FAKULTÄT FÜR INFORMATIK
Faculty of Informatics

Mining Query Logs to enhance Patent Searching

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doctor rerum socialum oeconumicarumque

by

Wolfgang Tannebaum
Registration Number 0727849

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Prof. Dr. Andreas Rauber

The dissertation has been reviewed by:

Prof. Dr. Allan Hanbury	Prof. Dr. Michail Salampasis

Wien, 25.10.2017

Wolfgang Tannebaum

Dissertation Committee

Prof. Dr. Merkl, Dieter	Vienna University of Technology, Austria
Prof. Dr. Salampasis, Michail	Alexander Technology Educational Institute (ATEI) of Thessaloniki, Greece
Prof. Dr. Andreas Rauber	Vienna University of Technology, Austria

Dissertation accepted on 25.10.2107

Research Advisor

Ao. univ. Prof. Dr. Andreas Rauber

Declaration

I certify that the work presented in this thesis does not incorporate any material previously submitted for a degree or a diploma in any university, and that, to the best of my knowledge it does not contain any material previously published or written by another person except where due reference is made in the text. The work in this thesis is my own except for the contributions made by others as described in the Acknowledgements.

Wolfgang Tannebaum
October 2017

I dedicate this dissertation to my family and especially to my grandmother, Frieda, who could not see the end. I know that she would be so proud of me.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Dr. Andreas Rauber for his continuous support, in particular for always being reachable to share his experience and knowledge. In particular, I thank him giving me precious advices in writing conference and journal papers. Also the readability of this thesis has greatly benefited from his feedback. Furthermore, I would like to thank him for his flexibility to complete the program on a part-time basis. I also would like to thank my employer, W&H Dentalwerk Bürmoos GmbH, for the financial support throughout the study. W&H supported the attendance of several conferences. Finally, I would like to thank my family and especially my wife Nina-Berrit for her constant encouragement throughout this entire time.

Abstract

A patent document is a legal title granting its holder the exclusive right to make use of an invention for a limited area and time by stopping others from making, using or selling it without authorization. In preparing a patent application or judging the validity of an applied patent based on novelty and inventiveness, an essential task is searching patent databases for related patents that may invalidate the invention. This task is usually performed by examiners in a patent office and patent searchers in private companies. Virtually all search systems of the patent offices and commercial operators process Boolean queries as these guarantee repeatability and allow clear tracking of results obtained. But despite the importance of Boolean retrieval, there is not much work in current research assisting patent experts in formulating such queries. Currently, these approaches are mostly limited to the usage of standard dictionaries, such as *WordNet*, or lexica, like *Wikipedia*, to provide synonymous expansion terms. But the highly specific vocabulary used in the settings of patent applications, where patent applicants are permitted to be their own lexicographers, is not included in these standard dictionaries.

In this thesis we investigate the problem of query term expansion (*QTE*) in the query generation step of patent searching with the goal of suggesting relevant expansion terms, in particular synonyms and equivalents, to a query term in a semi-automatic or fully automatic manner for Boolean retrieval. The first goal of this thesis is to analyse query logs of patent experts to gain insights into the search behaviour and characteristic of patent expert's queries. We use actual query logs of patent examiners of the United States Patent and Trademark Office (USPTO). We show that query generation in patent searching is highly domain specific and that the queries posed by the patent examiners can be valuable resources to provide lexical knowledge for the patent domain. The second contribution of this thesis is to extract lexical knowledge from the query logs to support *QTE* in patent searching. We detect keyword phrases and synonyms from the query logs based on Boolean and proximity operators in the text queries and build US class-specific, class-related and class-independent lexical databases from the query expansion sessions of patent examiners at the USPTO. We then show that the lexical databases can support patent searchers in the query generation process, in particular in formulating Boolean queries. The third contribution of this thesis is to improve precision in suggesting expansion terms in a semi-automatic or fully automatic manner by ranking the expansion terms. For that we consider the US patent classification, frequencies of the expansion terms, and the word senses. We perform an evaluation of our proposed query term expansion approach on real query sessions of patent examiners. Results show that the proposed domain-specific lexical databases achieve significantly better results than the baseline and other enhanced query term expansion approaches.

Finally, we study the impact of *QTE* using synonyms on patent document retrieval. Experiments on the CLEF-IP 2010 benchmark dataset show that automatic query expansion using synonyms from USPTO patent examiners tends to decrease or only slightly improve the retrieval effectiveness, with no significant improvement. But an analysis of the retrieval results shows that synonym expansion does not have generally a negative effect on the retrieval effectiveness. Recall is drastically improved for query topics, where the baseline queries achieve, on average, only low recall values. So the approach is a valuable *QTE* method for search systems, which support repeatability and allow tracking of the results, in particular for the search systems used by the patