hand, in Section 7.2.2, we propose the second approach that leverages the department-level evidence when estimating the relevance scores of the medical records before aggregating the relevance scores of patients using the voting techniques.

7.2.1 The Extended Voting Techniques for Department-Level Evidence

We first introduce our approach (namely, *the aggregate scoring approach*) that extends the voting techniques to take into account the department-level evidence. In particular, the aggregate scoring approach enables a voting technique to highly weight medical records from hospital departments that are expert in the medical conditions stated in a query, when aggregating the relevance scores of the patients. We

[2] TEE and ECHO are the abbreviations of ‘transesophageal echocardiography’ and ‘cardiology reports’, respectively.

extend the expCombSUM and expCombMNZ voting techniques from the Voting Model (see Equations (2.13) and (2.14) in Section 2.7, respectively) by allowing the setting of different weights on particular medical records, as they have been shown to be effective for the patient search task (see Tables 4.5 and 4.7 of Chapter 4). Hence, our extended *expCombSUMw* and *expCombMNZw* techniques can take into account the expertise of the department of each medical record (i.e. department-level evidence) to focus on medical records from hospital departments that have medical expertise relevant to the query when ranking patients.

In particular, we define $profile(p)$ to be the set of medical records associated with a patient p, while $R(q)$ is a ranking of all medical records with respect to query q. As each patient is represented by an aggregate of the associated medical records, each medical record retrieved in $R(q)$ is said to vote for the relevance of its associated patient. The proposed expCombSUMw and expCombMNZw voting techniques score a patient p with respect to a query q as follows:

$$score\_patient_{expCombSUMw}(p,q) = \sum_{d \in R(q) \cap profile(p)} w(d,q) \cdot e^{score_r(d,q)} \tag{7.1}$$

$$score\_patient_{expCombMNZw}(p,q) = \left[|R(q) \cap profile(p)| \cdot \sum_{d \in R(q) \cap profile(p)} w(d,q) \cdot e^{score_r(d,q)} \right] \tag{7.2}$$

where $R(q) \cap profile(p)$ is the set of medical records associated with the patient p that are also in the ranking $R(q)$; $|R(q) \cap profile(p)|$ is the number of medical records in this set; and $score_r(d,q)$ is the relevance score of medical record d for query q, as obtained from a standard weighting model such as BM25 (Equation (2.3)).

Within Equations (7.1) and (7.2), we draw attention to the addition of $w(d,q)$ to the expCombSUM and expCombMNZ voting techniques (Equations (2.13) and (2.14) in Section 2.7), which permits different weights for different medical records. As previously discussed in Section 4.2.3.3, the existing voting techniques (e.g. Equation (2.13)) do not take into account the information that the medical records of patients with a particular medical condition are likely to be predominantly from a particular hospital department that specialises in that condition. We use $w(d,q)$ to put emphasis on medical records associated with particular hospital departments that are relevant to query q, as follows:

$$dep = department(d) \tag{7.3}$$

$$w(d,q) = 1 + (\lambda \cdot score\_department(dep,q)) \tag{7.4}$$

where $department(d)$ returns the department dep that issued medical record d, and λ is a parameter that controls emphasis to place on the department-level evidence ($\lambda \geq 0$, where $\lambda = 0$ disables the use of department-level evidence). The relevance of a department dep to a query q, $score\_department(dep, q)$, allows the expCombSUMw and expCombMNZw to focus on medical records from particular hospital departments whose department-level evidence is relevant to the query q.

Later, in Section 7.3, we will introduce two techniques from a voting (Section 7.3.1) and a federated search paradigm (Section 7.3.2) to obtain department-level evidence and estimate the relevance score of a department ($score\_department(dep, q)$) in Equation (7.4).

7.2.2 Leveraging Department-Level Evidence when Ranking Medical Records

The department-level evidence can also be exploited when ranking medical records towards the queries, before aggregating the relevance scores to rank their associated patients. In this section, we discuss our approach (namely, *the record scoring approach*) to model the department-level evidence when ranking medical records based on their relevance towards a query. To take into account the hospital department that issues a medical record d, the record scoring approach calculates the relevance score of the medical record d, as a linear combination of the relevance score of the medical record d and the relevance score of the hospital department which d was issued from (i.e. $department(d)$), as follows:

$$score(d, q) = score_r(d, q) + (\lambda \cdot score\_department(department(d), q)) \tag{7.5}$$

where $department(d)$ is the department that issued the medical record d, $score_r(d, q)$ is the relevance score of the medical record d, which can be calculated using any standard weighting model, and λ is a parameter to specify the emphasis to place upon department-level evidence ($\lambda \geq 0$) in comparison to the relevance score of the medical record d. We discuss our techniques to estimate the score of the department-level evidence ($score\_department(department(d), q)$) in Section 7.3.

Then, when ranking patients the relevance scores ($score(d, q)$) of the medical records computed using Equation (7.5) are aggregated using a voting technique such as CombSUM (Equation (4.2)).

7.3 Obtaining Department-Level Evidence

In Sections 7.3.1 and 7.3.2, we discuss our techniques adapted from a voting and federated search paradigm, respectively, to capture the department-level evidence and estimate the relevance score of a department as used in Equations (7.4) and (7.5).

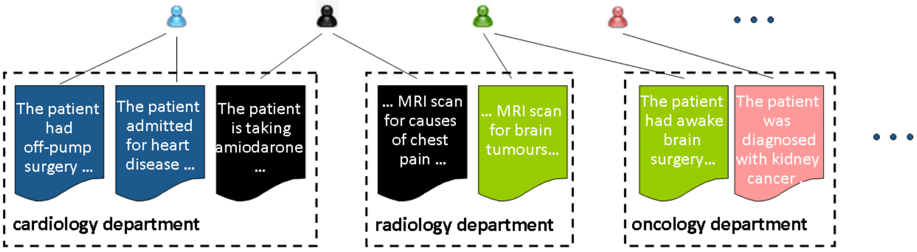

The patient had off-pump surgery …
The patient admitted for heart disease …
The patient is taking amiodarone …
cardiology department
… MRI scan for causes of chest pain …
… MRI scan for brain tumours…
radiology department
The patient had awake brain surgery…
The patient was diagnosed with kidney cancer
oncology department

Figure 7.2: Examples of medical records from hospital departments.

7.3.1 A Voting Paradigm for Modelling Department-Level Evidence

Within the voting paradigm, we introduce the first technique to represent the inherent implicit knowledge in the form of department-level evidence. In particular, we propose to aggregate the medical records from each hospital department to capture some of the implicit knowledge about the expertise of that department. This implicit knowledge may not be available in a traditional IR system, since such knowledge is not explicitly stated in a single medical record, but could be captured from the aggregates of medical records issued by particular hospital departments. Indeed, we argue that this implicit evidence about the hospital departments' expertise is useful for improving the retrieval performance.

Specifically, we estimate department-level evidence by using the medical records associated with individual departments. Figure 7.2 shows examples of the structure of medical records from each hospital department. For instance, we represent the department-level evidence of the cardiology department with all of the medical records issued by that department. This permits the IR system a high-level view of each hospital department's expertise that could not be captured in any individual medical record. For example, the expertise of the cardiology department captured in the department-level evidence may encompass evidence of its expertise in heart disease, heart failure, valvular disease, or off-pump surgery. Hence, for instance, the IR system can infer that a medical record from the cardiology department has at least a small probability to be about a heart condition.

We use this department-level evidence to estimate the relevance of a hospital department's expertise towards a query q, based on the relevance score of aggregates of medical records from the individual hospital departments. As the department-level evidence is represented by aggregates of their associated medical records, a voting technique from the Voting Model can be used to effectively rank departments with respect to a query. Recall that in Chapter 4 (e.g. Equation (4.2)), we use the voting techniques to aggregate the relevance scores of medical records in order to rank patients. On the other hand, in this section, we use the voting techniques to calculate the relevance scores of hospital departments. For instance, in Equation (4.2), we use $profile(dep)$, which is the set of medical records associated with the department dep, instead of $profile(p)$.

This relevance score of a hospital department is further used by the two approaches for leveraging department-level evidence introduced in Section 7.2 to put more emphasis on the medical records from particular hospital departments when ranking patients based on the relevance of their associated medical records towards a query.

7.3.2 A Federated Search Technique for Modelling Department-Level Evidence

The second technique to extract department-level evidence is inspired by the work of federated search of Callan (2000). We propose that federated search techniques could be deployed to rank hospital departments by representing those departments as databases containing their associated medical records. Specifically, to model the department-level evidence inherent within the medical records, we represent each database (i.e. resource) by the terms (and their frequencies) found in the medical records of the same hospital department. In particular, we build an index (i.e. a database) for the set of medical records from each hospital department. For instance, the database representing the cardiology department contains statistics of terms occurring in medical records issued from this department. This may allow each database to represent the expertise of the corresponding hospital department. For example, the medical records of patients having symptoms or treatments related to heart diseases are issued by the cardiology department, as shown in Figure 7.2.

Classical federated search approaches were typically designed for uncooperative environments of databases (Callan, 2000), and hence require the use of a query-based sampling technique to create a representation of each resource. Notably, we are working in a cooperative environment, hence resource sampling is not required. We use the CORI database selection algorithm (Callan, 2000) to calculate the relevance scores of databases (i.e. hospital departments), since this technique has been shown to be effective on different federated search tasks (Callan, 2000; Ogilvie & Callan, 2001; Si & Callan, 2002). In particular, the relevance score (i.e. belief) $p(t_i|dep)$ of the database representing a hospital department dep, according to a query term t_i is calculated as follows (Callan, 2000):

$$T = \frac{df}{df + 50 + 150 \cdot \frac{cw}{avg_{cw}}} \tag{7.6}$$

$$I = \frac{log|DB| + 0.5}{cf} \tag{7.7}$$

$$p(t_i|dep) = b + (1 - b) \cdot T \cdot I \tag{7.8}$$

where df is the number of medical records in the database representing the hospital department dep that contains term t_i, cf is the number of databases that contain t_i, $|DB|$ is the number of the databases in the collection, cw is the number of terms in database representing department dep, avg_{cw} is the average number of terms among the databases in the collection, and b is parameter, which is set to 0.4 as recommended by Callan (2000).

Next, the beliefs (i.e. relevance scores) based on each term in a query are combined into the final belief that a database representing department dep is relevant to the query (i.e. the relevance score of the department for the query) using belief operators (Turtle & Croft, 1991*b*). In particular, during our experiments, we combine beliefs using SUM, OR, and AND operators, as follows:

$$score\_department_{CORI\_SUM}(dep,q) = \frac{\sum_{t_i \in q} p(t_i|dep)}{|q|} \tag{7.9}$$

$$score\_department_{CORI\_OR}(dep,q) = 1 - \prod_{t_i \in q}(1 - p(t_i|dep)) \tag{7.10}$$

$$score\_department_{CORI\_AND}(dep,q) = \prod_{t_i \in q}(p(t_i|dep)) \tag{7.11}$$

where $p(t_i|dep)$ is the relevance score (i.e. belief) calculated using Equation (7.8) and $|q|$ is the number of query terms.

Generally in federated search systems, the 5 or 10 databases with the highest belief scores are selected and documents will be retrieved only from these databases. However, we use all of the databases' relevance scores, since we need the estimates of the relevance towards a given query for individual hospital departments. In particular, these database relevance scores are used by the two approaches for leveraging department-level evidence to take into account the expertise of hospital's departments, when ranking patients based on the relevance of their medical records (Equations (7.4) and (7.5) in Section 7.2).

7.4 Experiments

This section evaluates the retrieval effectiveness of the two proposed approaches for extracting and leveraging knowledge about hospital departments' expertise to enhance a patient search system. In particular, Section 7.4.1 describes the experimental setup we use to evaluate the two proposed approaches. Section 7.4.2 lists the research questions that are investigated in this chapter. In Sections 7.4.3 and 7.4.4, we

Table 7.1: Statistics of hospital departments in the collection.

Number of databases (i.e. hospital departments)	328
Minimum number of medical records per database	1
Maximum number of medical records per database	19,769
Average number of medical records in the databases	307.52
Standard deviation of the number of medical records in the databases	1397.44
Minimum number of terms per database	79
Maximum number of terms per database	2,723,596
Average number of terms in the databases	91,609.29
Standard deviation of the number of terms in the databases	332,880.76

evaluate the proposed approaches for leveraging the department-level evidence, using the techniques described in Sections 7.2.1 and 7.2.2, respectively. In particular, in Sections 7.4.3 and 7.4.4, we investigate the retrieval performance of the two approaches when uniformly setting the λ parameter (Equations (7.4) and (7.5)). Then, in Section 7.4.5, we discuss how we can automatically learn the λ value in the two approaches. Finally, we further analyse the retrieval performances achieved by the proposed approaches in Section 7.4.6.

7.4.1 Experimental Setup

As discussed in Section 7.2, we argue that modelling department-level evidence in patient search could leverage the inherent implicit knowledge within medical records, and hence could improve retrieval effectiveness. In particular, we argue that department-level evidence gained from the aggregates of medical records issued by particular departments could be used as novel evidence to infer the importance of a patient's medical record to a particular query when searching for relevant patients. We evaluate the two proposed approaches in the context of the TREC Medical Records track's test collection previously described and used in Section 4.3.1. Using the information of the structure of the collection, we define 328 hospital departments[3]. In particular, Table 7.1 shows statistical information about the collection of 328 hospital departments.

We deploy the same settings as in Section 4.3.1, where DFR DPH (Equation (2.8)) and BM25 (Equation (2.3)) are used to rank medical records, before aggregating the relevance scores of patients using the expCombSUM or expCombMNZ voting techniques. In addition, to estimate the relevance of hospital departments, we use either the voting techniques (Section 7.3.1) or the federated search techniques (Section 7.3.2). When computing the relevance scores of hospital departments using the voting techniques (Section 7.3.1), we deploy the same voting technique as the one used to rank patients.

[3] We define the department of a medical record automatically using the combination of its *type* and *subtype* tags (see Figure 7.1); however, this may allow sub-units of a department to be considered as departments.

For example, for the expCombMNZw approach, we use the expCombMNZ voting technique to estimate the relevance of hospital departments.

We compare the effectiveness of the two approaches to exploit the department-level evidence with the common baseline identified in Section 4.3 of Chapter 4, which does not consider the department-level evidence (i.e. setting λ in Equations (7.4) and (7.5) to 0).

7.4.2 Research Questions

We investigate the effectiveness of the two approaches to model department-level evidence when ranking patients based on the relevance of their medical records (Section 7.2) and our techniques to extract the department-level evidence (Section 7.3). In particular, in the remainder of this chapter, we examine the following research questions:

RQ 1. Can the aggregate scoring approach that leverages the department-level evidence extracted using the voting techniques improve the retrieval performance?

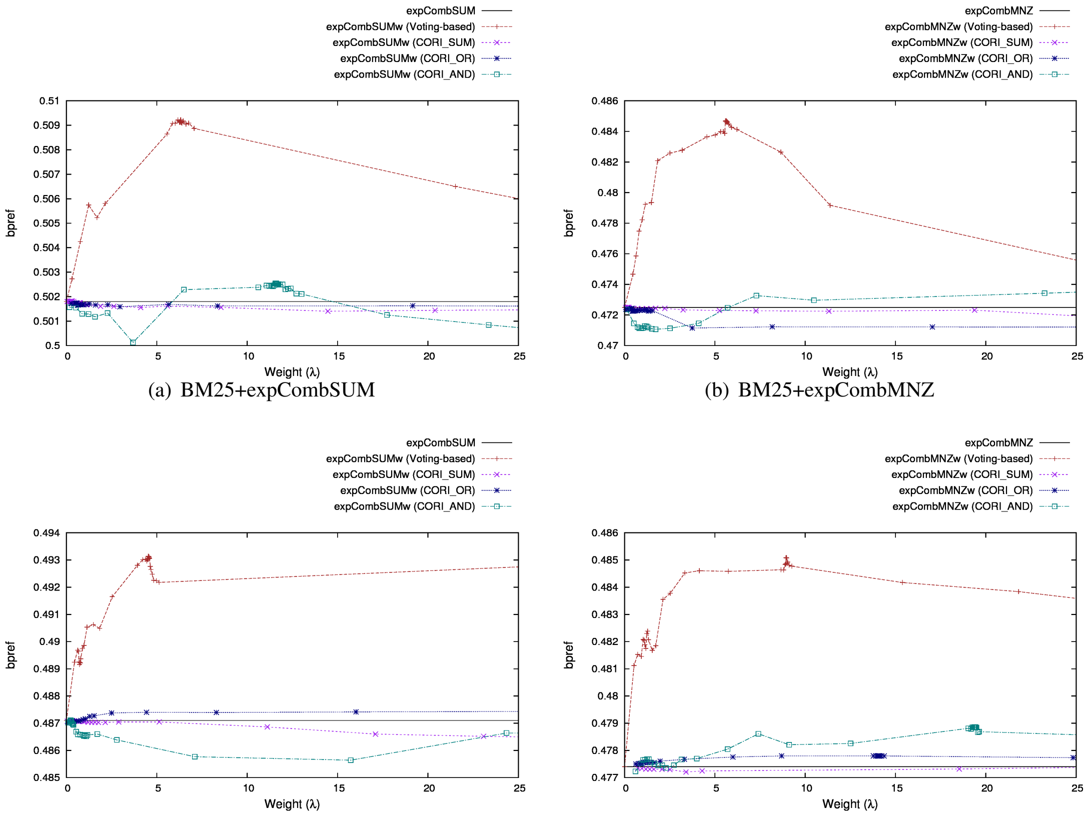
RQ 2. Can the aggregate scoring approach that leverages the department-level evidence extracted using the federated search techniques improve the retrieval performance?

RQ 3. Can the record scoring approach that leverages the department-level evidence extracted using the voting techniques improve the retrieval performance?

RQ 4. Can the record scoring approach that leverages the department-level evidence extracted using the federated search techniques improve the retrieval performance?

7.4.3 Experiments with Our Aggregate Scoring Approach

We firstly discuss the retrieval performance of the aggregate scoring approach that leverages the department-level evidence by extending a voting technique (see Section 7.2.1), when applied with a term-based representation. Recall that, in this section, we experiment with the uniform setting of the λ value. Figure 7.3 shows the retrieval effectiveness in terms of bpref of our approach compared with the common baseline (i.e. $\lambda = 0$) where the department-level evidence is not taken into account for the TREC 2011 queries. We observe that the aggregate scoring approach for leveraging the department-level evidence (i.e. expCombSUMw and expCombMNZw) outperforms the common baseline that does not consider the department-level evidence (i.e. expCombSUM and expCombMNZ, respectively). In particular, when leveraging the department-level evidence obtained using a voting technique (Section 7.3.1), both expCombSUMw and expCombMNZw are more effective than both the corresponding common



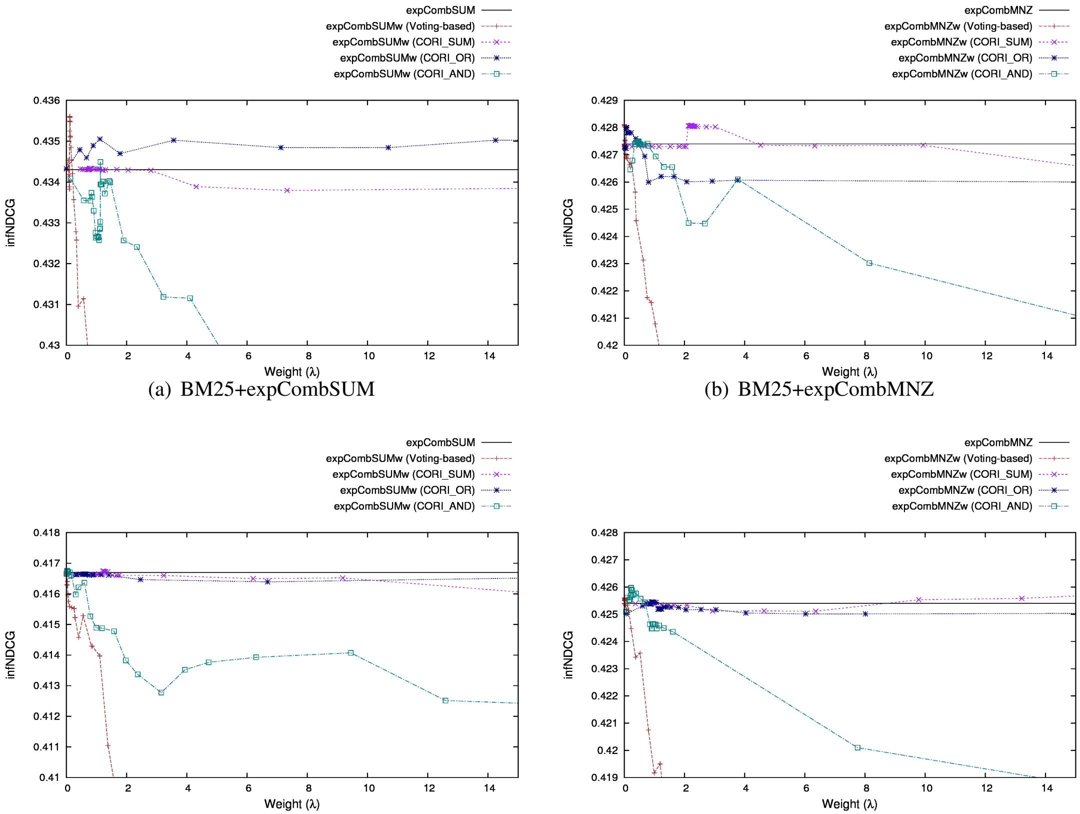
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 7.3: The bpref performances of the aggregate scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with a term-based representation.



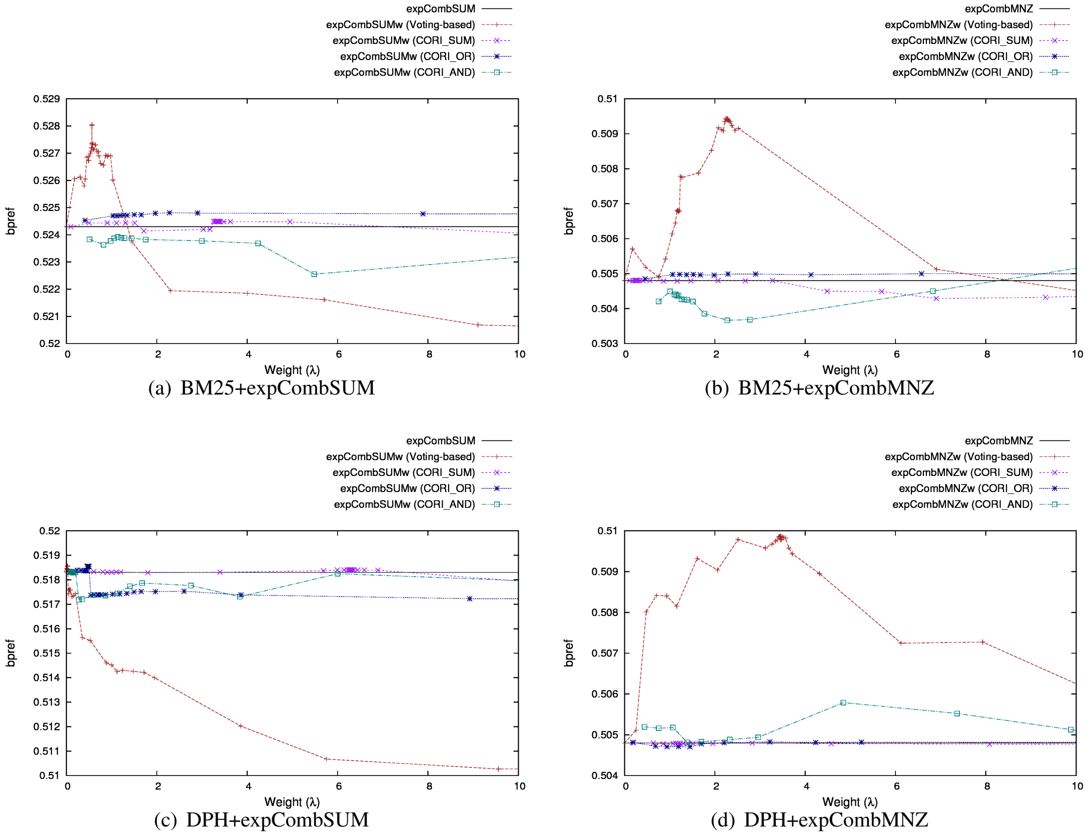
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 7.4: The infNDCG performances of the aggregate scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with a term-based representation.



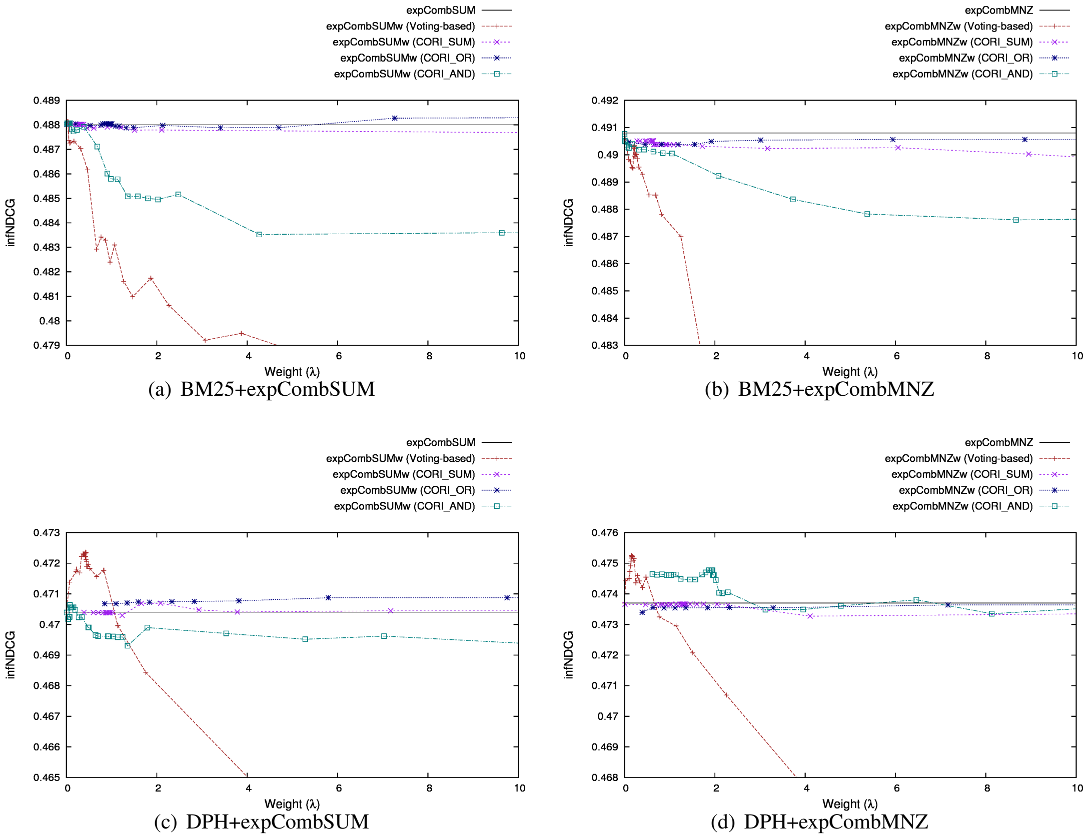
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 7.5: The bpref performances of the aggregate scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with the task-specific representation.



(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

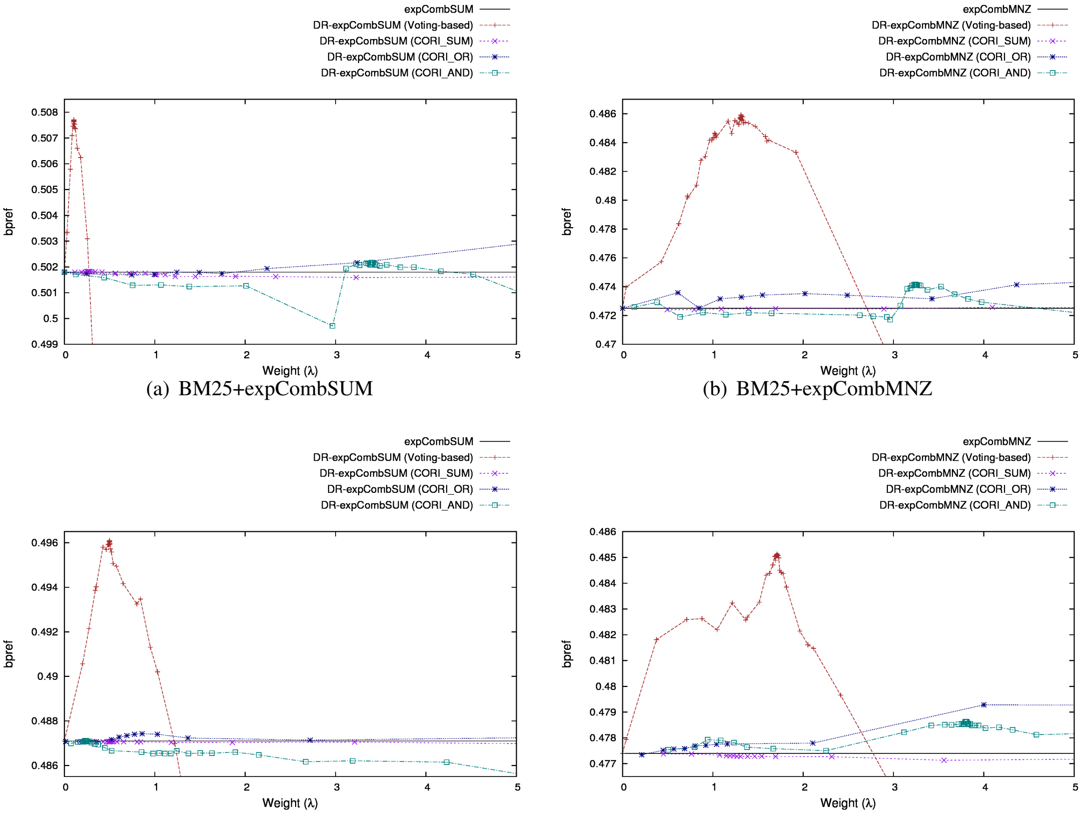
(d) DPH+expCombMNZ

Figure 7.6: The infNDCG performances of the aggregate scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with the task-specific representation.

baseline (i.e. expCombSUM and expCombMNZ, respectively) as well as the expCombSUMw and expCombMNZw approaches that use the department-level evidence obtained using the federated search techniques (i.e. CORI_SUM, CORI_OR, and CORI_AND) (Section 7.3.2). For instance, when applied with the BM25 weighting model, the expCombMNZw technique ($\lambda \approx 6$) achieves a bpref of 0.4848, while the bpref performance of the expCombMNZ baseline is 0.4725 (see Figure 7.3). However, note that these performance improvements are not statistically significant (paired t-test, $p > 0.05$). However, we find that the aggregate scoring approach is robust as it outperforms the common baseline (λ = 0) for a wide range of λ ($0 < \lambda \leq 25$). Meanwhile, when leveraging the department-level evidence obtained using the federated search techniques, both expCombSUMw and expCombMNZw marginally outperform the common baseline ($\lambda = 0$). However, as shown in Figure 7.4, the aggregate scoring approach could only perform comparably to the common baseline for TREC 2012 queries.

Next, in Figure 7.5, we show the retrieval effectiveness of the aggregate scoring approach when applied with the task-specific representation approach, which represents medical records and queries using medical concepts related to symptom, diagnostic test, diagnosis, and treatment (see Section 4.2.1 of Chapter 4). From Figure 7.5, we observe that except when applied with the DPH weighting model and the expCombSUM voting technique, the aggregate scoring approach (i.e. expCombSUMw and expCombMNZw) that leverages the department-level evidence obtained using the voting techniques outperforms the common baseline (i.e. expCombSUM and expCombMNZ, respectively), for the TREC 2011 queries. For example, when applied with the BM25 weighting model, the expCombSUMw technique ($\lambda \approx 2.3$) improves the retrieval performance from bpref 0.5243 to 0.5280. On the other hand, when using the department-level evidence obtained using the federated search techniques, the aggregate scoring approach performs comparably to the common baseline. In addition, we find that with the task-specific representation approach, the retrieval performance of the proposed aggregate scoring approach is more sensitive to the value of λ. For instance, expCombMNZw performs better than the baseline, when λ is set in the range of (0,6]. Meanwhile, we find that the aggregate scoring approach is less effective for the TREC 2012 queries, as shown in Figure 7.6.

Overall, we find that the aggregate scoring approach is more effective for the TREC 2011 queries than for the TREC 2012 queries. Meanwhile, the aggregate scoring approach is effective for both the term-based and the task-specific representation approaches. In addition, the aggregate scoring approach when leveraging the department-level evidence obtained using the voting techniques is more effective than when using the federated search techniques. In answer to the first and the second research questions, the aggregate scoring approach can effectively leverage the department-level evidence gained using



(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

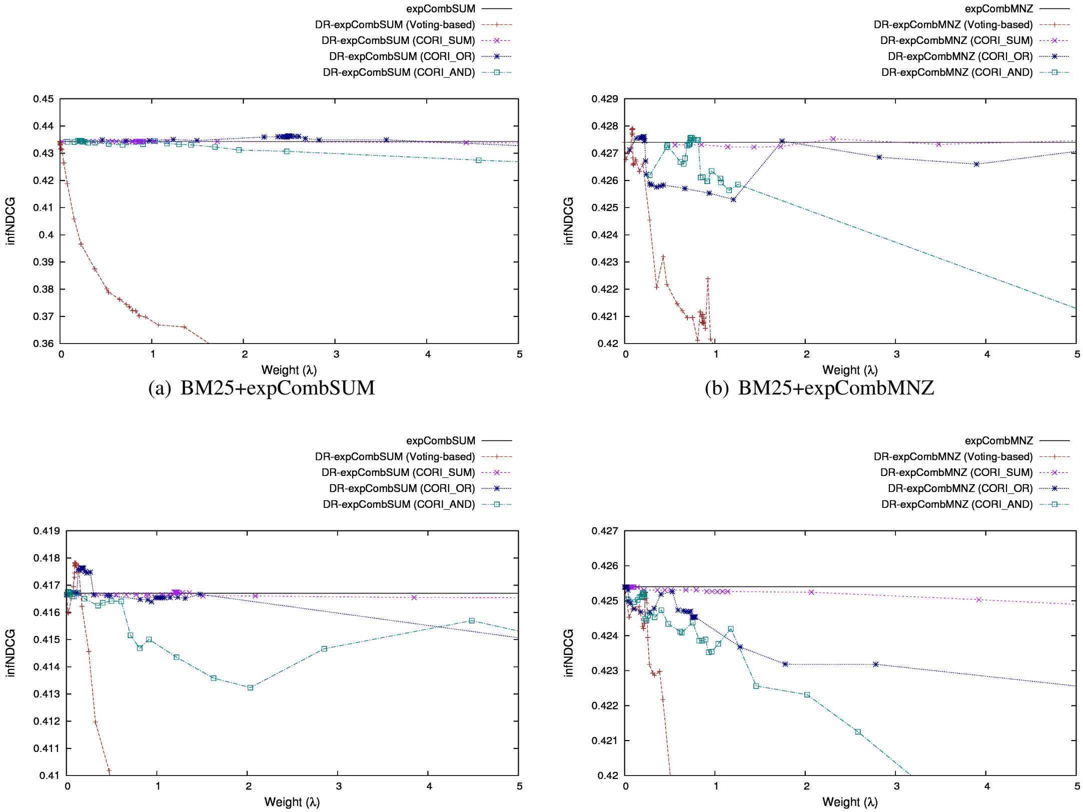
(d) DPH+expCombMNZ

Figure 7.7: The bpref performances of the record scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with a term-based representation.

either the voting techniques or the federated search techniques. However, using the department-level evidence obtained using the voting techniques leads to more effective retrieval performances.

7.4.4 Experiments with Our Record Scoring Approach

In this section, we evaluate the performance of the record scoring approach that models the department-level evidence when calculating the relevance scores of medical records, before aggregating the relevance scores for their associated patients (see Section 7.2.2). We compare the proposed record scoring approach (denoted *DR*) with the common baseline ($\lambda = 0$) used in Section 7.4.3. As shown in Figure 7.7, when applied with a term-based representation approach, the record scoring approach (i.e. *DR*) is able to leverage the department-level evidence obtained using the voting techniques and outperforms the common baseline ($\lambda = 0$) for all the variants of the used retrieval models and voting techniques, for the TREC 2011 queries. For instance, when applied with BM25 and expCombMNZ, the record



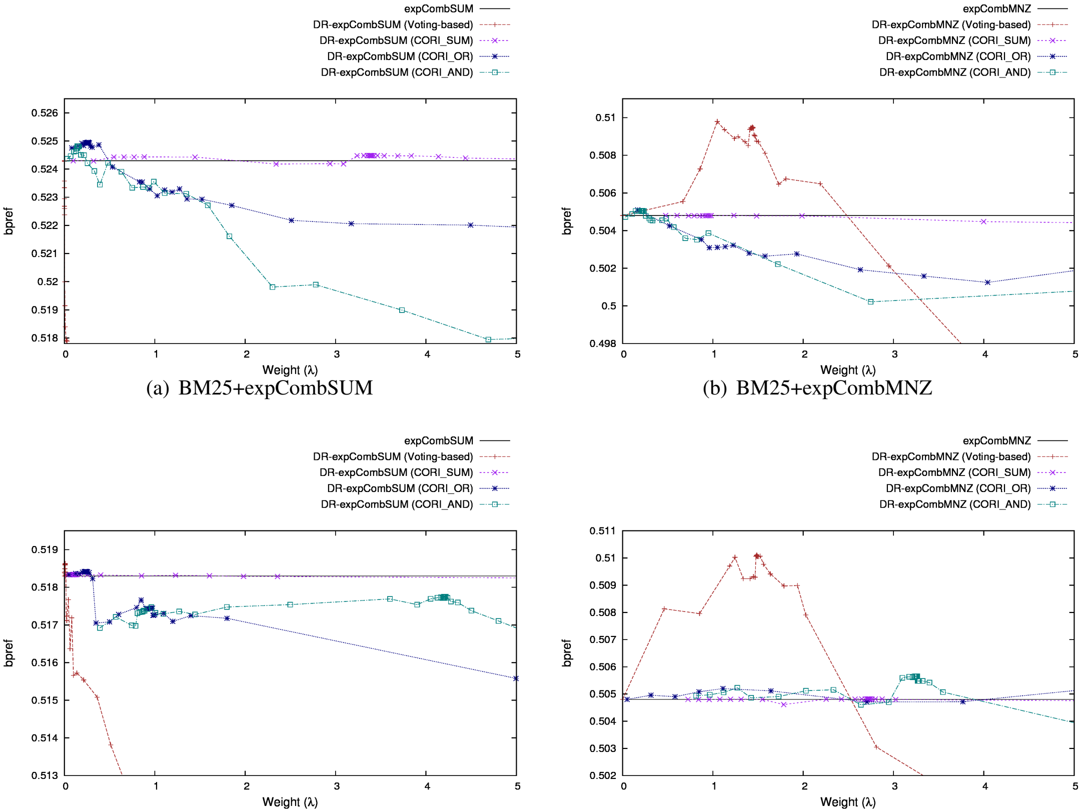
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 7.8: The infNDCG performances of the record scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with a term-based representation.



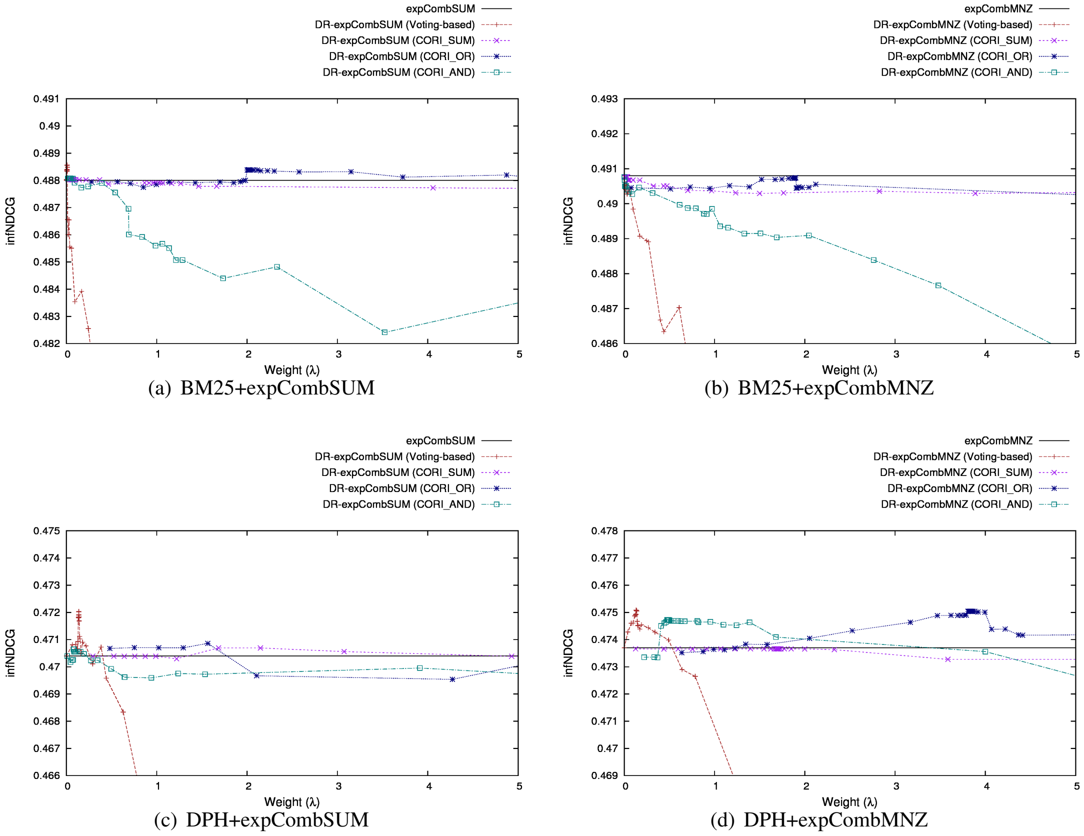
(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 7.9: The bpref performances of the record scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with the task-specific representation.



(a) BM25+expCombSUM

(b) BM25+expCombMNZ

(c) DPH+expCombSUM

(d) DPH+expCombMNZ

Figure 7.10: The infNDCG performances of the record scoring approach for leveraging the department-level evidence while varying the parameter λ, when applied with the task-specific representation.

scoring approach ($\lambda \approx 1.4$) improves the bpref retrieval performance to 0.4859, in comparison to the expCombMNZ baseline whose bpref is 0.4722. Meanwhile, the department-level evidence obtained using CORI is less effective. On the other hand, for the TREC 2012 queries, the record scoring approach performs comparably to the common baseline (see Figure 7.7).

In Figures 7.9 and 7.10, we compare the retrieval performance of the record scoring approach with the common baseline, when applied with the task-specific representation approach. From Figure 7.9, we observe the same pattern, i.e. that the record scoring approach when leveraging the department-level evidence obtained using the voting techniques outperforms the common baseline for the TREC 2011 queries. For example, when applied with BM25 and expCombMNZ, the record scoring approach ($\lambda \approx 1$) improves the retrieval performance from bpref 0.5044 to 0.5100. Meanwhile, the department-level evidence obtained using the federated search techniques is less effective. On the other hand, for the TREC 2012 queries, the retrieval performance of the record scoring approach is comparable to the common baseline. In answer to the third and the fourth research questions, the record scoring approach can effectively leverage the department-level evidence obtained using either the voting techniques or the federated search techniques. However, leveraging the department-level evidence obtained using the voting techniques is more effective than the one using the federated search techniques.

To summarise, we find that the record scoring approach is more effective for the TREC 2011 queries than for the TREC 2012 queries, which is in line with the performance of the aggregate scoring approach (see Section 7.4.3). However, the record scoring approach is effective for a narrower λ range. In addition, the obtained performance improvements are not statistically significant (paired t-test, $p > 0.05$). Meanwhile, we observe that both the aggregate scoring and the record scoring approaches are more effective when applied with the expCombMNZ voting technique than when applied with the expCombSUM voting technique. In addition, when comparing the two proposed approaches, we observe that the aggregate scoring approach is marginally more effective than the record scoring approach.

7.4.5 Learning to Leverage Department-Level Evidence

In this section, we investigate how to automatically set λ in Equations (7.4) and (7.5). Recall that λ is a parameter in the two proposed approaches (Equations (7.4) and (7.5)) that sets the level of emphasis on the department-level evidence when ranking patients. We argue that queries benefit from different levels of emphasis with respect to the department-level evidence. To effectively set the λ value, we use a regression technique to learn the effective λ values from the training queries using query features.

We first introduce our features used for choosing the effective parameter λ for an unseen query. We argue that if a query is difficult, it might be beneficial to take into account the department-level evidence. In particular, the department-level evidence is needed for the queries where the occurrences of query terms in the individual medical records are not enough for inferring relevance. In particular, we use the predicted difficulty of the query computed using 12 query performance predictors (Carmel & Yom-Tov, 2010; Cronen-Townsend *et al.*, 2002; Zhao *et al.*, 2008), which are also used and described in Section 6.5 of Chapter 6, as features (see Table 6.2).

In order to estimate an effective λ within our approaches for modelling the department-level evidence (i.e. Equations (7.4) and (7.5)), we identify the best λ that achieves the optimal retrieval performance in terms of a particular retrieval measure (i.e. bpref for TREC 2011 and infNDCG for TREC 2012) for each training query. Specifically, we sweep the λ parameter between 0 and 100 to find the best combination model in terms of the retrieval performance for each training query. Then, our learner uses the identified effective λ parameter to train a learned regression model from the retrieval performance prediction features.

7.4.5.1 Experimental Setup

To automatically learn the λ value for a given query using the technique described in Section 7.4.5, we use the Gradient Boosted Regression Trees (GBRT) learner, which is also used in Chapters 5 and 6. We use the root-mean-square error (RMSE) as the loss function when learning the λ value. We deploy 2 different learning regimes, namely 5-fold cross-validation (*5-fold*) and cross-collection validation (*x-collection*) (see Section 5.4.1), which are also used in the previous chapters.

We deploy the same experimental setup as discussed in Section 7.4.1. Due to its increased effectiveness as shown in Sections 7.4.3 and 7.4.4, we choose to conduct the experiments using the aggregate scoring approach (Section 7.2.2). In particular, we use the term-based representation approach (see Section 4.2.1), applying BM25 to rank medical records before using the expCombSUMw and expCombMNZw techniques to aggregate the relevance scores for patients. We extract the department-level evidence using a voting technique (see Section 7.3.1).

7.4.5.2 Experimental Results

Table 7.2 compares the aggregate scoring approach (denoted *expCombSUMw* or *expCombMNZw*), when using the learned approach presented in Section 7.4.5 to set the λ value, with the expCombSUM and the expCombMNZ baselines. Additionally, we also report the retrieval performance of the aggregate scoring approach when the λ parameter can be optimally set for every query (denoted *oracle*). From Table 7.2, we observe that the expCombSUMw technique using either the 5-fold or x-collection

Approach	bpref (TREC 2011)	infNDCG (TREC 2012)
expCombSUM	0.5018	**0.4343**
expCombMNZ	0.4725	0.4274
expCombSUMw (5-fold)	**0.5097**	0.4232
expCombSUMw (x-collection)	0.5032	0.4015
expCombMNZw (5-fold)	0.5037	0.4110
expCombMNZw (x-collection)	0.4984	0.3950
expCombSUMw (oracle)	$0.5385^{\oplus}$	$0.4583^{\oplus}$
expCombMNZw (oracle)	$0.5205^{\ominus}$	$0.4534^{\ominus}$

Table 7.2: Comparison of the retrieval performances of the aggregate scoring approach with the expCombSUM and expCombMNZ baselines on TREC 2011 and 2012 Medical Records track's queries. Statistical significance (paired t-test) at $p < 0.05$ over expCombSUM and expCombMNZ are denoted $^{\oplus}$ and $^{\ominus}$, respectively.

training regimes (bpref 0.5097 and 0.5032, respectively) outperforms both the expCombSUM and expCombMNZ baselines (bpref 0.5018 and 0.4725, respectively) for the TREC 2011 queries. Meanwhile, the expCombMNZw technique improves the retrieval performance over the corresponding expCombMNZ baseline. For example, using the 5-fold training regime, the expCombMNZw technique improves the retrieval performance over the expCombMNZ baseline by 6.6%. However, these performance improvements are not statistically significant. On the other hand, we observe that for the TREC 2012 queries, our approach could not outperform both baselines.

Comparing the retrieval performances of the expCombSUMw and expCombMNZw techniques, we observe that expCombSUMw achieves a better retrieval performance. However, for TREC 2011, compared to its corresponding baseline, expCombMNZw attains a better relative performance improvement than expCombSUMw. In particular, with the 5-fold training regime, expCombMNZw outperforms expCombMNZ by 6.6%, while expCombSUMw outperforms the expCombSUM baseline by 1.6%. In addition, we find that the 5-fold training regime is more effective than the x-collection training regime for both TREC 2011 and TREC 2012.

Next, we discuss the retrieval performance that our approach can achieve if the λ parameter can be set properly for every query (denoted *oracle*) from Table 7.2. For TREC 2011, we observe that the expCombSUMw technique (bpref 0.5385) significantly (paired t-test, $p < 0.05$) outperforms the expCombSUM baseline (bpref 0.5018). Meanwhile, the expCombMNZw technique also performs significantly (paired t-test, $p < 0.05$) better than the expCombMNZ baseline (bpref 0.5205 vs 0.4725). On the other hand, for TREC 2012, both the expCombSUMw (infNDCG 0.4583) and expCombMNZw (infNDCG 0.4534) techniques also significantly (paired t-test, $p < 0.05$) outperform the expCombSUM (infNDCG 0.4343) and expCombMNZ (infNDCG 0.4274) baselines, respectively. From these results, we conclude that there is scope to improve the learning of the λ parameter, as illustrated by the enhanced performance of the oracle. We leave the investigation of a more effective technique for automatically setting the λ parameter for future work.

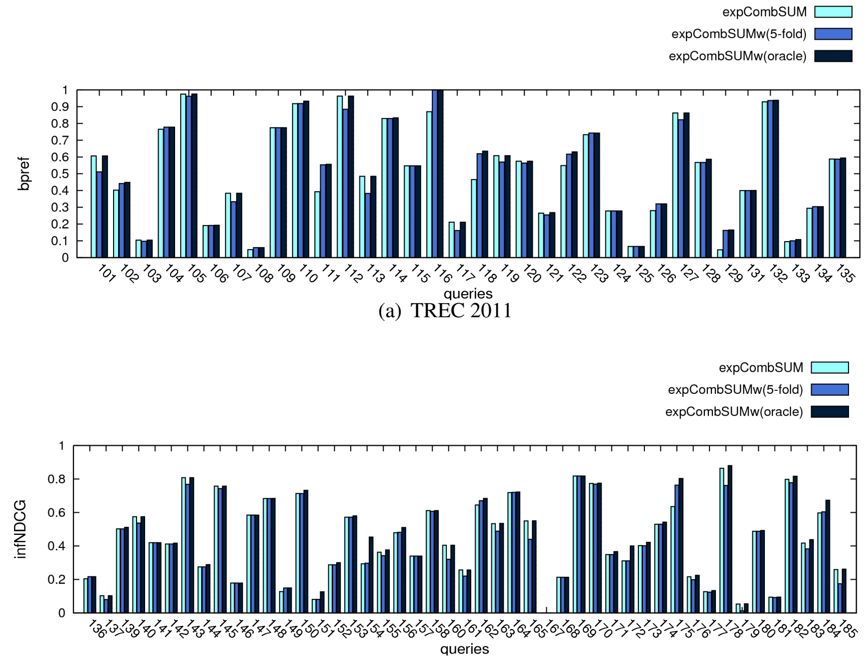
7.4.6 Per-Query Performance Analysis

In this section, we conduct a failure analysis to investigate the retrieval performance of the aggregate scoring approach that leverages department-level evidence to improve the performance of a patient search system. We analyse the retrieval performance of the aggregate scoring approach when the λ value within the aggregate scoring approach is learned using the x-collection training regime, as previously used in Section 7.4.5 (i.e. expCombSUMw (5-fold) in Table 7.2).

In Figure 7.11, we compare the retrieval performance on a per-query basis of *expCombSUMw (x-collection)* to the expCombSUM baseline. In addition, the best retrieval performance that our approach can obtain when the optimal λ value set for every query (*expCombSUMw (oracle)*) is also reported. In particular, Figures 7.11(a) and 7.11(b) show the retrieval performances in terms of bpref and infNDCG for the TREC 2011 and 2012 queries, respectively. For TREC 2011, we observe that *expCombSUMw (x-collection)* outperforms the expCombSUM baseline for 16 out of 34 queries, while it decreases the retrieval performance of 13 queries. Meanwhile, for TREC 2012, *expCombSUMw (x-collection)* improves the retrieval performance over the expCombSUM baseline for only 8 out of 47 queries. However, *expCombSUMw (x-collection)* performs worse than the expCombSUM baseline for 22 queries. In addition, we observe that the aggregate scoring approach is more sensitive to the used λ values for the TREC 2012 queries than for the TREC 2011 queries. In particular, we find that the mean of the best λ values for the TREC 2011 and TREC 2012 queries are 21.12 and 5.13, respectively. Meanwhile, the standard deviations are 29.94 and 15.09. On the other hand, we observe that the mean of the learned λ values are 20.56 and 5.12 for TREC 2011 and 2012, respectively (the standard deviations are 15.36 and 9.88, respectively). We can see that the effective range of λ values for TREC 2012 is narrower than TREC 2011 and the learned technique could set the λ values within that range. However, the performance achieved for TREC 2012 is overall less effective than TREC 2011.

7.5 Conclusions

We have introduced the third component of our framework (i.e. the Department Expertise component) that leverages the knowledge gained from aggregates of medical records associated with hospital departments (i.e. department-level evidence) to alleviate the issue of implicit knowledge in patient search. In particular, we proposed two approaches to leverage department-level evidence when ranking patients



(a) TREC 2011

(b) TREC 2012

Figure 7.11: Comparison of the retrieval performances of the aggregate scoring approach and the baseline, when applied with BM25 and the expCombSUM voting technique

based on the relevance of their medical records. The first approach (namely, the aggregate scoring approach) extends existing voting techniques to consider department-level evidence to better weight the importance of the individual medical records, based on the hospital departments that they are issued from, when ranking patients (Section 7.2.1). On the other hand, the second approach (namely, the record scoring approach) takes into account the department-level evidence when ranking the medical records before aggregating the relevance scores of the associated patients (Section 7.2.2). In addition, we proposed two techniques to build the department-level evidence from the medical records associated with particular hospital departments, based on a voting and a federated search paradigm, respectively (Sections 7.3.1 and 7.3.2). Our results show the potential of the aggregate scoring and the record scoring approaches. In particular, both approaches effectively leveraged the department-level evidence obtained using the voting techniques (Sections 7.4.3 and 7.4.4). In addition, these two approaches were effective when applied with both the term-based and the task-specific representation approaches. However, we found that both proposed approaches were more effective for the TREC 2011 queries than for the TREC 2012 queries. In addition, we found that the aggregate scoring approach is more effective than the record scoring approach. In Section 7.4.5, we introduced the technique for automatically setting the

λ parameter within both the aggregate scoring and the record scoring approaches. Our experimental results showed that the aggregate scoring approach improved the retrieval performance over a baseline that does not take into account the department-level evidence by up to 6.6% (see Table 7.2). We conduct a per-query performance analysis of the aggregate scoring approach in Section 7.4.6. We found that the aggregate scoring approach is more sensitive to the setting of the λ value for the TREC 2012 queries than for the TREC 2011 queries.

In the thesis statement (described in Section 1.3), we postulated that the problem introduced by the implicit knowledge could be handled at the inter-record level by focusing and reasoning on the medical conditions related to the medical decision process. From our experiments in this chapter, we conclude that the two proposed approaches are able to leverage the implicit knowledge at the inter-record level to improve the retrieval performance.

In the next chapter, we will introduce the final component of our framework, which models the relevance towards the multiple medical conditions stated in a given query, when ranking patients based on the relevance of their medical records.

Chapter 8

Modelling Relevance towards Multiple Inclusion Criteria

8.1 Introduction

In this chapter, we discuss the fourth component of our framework (namely, the Inclusion Criteria Coverage component), which uncovers implicit knowledge at the inter-record level. As discussed in Sections 3.4.2.1 and 4.2.3.4, existing search systems for retrieving patients may fail in identifying the cohorts whose medical histories cover all of the inclusion criteria specified in a query, which are often complex and include multiple medical conditions. For example, for the query “find patients who are suffering from lupus nephritis and thrombotic thrombocytopenic purpura”, healthcare practitioners implicitly recognise that the query aims to find patients with both ‘lupus nephritis’ and ‘thrombotic thrombocytopenic purpura’ (i.e. two inclusion criteria). Hence, healthcare practitioners search for patients who have both of the medical conditions, instead of those whose medical records have informative query terms matched as in a typical information retrieval (IR) system. Indeed, any patient exhibiting all of the inclusion criteria should naturally be ranked higher than a patient that only exhibits a subset, or none, of the criteria. However, none of the existing patient ranking approaches explicitly aims to rank patients based on the probability that they are relevant to all or most of the criteria (Edinger *et al.*, 2012) (see Section 3.4.2.1).

In this chapter, we propose a novel approach that extends the patient and the two-stage models for ranking patients (described in Section 4.2.2 of Chapter 4) to take into account the coverage of the inclusion criteria within the medical records of a patient by adapting a technique from recent research into coverage-based search result diversification. We propose to model the coverage of the inclusion criteria within the records of a particular patient, and thereby rank highly those patients whose medical records are likely to cover all of the criteria. In particular, our proposed approach estimates the relevance of a patient, based on the mixture probability of the relevance towards the query, and the likelihood that the patient's records cover the query criteria. The latter is measured using the relevance towards each of the criteria stated in the query, represented in the form of sub-queries.

The remainder of this chapter is structured as follows:

- Section 8.2 illustrates the problem that existing approaches could not effectively rank higher those patients who are relevant to more inclusion criteria stated in the query.
- Section 8.3 introduces our approach to model relevance towards multiple inclusion criteria for a particular patient by measuring the relevance towards each of the inclusion criteria using sub-queries. In addition, we discuss a technique for extracting the inclusion criteria from a given query using a well-established domain-specific resource.
- Section 8.4 discusses our experiments for evaluating the proposed approach for modelling relevance towards multiple inclusion criteria when ranking patients based on the relevance of their medical records, using the test collection previously used in Section 4.3.1.
- Section 8.5 provides concluding remarks on this chapter.

8.2 Motivation & Problem Definition

Recall from Section 4.2.2 that the *patient model* estimates the relevance of a patient p towards a query q, as follows:

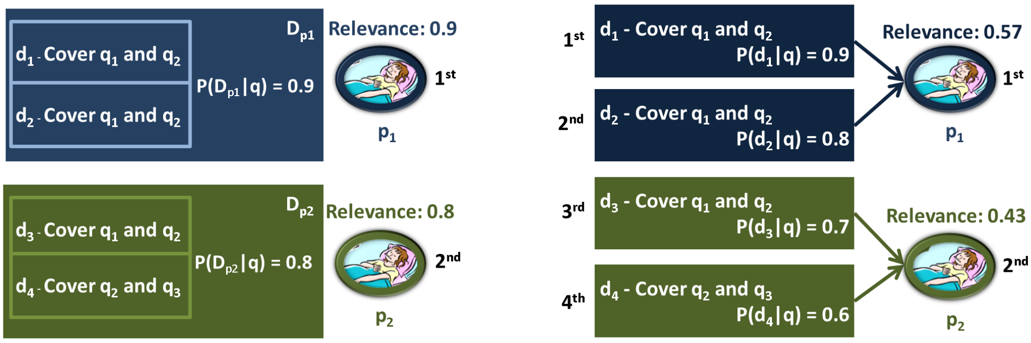
$$P(p|q) \propto P(D_p|q) \tag{8.1}$$

where the patient document D_p is created by concatenating the medical records associated to the patient p. $P(D_p|q)$, which is the probability that D_p is relevant to the query q, can be calculated using any probabilistic retrieval model, such as a language model. Note that Equation (8.1) is the probability estimate of Equation (4.1) of Chapter 4.

Alternatively, the *two-stage model* (see Section 4.2.2) estimates the relevance of a patient p, by suitably aggregating the relevance probabilities of the medical records associated to the patient p, as follows:

$$P(p|q) \propto aggregate_{d_i \in R_p}\Big[P(d_i|q)\Big] \tag{8.2}$$

where d_i is a medical record in R_p, which is the set of retrieved medical records that are also associated to the patient p. $P(d_i|q)$ is the probability that the medical record d_i is relevant to the query q (e.g.



(a) A patient model

(b) A two-stage model (CombSUM)

Figure 8.1: Illustrative examples of ranking patients using the two main existing approaches. Note that the relevance scores computed using CombSUM are normalised to maintain probability estimates.

estimated using a language model), while $aggregate_{d_i \in R_p}[\cdot]$ can be calculated using a voting technique (e.g. CombSUM – Equation (4.2) of Chapter 4).

As discussed in Section 8.1, both of the patient and two-stage models may fail in ranking the patients for a query searching for patients with multiple health conditions because they do not take into account the relevance towards each of the medical conditions stated in the query. We use Figure 8.1 to illustrate this problem. Assume that a query q is to find patients with 'heart disease' (i.e. criterion q_1), 'diabetes' (i.e. criterion q_2) and 'alzheimer's' (i.e. criterion q_3). Assume also that the medical records d_1 and d_2 are associated with the patient p_1, while the medical records d_3 and d_4 are related to the patient p_2. In Figure 8.1(a), the patient model (as in Equation (8.1)), which estimates the relevance of each patient using the concatenation of the medical records of that patient (e.g. D_{p1} and D_{p2}), ranks the patient p_1 higher than the patient p_2, according to their relevance probabilities, towards the query q (0.9 vs. 0.8). Meanwhile, in Figure 8.1(b), the two-stage model, which estimates the relevance of patients by suitably aggregating the relevance of their associated medical records (as in Equation (8.2)), also ranks the patient p_1 higher than the patient p_2. For instance, as described in Chapter 4, CombSUM estimates the relevance probability of a patient by summing up and normalising[1] the relevance probabilities of their associated medical records (e.g. after normalising, the relevance probabilities of patient p_1 and p_2 are 0.57 and 0.43, respectively). However, as previously discussed, an effective patient ranking approach should rank the patient p_2 higher than the patient p_1, as the patient p_2 is relevant to more criteria in the query q than the patient p_1 (q_1, q_2, q_3 vs. q_1, q_2).

[1] Note that there is a need to normalise the aggregated relevance probabilities to maintain probability estimates.

Denoting the probability that a patient p is relevant to a query q as $P(p|q)$, and the probability that p is relevant to the multiple inclusion criteria stated in the query q as $P_c(p|q)$, we argue that an effective patient ranking model, denoted by $F(p|q)$, must have the following two properties:

Property 1: If $P(p_1|q) = P(p_2|q)$, then patient p_1 should be ranked higher than patient p_2, $F(p_1|q) > F(p_2|q)$, when $P_c(p_1|q) > P_c(p_2|q)$.
Property 2: If $P(p_1|q) \neq P(p_2|q)$, then $F(p_1|q) > F(p_2|q)$ when $[P(p_1|q) \bigoplus P_c(p_1|q)] > [P(p_2|q) \bigoplus P_c(p_2|q)]$ where $\bigoplus$ is an appropriate mixture of the two probabilities.

In the next section, we discuss how to build an effective patient ranking model that satisfies the two above properties, which promotes patients who are relevant to more inclusion criteria, in a probabilistic manner.

8.3 An Approach for Modelling Relevance towards Multiple Inclusion Criteria

In this section, we propose to build a probabilistic model that satisfies the two properties previously discussed in Section 8.2. Specifically, the proposed approach models the mixture of the relevance towards the query and the likelihood of covering the inclusion criteria extracted from the query, to promote the patients whose medical records are relevant to a higher number of the inclusion criteria. Our proposed approach can be calculated, as follows:

$$F(p|q) \propto (1 - \lambda) \cdot P(p|q) + \lambda P_c(p|q) \tag{8.3}$$

where $P(p|q)$ is the probability that the patient p is relevant to the query q (i.e. *the relevance probability*), which can be measured using any existing patient ranking approach (e.g. Equations (8.1) or (8.2)); $P_c(p|q)$ is the probability that the medical records of the patient p cover the multiple inclusion criteria stated in the query q (we refer to this as the *coverage probability*), and λ ($0 \leq \lambda \geq 1$) is a mixture parameter to weight the importance of $P(p|q)$ and $P_c(p|q)$.

Given a set $Q = \{q_1, q_2, ..., q_n\}$ containing the inclusion criteria stated in the query q, we propose to estimate the *coverage probability* $P_c(p|q)$ as the combination of the beliefs that each criterion q_i in Q is covered by the medical records of the patient p, as follows:

$$P_c(p|q) = bel_{q_i \in Q}\left(P(p|q_i)\right) \tag{8.4}$$

where $P(p|q_i)$ is the probability that the patient p is relevant to the criterion q_i in Q. bel is a belief combination function, such as AND and OR, to combine the probabilities that the medical records of the patient p cover each inclusion criterion. In Section 7.3.2 of Chapter 7, we already used belief combination functions to calculate the relevance scores of hospital departments within a federated search technique. In this chapter, we use the belief combination functions in a different manner. Indeed, the belief combination functions have also been extensively deployed within search approaches (e.g. Metzler & Croft (2004); Ribeiro & Muntz (1996); Turtle & Croft (1991a)) to combine the probabilities that a particular document is relevant to each query term. For instance, bel^{AND} can be calculated as (Metzler & Croft, 2004):

$$bel^{AND}_{q_i \in Q} \left(P(p|q_i)\right) = \prod_{q_i \in Q} P(p|q_i) \tag{8.5}$$

In the remainder of this section, we discuss how the proposed approach can be applied within the existing patient and two-stage ranking models, respectively. Then, in Section 8.3.3, we describe our technique to extract the inclusion criteria from a query.

8.3.1 The Extended Patient Model

Within the patient model, the proposed approach can be adapted by inserting Equations (8.1) and (8.4) into Equation (8.3), as follows:

$$F(p|q) \propto (1-\lambda) \cdot P(D_p|q) + \lambda \cdot bel_{q_i \in Q} \left(P(D_p|q_i)\right) \tag{8.6}$$

where λ is a mixture parameter that weights the importance of the relevance and the coverage probabilities. Specifically, the first part of the equation, $(1-\lambda) \cdot P(D_p|q)$, focuses on the relevance probability of the patient document D_p towards the query q. The second part of the equation calculates the coverage probability of D_p. We use $P(D_p|q_i)$ to measure the probability that D_p covers a particular inclusion criterion q_i.

8.3.2 The Extended Two-Stage Model

As shown in Figure 8.1(b), the two-stage model (previously discussed in Section 4.2.2) firstly ranks the medical records, and then suitably aggregates their relevance probabilities to rank the associated patients. Hence, we can model the mixture of the relevance and the coverage probabilities either at the stage of ranking patients (Section 8.3.2.1), or at the stage of ranking medical records (Section 8.3.2.2).

8.3.2.1 Ranking Patients

First, we can model the mixture of the relevance and coverage probabilities at the patient ranking stage of a two-stage model by inserting Equations (8.2) and (8.4) into Equation (8.3), as follows:

$$\begin{aligned} F(p|q) \propto (1-\lambda) \cdot aggregate_{d_i \in R_p}\Big[P(d_i|q)\Big] \\ + \lambda \cdot bel_{q_i \in Q}\left(aggregate_{d_i \in R_p}\Big[P(d_i|q_i)\Big]\right) \end{aligned} \tag{8.7}$$

where the relevance towards query q and each of its criterion q_i is measured at the aggregated level of the medical records of particular patients.

8.3.2.2 Ranking Medical Records

In contrast, at the medical record ranking stage, the two-stage model considers each medical record individually, before suitably aggregating the relevance probabilities of the medical records to estimate the relevance of their associated patients. The existing aggregation techniques, such as the voting techniques used early in this thesis (e.g. Section 4.2.2), cannot take into account the coverage of the multiple inclusion criteria among the medical records of a particular patient. Indeed, without alteration, the medical record ranking stage of the two-stage model cannot examine the fact that a particular medical record may cover an inclusion criterion that the other medical records associated to the same patient do not cover. Thus, to highly rank the patients whose medical records cover the multiple inclusion criteria of the query, we need a mechanism to measure how well each of the inclusion criteria stated in the query is covered by different medical records of a particular patient. To achieve this, we introduce *the criterion novelty*, denoted by $\overline{P(R_p \setminus d_i|q_i)}$, which is the probability that a criterion q_i is not well covered by the other medical records that are also associated to the same patient. For instance, in the example of Figure 8.1(b), after considering the criterion novelty, the coverage probabilities of the medical records d_3 and d_4 are boosted, since both of them cover at least one new criterion that is not covered by the other medical records associated to the same patient. Consequently, the patient p_2, which is associated to the medical records d_3 and d_4, is likely to be ranked higher than the patient p_1. To integrate the criterion novelty at the medical record ranking stage of the two-stage model, we use Equations (8.3) and (8.4) to estimate the relevance and the coverage probabilities of the medical record d_i, before using Equation (8.2), as follows:

$$\begin{aligned} F(p|q) \propto aggregate_{d_i \in R_p}\Big[(1-\lambda) \cdot P(d_i|q) \\ + \lambda \cdot bel_{q_i \in Q}\Big(P(d_i|q_i) \cdot \overline{P(R_p \setminus d_i|q_i)}\Big)\Big] \end{aligned} \tag{8.8}$$

where $\overline{P(R_p \setminus d_i|q_i)}$ is the criterion novelty of q_i. To estimate the criterion novelty, we resort to techniques (e.g. Agrawal *et al.* (2009); Carbonell & Goldstein (1998)) from web search result diversification, which measure the novelty of a document within a set of web search results, based on the probability that the document covers an interpretation of the information need that is not well covered by the other documents in the result set. In this work, we adapt an existing state-of-the-art technique for search result diversification (namely, xQuAD (Santos *et al.*, 2010*a*)) to estimate the criterion novelty within the proposed approach, since it has been shown to be effective for diversifying web search results over several successive TREC tracks (e.g. Santos *et al.* (2010*a*,*b*)). However, other techniques that explicitly model the novelty of a document within a set of search results (e.g. IA-Select (Agrawal *et al.*, 2009)) can also be adapted.

Specifically, for a given query q and a patient p, we adapt xQuAD (Santos *et al.*, 2010*a*) to iteratively re-rank the medical records in R_p by maximising the mixture of the relevance and the coverage probabilities within Equation (8.8). By assuming the independence of the inclusion criteria within the query, $\overline{P(R_p \setminus d_i|q_i)}$ can be estimated as $\prod_{d_j \in {R_p}^*}(1 - P(d_j|q_i))$:

$$\begin{aligned} F(p|q) \propto aggregate_{d_i \in R_p} \Bigg[& (1-\lambda) \cdot P(d_i|q) \\ & + \lambda \cdot bel_{q_i \in Q} \bigg(P(d_i|q_i) \cdot \prod_{d_j \in {R_p}^*} (1 - P(d_j|q_i)) \bigg) \Bigg] \end{aligned} \tag{8.9}$$

where $\prod_{d_j \in {R_p}^*}(1 - P(d_j|q_i))$ estimates the probability that the criterion q_i is not well covered by any medical records in ${R_p}^*$, the set of medical records that are ranked higher than d_j and that are also associated to the patient p.

8.3.3 Inclusion Criteria Extraction

As discussed in Section 8.3, to measure the coverage probability, we need to extract the set of the inclusion criteria Q (e.g. in Equations (8.6), (8.7) and (8.9)) from a query q and use the extracted inclusion criteria as *sub-queries*. Importantly, the quality of the extracted sub-queries can affect the effectiveness of the proposed approach. For example, if the sub-queries are not a good representative of all of the inclusion criteria stated in the query, our approach may not be able to highly rank the patients whose medical records cover all or at least most of the inclusion criteria expected by the searchers. Several techniques (e.g. using domain knowledge, query log, or query suggestions (Agrawal *et al.*, 2009; Carbonell & Goldstein, 1998; Santos *et al.*, 2010*a*)) can be adapted to extract the inclusion criteria from a given query and to represent the extracted inclusion criteria as sub-queries. For example, when diversifying web search results, Santos *et al.* (2010*a*) used as sub-queries the recommended queries suggested by a commercial web search engine for the original query. However, we have shown in Section 4.3.4 of Chapter 4 that the medical conditions that healthcare practitioners focus on when searching the medical records often relate to four types of the medical conditions of patients, including symptom, diagnostic test, diagnosis and treatment. Consequently, we deploy MetaMap to extract the medical concepts related to these four types of medical conditions from the query (as described in Section 4.2.1 and Table 4.1). However, different from the approach in Section 4.2.1, we use the textual definitions of the extracted concepts, instead of the unique identifiers of the concepts, as sub-queries (e.g. using 'Thrombocytosis' instead of 'C0836924'). As shown in Figure 4.2 of Chapter 4, MetaMap can generate a number of candidate concepts when mapping a given phrase. We select only those that are defined as 'Meta Mapping' (i.e. the highest confidence) in order to avoid the redundancy of the extracted inclusion criteria. For example, in Figure 8.2, for the query "Patients with diabetes mellitus who also have thrombocytosis", MetaMap identifies four different concepts, three of which related to 'diabetes mellitus' and the other is related to 'thrombocytosis'. If we use all of the identified candidates, our approach may rank patients who are relevant to only 'diabetes mellitus' higher than those who are relevant to both 'diabetes mellitus' and 'thrombocytosis', as the proposed approach is modelled based on the relevance towards multiple identified concepts.

8.4 Experiments

This section examines the retrieval effectiveness of the proposed Inclusion Criteria Coverage (IC-Cover) approach to model the relevance towards multiple inclusion criteria when ranking patients based on the relevance of their medical records. In Section 8.4.1, we describe our experimental setup for evaluating the IC-Cover approach by comparing to the common baseline previously defined in Chapter 4. Section 8.4.2 lists the research questions to be investigated in this chapter. Sections 8.4.3 and 8.4.4 discuss the experimental results of the IC-Cover approach when uniformly setting the parameter λ and when deploying a learned approach to set λ on a per-query basis, respectively. Finally, we further analyse the retrieval performance of the IC-Cover approach in Section 8.4.5.

8.4.1 Experimental Setup

In this section, we discuss the experimental setup we use to evaluate the effectiveness of the proposed approach for modelling the coverage of the inclusion criteria.

As used in previous chapters and outlined in Section 4.3.1, we evaluate the IC-Cover approach using the TREC Medical Records track's test collection. Table 8.1 shows the statistics of the inclusion criteria extracted from the queries using the MetaMap-based technique discussed in Section 8.3.3. For example,

```
Input: "Patients with diabetes mellitus who also
        have thrombocytosis"

Phrase: "Patients"

Phrase: "with diabetes mellitus"
Meta Candidates (4):
  1000 C0011849:Diabetes Mellitus [Disease or Syndrome]
          Diabetes
   861 C0011847:Diabetes [Disease or Syndrome]
   789 C0241863:DIABETIC [Finding]
Meta Mapping (1000):
  1000 C0011849:Diabetes Mellitus [Disease or Syndrome]

Phrase: "who also"

Phrase: "have"

Phrase: "thrombocytosis."
Meta Candidates (1):
  1000 C0836924:Thrombocytosis [Disease or Syndrome]
Meta Mapping (1000):
  1000 C0836924:Thrombocytosis [Disease or Syndrome]
```

Figure 8.2: Medical concepts extracted by the MetaMap tool from query 102: ‘Patients with diabetes mellitus who also have thrombocytosis’, as the query inclusion criteria.

we find that on average we extract more inclusion criteria from the queries from TREC 2011 than those from TREC 2012 (i.e. on average, 4.32 vs. 3.36 inclusion criteria per query). The highest numbers of inclusion criteria extracted from a query are 13 (2 queries) and 17 (1 query), for TREC 2011 and TREC 2012, respectively.

We compare the performance of the IC-Cover approach to the common baseline identified in Chapter 4. In particular, we use the same setting identified in Section 4.3.1, and used in the subsequent chapters. Another possible baseline is to use a Boolean model to retrieve patients whose medical records contain all of the extracted inclusion criteria (i.e. the textual definitions of the medical concepts extracted

# of Criteria	TREC 2011	TREC 2012
Average	4.32	3.36
Standard deviation	2.90	3.06
Minimum	1	1
Maximum	13	17

Table 8.1: Statistics of the inclusion criteria extracted from the TREC queries

from the query, as described in Section 8.3.3); however, we find that the Boolean model is not effective, as it retrieves patients for only 6 out of the 34 queries of TREC 2011 and 16 out of the 47 queries of TREC 2012. Hence, we exclude it from our experiments. In addition, to investigate the impact of the use of negated language in the medical records and queries on the IC-Cover approach, we also compare the performances of the IC-Cover approach when the NegFlag approach (introduced Section 5.2.1) is either deployed or not deployed. Recall that the NegFlag approach tokenises term occurrences with a positive (e.g. patient has nausea) or negative context (e.g. patient has no nausea) differently, so that our search system can distinguish between terms with positive and negative contexts in both the medical records and the queries. This may help to more accurately measure whether a patient has the medical conditions stated in the query. Note that we have not investigated the impact of the negated language on the Conceptual Reasoning (Chapter 6) and the Department Expertise (Chapter 7) components. In particular, for the Conceptual Reasoning component, taking into account negated language can add more uncertainty when inferring the relationships of the medical conditions. Meanwhile, considering the use of negated language in the Department Expertise component may prevent the effective identification of the medical conditions that each hospital department has expertise in. Later in Chapter 9, we investigate approaches to combination all of the four components of the proposed framework.

We evaluate the IC-Cover approach within both the patient and the two-stage models (see Section 4.2.2) by deploying the same ranking techniques as those of the corresponding baseline. Indeed, we apply the IC-Cover approach within the patient model (denoted *IC-Cover-P*) as in Equation (8.6) using either BM25 or DFR DPH to estimate the relevance of a patient (as in Section 4.3.1 of Chapter 4). Next, for the two-stage model, we apply the proposed approach either within the patient ranking stage (denoted *IC-Cover-2P*), as in Equation (8.7), or within the medical record ranking stage (denoted *IC-Cover-2R*), as in Equation (8.9). Using the same settings as in Section 4.3.1, we deploy either BM25 or DFR DPH to estimate the relevance of the medical records, while using the CombSUM, expCombSUM, or expCombMNZ voting techniques to estimate the relevance of a particular patient.

Mixture Parameter Setting: Initially, we uniformly set the mixture parameter λ in Equations (8.6), (8.7) and (8.9) to 0.5 for every query. This gives an equal emphasis on both the relevance and the coverage probabilities of a particular patient. Later, in Section 8.4.4, we discuss and evaluate a technique to learn a suitable λ value for a given query using training data.

Belief Combination Function: We evaluate the IC-Cover approach using three different belief combination functions (namely, *AND*, *OR* and *SUM*), which have been shown to be effective for different search tasks (Metzler & Croft, 2004; Ribeiro & Muntz, 1996; Turtle & Croft, 1991*a*). Specifically, the belief combination functions AND, OR and SUM combine the probabilities as follows (Metzler & Croft, 2004; Turtle & Croft, 1991*a*):

$$bel^{AND}_{q_i \in Q}\left(P(p|q_i)\right) = \prod_{q_i \in Q} P(p|q_i) \tag{8.10}$$

$$bel^{OR}_{q_i \in Q}\left(P(p|q_i)\right) = 1 - \prod_{q_i \in Q} (1 - P(p|q_i)) \tag{8.11}$$

$$bel^{SUM}_{q_i \in Q}\left(P(p|q_i)\right) = \sum_{q_i \in Q} \frac{P(p|q_i)}{|Q|} \tag{8.12}$$

where $|Q|$ is the number of inclusion criteria in the set $Q = \{q_1, q_2, ..., q_n\}$.

8.4.2 Research Questions

Our aim is to investigate the effectiveness of the proposed approach to model the relevance towards multiple inclusion criteria when ranking patients based on their medical records. Specifically, in the remainder of this chapter, the following research questions are thoroughly investigated:

RQ 1. Does the IC-Cover approach to model the relevance of the patients' medical records to each of the inclusion criteria in the query improve retrieval performance?

RQ 2. What are the belief functions that could be effectively deployed in the IC-Cover approach in order to combine the relevance towards each of the inclusion criteria?

RQ 3. Is it more effective to also deal with negated language when using the IC-Cover approach?

RQ 4. How to effectively weight the importance of the relevance probability and the coverage likelihood of our approach?

RQ 5. What is the robustness with respect to the impact of the mixture parameter of the proposed IC-Cover approach?

RQ 6. What types of queries benefit from the proposed approach for modelling relevance towards the query inclusion criteria?

8.4.3 Experiments when Setting λ Uniformly

We first compare the retrieval performance of the IC-Cover approach, when the mixture parameter λ is uniformly set to 0.5 to balance the importance of the relevance and the coverage probabilities, with the common baseline ($\lambda = 0$). Tables 8.2 and 8.3 compare the retrieval effectiveness of the IC-Cover approach with the baseline, in terms of bpref for TREC 2011 and infNDCG & infAP for TREC 2012. In particular, in Tables 8.3 and 8.2, we apply and do not apply the NegFlag approach to accurately represent the contexts of the medical conditions, respectively. Moreover, the number of queries that the IC-Cover approach improves or harms compared to the corresponding baseline is also reported. The remainder of the queries are unaffected. For ease of notation, in both Tables 8.2 and 8.3, the used belief combination function along the IC-Cover approach is indicated between parentheses. For instance, for IC-Cover-P(AND), we apply IC-Cover within a patient model and use the belief combination function AND to combine the probabilities that the medical records of a patient cover the multiple inclusion criteria extracted from the query.

From Table 8.2, we observe that applying the IC-Cover approach within the medical record ranking stage of the two-stage model (i.e. IC-Cover-2R) is effective. For example, using the CombSUM voting technique with either DFR DPH or BM25, IC-Cover-2R(OR) and IC-Cover-2R(SUM), which use the belief combination functions OR and SUM, respectively, significantly (paired t-test, $p < 0.05$) outperform the CombSUM-based baseline, for every reported measure across both TREC 2011 and 2012. In addition, using the expCombSUM voting technique with either DFR DPH or BM25, IC-Cover-2R(OR) and IC-Cover-2R(SUM) improve the retrieval performances over the expCombSUM-based baseline by up to 4.09% in terms of bpref for TREC 2011. Indeed, using the expCombSUM voting technique and BM25, the performance improvement is statistically significant (paired t-test, $p < 0.05$). On the other hand, we find that the belief combination function AND is in general less effective than the belief combination functions SUM and OR (see Table 8.2). We also observe that the expCombMNZ voting technique is not effective when used in conjunction with our approach. This is partly due to the fact that the expCombMNZ voting technique also considers the number of retrieved medical records associated with a particular patient when estimating the relevance of that patient, which could outweigh the coverage probability within our approach.

When applied within the patient model (i.e. IC-Cover-P), IC-Cover improves the retrieval performance by up to 4.95% compared to the corresponding baseline, as shown in Table 8.2. For example, when using DPH, the retrieval performances of IC-Cover-P(SUM) are bpref 0.4767 and infNDCG 0.3761, while the retrieval performances of the DPH baseline are bpref 0.4542 and infNDCG 0.3734. In particular, IC-Cover-P(SUM) outperforms the DPH baseline for 20 out of 34 queries for TREC 2011.

When applied within the patient ranking stage of the two-stage model, we observe that our approach (i.e. IC-Cover-2P) is not effective. For example, when using the CombSUM voting technique to aggregate the relevance scores of medical records calculated using DFR DPH, our IC-Cover-2P(OR) performs only comparably to the CombSUM-based baseline (e.g. bpref 0.3633 vs. 0.3656) in Table 8.2. This is

Approach	2011 (34 queries)			2012 (47 queries)					
	bpref	△	▽	infNDCG	△	▽	infAP	△	▽
Patient Models									
DPH	0.4542			0.3734			0.1713		
+IC-Cover-P(AND)	0.4549	2	1	0.3727	4	4	0.1718	4	3
+IC-Cover-P(OR)	**0.4767**	20	9	0.3761	19	19	0.1704	19	19
+IC-Cover-P(SUM)	**0.4767**	20	9	0.3761	19	19	0.1704	19	19
BM25	0.4539			**0.3916**			**0.1855**		
+IC-Cover-P(AND)	0.4531	2	1	0.3890	4	6	0.1818	4	5
+IC-Cover-P(OR)	0.4733	19	10	0.3891	18	19	0.1737	17	18
+IC-Cover-P(SUM)	0.4733	19	10	0.3891	18	19	0.1737	17	18
Two-Stage Models (DFR DPH)									
CombSUM	0.3656			0.3367			0.1026		
+IC-Cover-2P(AND)	0.3667	2	0	0.3367	0	0	0.1026	0	0
+IC-Cover-2P(OR)	0.3633	14	15	0.3367	0	0	0.1026	0	0
+IC-Cover-2P(SUM)	0.3633	14	15	0.3367	0	0	0.1026	0	0
+IC-Cover-2R(AND)	0.3616	5	24	0.3421	17	27	0.1076	13	31
+IC-Cover-2R(OR)	0.3743$^{\blacktriangle}$	21	13	0.3515$^{\blacktriangle}$	31	12	0.1146$^{\blacktriangle}$	31	12
+IC-Cover-2R(SUM)	0.3743$^{\blacktriangle}$	21	13	0.3515$^{\blacktriangle}$	31	12	0.1146$^{\blacktriangle}$	31	12
expCombMNZ	0.4774			0.4254			0.1698		
+IC-Cover-2P(AND)	0.4764	4	1	0.4252	4	5	0.1692	5	4
+IC-Cover-2P(OR)	0.3410	9	22	0.3916	12	30	0.1458	14	27
+IC-Cover-2P(SUM)	0.3410	9	22	0.3915	12	30	0.1458	14	14
+IC-Cover-2R(AND)	0.4472	12	19	0.4116	20	25	0.1545	20	25
+IC-Cover-2R(OR)	0.4792	20	12	0.4267	22	23	0.1712	23	22
+IC-Cover-2R(SUM)	0.4792	20	12	0.4267	22	23	0.1712	23	22
expCombSUM	0.4871			0.4167			0.1703		
+IC-Cover-2P(AND)	0.4853	2	1	0.4179	5	4	0.1695	6	3
+IC-Cover-2P(OR)	0.3440	11	21	0.3962	16	22	0.1530	14	23
+IC-Cover-2P(SUM)	0.3441	11	21	0.3962	16	22	0.1530	14	23
+IC-Cover-2R(AND)	0.4821	18	14	0.4210	22	21	0.1667	23	21
+IC-Cover-2R(OR)	**0.5070**	21	12	**0.4257**	19	24	**0.1772**	22	22
+IC-Cover-2R(SUM)	**0.5070**	21	12	**0.4257**	19	24	**0.1772**	22	22
Two-Stage Models (BM25)									
CombSUM	0.3663			0.3338			0.1027		
+IC-Cover-2P(AND)	0.3672	2	1	0.3338	0	0	0.1027	0	0
+IC-Cover-2P(OR)	0.3640	15	16	0.3338	0	0	0.1027	0	0
+IC-Cover-2P(SUM)	0.3640	15	16	0.3338	0	0	0.1027	0	0
+IC-Cover-2R(AND)	0.3595	6	19	0.3323	16	28	0.1044	13	31
+IC-Cover-2R(OR)	0.3737	23	7	0.3459$^{\blacktriangle}$	30	15	0.1137$^{\blacktriangle}$	30	14
+IC-Cover-2R(SUM)	0.3737	23	7	0.3459$^{\blacktriangle}$	30	15	0.1137$^{\blacktriangle}$	30	14
expCombMNZ	0.4725			0.4274			0.1730		
+IC-Cover-2P(AND)	0.4728	2	2	0.4261	1	7	0.1730	3	5
+IC-Cover-2P(OR)	0.3387	7	25	0.4114	9	28	0.1607	11	26
+IC-Cover-2P(SUM)	0.3387	7	25	0.4114	9	28	0.1607	11	26
+IC-Cover-2R(AND)	0.4234	10	25	0.3941	14	28	0.1475	15	31
+IC-Cover-2R(OR)	0.4602	14	22	0.4186	20	22	0.1651	18	24
+IC-Cover-2R(SUM)	0.4602	14	22	0.4186	20	22	0.1651	18	24
expCombSUM	0.5018			**0.4343**			0.1828		
+IC-Cover-2P(AND)	0.5006	2	1	0.4341	1	2	0.1833	2	1
+IC-Cover-2P(OR)	0.3188	8	25	0.4276	11	17	0.1776	11	16
+IC-Cover-2P(SUM)	0.3188	8	25	0.4276	11	17	0.1776	11	16
+IC-Cover-2R(AND)	0.4806	17	14	0.4254	19	25	0.1746	25	22
+IC-Cover-2R(OR)	**0.5215**$^{\blacktriangle}$	23	9	0.4337	25	22	**0.1849**	23	22
+IC-Cover-2R(SUM)	**0.5215**$^{\blacktriangle}$	23	9	0.4337	25	22	**0.1849**	23	22

Table 8.2: Comparison of the retrieval performances of the IC-Cover approach (λ = 0.5) against the common baseline (λ = 0) on the TREC Medical Records Track 2011 and 2012. Statistical significance (paired t-test, $p < 0.05$) over the corresponding baseline is denoted $^{\blacktriangle}$. The column denoted $\triangle$ (resp. $\bigtriangledown$) shows the number of queries improved (resp. harmed) in relation to the corresponding baseline.

because when applied within the patient ranking stage of the two-stage model, the voting techniques used tend to give very high relevance estimates to the patients who are relevant to a single particular criterion. Hence, when using a belief function to combine the estimates that the medical records of the patients cover the multiple inclusion criteria, the highly ranked patients could be relevant to only one or a few criteria.

Next, in Table 8.3, we show the retrieval performance when applying the NegFlag approach (introduced in Section 5.2.1) with both the IC-Cover approach and the baseline. From Table 8.3, we observe that when using with the NegFlag approach, the retrieval performance of both the IC-Cover approach and the baseline further improve, except for the infAP performance of the IC-Cover approach and the baseline that use the CombSUM voting technique. In addition, we observe the same patterns of the performance improvement of the IC-Cover approach over the baseline as shown in Table 8.2 but with a higher magnitude.

In particular, we observe that applying the IC-Cover approach within the medical record ranking stage of the two-stage model (i.e. IC-Cover-2R) is effective. For example, using CombSUM with either DFR DPH or BM25, IC-Cover-2R(OR) and IC-Cover-2R(SUM), significantly (paired t-test, $p < 0.05$) outperform the CombSUM-based baseline, for both TREC 2011 and 2012. Meanwhile, using the expCombSUM voting technique with either DFR DPH or BM25, IC-Cover-2R(OR) and IC-Cover-2R(SUM) improve the retrieval performances over the expCombSUM-based baseline by up to 4.77% in terms of both bpref for TREC 2011 and infNDCG for TREC 2012. The performance improvement is statistically significant (paired t-test, $p < 0.05$) for TREC 2011.

Moreover, the IC-Cover approach, when applied within the patient model (i.e. IC-Cover-P), markedly improves the retrieval performance by up to 9.13% compared to the corresponding baseline (see Table 8.3). For example, when using BM25, the retrieval performances of IC-Cover-P(SUM) are bpref 0.5315, infNDCG 0.4286 and infAP 0.1959, while the retrieval performances of the BM25 baseline are bpref 0.4870, infNDCG 0.4080 and infAP 0.1922. Even though the performance improvements are not statistically significant, our approach improves the retrieval performances for about half of the queries. For example, when using BM25, IC-Cover-P(SUM) improves the retrieval performance over the BM25 baseline, in terms of bpref, for 19 out of 34 queries from TREC 2011, while for TREC 2012 it benefits 24 and 23 out of 47 queries, in terms of infNDCG and infAP, respectively.

These performance improvements are in line with the retrieval performance obtained when we do not apply the NegFlag approach; however, the magnitude of the performance improvement is higher. We also find that the belief combination function AND and the expCombMNZ voting technique could

Approach	2011 (34 queries)			2012 (47 queries)					
	bpref	△	▽	infNDCG	△	▽	infAP	△	▽
Patient Models									
DPH+NegFlag	0.4968			0.4392			0.1845		
+IC-Cover-P(AND)	0.4985	1	1	0.4381	4	5	0.1831	4	5
+IC-Cover-P(OR)	0.5165	19	13	**0.4507**	23	18	0.1909	22	19
+IC-Cover-P(SUM)	0.5165	19	13	**0.4507**	23	18	0.1909	22	19
BM25+NegFlag	0.4870			0.4080			0.1922		
+IC-Cover-P(AND)	0.4879	2	1	0.4092	5	4	0.1916	5	3
+IC-Cover-P(OR)	0.5227	19	13	0.4269	24	13	0.1958	23	14
+IC-Cover-P(SUM)	**0.5315**	19	13	0.4286	24	13	**0.1959**	23	14
Two-Stage Models (DFR DPH)									
CombSUM+NegFlag	0.3771			0.3304			0.0969		
+IC-Cover-2P(AND)	0.3769	2	1	0.3389	8	1	0.1008	8	1
+IC-Cover-2P(OR)	0.3734	13	17	0.3361	18	22	0.0983	16	23
+IC-Cover-2P(SUM)	0.3735	13	17	0.3361	18	22	0.0983	16	23
+IC-Cover-2R(AND)	0.3731	5	21	0.3342	16	28	0.1005	13	28
+IC-Cover-2R(OR)	0.3859$^{\blacktriangle}$	20	11	0.3496$^{\blacktriangle}$	28	16	0.1098$^{\blacktriangle}$	30	13
+IC-Cover-2R(SUM)	0.3859$^{\blacktriangle}$	20	11	0.3496$^{\blacktriangle}$	28	16	0.1098$^{\blacktriangle}$	30	13
expCombMNZ+NegFlag	0.5007			0.4506			0.1822		
+IC-Cover-2P(AND)	0.4989	3	2	0.4484	4	6	0.1775	5	5
+IC-Cover-2P(OR)	0.3582	9	23	0.3445	6	36	0.1218	10	32
+IC-Cover-2P(SUM)	0.3582	9	23	0.3445	6	36	0.1218	10	32
+IC-Cover-2R(AND)	0.4602	13	21	0.4264	17	28	0.1620	17	28
+IC-Cover-2R(OR)	0.5015	17	12	0.4469	21	23	0.1819	20	24
+IC-Cover-2R(SUM)	0.5015	17	12	0.4469	21	23	0.1819	20	24
expCombSUM+NegFlag	0.5055			0.4355			0.1833		
+IC-Cover-2P(AND)	0.5100	3	4	0.4382	12	6	0.1738	10	6
+IC-Cover-2P(OR)	0.3738	12	21	0.3215	11	33	0.1129	11	32
+IC-Cover-2P(SUM)	0.3738	12	21	0.3215	11	33	0.1129	11	33
+IC-Cover-2R(AND)	0.5080	18	13	0.4453	23	21	0.1785	22	22
+IC-Cover-2R(OR)	**0.5296**$^{\blacktriangle}$	21	9	**0.4515**	22	22	**0.1913**	23	21
+IC-Cover-2R(SUM)	**0.5296**$^{\blacktriangle}$	21	9	**0.4515**	22	22	**0.1913**	23	21
Two-Stage Models (BM25)									
CombSUM+NegFlag	0.3689			0.3278			0.0967		
+IC-Cover-2P(AND)	0.3689	1	1	0.3313$^{\blacktriangle}$	6	2	0.0986	7	2
+IC-Cover-2P(OR)	0.3670	14	16	0.3238	17	24	0.0957	16	25
+IC-Cover-2P(SUM)	0.3670	14	16	0.3238	17	24	0.0957	16	25
+IC-Cover-2R(AND)	0.3641	6	21	0.3279	15	29	0.0979	13	30
+IC-Cover-2R(OR)	0.3793$^{\blacktriangle}$	23	5	0.3407$^{\blacktriangle}$	23	21	0.1065$^{\blacktriangle}$	25	19
+IC-Cover-2R(SUM)	0.3793$^{\blacktriangle}$	23	5	0.3407$^{\blacktriangle}$	23	21	0.1065$^{\blacktriangle}$	25	19
expCombMNZ+NegFlag	0.4910			0.4410			0.1810		
+IC-Cover-2P(AND)	0.4900	2	2	0.4389	4	6	0.1774	5	5
+IC-Cover-2P(OR)	0.3503	9	22	0.3391	8	32	0.1185	7	33
+IC-Cover-2P(SUM)	0.3503	9	22	0.3391	8	32	0.1185	7	33
+IC-Cover-2R(AND)	0.4290	8	25	0.3970	14	31	0.1482	15	30
+IC-Cover-2R(OR)	0.4698	12	22	0.4240	20	25	0.1690	18	27
+IC-Cover-2R(SUM)	0.4698	12	22	0.4240	20	25	0.1689	18	27
expCombSUM+NegFlag	0.5180			0.4532			**0.1970**		
+IC-Cover-2P(AND)	0.5165	3	2	0.4429	4	7	0.1855	4	7
+IC-Cover-2P(OR)	0.3345	8	25	0.3100	9	33	0.1082	9	33
+IC-Cover-2P(SUM)	0.3345	8	25	0.3099	9	33	0.1082	9	33
+IC-Cover-2R(AND)	0.4973	14	16	0.4401	19	26	0.1823	21	24
+IC-Cover-2R(OR)	**0.5371**$^{\blacktriangle}$	20	11	**0.4543**	25	21	0.1959	21	24
+IC-Cover-2R(SUM)	**0.5371**$^{\blacktriangle}$	20	11	**0.4543**	25	21	0.1959	21	24

Table 8.3: Comparison of the retrieval performances of the IC-Cover approach ($\lambda = 0.5$) against the common baseline ($\lambda = 0$) on the TREC Medical Records Track 2011 and 2012, when applied with the NegFlag approach. Statistical significance (paired t-test, $p < 0.05$) over the corresponding baseline is denoted $^{\blacktriangle}$. The column denoted △ (resp. ▽) shows the number of queries improved (resp. harmed) in relation to the corresponding baseline.

not be effectively deployed with the IC-Cover approach. Meanwhile, the IC-Cover approach (i.e. IC-Cover-2P) is not effective when applied within the patient ranking stage of the two-stage model.

Overall, we conclude that the IC-Cover approach to model the relevance towards multiple inclusion criteria applied either within the patient model or within the medical record ranking stage of the two-stage model is effective for the patient ranking task, answering the first research question (Section 8.4.2). In addition, in answer to the second research question, we find that the belief combination functions SUM and OR are effective for combining the probabilities that the medical records of a patient cover each of the inclusion criteria stated in the query. Moreover, applying the NegFlag approach with the IC-Cover approach further improves the retrieval performance markedly, answering third research question.

8.4.4 Experiments with the Learned λ Values

This section discusses an automatic technique to set the mixture parameter λ in Equations (8.6), (8.7) or (8.8). We argue that different queries might benefit from a particular level of emphasis on the importance of the relevance and the coverage probabilities. For example, a query that contains several inclusion criteria may benefit from more emphasis on the coverage probability. Within the xQuAD framework, Santos *et al.* (2010*b*) also suggested to selectively set the level of diversification based on the ambiguity of the query.

In this work, we deploy the Gradient Boosted Regression Trees (GBRT) technique (Tyree *et al.*, 2011), which was previously used in previous chapters (e.g. Section 5.3 of Chapter 5), to learn the parameter λ from a set of training queries. However, any regression technique can be deployed. To train a regression model, the root-mean-square error (RMSE) is used as a loss function when learning λ.

8.4.4.1 Estimating an Effective Mixture Parameter

To estimate an effective mixture parameter λ, we identify the λ that attains the best retrieval performance in terms of a particular retrieval measure (e.g. bpref or infNDCG) for each training query. In particular, for a given query, we sweep the λ values between 0 and 1 (with an interval of 0.1) to find the best setting of λ. Then, the set of the identified λ values from the training queries are used as the labelled data to train the regression model for choosing a λ value for an unseen query. We learn the λ values using the *5-fold* cross-validation training regime as previously defined and used in Section 5.4.1 of Chapter 5.

8.4.4.2 Learning Features

Next, we define the features used for choosing the effective parameter λ for an unseen query. An effective feature should indicate the level of emphasis on the relevance and the coverage probabilities

Query Performance Predictors	
Clarity Score (Cronen-Townsend *et al.*, 2002)	AvIDF (Carmel & Yom-Tov, 2010)
Query Scope (He & Ounis, 2006)	EnIDF (Carmel & Yom-Tov, 2010)
AvICTF (Carmel & Yom-Tov, 2010)	γ_1 (He & Ounis, 2006)
AvPMI (Carmel & Yom-Tov, 2010)	γ_2 (He & Ounis, 2006)
MAXCQ (Zhao *et al.*, 2008)	SCQ (Zhao *et al.*, 2008)
NSCQ (Zhao *et al.*, 2008)	# of medical concepts
Inclusion Criteria Similarity	
WUPALMER (Garla & Brandt, 2012)	PATH (Garla & Brandt, 2012)
INTRINSIC_PATH (Garla & Brandt, 2012)	RADA (Garla & Brandt, 2012)
INTRINSIC_LIN (Garla & Brandt, 2012)	LCH (Garla & Brandt, 2012)
INTRINSIC_LCH (Garla & Brandt, 2012)	SOKAL (Garla & Brandt, 2012)
INTRINSIC_RADA (Garla & Brandt, 2012)	LIN (Garla & Brandt, 2012)
JACCARD (Garla & Brandt, 2012)	

Table 8.4: List of the used features.

for each query. In this work, we use 23 features, which measure the predicted difficulty of the query. A query with multiple inclusion criteria tends to be complex and long; hence, it can be assumed to be difficult. If the query is difficult, then it might be beneficial to focus on the coverage probability. Table 8.4.4.2 lists the two groups of features used in this experiments. The first group of features are the 12 query performance predictors computed on the original query. 11 of these features were previously used in Section 6.5 of Chapter 6, while the feature '# of medical concepts' is the number of inclusion criteria that we extract from the query using the MetaMap-based technique discussed in Section 8.3.3. The query is likely to be difficult if it contains several inclusion criteria. The second group of features are the 11 semantic similarities (Garla & Brandt, 2012), which can estimate the similarity between the inclusion criteria extracted from the original query. The more dissimilar the inclusion criteria in the query, the more difficult the query is likely to be, since it may be difficult to find a patient with such unrelated conditions (i.e. inclusion criteria). We use YTEX[1] to calculate 11 recent semantic similarity measures (Garla & Brandt, 2012) and average the similarity scores among every pair of the medical concepts (i.e. the extracted inclusion criteria, using the approach described in Section 8.3.3).

8.4.4.3 Experimental Results

In this section, we compare the retrieval performance of the IC-Cover approach when using the cross-validation setting to learn the mixture parameter λ (i.e. 5-fold) with when the λ is set uniformly, including λ = 0 (i.e. the focus is only on the relevance probability towards the query), λ = 1 (i.e. the focus is only on the coverage probability) and λ = 0.5 (i.e. equally weight the importance of both the relevance and the coverage probabilities). In addition, the best possible retrieval performance, when the mixture

[1] `http://code.google.com/p/ytex/wiki/SemanticSim_V06`

Approach	TREC 2011 (34 queries)			TREC 2012 (47 queries)					
	bpref	△	▽	infNDCG	△	▽	infAP	△	▽
IC-Cover-P(SUM)									
+$\lambda = 0$	0.4539			**0.3961**			**0.1855**		
+$\lambda = 1$	0.4607	18	12	0.3329	12	27	0.1393	14	25
+$\lambda = 0.5$	**0.4733**	19	10	0.3891	18	19	0.1737	17	18
+5-fold	0.4674	18	10	0.3952	17	12	0.1830	16	12
+oracle	$0.5128^{\oplus,\ominus,\odot,\otimes}$	23	0	$0.4251^{\oplus,\ominus,\odot,\otimes}$	25	0	$0.2005^{\oplus,\ominus,\odot,\otimes}$	25	0

Table 8.5: Comparison of the retrieval performances using various λ values within the proposed IC-Cover approach. Statistical significances (paired t-test, $p < 0.05$) over the settings when $\lambda = 0$, $\lambda = 1$, $\lambda = 0.5$, and when using a learned technique (5-fold) are denoted $^{\oplus}$, $^{\ominus}$, $^{\odot}$ and $^{\otimes}$, respectively. The column denoted △ (resp. ▽) shows the number of queries improved (resp. harmed) in relation to the baseline where $\lambda = 0$.

Approach	TREC 2011 (34 queries)			TREC 2012 (47 queries)					
	bpref	△	▽	infNDCG	△	▽	infAP	△	▽
IC-Cover-P(SUM)+NegFlag									
+$\lambda = 0$	0.4870			0.4080			0.1922		
+$\lambda = 1$	0.5098	17	15	0.3764	20	21	0.1585	15	24
+$\lambda = 0.5$	0.5315	19	13	$0.4286^{\ominus}$	24	13	$0.1959^{\ominus}$	23	14
+5-fold	**0.5346**	18	11	**0.4384**$^{\oplus,\ominus}$	24	11	**0.2066**$^{\ominus}$	23	14
+oracle	$0.5872^{\oplus,\ominus,\odot,\otimes}$	24	0	$0.4637^{\oplus,\ominus,\odot,\otimes}$	30	0	$0.2208^{\oplus,\ominus,\odot,\otimes}$	29	0

Table 8.6: Comparison of the retrieval performances using various λ values within the proposed IC-Cover approach, when applied with the NegFlag approach. Statistical significances (paired t-test, $p < 0.05$) over the settings when $\lambda = 0$, $\lambda = 1$, $\lambda = 0.5$, and when using a learned technique (5-fold) are denoted $^{\oplus}$, $^{\ominus}$, $^{\odot}$ and $^{\otimes}$, respectively. The column denoted △ (resp. ▽) shows the number of queries improved (resp. harmed) in relation to the baseline where $\lambda = 0$.

parameter λ is optimally set for each query (i.e. an oracle), is also reported. We experiment only with the IC-Cover approach when applied within the patient model (i.e. BM25) and using the belief combination function SUM, since it is effective when applied with or without the NegFlag approach, as shown in Tables 8.3 and 8.3.

We first compare the retrieval performance of the IC-Cover approach when not using the NegFlag approach in Table 8.5. We observe that, for TREC 2011, with the 5-fold cross-validation training regime (5-fold), the IC-Cover approach improves the retrieval performance over the baseline where $\lambda = 0$ by 2.98%. In particular, using the 5-fold training regime, the IC-Cover approach outperforms the baseline where $\lambda = 0$ for 18 out of 34 queries, while decreasing the retrieval performance for 10 out of 34 queries. However, we find that it is less effective than the baseline where $\lambda = 0.5$. On the other hand, for TREC 2012, even though the IC-Cover approach with the 5-fold setting outperforms the baseline where $\lambda = 0$ for many queries (i.e. 17 and 16 out of 47, in terms of infNDCG and infAP, respectively), on average it could not outperform this baseline.

Next, we discuss the retrieval performance of the IC-Cover approach when applied with the NegFlag approach. From Table 8.6, we observe that the cross-validation setting (5-fold) is effective, as it outperforms all of the uniform setting baselines (i.e. when λ is set to 0, 1, or 0.5). In particular, for TREC 2011, our 5-fold cross-validation setting improves the retrieval performance over the baseline where λ = 0 by up to 9.8% (bpref 0.5346 vs. 0.4870). Indeed, it improves the retrieval performance for the majority of the queries (i.e. 18 of 34 queries). For TREC 2012, in terms of infNDCG, the 5-fold cross-validation setting significantly outperforms the settings where λ = 0 and λ = 1 (paired t-test, $p < 0.05$). The performance improvements are up to 7.5% and 16.5%, respectively. Specifically, the 5-fold cross-validation setting outperforms the baseline that does not take into account the coverage probability (i.e. λ = 0) for 24 out of 47 queries. Meanwhile, in terms of infAP, the 5-fold cross-validation setting (infAP 0.2066) significantly outperforms the baseline where λ = 1 (infAP 0.1585). However, even though the 5-fold cross-validation setting outperforms the λ = 0.5 setting for all of the reported measures, the improvements are not statistically significant. We find that λ = 0.5 is an effective baseline, since it can improve the retrieval performance for most of the queries improved by the oracle. For instance, for TREC 2012, $\lambda = 0.5$ improves the retrieval performance, in terms of infNDCG, over the baseline where $\lambda = 0$ for 24 out of 47 queries, while the oracle setting improves the retrieval performance for 30 out of 47 queries. In answer to the fourth research question, either fixing the λ value to 0.5 or learning the λ value on a per-query basis by using query performance predictors as learning features is effective.

Next, we discuss the retrieval performances assuming we can effectively set the λ parameter for each query (i.e. oracle). We observe that with the oracle setting, our approach further improves the retrieval performances markedly. It significantly (paired t-test, $p < 0.05$) outperforms all other settings in Table 8.6. In addition, we find that it improves the retrieval performances over the corresponding baseline where the coverage probability is not taken into account (i.e. λ = 0) for the majority of the queries (i.e. 24 out of 34 queries for TREC 2011, 30 out of 47 queries for infNDCG TREC 2012, and 29 out of 47 queries for infAP TREC 2012). We find that with the oracle setting could not improve the retrieval performance of all of the queries because the MetaMap tool could not effectively extract inclusion criteria from some queries, as will be discussed later in Section 8.4.5.2. We leave for future works the investigation of more effective inclusion criteria extraction techniques.

8.4.5 Analysis and Discussion

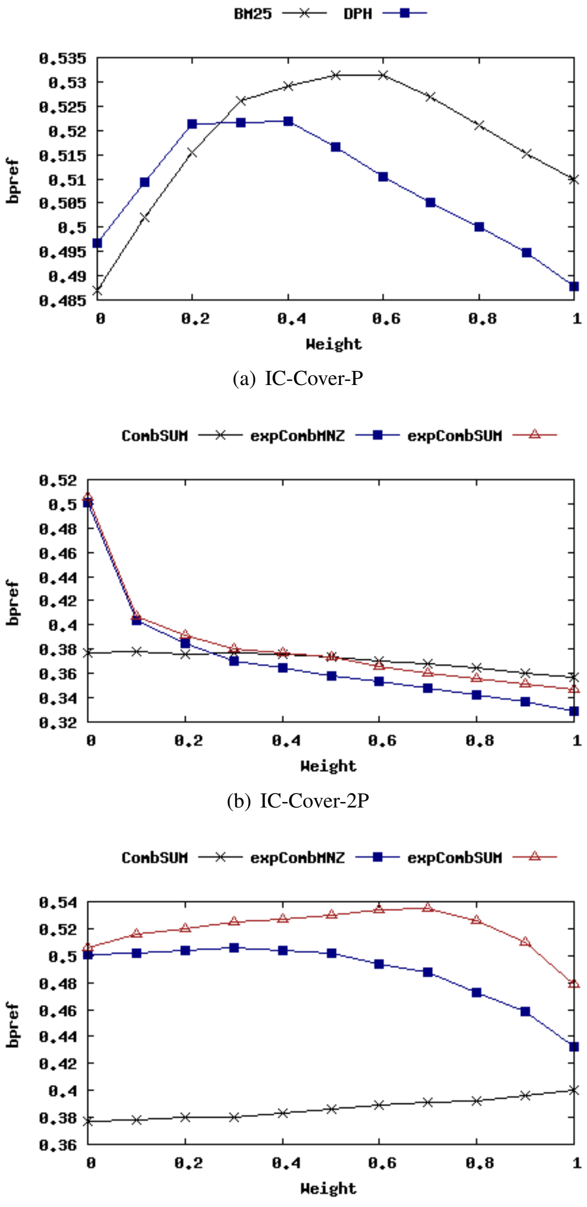
In the previous section, we show that the IC-Cover approach could effectively measure the relevance towards multiple inclusion criteria stated in the query. In addition, we show that applied with the NegFlag approach, the retrieval performance of the IC-Cover approach further improves, as the NegFlag approach helps to differentiate contexts of medical conditions when measuring the relevance towards the medical conditions. This section further discusses the retrieval performance of the IC-Cover approach to model the relevance towards multiple inclusion criteria in patient search when applied with the NegFlag approach. Specifically, Section 8.4.5.1 discusses the impact of the choice of the mixture parameter λ on the robustness of the IC-Cover approach. Section 8.4.5.2 analyses when the IC-Cover approach performs particularly effectively.

8.4.5.1 Model Robustness

This section investigates the robustness of the IC-Cover approach by varying the mixture parameter λ that weights the importance of the relevance and the coverage probabilities. To analyse the impact of the parameter λ within our approach, we experiment with setting λ within a range of values between 0 and 1, with an interval of 0.1. When λ = 0, the IC-Cover approach considers only the relevance probability towards the query, while when λ = 1 our approach takes into account only the coverage probability. In Figure 8.3, we report only the experiment on TREC 2011. We find that the experimental results on TREC 2012 follow the same pattern. In addition, for readability purposes, while we only show the retrieval performances of our approach using the most effective belief combination function (namely SUM as shown in Table 8.3), the results with the belief combination function OR are consistently similar, although slightly less effective in magnitude. For IC-Cover-P, we use BM25 and DPH. Meanwhile, for both IC-Cover-2P and IC-Cover-2R, we use BM25 to rank medical records before aggregating the relevance scores of the patients using the CombSUM, expCombMNZ or expCombSUM voting techniques.

From Figure 8.3(a), we observe that with BM25 or DPH, IC-Cover-P, which is applied within the patient model, performs better than the baseline (i.e. λ = 0), when $0.1 \leq \lambda \leq 0.8$. Hence, IC-Cover-P is robust, as it is more effective than the baseline for a very wide range of λ values (in fact for $\lambda > 0$ when using BM25). In addition, the most effective performance is achieved when λ is set to 0.2 and 0.6 when using BM25 and DPH, respectively. Next, Figure 8.3(b) shows the retrieval performances of our approach when applied within the patient ranking stage of the two-stage model (i.e. IC-Cover-2P). We observe that IC-Cover-2P is not effective, which is in line with the observations in Section 8.4.3. Meanwhile, in Figure 8.3(c), when applied within the medical record ranking stage (i.e. IC-Cover-2R), the proposed approach outperforms the baseline (λ = 0) for a wide range of λ values. For CombSUM, when $0 < \lambda \leq 1$, our approach outperforms the baseline, especially for λ values closer to 1. For expCombSUM, our approach performs better than the baseline when $0.1 \leq \lambda \leq 0.9$, while the most effective performance is obtained when $\lambda = 0.7$. Nevertheless, Figure 8.3(c) shows that applying expCombMNZ in conjunction with our approach is not effective, which supports the finding discussed in Section 8.4.3.

To summarise, we find that the IC-Cover approach is effective and robust when applied within the patient model or when applied within the medical record ranking stage of the two-stage model. Specif-



(a) IC-Cover-P

(b) IC-Cover-2P

(c) IC-Cover-2R

Figure 8.3: The retrieval performances of our proposed approach when applied within the patient model (IC-Cover-P), within the patient ranking stage (IC-Cover-2P) and within the medical record ranking stage (IC-Cover-2R) of the two-stage model, in terms of bpref for TREC 2011, as we vary the mixture parameter λ between 0 and 1.

ically, as shown in Figure 8.3, our approach is effective when setting λ uniformly for a wide range of values (particularly $0.2 \leq \lambda \leq 0.7$). Therefore, in answer to the fifth research question, these experimental results show that the proposed IC-Cover approach is robust as it is effective when the λ parameter is uniformly set within a wide range of values.

8.4.5.2 Failure Analysis

In this section, we analyse the retrieval performance of our approach for each query, when using the cross-validation setting discussed in Section 8.4.4. In particular, we discuss the performance of IC-Cover-P(SUM) when applied with BM25 and the NegFlag approach. Table 8.7 shows, for various numbers of extracted inclusion criteria, the numbers and percentages of the queries impacted (either positively or negatively) by the proposed approach. The impacted performances are measured based on bpref and infNDCG for TREC 2011 and TREC 2012, respectively. We divide the 81 queries from TREC 2011 (34 queries) and 2012 (47 queries) into 4 groups, according to the number of the extracted inclusion criteria in the queries, and report the percentages of queries for each group. The sizes of the groups vary from 14 to 30 queries (average of 20.25). From the cross-validation setting, we observe that our approach is most likely to benefit (63%) the queries with 3 inclusion criteria, followed by the queries having a number of inclusion criteria between 4 and 17 (57%). This is in line with the oracle setting where the queries for which the number of extracted criteria is at least 3 are most likely (more than 70%) to benefit. On the other hand, for the queries with 1 or 2 inclusion criteria, our approach is less likely to be beneficial (e.g. 43% and 39% for the queries with 1 and 2 inclusion criteria, respectively), which is intuitive. Indeed, as our approach aims to promote the relevance towards multiple inclusion criteria, queries that contain only very few criteria may not benefit much from our approach (e.g. queries 109, 143, 147 and 154). However, we also observe improvements even for queries with one inclusion criterion, since our approach enables the retrieval system to focus on the inclusion criterion instead of the non-important terms in the queries. In contrast, the proposed approach tends to be effective for long and complex queries (i.e. the queries with at least 3 inclusion criteria), such as queries 111, 113, 121 and 176. For example, for query# 121:‘patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and were given Plavix’, IC-Cover-P can improve the retrieval performance to bpref 0.4088, while the performance of the patient model baseline is bpref 0.1869. This is because the patient model, which uses a retrieval model to estimate the relevance probability of the patients, tends to focus on the occurrences of informative terms (e.g. Plavix) within the medical records. However, the relevant patients are also required to be relevant to the other conditions, indicated by theoccurrences of other query terms, such as CAD, acute, coronary and syndrome. Meanwhile, the proposed approach

# of Criteria	Cross-validation			Oracle[1]	
	#Benefiting	#Harmed	#Stable	#Benefiting	#Stable
1 (14 queries)	6 (43%)	3 (21%)	5 (36%)	8 (57%)	6 (43%)
2 (18 queries)	7 (39%)	5 (28%)	6 (33%)	10 (56%)	8 (44%)
3 (19 queries)	12 (63%)	6 (32%)	1 (5%)	15 (79%)	4 (21%)
4-17 (30 queries)	17 (57%)	12 (40%)	1 (3%)	21 (70%)	9 (30%)

Table 8.7: Analysis of IC-Cover-P(SUM) when applied with BM25 and the NegFlag approach w.r.t. the number of inclusion criteria extracted from the queries. The numbers between the parentheses indicate the percentage compared to the total number of queries.

is effective since it promotes the patients who are relevant to multiple inclusion criteria (e.g. ‘CAD’, ‘Acute Coronary Syndrome’ and ‘Plavix’). In answer to our sixth research question, our approach is likely to benefit complex queries that contain several (i.e. at least 3) inclusion criteria.

On the other hand, when looking at the harmed queries, we observe that the queries with 3 and 4-17 inclusion criteria are also more likely to be harmed by the proposed approach than the queries with 1 or 2 inclusion criteria. Due to the limited number of queries (i.e. we use a 5-fold cross-validation on two separate sets of only 34 and 47 queries when learning the regression model), we find that the learned model could not always generalise, as it tends to favour the coverage probability for the queries with several inclusion criteria, while focusing on the relevance probability for the queries with a very few inclusion criteria.

Moreover, when investigating the effectiveness of each query, we find that some queries do not benefit from our approach, because the used MetaMap tool could not effectively extract the inclusion criteria (e.g. in the cases of queries 104, 107, 146 and 149). For example, instead of only 2, 13 inclusion criteria are incorrectly extracted from the query 104:‘patients diagnosed with localized prostate cancer and treated with robotic surgery’. Consequently, our approach could not effectively measure the coverage of the expected inclusion criteria from the query. We leave for future work the investigation of a more effective inclusion criteria extraction technique.

8.5 Conclusions

We have discussed the final component (i.e. the Inclusion Criteria Coverage component) of our framework that uncovers implicit knowledge and promotes patients whose medical records are relevant to several medical conditions associated to symptom, diagnostic test, diagnosis and treatment (see Section 4.2.3). We proposed to rank highly patients whose medical records cover the multiple inclusion

[1] Note that there is no harmed query for the oracle setting because the λ is set to 0, which is the baseline, if the coverage probability could not improve the retrieval performance.

criteria stated in the query. In Section 8.3, we introduced our approach for modelling the mixture of the relevance probability towards the query and the probability that the medical records of a patient are relevant to the multiple inclusion criteria occurring in the query (i.e. the coverage probability) as measured in Equation (8.3)). The coverage probability is defined in terms of the relevance probabilities towards each of the inclusion criteria (Sections 8.3.1 and 8.3.2). The inclusion criteria are the medical concepts extracted from the query using an existing medical resource (Section 8.3.3). Then, in Section 8.4, we evaluated the proposed IC-Cover approach by comparing to the common baseline defined in Chapter 4. We show that the IC-Cover approach is particular effective when applied with the NegFlag approach (introduced in Section 5.2.1 of Chapter 5), as the NegFlag approach enables the IC-Cover approach to measure the relevance towards each of the inclusion criteria more accurately. In particular, when applied within the patient model (IC-Cover-P) or at the medical record ranking stage of the two-stage model (IC-Cover-2R), the approach significantly (paired t-test, $p < 0.05$) improved the retrieval performance over both the patient and the two-stage model baselines (Tables 8.3 and 8.6). Moreover, we showed that the proposed approach was effective, either when weighting equally the importance of the relevance and the coverage probabilities or when deploying a regression technique to learn an effective setting of the mixture parameter (Tables 8.3 and 8.6 in Sections 8.4.3 and 8.4.4, respectively).

When analysing the retrieval performance in Section 8.4.5.1, we observed that the proposed approach was robust, as it improved the retrieval performances for a wide range of the mixture parameter's values, especially when the λ parameter is set between 0.2 and 0.7 (see Figure 8.3). In addition, when analysing the retrieval performances for each query, we found that the approach tended to particularly improve the retrieval performances for queries from which at least 3 inclusion criteria could be extracted (see Table 8.7 of Section 8.4.5.2).

These experimental results support the thesis statement described in Section 1.3, in that we can effectively uncover the implicit knowledge that healthcare practitioners use to find patients having several query medical conditions by recognising the medical conditions stated in the query (i.e. inclusion criteria) and modelling the relevance with respect to each of the conditions among different medical records of the patients.

In the next chapter, we investigate how the four components of the proposed framework that tackle the implicit knowledge at the different levels of the retrieval process (described in Chapters 5, 6, 7 and 8), can be combined to further improve the retrieval performance of a patient search system.

Chapter 9

Combination of Approaches in Patient Search

9.1 Introduction

Recall that in Chapter 4 we introduced our patient search system (see Figure 4.1) that consists of the query and medical record representation unit (Section 4.2.1), the patient ranking unit (Section 4.2.2), and our proposed framework for enhancing the representations of the queries and medical records (Section 4.2.3). In the previous four chapters, we have discussed each of the four components of this framework, which uncovers implicit knowledge at different levels of the retrieval process. In this chapter, we first investigate approaches to combine the relevance scores of patients computed using the four different components of our framework to generate a more effective ranking of patients. Indeed, we investigate how to effectively combine rankings of patients using techniques from three main research areas for combining search results in information retrieval (IR), including, data fusion (Shaw & Fox, 1994), learning to rank (Section 2.6.3) and selective query dependent ranking (Geng *et al.*, 2008).

In addition, as discussed in Sections 4.2.1 and 4.2.2, several approaches can instantiate the query and medical record representation unit or the patient ranking unit. In the previous chapters, we have instantiated these units by using only one of the available approaches at a time. For example, to instantiate the patient ranking unit, we used either the patient or the two-stage models in Chapter 4. However, in this chapter, we show that instantiating these units by combining several approaches leads to a more effective retrieval performance[1].

The remainder of this chapter is organised as follows:

[1] Note that we separate the combination approaches for instantiating the two units from the approaches that instantiate the four components of our framework in order to control the confounding variables in our system.

- Section 9.2 investigates the use of data fusion techniques to combine the relevance scores of patients, which are previously computed using the components of our framework to generate a final ranking.
- In Section 9.3, we investigate the deployment of learning to rank techniques to effectively combine the relevance scores of patients using training data.
- In Section 9.4, we introduce a learned approach that selects which components to apply for a given query, by using query performance predictors as features for a regression-trees learner.
- In Section 9.5, we conduct the experimental evaluation of the combination approaches introduced in Sections 9.2, 9.3 and 9.4.
- In Section 9.6, we investigate novel combination approaches for instantiating the query and medical record representation unit (Section 9.6.1), and the patient ranking unit (Section 9.6.2), respectively.
- In Section 9.7, we provide a summary of our findings and the conclusions for this chapter.

9.2 Data Fusion Approaches for Combining Relevance Scores

Data fusion techniques (Shaw & Fox, 1994) (i.e. metasearch) have been used to combine several rankings of documents retrieved from the same collection into a single ranking. The relevance scores of each document, which are computed by different IR systems, are combined using an aggregate function as the final relevance score of that document. Data fusion was firstly used by Shaw & Fox (1994) to combine the rankings produced by several participating IR systems at TREC. In particular, Shaw & Fox (1994) defined various data fusion techniques to combine the relevance scores computed by different IR systems when calculating the final relevance score of each document. For example, CombSUM adds up the relevance scores from different IR systems, when calculating the final relevance score of a document d, as follows:

$$score_{CombSUM}(d,q) = \sum_{r \in R} score_r(d,q) \tag{9.1}$$

where r is a ranking in R, which is a set of rankings generated by different IR systems. $score_r(d,q)$ is the relevance score of the document d towards the query q in the ranking r. If d is not in a ranking r, $score_r(d,q)$ is set to 0. Note that the voting techniques discussed in Section 2.7, which were used to aggregate the relevance scores of patients (e.g. in Section 4.3.3) were inspired by the data fusion techniques. However, they are different in that instead of aggregating the relevance scores computed by different retrieval systems to estimate the relevance of a particular document (e.g. Equation (9.1)), the voting techniques aggregate the relevance scores of documents ranked by a single system to estimate the relevance of the entities (e.g. patients) associated to such documents (e.g. Equation (2.11)).

In this chapter, we use data fusion techniques, such as CombSUM and CombMNZ to combine the relevance scores computed by the four components of our framework for a particular patient when calculating the final relevance score of that patient. Specifically, we investigate existing data fusion techniques including CombMAX, CombSUM, CombMNZ, CombMED, expCombSUM and expCombMNZ, which respectively calculate the relevance score of a patient d towards a query q as follows:

$$score_{CombMAX}(d,q) = \max_{r \in R}(score_r(d,q)) \tag{9.2}$$

$$score_{CombSUM}(d,q) = \sum_{r \in R}(score_r(d,q)) \tag{9.3}$$

$$score_{CombMNZ}(d,q) = N(d,R) \cdot \sum_{r \in R}(score_r(d,q)) \tag{9.4}$$

$$score_{CombMED}(d,q) = \frac{\sum_{r \in R}(score_r(d,q))}{|R|} \tag{9.5}$$

$$score_{expCombSUM}(d,q) = \sum_{r \in R} e^{score_r(d,q)} \tag{9.6}$$

$$score_{expCombMNZ}(d,q) = N(d,R) \cdot \sum_{r \in R} e^{score_r(d,q)} \tag{9.7}$$

where $score_r(d,q)$ is the relevance score of the patient d in the ranking r in the set of rankings R[1]. $N(d,R)$ returns the number of rankings in R that contains the patient d. $|R|$ is the number of rankings. In this chapter, R consists four approaches, each of which is an instantiation of each component of our framework.

In addition, it is intuitive that the more the number of systems retrieving a particular patient, the more likely that the patient is relevant. Hence, we introduce a variant of CombMAX and CombMED that considers the number of rankings returning a given patient when calculating the relevance score of that patient, as follows:

$$score_{CombMAX_{MNZ}}(d,q) = N(d,R) \cdot \max_{r \in R}(score_r(d,q)) \tag{9.8}$$

$$score_{CombMED_{MNZ}}(d,q) = N(d,R) \cdot \frac{\sum_{r \in R}(score_r(d,q))}{|R|} \tag{9.9}$$

[1] Note that we follow Montague & Aslam (2001) and normalise the relevance scores in each ranking into the range [0,1].

9.3 Learning to Rank for Combining the Components of Our Framework

In this section, we investigate the use of learning to rank techniques to learn an effective combination of rankings. As discussed in Section 2.6.3, learning to rank is an IR research area that aims to improve retrieval performance by learning the effective combination of a set of features using training data. Features are typically the relevance scores of documents computed using retrieval models, such as BM25 (i.e. Equation (2.3) in Section 2.4.1). In this work, we use three existing learning to rank techniques that aim to optimise a targeted retrieval measure during training (namely, AdaRank, AFS and LambdaMART), as they have been shown to be effective (Macdonald *et al.*, 2013; Wu *et al.*, 2008; Xu & Li, 2007). Indeed, these three techniques deploy different types of algorithms to learn the optimal combination of multiple features. Firstly, AdaRank (Xu & Li, 2007) applies a boosting technique to optimise the targeted retrieval measure (e.g. MAP or nDCG) by considering each feature as a weak ranker. For each iteration of the training process, the algorithm re-weights the training data and linearly combines the weak rankers to create a learned model. Therefore, AdaRank is focused more on difficult queries. Secondly, AFS (Metzler, 2007) applies a greedy algorithm to learn an effective linear combination of features. In particular, for each iteration, AFS selects a feature and estimates its weight that optimises the targeted evaluation measure on the training data. The algorithm terminates when none of the remaining features could improve the performance or when it reaches a certain number of iterations. Lastly, LambdaMART (Wu *et al.*, 2008) deploys boosted regression trees to find an effective combination of features that optimises a targeted evaluation measure. For each iteration, the algorithm builds a regression tree to model the gradient of the targeted measure (e.g. MAP) using an approximation (called the λ-gradient).

In this chapter, we use as features the relevance scores of patients computed by the four mentioned components since we aim to deploy learning to rank techniques to combine the rankings produced by these components. However, note that features for a learning to rank technique are not limited to the relevance scores. For example, the predicted difficulty of each query can also be used as a feature.

9.4 Selective Query Dependent Retrieval

Besides combining rankings into one final ranking, we can leverage a set of rankings by selecting to use the ranking that is predicted to be the most effective for a particular query. In this section, we propose a novel selective approach for ranking patients that applies one of the four components on a per-query basis. In particular, we argue that some queries are better served by different components of our framework. For example, some queries may require the negated language to be handled, while others may need inference to improve the representations of medical conditions in the queries. We propose an automatic mechanism that selectively apply one of the four components for each query. Indeed, we adapt a multiclass classification technique to select when to apply a given component of our framework, using query performance predictors as learning features. Next, we discuss the learning features and the learning procedure for our selective approach.

9.4.1 Learning Features

We define the features used for choosing a component to apply for an unseen query. We argue that different components should be selectively used for a given query, based on the perceived difficulty of the query. In this work, we use the same set of 23 features previously used and defined in Table 8.4.4.2 of Section 8.4.4.2. These features include 12 query performance predictors and 11 semantic similarity measures.

9.4.2 The Learning Procedure

Next, we describe our learning procedure, which is based on multiple classifiers. Specifically, our selective approach deploys a multiclass classification technique. For each component of our framework, we train a classifier to indicate whether we should deploy that component when retrieving patients for a particular query.

In particular, to train each classifier, we label each query in the training set with 1, if the corresponding component can achieve the highest retrieval performance among the other components; otherwise, it will be labelled -1. This is to indicate which component is the most effective for each query in the training set. Then, we train each classifier using the obtained accuracy as the loss function. In this work, we use the Gradient Boosted Regression Trees (GBRT) learner that is also used in previous chapters (e.g. Section 5.3.4.3 of Chapter 5), as a classifier. However, any classifier can be used. We use the predicted difficulty of each query (discussed in Section 8.4.4.2) as learning features.

The learned mechanism decides which component to be applied for an unseen query by selecting the component associated with the classifier that produces the highest prediction score, as follows:

$$\hat{y} = \arg\max_{k \in K} f_k(\mathbf{x}) \tag{9.10}$$

where $\hat{y}$ is a classifier associated to $f_k(\mathbf{x})$, which returns the prediction score of the classifier k, given a set of features $\mathbf{x}$ associated to the query. K is the set of classifiers corresponding to each of the four components.

The Four Components of Our Framework	Obtained Performances	
	bpref (TREC 2011)	infNDCG (TREC 2012)
Negation Handling	0.5337	0.4957
Conceptual Reasoning	0.5474	0.5133
Department Expertise	0.5097	0.4232
Inclusion Criteria Coverage	0.4674	0.3952

Table 9.1: The components of our framework and their retrieval performances.

9.5 Experiments

In this section, we evaluate the retrieval effectiveness of the three discussed combination approaches for combining the four components of our framework, which were introduced in Chapters 5, 6, 7 and 8, to further enhance the retrieval performance. In Section 9.5.1, we describe our experimental setup for evaluating the three proposed combination approaches. Section 9.5.2 lists all the research questions to be investigated in this chapter. Then, in Section 9.5.3, we discuss the obtained experimental results. Finally, we provide further analysis of the performances of the combination approaches in Section 9.5.4.

9.5.1 Experimental Setup

To evaluate the three proposed approaches that combine the four components of our framework, we also use the TREC Medical Records track's test collection previously discussed and used in Section 4.3.1 of Chapter 4. We use the best rankings obtained from each component of our framework as inputs for the combination approaches. We list in Table 9.1 the components of our framework and their individual retrieval performances, including the Negation Handling component, the Conceptual Reasoning component, the Department Expertise component and the Inclusion Criteria Coverage component, which were introduced in Chapters 5, 6, 7 and 8, respectively. In particular, for the Negation Handling component, we use our learned approach for handling negation (Table 5.8)). We instantiate the Conceptual Reasoning component using the Bayesian-based approach (Table 6.6). For the Department Expertise component and the Inclusion Criteria Coverage component, we apply the aggregate scoring approach (Table 7.2) and the inclusion criteria coverage approach (Table 8.5), respectively. Our combination approaches aim to generate a ranking that is better than these individual components.

As previously discussed in Section 9.3, we aim to examine data fusion techniques, including CombMAX, CombSUM, CombMNZ, CombMED, expCombSUM, expCombMNZ, CombMAX$_{MNZ}$ and CombMED$_{MNZ}$. Note that the relevance scores within individual rankings produced by each component are normalised to be in the range $[0, 1]$, as suggested by Montague & Aslam (2001).

Meanwhile, for the approach based on the learning to rank techniques, we investigate the effectiveness of deploying the AdaRank, Automatic Feature Selection (AFS) and LambdaMART techniques, respectively (see Section 9.3). For AdaRank and LambdaMART, we use the implementation of the RankLib library[1] with the default settings. For AFS, our implementation uses simulated annealing (Kirkpatrick *et al.*, 1983) to find the combination of features that maximises a target evaluation measure. To train these learning to rank techniques, we follow Zhu & Carterette (2013) and optimise the MAP measure.

Next, to evaluate the selective query dependent approach (see Section 9.4), we use GBRT as a learner. To label the training queries, we use the bpref and infNDCG measures for TREC 2011 and 2012, respectively. To learn the combination models, we use two different training regimes (namely, *5-fold* and *x-collection validations*) previously defined and used in Section 5.4.1 of Chapter 5.

9.5.2 Research Questions

In order to combine the components of the framework, we investigate approaches, based on data fusion (Section 9.2), learning to rank (Section 9.3) and selective retrieval (Section 9.4), to combine the relevance scores computed by each of the four components of the framework. For the remainder of this chapter, we examine the following research questions:

RQ 1. Can data fusion techniques be deployed to effectively combine the components of the framework?

RQ 2. Can learning to rank approaches be effectively exploited to combine the components of the framework?

RQ 3. Will the retrieval performance further improve, if we selectively apply one of the framework components for some queries?

9.5.3 Experiments with Our Approaches for Combining Relevance Scores

In this section, we evaluate the three proposed approaches for combining rankings produced by different components of the framework into a final ranking. We compare the combination approaches previously discussed in this chapter with the four baselines where each of the component of the framework is individually used to retrieve patients, denoted ‘Negation Handling’, ‘Conceptual Reasoning’, ‘Department Expertise’ and ‘Inclusion Criteria Coverage’, respectively.

Firstly, we consider the approach based on the data fusion techniques, previously described in Section 9.2. As shown in Table 9.2, we compare the retrieval performances, in terms of bpref and

[1] `http://sourceforge.net/p/lemur/wiki/RankLib/`

Approach	bpref (TREC 2011)	infNDCG (TREC 2012)
Negation Handling	0.5337^{i}	$0.4957^{d,i}$
Conceptual Reasoning	**0.5474**^{i}	**0.5133**^{d,i}
Department Expertise	0.5097^{i}	0.4232
Inclusion Criteria Coverage	0.4674	0.3952
Data fusion		
CombSUM	0.5152^{i}	0.4392^{i}
CombMNZ	**0.5404**^{i}	$0.4786^{d,i}$
CombMAX	0.4770^{i}	0.4070
CombMAX$_{MNZ}$	0.5107^{i}	0.4613^{i}
CombMED	0.5041^{i}	0.4248^{i}
CombMED$_{MNZ}$	0.5317^{i}	$0.4731^{d,i}$
expCombSUM	0.5087^{i}	0.4327^{i}
expCombMNZ	0.5370^{i}	**0.4798**^{d,i}
Learning to rank		
AdaRank (5-fold)	$0.5194^{d,i}$	0.4275^{d}
AdaRank (x-collection)	0.5138^{i}	0.4257
AFS (5-fold)	$0.5510^{n,i}$	**0.5073**^{d,i}
AFS (x-collection)	**0.5525**^{n,i}	$0.4569^{d,i}$
LambdaMART (5-fold)	0.4775	0.0619
LambdaMART (x-collection)	0.2191	0.3813
Selective retrieval		
Our selective approach (5-fold)	**0.5470**^{i}	0.4425
Our selective approach (x-collection)	0.5271^{i}	**0.4971**^{d,i}
Our selective approach (oracle)	$0.6273^{n,c,d,i}$	$0.5678^{n,c,d,i}$

Table 9.2: The comparison of the retrieval performances of the proposed approaches for combining the components of our framework on TREC 2011 and 2012 Medical Records track's queries.

infNDCG for TREC 2011 and 2012, respectively, when combining relevance scores using different data fusion techniques, including CombSUM, CombMNZ, CombMAX, CombMAX$_{MNZ}$, CombMED, CombMED$_{MNZ}$, expCombSUM, and expCombMNZ, against the relevance scores that are computed using each of the components discussed in Chapters 5, 6, 7, and 8, individually.

For TREC 2011, we observe that the CombMNZ, CombMED$_{MED}$ and expCombMNZ are the most effective among the 8 examined data fusion techniques for both TREC 2011 and TREC 2012. In addition, we find that all of the 8 examined data fusion techniques significantly (paired t-test, $p < 0.05$) outperform 'Inclusion Criteria Coverage'. However, all of the 8 techniques could not outperform the 'Conceptual Reasoning' baseline, which achieved the best retrieval performance (bpref 0.5474) among the four components of the framework. Moreover, we observe that the CombMNZ technique is the most effective data fusion techniques for TREC 2011 (bpref 0.5404). On the other hand, for the TREC 2012 queries, we also observe that all of the used data fusion techniques outperform the 'Inclusion Criteria Coverage' baseline. However, all of the data fusion techniques could not outperform the most effective baseline (i.e. 'Conceptual Reasoning', infNDCG 0.5133). In answer to the first research question, the data fusion techniques (e.g. CombMNZ) could not effectively combine the four components of the framework. Later in Section 9.5.4, we further analyse and discuss the performances of the data fusion techniques.

Secondly, we discuss the retrieval performance of the approach based on the learning to rank techniques (i.e. AdaRank, AFS and LambdaMART) described in Section 9.3. As previously discussed in Section 9.5.1, we use 2 different training regimes (namely, *5-fold* and *x-collection*) when evaluating the three learning to rank techniques. As shown in Table 9.2, overall AFS is more effective than AdaRank and LambdaMART. In particular, using either the 5-fold or the x-collection training regimes, AFS (bpref 0.5510 and 0.5525, respectively) outperforms the use of the individual components of the framework for TREC 2011. In particular, AFS significantly (paired t-test, $p < 0.05$) outperforms the 'Negation Handling' and the 'Inclusion Criteria Coverage' baselines. Meanwhile, for TREC 2012, AFS significantly (paired t-test, $p < 0.05$) outperform the 'Department Expertise' and the 'Inclusion Criteria Coverage' baselines. However, the AFS technique could not outperform the 'Conceptual Reasoning' baseline. In addition, we find that LambdaMART is not effective. This is partially due to the nature of the learner, as LambdaMART is based on boosted regression trees, which prefers several features and many training queries when training a learned model. However, we use only 4 features for 34 queries (resp. 47 queries) for TREC 2011 (resp. TREC 2012). Hence, LambdaMART could not achieve its optimal performance. In answer to the second research question, the AFS technique can effectively combine the four components of the framework and hence further improve the retrieval performance, especially on TREC 2011.

Thirdly, we evaluate the performance of our selective ranking approach (described in Section 9.4). We observe that for both TREC 2011 and TREC 2012, the selective ranking approach could not outperform the best baseline (i.e. 'Conceptual Reasoning'). In particular, the selective ranking approach with the 5-fold training regime performs better than 3 out of the 4 baselines for TREC 2011. Meanwhile, for TREC 2012 the selective ranking approach with the x-collection training regime outperforms 3 out of the 4 baselines. Indeed, a challenge faced by the selective ranking approach is the imbalanced class distributions, where some classes (i.e. components) are underrepresented compared to other classes. For example, when identifying the most effective components for the TREC 2011 queries, the Conceptual Reasoning component is the most effective component for 15 queries, while the Inclusion Criteria Coverage component is the most effective for 9 queries. In addition, we observe that, with the oracle setting[1], the selective approach could further improve the retrieval performances to bpref 0.6273 and infNDCG 0.5678, respectively. This shows the potential of the selective approach if we could more effectively train the learned model. We leave for future work the investigation of a more effective technique to deal with the multiclass imbalanced problem. In answer to the third research question, these experimental results show the potential of the selective ranking approach to effectively combine the relevance scores computed by the four components of our framework.

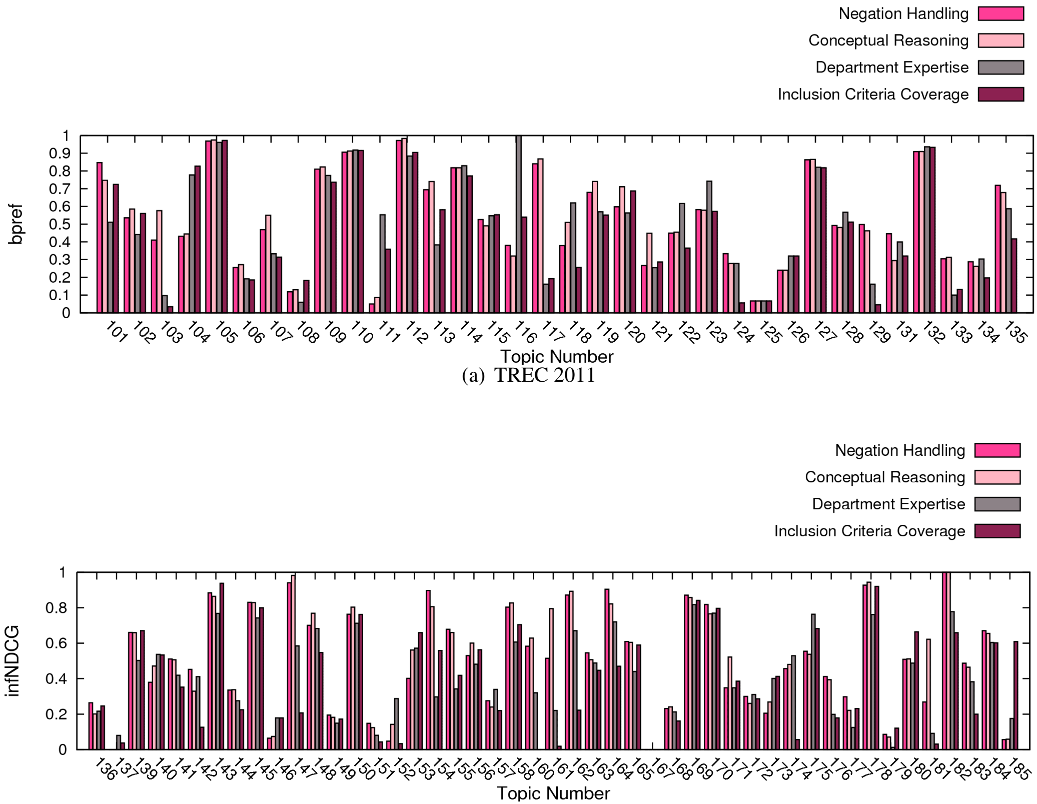
Overall, from our experiment, we observe that the learning to rank approach (especially, AFS) is more effective than the other two combination approaches. Meanwhile, the selective ranking approach could outperform all of the used data fusion techniques for both TREC 2011 and TREC 2012. Next, in Section 9.5.4, we provide a further analysis and discussion of the reported retrieval performances.

9.5.4 Analysis and Discussion

This section further discusses the performances of the three approaches to combine relevance scores computed using the four components of our framework.

We first compare the retrieval performances achieved by each of the components of our framework on a per-query basis. As shown in Figure 9.1, we observe that each component benefits different queries. For example, 'Negation Handling' and 'Conceptual Reasoning' perform effectively on query 117, while 'Department Expertise' and 'Inclusion Criteria Coverage' are more effective for query# 104. Overall, 'Conceptual Reasoning' is the most effective component for both TREC 2011 and TREC 2012. In particular, we find that 'Conceptual Reasoning' performs better than the other components for 15 out of 34

[1] We manually choose the component that achieves the highest retrieval performance for individual queries.



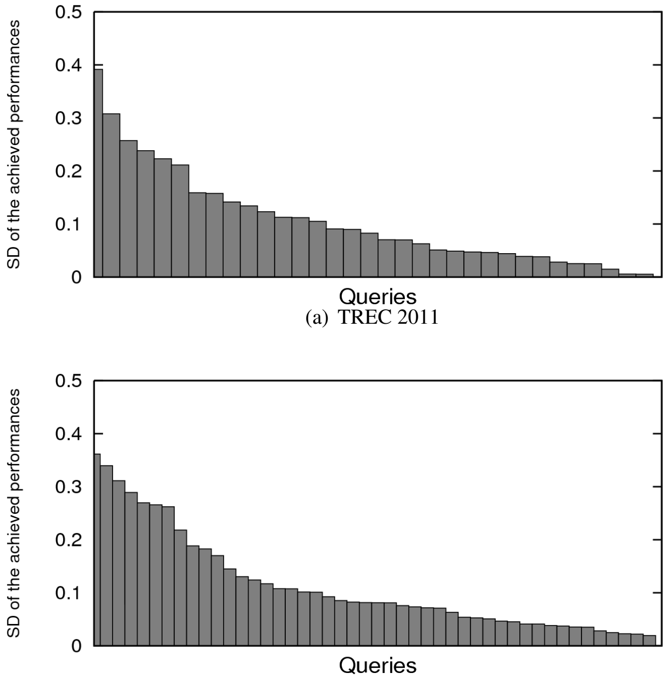
(a) TREC 2011

(b) TREC 2012

Figure 9.1: The comparison of the performances of our four components on the TREC 2011 and 2012 Medical Records track, on a per-query basis.

and 47 queries for TREC 2011 and 2012, respectively. ‘Negation Handling’ outperforms the other three components for 6 and 18 queries for TREC 2011 and 2012, respectively. Next, ‘Department Expertise’ outperforms the other components for 12 and 9 queries for TREC 2011 and 2012, respectively. Finally, ‘Inclusion Criteria Coverage’ achieves the best retrieval performance among the other components for 5 and 9 queries for TREC 2011 and TREC 2012, respectively. This result shows that different queries benefit more from some particular components of our framework.

Next, we investigate the performances of the three combination approaches. As previously discussed in Section 9.5.3, the combination approaches are effective for TREC 2011 than for TREC 2012 (see Table 9.2). We find that this is partly because the retrieval performances achieved by each of the four components of our framework are more different between one another for the TREC 2012 queries than for the TREC 2011 queries. In particular, in Figure 9.2, we show the standard deviation (SD) of the retrieval performances achieved by each of the four components for each query. We find that, on average,



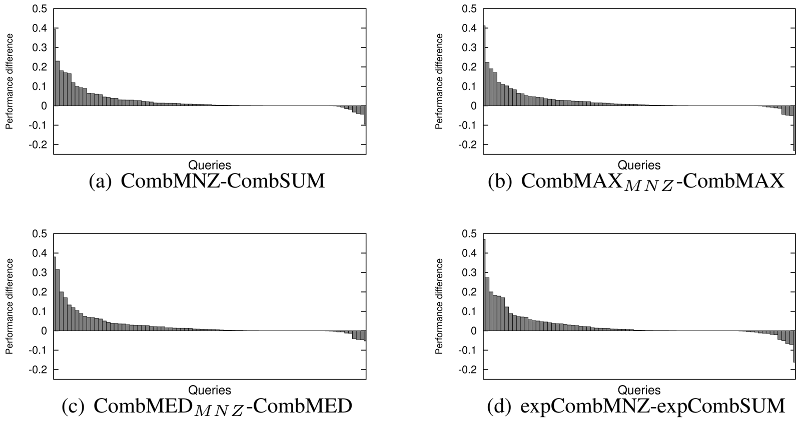
(a) TREC 2011

(b) TREC 2012

Figure 9.2: The standard deviations of the retrieval performances achieved by each component of our framework on a per-query basis, on the TREC 2011 and 2012 Medical Records track.

the SD of the TREC 2012 queries are higher than that of the TREC 2011 queries. Consequently, it appear to be more difficult for the combination approaches to combine the four components.

Next, we analyse the effect of taking into account the number of the components that retrieved a particular patient when estimating the final relevance score of that patient (referred to as *the agreement between the components*) by the data fusion techniques. In Figure 9.3, we show the difference of the achieved retrieval performances on a per-query basis between two corresponding data fusion techniques (e.g. CombSUM vs CombMNZ): one that takes into account the agreement between the components and one that does not, sorted by the margin of the difference. From Figure 9.3, we observe that the data fusion techniques that take into account the number of components that retrieve a patient when estimating the relevance of that patient outperform their corresponding data fusion techniques (e.g. CombSUM vs CombMNZ and CombMAX vs CombMAX$_{MNZ}$) for most of the queries (e.g. queries# 101, 102, 103, 116). In particular, the improved retrieval performances are statistically significant (paired t-test, $p < 0.05$) for both TREC 2011 and TREC 2012. These results shows the importance of taking into account the agreement between the components when combining the relevance scores.



(a) CombMNZ-CombSUM

(b) CombMAX$_{MNZ}$-CombMAX

(c) CombMED$_{MNZ}$-CombMED

(d) expCombMNZ-expCombSUM

Figure 9.3: The per-query performance difference between of the data fusion approaches that consider and do not consider *the agreement between the components*.

9.6 Combination Approaches for the Two Units of Our Search System

In this section, we introduce our combination approaches for instantiating the two units of our patient search system (see Sections 4.2.1 and 4.2.2). Specifically, we discuss our learned approach for combining the term-based and the conceptual representations in Section 9.6.1. Then in Section 9.6.2, we propose an approach for selectively applying either the patient or the two-stage models when ranking patients.

9.6.1 Learning to Combine Representations

As discussed in Section 4.2.1, a patient search system typically deploys one of the two main approaches for representing queries and medical records: the term-based representation and the conceptual representation approaches. Recall that the term-based representation (i.e. the *bag-of-words* representation (BoW)) approaches use terms to represent the queries and the medical records. On the other hand, the conceptual representation (i.e. the *bag-of-concepts* representation (BoC)) approaches represent medical records and queries using concepts from medical resources, such as MeSH[1] and UMLS Metathesaurus[2] (see Section 3.3.1).

As discussed in Section 3.3.1, prior works (e.g. Srinivasan (1996); Trieschnigg *et al.* (2010)) showed that combining the relevance scores of both BoW and BoC when inferring the relevance of a document

[1] http://www.ncbi.nlm.nih.gov/mesh

[2] http://www.nlm.nih.gov/research/umls/

was effective. For example, Srinivasan (1996) linearly combined the relevance scores from both BoW and BoC representations, when inferring the relevance of a document.

However, we argue that the retrieval performance will further improve if we learn a weight for combining the BoW and BoC representations on a per-query basis. We propose a novel learned approach to model the importance of the BoW and BoC representations, when inferring the relevance of medical records before aggregating the relevance scores for their related patients using a voting technique (e.g. using CombSUM as in Equation (4.2)). Our proposed regression-based learning approach leverages the retrieval performance predictors, such as the clarity score (Cronen-Townsend *et al.*, 2002) and the query scope (He & Ounis, 2006), computed on both the BoW and BoC representations as features, to learn an effective combination model on a per-query basis.

To take advantage of both the BoW and BoC representations, we follow Srinivasan (1996) (see Section 3.3.1) and combine the relevance scores of a medical record d towards a query q as follows:

$$score(d, q) = \lambda_q \cdot score_{BoW}(d, q) + (1 - \lambda_q) \cdot score_{BoC}(d, q) \tag{9.11}$$

where λ_q ($0 \leq \lambda_q \leq 1$) is a per-query parameter to estimate the importance of the relevance scores computed using the bag-of-words (BoW) and bag-of-concepts (BoC) representations. The higher the λ_q, the more the relevance score depends on the BoW representation. Indeed, to generalise the model, we introduce a modification to Equation (3.1) (Section 3.3.1 of Chapter 3) with respect to the weighting between the relevance scores of the BoW and BoC representations, so that our combination model can take into account the situation where only BoW (λ_q = 1) or BoC (λ_q = 0) is individually effective. In addition, when $\lambda_q = 0.667$, our model could produce the same list of medical records as Equation (3.1) with the recommended setting (i.e. $\delta = 2.00$), since the proportion of the relevance scores from BoW and BoC computed by Equations (3.1) and (9.11) are equal.

In order to estimate an effective λ_q of the combination model in Equation (9.11), on the training set, we identify the best λ_q that achieves the optimal retrieval effectiveness in terms of a particular retrieval measure (e.g. infNDCG) for each training query. Indeed, for each query, we sweep the λ_q parameter between 0 and 1 to find the best combination model in terms of the retrieval performance for that query. The identified effective λ_q parameter is used as the weight for our learner to learn an effective combination model from our features.

We use the Gradient Boosted Regression Trees (GBRT) technique, which were used in the previous chapters (e.g. Section 5.3.4.3 of Chapter 5) to learn the λ_q value in Equation (9.11). We view the task of estimating the importance of different representation approaches as a supervised regression problem, where the objective is to predict a proper weight (λ_q) for each query, based on the effective weights for

ID	Feature – Ratio (BoW/BoC)
1	Clarity Score (Cronen-Townsend *et al.*, 2002)
2	SCQ (Zhao *et al.*, 2008)
3	MAXCQ (Zhao *et al.*, 2008)
4	NSCQ (Zhao *et al.*, 2008)
5	AvICTF (Carmel & Yom-Tov, 2010)
6	AvIDF (Carmel & Yom-Tov, 2010)
7	EnIDF (Carmel & Yom-Tov, 2010)
8	Query Scope (ω) (He & Ounis, 2006)
9	AvPMI (Carmel & Yom-Tov, 2010)
10	γ_1 (He & Ounis, 2006)
11	γ_2 (He & Ounis, 2006)
12	Query length (He & Ounis, 2006)

Table 9.3: List of the learning features used to predict the importance of the relevance scores from the bag-of-words (BoW) and bag-of-concepts (BoC) representations.

similar training queries. By doing so, we would benefit from the fact that several retrieval performance predictors of the two representation approaches can be used as learning features, when combining the relevance scores. We use the root-mean-square error (RMSE) as the loss function when learning a combination model.

We next identify the features that we will use to choose the weight for an unseen query. We propose to use existing retrieval performance predictors to estimate the retrieval performance of the BoW and BoC representations. We use the set of 12 query performance predictors that were previously defined and used in Section 6.5 of Chapter 6 (Table 6.2). However, we use these retrieval performance predictors in a different manner. In particular, we use the *ratio* between the retrieval performance predictors computed on the BoW and BoC representations, as the learning features (see Table 9.3). This will allow the learner to know based on a retrieval performance predictor, whether the BoW or the BoC representation is likely more likely to better serve a given query.

9.6.1.1 Experimental Setup: the Combined Representation

We evaluate our proposed learned approach using the same settings as in Section 4.3.1. In particular, we use the TREC Medical Records track test collection (see Section 3.4.1). We use the parameter-free DPH term weighting model (Equation (2.8) in Section 2.4.2) to rank medical records before aggregating the relevance scores of their associated patients using the expCombSUM voting technique (Equation (2.13) in Section 2.7). To learn the combination model, we use the 5-fold cross validation regime, previously used and described in Section 5.4.1.

We compare the retrieval performance of our approach with three baselines:

Approaches	2011	2012	
	bpref	infNDCG	infAP
Bag-of-words representation (BoW)	0.4871	0.4167	0.1703
Task-specific representation (BoC)	0.4929	0.4218	0.1920
Score Combination (Srinivasan, 1996) ($\delta = 2$)	**0.5118**	0.4557^{11}	0.1975^{1}
Our learned approach (5-fold)	0.5078	**0.4723**^{11,22,3}	**0.2133**^{1,2}
Our learned approach (oracle)	**0.5796**^{111,222,333,444}	**0.5130**^{111,222,333,444}	**0.2381**^{111,222,33,444}

Table 9.4: The retrieval performances of different representation approaches on the TREC 2011 and 2012 Medical Records track test collection. Statistical significance (paired t-test) at $p < 0.05$, at $p < 0.01$, and at $p < 0.001$ over a baseline are denoted ^{a}, ^{aa} and ^{aaa}, respectively. ^{a} is ^{1}, ^{2}, ^{3} or ^{4} to represent the bag-of-words representation (BoW), the task-specific representation (BoC), the score combination ($\delta = 2$) baselines, or our learned approach (5-fold) approach, respectively.

- A traditional bag-of-words representation (BoW) approach
- The task-specific representation (BoC) approach
- An existing score combination approach (Srinivasan, 1996) (i.e. Equation (3.1)) with the suggested setting of $\delta = 2$.

9.6.1.2 Experimental Results: the Combined Representation

Table 9.4 compares the retrieval effectiveness of the proposed approach with the mentioned three baselines on the TREC 2011 and 2012 Medical Records track test collection. In addition, to evaluate the optimal potential effectiveness, the best retrieval performance that our learned approach could achieve is also reported (denoted oracle[1]).

From Table 9.4, we observe the following. First, we see that for both TREC 2011 and TREC 2012, both the proposed learned approach and the existing score combination approach markedly outperform the baselines where either of the BoW/BoC representations are taken into account. This shows that combining the relevance scores from BoW and BoC is effective for patient search. Next, for the TREC 2012 queries, the retrieval performances of our learned approach (5-fold) markedly outperform those of the score combination baseline ($\delta = 2$). In particular, in terms of the infNDCG retrieval performance, our learned approach (infNDCG 0.4723) significantly outperforms (paired t-test, $p < 0.05$) the existing score combination baseline (infNDCG 0.4557). For the infAP measure, the proposed learned approach performs markedly better than the score combination baseline (+6.5% improvement, from 0.1975 to 0.2133). However, for the TREC 2011 queries, our learned approach (5-fold) could not outperform the score combination ($\delta = 2$) baseline (bpref 0.5078 vs. 0.5118). This is partially due to the fact that the TREC 2011 queries contains only 34 queries; hence, with a small number of queries, when we conduct a 5-fold cross validation, the training and test sets could not generalise.

[1] We sweep the λ_q value between 0 and 1 to identify the oracle setting for each query.

Finally, we discuss the optimal retrieval performance that our proposed learned approach could achieve to evaluate the potential effectiveness of our learned approach, if we could better train the learner. As expected, we observe that, with the best setting, the oracle approach significantly ($p < 0.01$) outperforms all of the approaches discussed in this experiment. This shows some particular queries differently benefit from the BoW and BoC representations. In particular, the retrieval performance of our approach with the best setting (oracle) is up to +17.06% better than the 5-fold cross validation. Importantly, we find that the mean of the effective weights (λ_q with the best possible setting) across the two collections is 0.48459 ($0 \leq \lambda_q \leq 1$), while the standard deviation is 0.38085, which suggests that the effective weight should indeed vary across queries. From this experimental result, we conclude that there is no one combination of BoW and BoC that is effective for all queries. Hence, per-query prediction approaches, like the ones deployed here, have great potential to improve patient search. However, there is still an open research area to explore effective features and learners to close the performance gap between the cross-validation and oracle regimes, even though by deploying the existing learner and features, our learned approach could in general markedly and significantly outperform the existing score combination approach of Srinivasan (1996).

9.6.2 Learning to Selectively Rank the Patients' Medical History

Next, we describe our approach for selectively applying a patient ranking model. As discussed in Sections 4.2.2, two main types of approaches (namely, the patient and the two-stage models) are used to rank patients based on the relevance of their medical history. As discussed in Section 3.4.2.1, Zhu & Carterette (2012) recently proposed to use a data fusion technique (Shaw & Fox, 1994), such as CombSUM and CombMAX, to merge the relevance scores from both the patient model and the two-stage model in order to exploit the effectiveness of both patient ranking approaches (see Section 3.4.2.1). Instead, we argue that a selective approach that can appropriately identify which of the two ranking approaches is more effective for a given query can further improve retrieval performance.

Specifically, we propose a novel selective approach for ranking patients based on the relevance of their medical history. In particular, we postulate that some queries are better served with different patient ranking approaches. Therefore, our proposed approach aims to effectively apply a ranking approach that can accomplish a better retrieval performance for a particular query. Specifically, we deploy a regression-trees classifier to learn how to select a ranking approach using query difficulty measures such as AvIDF (Carmel & Yom-Tov, 2010). We choose to use the query difficulty measures as our learning features, since we argue that to attain an effective retrieval performance, a search system should deploy

ID	Feature
1	Clarity Score[1] (Cronen-Townsend *et al.*, 2002)
2	SCQ[1] (Zhao *et al.*, 2008)
3	MAXCQ[1] (Zhao *et al.*, 2008)
4	NSCQ (Zhao *et al.*, 2008)
5	AvICTF[1] (Carmel & Yom-Tov, 2010)
6	AvIDF[1] (Carmel & Yom-Tov, 2010)
7	EnIDF[1] (Carmel & Yom-Tov, 2010)
8	Query Scope (ω)[1] (He & Ounis, 2006)
9	γ_1[1] (He & Ounis, 2006)
10	γ_2[1] (He & Ounis, 2006)
11	Query length (He & Ounis, 2006)
12	Number of detected medical concepts in the query
13	Occurrence probability of symptom concepts in the query
14	Occurrence probability of diagnostic-test concepts in the query
15	Occurrence probability of diagnosis concepts in the query
16	Occurrence probability of treatment concepts in the query

Table 9.5: List of the query features used by our classifier to decide on when to apply either the patient model or the two-stage model.

the ranking approach that finds the query the least difficult. The intuition is that an easy query leads to a better retrieval performance.

We first describe the learning features used in our selective approach. To effectively select between the patient model and the two-stage model, we learn a classifier using the query features listed in Table 9.5. Indeed, these features, which measure the difficulty of a query, can be categorised into two groups. The first group (Features 1-10) measures the *relative difference* of the query performance predictor scores between the patient model and the two-stage model. In particular, these 10 query performance predictors were previously defined and used in Section 6.5 (Table 6.2).

On the other hand, the second group of features (Features 11-16) measures the difficulty of a query by using the information from the query itself. Feature 11 is the number of non-stopword query terms. Moreover, since medical queries normally focus on the four aspects of the medical decision process, namely, symptoms, diagnostic tests, diagnoses and treatments (see Section 4.2.1), Features 12-16 are based on the occurrences of the medical concepts associated with these four aforementioned aspects. In particular, Feature 12 is the number of the concepts related to the medical decision process, which can be extracted from the query. Specifically, we deploy MetaMap to identify those medical concepts in the query, as discussed in Section 4.2.1[2]. Features 13, 14, 15 and 16 estimate the probability that the medical concepts detected in a query are related to symptom, diagnostic test, diagnosis and treatment,

[1] The difference between the values of the feature computed on the patient model and the two-stage model.

[2] We only use the concepts with the highest scores from MetaMap (i.e. indicated as ‘Meta Mapping’).

Approaches	2011	2012	
	bpref	infNDCG	infAP
Patient model	0.5006	0.4459	0.1865
Two-stage model	0.5141	0.4481	0.1857
CombMAX	0.4968	0.4459	0.1865
CombSUM	0.4978	0.4453	0.1866
Our selective ranking approach (5-fold)	**0.5261**$^{\oplus,\otimes,\oslash}$	**0.4500**	**0.1906**
Our selective ranking approach (oracle)	0.5368$^{\oplus\oplus\oplus,\ominus\ominus,\otimes\otimes\otimes,\oslash\oslash\oslash,\odot}$	0.4830$^{\oplus\oplus,\ominus\ominus\ominus,\otimes\otimes,\oslash\oslash,\odot\odot\odot}$	0.2037$^{\oplus\oplus,\ominus\ominus\ominus,\otimes\otimes,\oslash\oslash,\odot\odot}$

Table 9.6: The retrieval performances of different patient ranking approaches on the TREC Medical Records track's collection. Statistical significance (paired t-test) at $p < 0.05$, at $p < 0.01$, and at $p < 0.001$ over an alternative approach are denoted ^{x}, ^{xx} and ^{xxx}, respectively. Where ^{x} is $^{\oplus}$, $^{\ominus}$, $^{\otimes}$, $^{\oslash}$ or $^{\odot}$ and refers to the patient model, the two-stage model, CombMAX, CombSUM, and our selective approach (5-fold), respectively.

respectively. The probability is estimated using the maximum likelihood by counting the number of medical concepts detected in the query. The more medical concepts are detected in the query, the more likely that the query is difficult.

Next, we describe our decision mechanism, which is based on a learned classifier. In particular, the classifier decides to apply the patient model or the two-stage model for a particular query using the introduced query features. By doing so, we benefit from the fact that several query difficulty measures, which are used as learning features, can be taken into account when choosing an effective patient ranking approach. We use the GBRT as a classifier (see Section 5.3.4.3).

To effectively train the classifier, on a training set of queries, we label each query with 1 if the patient model can achieve a better retrieval performance on a particular target measure; otherwise, it is labelled -1. This allows the classifier to learn which ranking approach is more effective for a particular query. Next, when training the GBRT classifier, we use the obtained accuracy as the loss function. We train the classifier using the features listed in Table 9.5.

9.6.2.1 Experimental Setup: Our Selective Query Dependent Approach

We evaluate our selective ranking approach using the same settings as in Section 4.3.1 of Chapter 4. Recall that we use the TREC Medical Records track's test collection. For the patient model, we use the parameter-free DPH weighting model to rank the patient documents (i.e. the concatenation of the medical records associated to the same patient) described in Section 4.3.1. For the two-stage model, we deploy the expCombSUM voting technique to aggregate the relevance scores of medical records when calculating the relevance scores of their associated patients. To evaluate the proposed selective ranking approach, we use the 5-fold cross-validation regime defined and used in Section 5.4.1 of Chapter 5. When labelling the training set, we target the retrieval performance in terms of bpref and infNDCG for TREC 2011 and 2012, respectively.

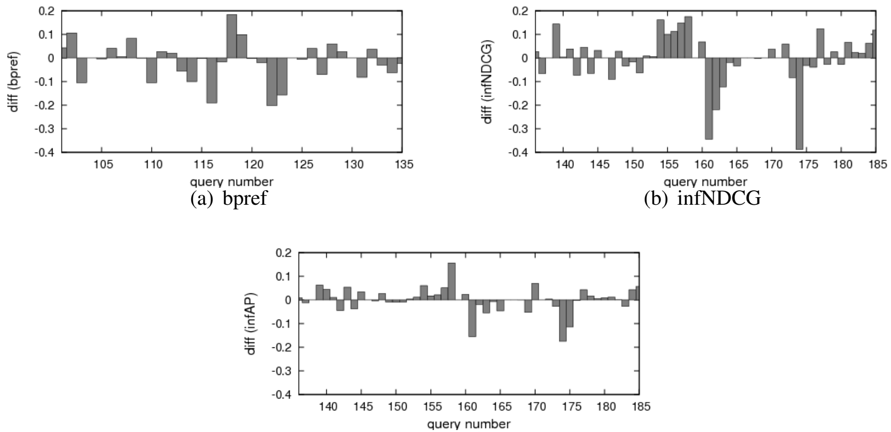
We compare the retrieval performance of our selective ranking approach with the baselines, namely:

- The patient model (i.e. using DPH to rank the patient documents)
- The two-stage model (i.e. using DPH to rank the medical records before aggregating the relevance scores of patients using expCombSUM)
- CombSUM (i.e. using CombSUM to combine the relevance score of the patient and the two-stage models) as suggested by Zhu & Carterette (2012)
- CombMAX (i.e. using CombMAX to combine the relevance score of the patient and the two-stage models) as suggested by Zhu & Carterette (2012)

9.6.2.2 Experimental Results: Our Selective Query Dependent Approach

Table 9.6 compares the retrieval performance of our selective ranking approach with the aforementioned baselines, in terms of bpref, infNDCG, and infAP. In addition, to gauge the potential effectiveness of our deployed classifier, we also report the best possible retrieval performances that could be attained by our approach (i.e. an oracle[1]), when the classifier correctly identifies an effective patient ranking approach for all of the queries. Firstly, we observe that the patient model and the two-stage model attain a comparable retrieval effectiveness. Indeed, the two-stage model outperforms the patient model in terms of the bpref and infNDCG measures, for TREC 2011 and 2012, respectively (bpref 0.5141 vs. 0.5006 and infNDCG 0.4481 vs. 0.4459), while the patient model performs better in terms of infAP for TREC 2012 (0.1865 vs. 0.1857). Moreover, we find that applying the data fusion techniques, including CombMAX and CombSUM, as suggested in Zhu & Carterette (2012), does not in general improve the retrieval performance over both the patient and the two-stage models. On the other hand, we find that, with the 5-fold cross-validation setting, our selective approach outperforms all of the four baselines. Specifically, in terms of bpref, our approach (bpref 0.5261) significantly (paired-t test, $p < 0.05$) outperforms the patient model, the CombSUM, and the CombMAX baselines for up to 5.9%. However, in terms of infNDCG and infAP, the increased performances are not statistically different. The observed results demonstrate the effectiveness of our deployed approach in selecting the right ranking approach for a particular query. Note that the CombSUM and CombMAX baselines do not perform as effectively as in Zhu & Carterette (2012), partially because we use DPH to rank documents in *the patient model* and to rank medical records at *the first phase of the two-stage model*, instead of a language model. In addition, when aggregating the relevance scores of the medical records in the two-stage model, we

[1] The oracle setting is obtained by manually choosing between the patient and the two-stage models, the one that achieves the better retrieval performance for the individual queries.



(a) bpref

(b) infNDCG

(c) infAP

Figure 9.4: The difference between the retrieval performance obtained using the patient model and the two-stage model on each query, in terms of bpref, infNDCG and infAP, respectively.

use the expCombSUM voting technique, instead of just summing up the relevance scores as in Zhu & Carterette (2012). This suggests that the performances of CombSUM and CombMAX depend on the underlying used retrieval models. In contrast, the proposed approach overcomes this problem by appropriately learning from the features when selecting a patient ranking model. Furthermore, we find that if our classifier could make the correct decisions for each of the queries, the retrieval performances can further improve (see the oracle row in Table 9.6).

Next, in Figure 9.4, we compare the difference between the retrieval performances obtained using the patient model and the two-stage model on every query. Note that the difference in the retrieval performance is positive when the patient model is more effective. From this figure, we find that neither the patient model nor the two-stage model is consistently more effective for all of the queries. We also find that the most effective patient ranking approach depends on the used measure (e.g. for query# 183, the patient model is more effective on infNDCG, while for the same query, the two-stage model is more effective for infAP). This confirms the importance of deploying selective ranking approaches when ranking patients.

9.7 Conclusions

In this chapter, we have investigated three approaches for combining the four components of our framework, including the Negation Handling component (see Chapter 5), the Conceptual Reasoning component (see Chapter 6), the Department Expertise component (see Chapter 7) and the Inclusion Criteria Coverage component (see Chapter 8).

We have used techniques from data fusion and machine learning to combine the relevance scores of patients computed by each of the four components of the framework into a final ranking. In particular, in Section 9.2, we described the used data fusion-based techniques including CombSUM, CombMNZ, CombMAX, CombMAX$_{MNZ}$, CombMED, CombMED$_{MNZ}$, expCombSUM, and expCombMNZ. Section 9.3 discussed the use of AdaRank, AFS and LambdaMART for combining the components of our framework. Section 9.4 introduced our selective query dependent approach to selectively apply one of the four components when ranking patients for a particular query by taking into account the predicted difficulty of the query.

Our experimental results in Section 9.5, showed the potential of our combination approaches as they outperformed several baselines computed using each of the four components of the framework individually. In particular, from Table 9.2, we found that the AFS learning to rank technique was more effective than the other examined combination approaches for both TREC 2011 and TREC 2012. Indeed, the AFS-based combination technique outperformed all of the individual components of our framework for TREC 2011. However, for TREC 2012, it could not outperform the most effective component (i.e. the Conceptual Reasoning component). Meanwhile, as shown in Table 9.2, we found that the selective query dependent approach was effective for TREC 2011, while it was less effective for TREC 2012. Next, we found that CombMNZ, CombMED$_{MNZ}$ and expCombMNZ are the most effective data-fusion techniques (see Table 9.2), when combining the components of our framework. In addition, we found that the data fusion techniques that take into account *the agreement between the components of our framework* (i.e. CombMNZ, CombMAX$_{MNZ}$, CombMED$_{MNZ}$ and expCombMNZ, respectively) performed significantly better than their corresponding techniques, which do not consider this evidence (i.e. CombSUM, CombMAX, CombMED and expCombMNZ, respectively).

In the thesis statement (Section 1.3), we hypothesised that uncovering and leveraging implicit knowledge at the three different levels could improve the retrieval performance of a patient search system. Our experimental results in Section 9.5 supports the thesis statement, as the combination of the four components of the framework can significantly improve the retrieval performance over using the individual components.

In addition, in Section 9.6, we have introduced two combination approaches to instantiate the query and patient representation unit or the patient ranking unit of the patient search system, respectively. In particular, Section 9.6.1 discussed our approach for combining the BoW and BoC representations of the medical records and queries. Section 9.6.2 described our approach for selectively applying either the patient or the two-stage models for ranking patients for a particular query, as an instance of the patient ranking unit. Our experimental results showed that combining the BoW and BoC representations lead to a significant improvement of the retrieval performance (Table 9.4). Meanwhile, our approach for selectively applying the patient or the two-stage models significantly outperformed the use of the individual models (Table 9.6). We leave for future work, the investigation of an approach that combines the four components of the framework with different instantiations of both the representation unit and the patient ranking unit at the same time.

Next, in Chapter 10, we will close this thesis by summarising the conclusions and contributions from each of the chapters in this thesis. Moreover, we will discuss possible directions for future research uncovered by this thesis.

Chapter 10

Conclusions and Future Work

10.1 Contributions and Conclusions

This thesis proposed a framework that uncovers the implicit knowledge in patient search. In particular, we proposed to model and reason upon medical information related to four aspects of the medical decision process (i.e. symptom, diagnostic test, diagnosis and treatment) at the three levels of the retrieval process (i.e. sentence level, record level and inter-record level), in order to extract possible implicit knowledge from medical records and queries. Through thorough experiments, we have drawn insights and concluded that effective patient search could be achieved by dealing with the implicit knowledge at the three levels of the retrieval process. The remainder of this section discusses the contributions and conclusions of this thesis.

10.1.1 Contributions

The main contributions of this thesis are as follows:

- In Chapter 4, we proposed a novel framework that uncovers the implicit knowledge by modelling the medical conditions related to the four medical aspects. We introduced four components for this framework, each of which extracts implicit knowledge at different levels of the retrieval process. Specifically, at the sentence level, the first framework component (i.e. *the Negation Handling component*) deals with the negated language in medical records and queries to prevent the retrieval of patients that do not have the medical conditions that the query searches for. (see Section 4.2.3.1). At the record level, the second component (i.e. *the Conceptual Reasoning component*) uncovers the implicit knowledge by inferring the relationships between medical conditions, e.g. patients with 'heart disease' are likely to have 'high blood pressure', in order to deal with the complexity of the medical terminology (see Section 4.2.3.2). At the inter-record level, the third component (i.e. *the Department Expertise component*) leverages the inferred expertise of hospital departments to highly rank patients who have medical records from the hospital departments that have expertise in the medical conditions stated in the query (see Section 4.2.3.3). Finally, also at the inter-record level, the final component of the proposed framework (i.e. *the Inclusion Criteria Coverage component*) models the relevance with respect to all of the medical conditions in the query so that patients who have medical records relevant to many query medical conditions are ranked highly (see Section 4.2.3.4). Our proposed framework is the key contribution of this thesis, in that it defines several approaches that uncover implicit knowledge in patient search.

- In Chapter 5, the Negation Handling component, which deals with the negated language in medical records and queries, was presented and investigated. We contributed the NegFlag approach that leverages the NegEx algorithm to detect negated phrases in medical records and queries (see Section 5.2.1), representing terms with positive and negative contexts differently. Building upon the NegFlag approach, we introduced two novel approaches to prevent patients whose medical records contain the query terms in an opposite context from being retrieved (e.g. preventing patients with diabetes from being retrieved for a query to find patients who have no history of diabetes). The first approach leveraged the term dependence (i.e. co-occurrence statistics) of query terms when demoting patients whose medical records contain query terms with an opposite context (see Section 5.2.2). Meanwhile, for the second approach (see Section 5.3), we proposed a learning mechanism that learns to give a negative score to a patient based on the occurrence of a query term with an opposite context to the query. We thoroughly evaluated our approaches using the test collections provided by the TREC Medical Records track to determine how effectively our three proposed approaches can deal with the negated language in the medical records and queries (see Section 3.4.1). In addition, we also investigated which types of queries benefit from the negated language handling (see Section 5.4.5).

- In Chapter 6, we proposed the Conceptual Reasoning component, which infers relationships between medical conditions. We introduced two novel approaches to deal with the complexity of medical terminology when searching patients based on the relevance of their medical records towards a given query. As discussed in Section 3.4.2.3, the existing approaches suffer from the complexity of the medical terminology. Our two proposed approaches tackled this problem by exploiting associations between medical conditions (i.e. medical concepts). In particular, we built a database of association rules between medical conditions extracted from existing medical resources (e.g. ontologies and health-related websites) (see Section 6.2). Then, our two proposed approaches leveraged this association rule database using Bayes' theorem (see Section 6.3) and a stochastic analysis (see Section 6.4), respectively, to infer relationships between medical conditions and to allow the retrieval of patients whose medical records do not explicitly contain the query terms. We also analysed which types of queries our two proposed approaches improve in terms of retrieval performance (see Section 6.6.5).

- Next, in Chapter 7, we introduced the Department Expertise component, which measures and leverages the expertise in particular medical conditions of different hospital departments. Specifically, we proposed two novel approaches (see Section 7.2) that instantiate this component by promoting the patients who have medical records issued from the hospital departments that are expert in the medical conditions stated in the query. We proposed two novel techniques to measure the expertise of hospital departments using the aggregates of medical records issued from individual hospital departments. The two techniques were based on the federated search and voting paradigms (see Sections 7.3.1 and 7.3.2, respectively). Our contributions are the two novel approaches that highly weight medical records issued from those hospital departments when aggregating the scores of the retrieved medical records (Section 7.2.1) and when calculating the relevance scores of the retrieved medical records before ranking the patients (Section 7.2.2), respectively.

- In Chapter 8, we introduced the final component of our framework (i.e. the Inclusion Criteria Coverage component), which estimates the relevance with respect to each of the medical conditions stated in a given query. We contributed a novel approach that promotes the patients whose medical records are relevant to multiple query inclusion criteria (see Section 8.3). Given a query, we used the MetaMap tool to extract the medical conditions (i.e. inclusion criteria) from the query, representing each of them as a sub-query (Section 8.3.3). Our novel approach used these sub-queries to measure the relevance towards the query medical conditions promoting those patients that are relevant to most of the medical conditions. Our approach can be applied with the two existing families of patient ranking models (namely, the patient model (see Section 8.3.1) and the two-stage model (see Section 8.3.2)). We analysed the types of queries that benefited from our approach in Section 8.4.5.2.

- In Chapter 9, we investigated several approaches to combine the four components of the proposed framework. We combined the relevance scores computed by the components of our framework by deploying approaches from three well-established research areas (namely, data fusion, learning to rank, and selective query dependent retrieval). Firstly, we examined several data fusion techniques (e.g. CombSUM and CombMNZ) to aggregate those relevance scores (see Section 9.2). Secondly, we applied learning to rank techniques, including AdaRank, AFS and LambdaMART, to learn an effective combination of the relevance scores computed using the four components of the framework (see Section 9.3). Thirdly, we introduced a selective query dependent retrieval approach that learns when to apply one of the components of the framework for a particular query, based on the predicted difficulty of that query (see Section 9.4). Our selective query dependent retrieval approach deployed multiple classifiers to identify which of the components should be used for a given query. In addition, in Section 9.6, we investigated a learned approach that combines the relevance scores computed using a term-based and a task-specific representation on a per-query basis. Moreover, we also proposed an automatic approach to selectively apply either the patient model or the two-stage model for a given query.

10.1.2 Conclusions

In this section, we summarise the main conclusions and achievements of this work. In particular, these conclusions validate the statement of this thesis proposed in Section 1.3 using the test collection provided by the TREC Medical Records track.

Effectiveness of Focusing the Search System on the Medical Decision Process From the experiments conducted in Chapter 4, we examined the effectiveness of focusing only on the medical conditions related to the four aspects of the medical decision process (namely, symptom, diagnostic test, diagnosis, and treatment) when retrieving patients based on the relevance of their medical records towards a query. Compared to the two baselines where all terms or all medical conditions (i.e. medical concepts) are used to represent the medical records and queries, our results showed that focusing on the medical conditions related to the medical decision process is more effective (see Section 4.3.4). Specifically, from Table 4.7, we observed that a retrieval system that focuses only on the medical conditions related to the medical decision process significantly outperforms the retrieval system that uses either all of the terms or all the medical conditions extracted from the medical records and queries.

This thesis postulated that a patient search system should focus on the medical conditions related to the four medical aspects, which are the key information that healthcare practitioners take into account when consulting patients, and allowing inferences to uncover the implicit knowledge (see Section 1.3). From our experimental results in Chapter 4, we conclude that focusing a search system on the medical conditions related to the medical decision process is an effective approach in patient search.

Effective Negation Handling In Chapter 5, we examined whether explicitly dealing with the negated language in medical records and queries could improve retrieval performance. We proposed three approaches for dealing with the negated language in medical records and queries, including the NegFlag approach, the term dependence approach and a learned approach. Our experimental results showed that the NegFlag approach significantly improved the retrieval performance by up to 7.07% over an effective baseline that does not explicitly deal with the negated language (see Table 5.5). Meanwhile, our term dependence and our learned approaches improved the retrieval performance by 7.71% (see Table 5.7).

In relation to the thesis statement (Section 1.3), our experiments in Chapter 5 confirmed that uncovering implicit knowledge at the sentence level of the retrieval process by recognising the contexts of the medical conditions in the medical records and queries improves the retrieval effectiveness of a patient search system.

Effectiveness of Inferring Relationships between Medical Conditions In Chapter 6, we examined two approaches (namely, the Bayesian-based and the stochastic approaches) to infer relationships between the medical conditions in the medical records and queries. Our experimental results in Sections 6.6.3 and 6.6.4, showed that exploiting the relationships between medical conditions significantly improved the retrieval performance. In particular, as shown in Table 6.6 and 6.7, the Bayesian-based and the stochastic approaches outperformed an effective baseline that does not infer these relationships by up to 5.2% and 2.5%, respectively.

According to our thesis statement (see Section 1.3), we hypothesised that relationships between medical conditions could be inferred to uncover implicit knowledge. Based upon our experimental results in Sections 6.6.3 and 6.6.4, we conclude that relationships between medical conditions can be effectively inferred to improve the retrieval performance of a patient search system.

Effectiveness when Leveraging the Extracted Hospital Departments' Expertise In Chapter 7, We proposed two approaches (namely, the aggregate scoring and the record scoring approaches) that make use of the expertise of hospital departments when ranking patients based on the relevance of their medical records. Our experimental results in Sections 7.4.3 and 7.4.4 showed the potential of our two proposed approaches as they outperformed an effective baseline that does not take into account the expertise of hospital departments. In addition, in Section 7.4.5.2, we showed that we can automatically set the parameter within our aggregate scoring approach and improved the retrieval performance by 6.6% over the baseline that does not take into account the expertise of hospital departments when ranking patients (see Table 7.2).

In Section 1.3, our thesis statement postulated that the implicit knowledge could be uncovered at the inter-record level of the retrieval process. From the experiments of Section 7.4, we conclude that using the expertise of hospital departments estimated from aggregates of the medical records issued by the individual departments help improve the effectiveness of a patient search system.

Importance of Modelling Relevance towards Multiple Inclusion Criteria in the Medical Queries In Chapter 8, we introduced our inclusion criteria coverage approach to highly rank patients whose medical records are relevant to multiple medical conditions stated in a given query. Our experimental results in Sections 8.4.3 and 8.4.4 showed that the inclusion criteria coverage approach significantly outperformed an effective baseline by up to 9.13% (see Table 8.6).

In Section 1.3, our thesis statement hypothesised that uncovering the implicit knowledge at the inter-record level enhances retrieval performance. Our inclusion criteria coverage approach uncovers the implicit knowledge about the medical conditions that the query focuses on and promotes the patients who are relevant to several of those medical conditions. Our experimental results showed the significant improvement achieved by the inclusion criteria coverage approach over an effective baseline, validating the thesis statement.

Effectiveness of Combining Approaches within Our Framework In Chapter 9, we examined whether the combination of the components of our framework that deal with the implicit knowledge at different levels of the retrieval process could lead to a better retrieval effectiveness. Through the experiments in Section 9.5.3, we compare three combination approaches based on data fusion and machine learning with the retrieval performance achieved by each of the components of our framework. The AFS learning to rank technique performed better than all of the examined combination approaches. In particular, for TREC 2011, AFS achieved bpref 0.5525, which is better than all of the four component of our framework (see Table 9.2). This performance is significantly better the Negation Handling and the Inclusion Criteria Coverage components. However, for TREC 2012, AFS could not outperform the most effective component (i.e. the Conceptual Reasoning component), as shown in Table 9.2.

Our results showed that the combination of the four components of our framework performed better than the use of individual components of framework alone, especially the AFS learning to rank technique.

10.2 Directions for Future Work

This section discusses possible directions for future research, related to or inspired by this thesis.

Learning to Focus on Particular Types of Medical Conditions Our task-specific representation approach proposed in Chapter 4 improved the representations of medical records and queries by focusing on the medical conditions related the medical decision process (namely, symptom, diagnostic test, diagnosis, and treatment). Experiments in Section 4.3.4 showed that the approach significantly outperformed both a term-based and a conceptual representation approaches. However, it is also interesting to explore a learned approach that weights the medical conditions related to each of the aspects (e.g. symptom) of the medical decision process differently, so that a patient search system could more appropriately promote the patients who are relevant to the medical conditions that are related to a particular aspect during retrieval. Learning techniques, such as Gradient Boosted Regression Trees (GBRT) (Tyree *et al.*, 2011), could be used to learn the weights of the medical conditions related to different aspects of the medical decision criteria.

Dealing with the Other Contexts of Medical Conditions We have shown in Chapter 5 that dealing with negation could significantly improve the retrieval performance of a patient search system, as medical records and queries are often written using a negated language. However, there are other types of contexts that may be used in the medical records and queries. For example, a healthcare practitioner may state in a medical record that the patient had a mother who has diabetes. This does not explicitly confirm that the patient is also suffering from this disease. Hence, we could not demote this patient for the query searching for diabetic patients. In Section 5.4.5.5, we have shown that applying a simple technique similar to the NegFlag approach (Section 5.2.1) effectively represented these contexts. However, it may be worth investigating other approaches, such as the one similar to our learned approach to handle negation introduced in Section 5.3 to deal with those contexts.

Considering the Context of Medical Conditions during Inference Our two proposed approaches (introduced in Sections 6.3 and 6.4) make inference on the relationships between medical conditions without considering the context (i.e. positive or negative) of the medical conditions. It is interesting to further investigate whether the retrieval performance could improve if we consider the context of the inferred medical conditions. For example, assuming that *symptom A* is one of the symptoms of heart disease, when searching for patients with heart disease, the patients whose medical records state that the patient does not have *symptom A* may be demoted in the ranking.

The Application of the Proposed Framework on Other Medical Search Tasks This thesis aims to tackle the patient search task by uncovering the implicit knowledge. However, other search scenarios in the medical domain are also hindered with the implicit knowledge problem. For example, the searching of web pages that provide a simple source for patients with particular health conditions confront with the situations where the medical web pages are typically written by healthcare experts, while the search users are normally novices. Hence, an effective medical web page search system requires a mechanism to deal with the knowledge that is not explicitly available of the search users. The information retrieval task of the ShARe/CLEF eHealth Evaluation Lab (Suominen *et al.*, 2013) simulates such situations. In future work, we could investigate the application of the framework proposed in this thesis on these search scenarios.

10.3 Closing Remarks

In this thesis, we argued that the representations of medical records and queries can be improved by uncovering the implicit knowledge at different levels of the retrieval process. We proposed a novel framework for uncovering the implicit knowledge by taking into account and applying inferences on the medical conditions related to the four aspects of medical decision process (namely, symptom, diagnostic test, diagnosis and treatment). From a thorough empirical investigation using the test collection provided by the TREC Medical Records track, we showed that our framework is effective for the patient search task. In particular, recognising the medical conditions related to the four aforementioned aspects and representing the medical records and queries using only those medical conditions leads to a more effective patient search system. Detecting the contexts of the medical conditions and preventing patients whose medical records do not have the medical conditions stated in the query leads to an effective retrieval performance. The relationships between medical conditions should be inferred to improve the retrieval performance. The medical records issued from particular hospital departments that are expert in the medical conditions stated in the query should be highly weighted when ranking patients. Furthermore, it is important to promote the patients whose medical records are relevant to multiple query medical conditions. Finally, we showed that under some conditions, the improvements obtained by the various deployed approaches were additive, leading to an effective patient search system.

Bibliography

Abhyankar, S., Demner-Fushman, D., Callaghan, F. M. & McDonald, C. J. (2014). Research and applications: Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *Journal of the American Medical Informatics Association*. 3.4.2.2

Agrawal, R., Gollapudi, S., Halverson, A. & Ieong, S. (2009). Diversifying search results. *In* 'Proceedings of the Second ACM International Conference on Web Search and Data Mining'. 1.6, 3.4.2.1, 8.3.2.2, 8.3.3

Aha, D. W., ed. (1997). *Lazy Learning*. Kluwer Academic Publishers. Norwell, MA, USA. 2.6.1

Amati, G. (2003). Probabilistic Models for Information Retrieval based on Divergence from Randomness. PhD thesis. University of Glasgow. 2.4.2, 2.5, 2.5.1, 2.5.1, 5.4.4.1, 6.3

Aronson, A. R. (1994). Exploiting a large thesaurus for information retrieval. *In* 'Proceedings of RIAO 1994'. 2.2.4, 3.3.1

Aronson, A. R. & Lang, F. M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*. 2.2.4, 3.3.1

Aronson, A. R. & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. *Proceedings of the American Medical Informatics Association (AMIA) Symposium*. 1.1

Averbuch, M., Karson, T., Ben-Ami, B., Maimond, O. & Rokachd, L. (2004). Context-sensitive medical information retrieval. *Studies in Health Technology and Informatics*. 3.4, 3.4.2.4

Baeza-Yates, R. A. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc.. Boston, MA, USA. 2.2, 2.2.2, 2.2.3, 2.2.4

Bai, J., Nie, J.-Y., Cao, G. & Bouchard, H. (2007). Using query contexts in information retrieval. *In* 'Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 5.3.4.1

Balog, K., Azzopardi, L. & de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. *In* 'Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.7, 2.7, 3.4.2.1, 4.3.3

Bendersky, M. & Croft, W. B. (2008). Discovering key concepts in verbose queries. *In* 'Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval'. 2.5

Blanco, R. & Lioma, C. (2012). Graph-based term weighting for information retrieval. *Information Retrieval*. 6.4

Boer, P. T. D., Kroese, D., Mannor, S. & Rubinstein, R. (2002). A tutorial on the cross-entropy method. *Annals of Operations Research*. 2.6.3

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *In* 'Proceedings of the Seventh International Conference on World Wide Web 7'. 3.4.2.3, 6.4

Broder, A., Gabrilovich, E., Josifovski, V., Mavromatis, G., Metzler, D. & Wang, J. (2010). Exploiting site-level information to improve web search. *In* 'Proceedings of the 19th ACM International Conference on Information and Knowledge Management'. 1.2, 4.2.3.3

Buckley, C. & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *In* 'Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.8.2

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. & Hullender, G. (2005). Learning to rank using gradient descent. *In* 'Proceedings of the 22nd International Conference on Machine Learning'. 2.6.3

Büttcher, S., Clarke, C. A. & Cormack, G. V. (2004). Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval (MultiText Experiments for TREC 2004). *In* 'Proceedings of the 13th Text REtrieval Conference'. 3.3, 3.3.2

Callan, J. (2000). Distributed information retrieval. *In* W. Croft, ed., 'Advances in Information Retrieval'. Vol. 7 of *The Information Retrieval Series*. Springer US. pp. 127–150. 7.3.2, 7.3.2

Cao, G., Nie, J.-Y., Gao, J. & Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. *In* 'Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 5.3.4.2

Carbonell, J. & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. *In* 'Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 1.6, 3.4.2.1, 8.3.2.2, 8.3.3

Carmel, D. & Yom-Tov, E. (2010). *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers. 6.5, 7.4.5, 8.4.4.2, 9.6.1, 9.6.2

Chang, J. T., Schtze, H. & Altman, R. B. (2002). Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*. 3.3.2

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*. 3.4.2.4, 3.4.2.4, 4.2.3.1

Chapman, W. W., Chu, D. & Dowling, J. N. (2007). Context: An algorithm for identifying contextual features from clinical text. *In* 'Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing'. 5.4.5.5

Cleverdon, C. W. (1962). Aslib Cranfield Research Project: Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Technical report. 2.8.1

Cormack, G. V., Smucker, M. D. & Clarke, C. L. A. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*. 2.6

Croft, W. B. & Harper, D. J. (1988). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation* (35), 285–295. 2.4.1

Croft, W. B., Metzler, D. & Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. 1st edn. Addison-Wesley Publishing Company. USA. 2.2.2, 2.2.3

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'. 6.5, 7.4.5, 8.4.4.2, 9.6.1, 9.6.1, 9.6.2

Demner-Fushman, D., Abhyankar, S., Jimeno-Yepes, A., Loane, R., Rance, B., Lang, F., Ide, N., Apostolova, E. & Aronson, A. R. (2011). A knowledge-based approach to medical records retrieval. *In* 'Proceedings of the 20th Text REtrieval Conference'. 1.1, 1.2, 2.2.4, 3.4.2, 3.4.2.1, 3.4.2.3, 3.4.2.4, 3.4.2.4, 5.2

Diaz, F. & Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. *In* 'Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.5, 3.4.2.3

Edinger, T., Cohen, A. M., Bedrick, S., Ambert, K. & Hersh, W. (2012). Barriers to retrieving patient information from electronic health record data: Failure analysis from the trec medical records track. *In* 'Proceedings of the American Medical Informatics Association (AMIA) Symposium'. 1.1, 1.2, 3.4.2, 3.4.2.1, 8.1

Fox, C. (1992). Information retrieval. Prentice-Hall, Inc.. chapter Lexical Analysis and Stoplists. 2.2.2

Fuhr, N. (1989). Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems*. 2.6.3

Fujita, S. (2004). Revisiting again document length hypotheses trec 2004 genomics track experiments at patolis. *In* Voorhees & Buckland (2004). 3.3.2

Ganjisaffar, Y., Caruana, R. & Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models. *In* 'Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.6.1

Garla, V. & Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinformatics*. 8.4.4.2

Geng, X., Liu, T., Qin, T., Arnold, A., Li, H. & Shum, H. (2008). Query dependent ranking using k-nearest neighbor. *In* 'Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 9.1

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*. 1.2

Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing. *Journal of the Association for Information Science and Technology*. 2.4.1

He, B. & Ounis, I. (2006). Query performance prediction. *Information Systems*. 6.5, 8.4.4.2, 9.6.1, 9.6.1, 9.6.2

Hersh, W. (2004). Health care information technology: Progress and barriers. *JAMA: The Journal of the American Medical Association*. 3.2

Hersh, W. (2008*a*). *Information Retrieval: A Health and Biomedical Perspective (Health Informatics)*. 3rd edn. Springer. 1.1, 1.2, 3.1, 3.2, 3.4, 3.4.2.2

Hersh, W. (2008*b*). Ubiquitous but unfinished: grand challenges for information retrieval. *In* 'Health Information and Libraries Journal'. 3.3

Hersh, W. & Voorhees, E. (2009). Trec genomics special issue overview. *Information Retrieval*. 3.3

Hersh, W., Bhupatiraju, R., Ross, L., Cohen, A. M., Kraemer, D. & Johnson, P. (2004). Trec 2004 genomics track overview. *In* Voorhees & Buckland (2004). 1, 3.3.2

Hersh, W., Buckley, C., Leone, T. J. & Hickam, D. (1994). Ohsumed: An interactive retrieval evaluation and new large test collection for research. *In* 'Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.5.2

Hersh, W., Cohen, A., Yang, J., Bhupatirajuand, R. & Hearst, M. (2005). Trec 2005 genomics track overview. *In* 'Proceedings of the 14th Text REtrieval Conference'. 1, 3.3.2, 5.4.4.1

Hersh, W., Hickam, D., Haynes, R. & McKibbon, K. (1994). A performance and failure analysis of saphire with a medline test collection. *Journal of the American Medical Informatics Association*. 1.1, 1.2, 3.3.1

Hersh, W., Price, S. & Donohoe, L. (2000). Assessing thesaurus-based query expansion using the umls metathesaurus. *In* 'Proceedings of the American Medical Informatics Association (AMIA) Symposium'. 2.5.2

Hersh, W. R., Cohen, A. M., Ruslen, L. & Roberts, P. M. (2007). TREC 2007 Genomics Track Overview. *In* 'Proceedings of the 16th Text REtrieval Conference'. 2, 3.4.2.3

Hersh, W., Weiner, M., Embi, P. J., Logan, J. R., Payne, P. R. O., Bernstam, E. V., Lehmann, H. P., Hripcsak, G., Hartzog, T. H., Cimino, J. & Saltz, J. H. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*. 3.2

Hiemstra, D. (2001). *Using language models for information retrieval*. University of Twente. 2.7, 3.4.2.2

Huang, X., Zhong, M. & Si, L. (2005). York university at trec 2005: Genomics track. *In* 'Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, November 15-18, 2005'. 3.3.2, 3.3.2

King, B., Wang, L., Provalov, I. & Zhou, J. (2011). Cengage learning at trec 2011 medical track. *In* 'Proceedings of the 20th Text REtrieval Conference'. 1.1, 1.2, 2.2.4, 2.5.2, 3.4.2, 3.4.2.1, 3.4.2.2, 3.4.2.3, 3.4.2.4, 3.4.2.4, 5.2, 6.6.3

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *SCIENCE*. 9.5.1

Koopman, B. & Zuccon, G. (2014). Understanding negation and family history to improve clinical information retrieval. *In* 'Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval'. 3.4.2.4, 3.4.2.4

Koopman, B., Bruza, P. D., Sitbon, L. & Lawley, M. (2011). Towards semantic search and inference in electronic medical records : an approach using concept-based information retrieval. *In* S. Khanna, A. Sattar & D. Hansen, eds, 'First Australian Workshop on Artificial Intelligence in Health 2011'. 3.4.2.3

Koopman, B., Bruza, P. D., Sitbon, L. & Lawley, M. J. (2010). Analysis of the effect of negation on information retrieval of medical data. *In* '15th Australasian Document Computing Symposium (ADCS)'. 1.1, 3.4.2.4, 3.4.2.4

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L. & Lawley, M. (2012). Graph-based concept weighting for medical information retrieval. *In* 'Proceedings of the Seventeenth Australasian Document Computing Symposium'. 3.4.2.3, 4.2

Kotsiopoulos, I. A., Keane, J. A., Turner, M., Layzell, P. J. & Zhu, F. (2003). Ibhis: Integration broker for heterogeneous information sources. *In* 'Proceedings of the 27th Annual International Computer Software and Applications Conference'. 1.1

Lease, M., Allan, J. & Croft, W. B. (2009). Regression rank: Learning to meet the opportunity of descriptive queries. *In* 'Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval'. 2.6.2, 5.3.4.1

Lempel, R. & Moran, S. (2001). Salsa: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*. 6.4, 6.4

Leveling, J., Goeuriot, L., Kelly, L. & Jones, G. J. F. (2012). Dcu@trecmed 2012:using adhoc baselines for domain-specific retrieval. *In* 'Proceedings of the 21st Text REtrieval Conference'. 1.1, 3.4.2

Limsopatham, N. (2013). A query and patient understanding framework for medical records search. *In* 'Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 1.5

Limsopatham, N., Macdonald, C. & Ounis, I. (2013*a*). Aggregating evidence from hospital departments to improve medical records search. *In* 'Proceedings of the 35th European Conference on Information Retrieval'. 1.2, 1.5

Limsopatham, N., Macdonald, C. & Ounis, I. (2013*b*). Inferring conceptual relationships to improve medical records search. *In* 'Proceedings of the 10th Open research Areas in Information Retrieval'. 1.5

Limsopatham, N., Macdonald, C. & Ounis, I. (2013*c*). Learning to handle negated language in medical records search. *In* 'Proceedings of the 22nd ACM International Conference on Information and Knowledge Management'. 1.5

Limsopatham, N., Macdonald, C. & Ounis, I. (2013*d*). Learning to selectively rank patients' medical history. *In* 'Proceedings of the 22nd ACM International Conference on Information and Knowledge Management'. 1.5

Limsopatham, N., Macdonald, C. & Ounis, I. (2013*e*). A task-specific query and document representation for medical records search. *In* 'Proceedings of the 35th European Conference on Information Retrieval'. 1.5

Limsopatham, N., Macdonald, C. & Ounis, I. (2014). Modelling relevance towards multiple inclusion criteria when ranking patients. *In* 'Proceedings of the 23rd ACM International Conference on Information and Knowledge Management'. 1.5

Limsopatham, N., Macdonald, C., McCreadie, R. & Ounis, I. (2012). Exploiting term dependence while handling negation in medical search. *In* 'Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval'. 1.5

Limsopatham, N., Macdonald, C., McCreadie, R. & Ounis, I. (2013). Learning to combine representations for medical records search. *In* 'Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 1.5

Limsopatham, N., Macdonald, C., Ounis, I., McDonald, G. & Bouamrane, M. (2011). University of glasgow at medical records track 2011: Experiments with terrier. *In* 'Proceedings of the 20th Text REtrieval Conference'. 1.5, 3.4.2.1

Lin, J. & Demner-Fushman, D. (2006). The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. *In* 'Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 3.4.2.2

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*. 1.6, 2.6, 2.6.3

Liu, Z. & Chu, W. W. (2007). Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*. 3.3

Lo, R. T., He, B. & Ounis, I. (2005). Automatically building a stopword list for an information retrieval system. *Journal of Digital Information Management*. 2.2.2

Lu, Z., Kim, W. & Wilbur, W. (2009). Evaluation of query expansion using mesh in pubmed. *Information Retrieval*. 3.3.2

Macdonald, C. (2009). The Voting Model for People Search. PhD thesis. University of Glasgow. 2.1, 2.7, 4.3.1

Macdonald, C. & Ounis, I. (2006). Voting for candidates: adapting data fusion techniques for an expert search task. *In* 'Proceedings of the 15th ACM international conference on Information and knowledge management'. 2.7, 2.7, 3.4.2.1, 4.3.1

Macdonald, C. & Ounis, I. (2011). Learning models for ranking aggregates. *In* 'Proceedings of the 33rd European Conference on Advances in Information Retrieval'. 2.7

Macdonald, C., Santos, R. L. & Ounis, I. (2013). The whens and hows of learning to rank for web search. *Information Retrieval*. 2.6.3, 9.3

Metzler, D. (2007). Automatic feature selection in the markov random field model for information retrieval. *In* 'Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management'. 2.6.3, 9.3

Metzler, D. & Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval. *Information Processing and Management*. 8.3, 8.4.1

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'. 5.2.2, 5.2.2

Metzler, D., Novak, J., Cui, H. & Reddy, S. (2009). Building enriched document representations using aggregated anchor text. *In* 'Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 4.2.3.3

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*. 2.5.2

Montague, M. & Aslam, J. A. (2001). Relevance score normalization for metasearch. *In* 'Proceedings of the Tenth International Conference on Information and Knowledge Management'. 1, 9.5.1

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press. 2.6

Mutalik, P. G., Deshpande, A. & Nadkarni, P. M. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *Journal of the American Medical Informatics Association*. 3.4.2.4

Ogilvie, P. & Callan, J. (2001). The effectiveness of query expansion for distributed information retrieval. *In* 'Proceedings of the Tenth International Conference on Information and Knowledge Management'. 7.3.2

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. & Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. *In* 'Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)'. 2.3, 4.3.1

Peng, J., Macdonald, C., He, B., Plachouras, V. & Ounis, I. (2007). Incorporating term dependency in the dfr framework. *In* 'Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 5.2.2

Porter, M. F. (1997). Readings in information retrieval. Morgan Kaufmann Publishers Inc.. San Francisco, CA, USA. chapter An Algorithm for Suffix Stripping. 2.2.3

Qi, Y. & Laquerre, P. F. (2012). Retrieving medical records with "sennamed": Nec labs america at trec 2012 medical record track. *In* 'Proceedings of the 21st Text REtrieval Conference'. 2.2.4, 2.5.2, 3.4.2.3, 4.3.4

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.. San Francisco, CA, USA. 2.6.1

Ribeiro, B. A. N. & Muntz, R. (1996). A belief network model for ir. *In* 'Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 8.3, 8.4.1

Rijsbergen, C. J. V. (1979). *Information Retrieval*. 2nd edn. Butterworth-Heinemann. Newton, MA, USA. 2.3

Robertson, S. E., 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'. 2.4, 2.4.1, 2.4.1

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. & Gatford, M. (1994). Okapi at trec-3. *In* 'Proceedings of the 13th Text REtrieval Conference'. 1.1, 2.4.1

Robertson, S., Zaragoza, H. & Taylor, M. (2004). Simple bm25 extension to multiple weighted fields. *In* 'Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management'. 1.2

Salles, T., Rocha, L., Pappa, G. L., Mourão, F., Meira, Jr., W. & Gonçalves, M. (2010). Temporally-aware algorithms for document classification. *In* 'Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.6

Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc.. Upper Saddle River, NJ, USA. 2.4.1

Salton, G. & Buckley, C. (1991). Automatic text structuring and retrieval-experiments in automatic encyclopedia searching. *In* 'Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 1.2

Salton, G. & McGill, M. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.. New York, NY, USA. 3.4.2.4

Salton, G., Allan, J. & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. *In* 'Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 1.2

Santos, R. L., Macdonald, C. & Ounis, I. (2010*a*). Exploiting query reformulations for web search result diversification. *In* 'Proceedings of the 19th International Conference on World Wide Web'. 1.6, 3.4.2.1, 8.3.2.2, 8.3.3

Santos, R. L., Macdonald, C. & Ounis, I. (2010*b*). Selectively diversifying web search results. *In* 'Proceedings of the 19th ACM International Conference on Information and Knowledge Management'. 8.3.2.2, 8.4.4

Shaw, J. A. & Fox, E. A. (1994). Combination of multiple searches.. *In* 'Proceedings of the 13th Text REtrieval Conference'. 1.6, 3.4.2.1, 9.1, 9.2, 9.6.2

Si, L. & Callan, J. (2002). Using sampled data and regression to merge search engine results. *In* 'Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 7.3.2

Silfen, E. (2006). Documentation and coding of ed patient encounters: an evaluation of the accuracy of an electronic medical record. *The American Journal of Emergency Medicine*. 1.2, 3.4.2.2

Sparck Jones, K. & van Rijsbergen, C. J. (1975). Report on the need for the provision of an 'ideal' information retrieval test collection. Technical report. 2.8.1

Srinivasan, P. (1996). Optimal document-indexing vocabulary for medline. *Information Processing and Management*. 1.1, 1.2, 3.3.1, 9.6.1, 9.6.1.2, 9.6.1.1, 9.6.1.2

Stokes, N., Li, Y., Cavedon, L. & Zobel, J. (2007). Exploring abbreviation expansion for genomic information retrieval. *In* 'Proceedings of the Australasian Language Technology Workshop 2007'. 3.3.2

Stokes, N., Li, Y., Cavedon, L. & Zobel, J. (2009). Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval* **12**, 17–50. 10.1007/s10791-008-9073-9. 2.5.2

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G. K., Elhadad, N., Pradhan, S., South, B. R., Mowery, D., Jones, G. J. F., Leveling, J., Kelly, L., Goeuriot, L., Martínez, D. & Zuccon, G. (2013). Overview of the share/clef ehealth evaluation lab 2013. *In* 'Proceedings of the 4th International Conference of the CLEF Initiative'. Lecture Notes in Computer Science. 3.3, 10.2

Tambouris, E. & Makropoulos, C. (1999). Hin6/427: Co-operative health information network in europe: The greek experience. *Journal of Medical Internet Research*. 1.1

Trieschnigg, D., Hiemstra, D., de Jong, F. & Kraaij, W. (2010). A cross-lingual framework for monolingual biomedical information retrieval. *In* 'Proceedings of the 19th ACM International Conference on Information and Knowledge Management'. 1.1, 3.3.1, 9.6.1

Turtle, H. & Croft, W. B. (1991*a*). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*. 8.3, 8.4.1

Turtle, H. R. & Croft, W. B. (1991*b*). Efficient probabilistic inference for text retrieval. *In* 'Proceedings of RIAO 1991'. 7.3.2

Tyree, S., Weinberger, K. Q., Agrawal, K. & Paykin, J. (2011). Parallel boosted regression trees for web search ranking. *In* 'Proceedings of the 20th International Conference on World Wide Web'. 2.6.1, 8.4.4, 10.2

van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*. 5.3.4.1

Voorhees, E. & Hersh, W. (2012). Overview of the TREC 2012 medical records track. *In* 'Proceedings of the 21st Text REtrieval Conference'. 1.1, 2.8.1, 2.8.2, 3.4, 3.4.1, 3.4.2, 3.4.2.3

Voorhees, E. & Tong, R. (2011). Overview of the TREC 2011 medical records track. *In* 'Proceedings of the 20th Text REtrieval Conference'. 1.1, 2.8.1, 2.8.2, 3.4, 3.4.1, 3.4.1, 3.4.2

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. *In* 'Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.5

Voorhees, E. M. & Buckland, L. P., eds (2004). *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, November 16-19, 2004*. Vol. Special Publication 500-261. National Institute of Standards and Technology (NIST). 10.3

Voorhees, E. M. & Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press. 2.1, 2.4, 2.8.1

Widdows, D. (2003). Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. *In* 'Proceedings of the 41st Annual Meeting on Association for Computational Linguistics'. 3.4.2.4, 5.4.3

Wilkinson, R. (1994). Effective retrieval of structured documents. *In* 'Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 1.2

Wu, Q., Burges, C. J., Svore, K. & Gao, J. (2008). Ranking, boosting, and model adaptation. Technical Report MSR-TR-2008-109. Microsoft Research. 2.6.3, 9.3

Wu, Q., Burges, C. J., Svore, K. M. & Gao, J. (2010). Adapting boosting for information retrieval measures. *Information Retrieval*. 2.6.3

Xu, J. & Li, H. (2007). Adarank: A boosting algorithm for information retrieval. *In* 'Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.6.3, 9.3

Yilmaz, E., Kanoulas, E. & Aslam, J. A. (2008). A simple and efficient sampling method for estimating ap and ndcg. *In* 'Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. 2.8.2, 2.8.2

Zhao, Y., Scholer, F. & Tsegay, Y. (2008). Effective pre-retrieval query performance prediction using similarity and variability evidence. *In* 'Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval'. 6.5, 7.4.5, 8.4.4.2, 9.6.1, 9.6.2

Zhong, M. & Huang, X. (2006). Concept-based biomedical text retrieval. *In* 'Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval'. 3.3.2

Zhou, W., Yu, C. & Meng, W. (2008). A system for finding biological entities that satisfy certain conditions from texts. *In* 'Proceedings of the 17th ACM conference on Information and knowledge management'. 2.5.2, 3.3.1, 3.3.2

Zhu, D. & Carterette, B. (2012). Combining multi-level evidence for medical record retrieval. *In* 'Proceedings of the 2012 International Workshop on Smart Health and Wellbeing'. 3.4.2.1, 3.4.2.3, 3.4.2.4, 3.4.2.4, 5.2, 9.6.2, 9.6.2.1, 9.6.2.2

Zhu, D. & Carterette, B. (2013). An adaptive evidence weighting method for medical record search. *In* 'Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval'. 1.1, 9.5.1

Zhu, D., Wu, S., B.Carterette & Liu, H. (2014). Using large clinical corpora for query expansion in text-based cohort identification. *Journal of Biomedical Informatics*. 3.4.2.3

Zuccon, G., Koopman, B., Nguyen, A., Vickers, D. & Butt, L. (2012). Exploiting medical hierarchies for concept-based information retrieval. *In* 'Proceedings of the Seventeenth Australasian Document Computing Symposium'. 1.2, 2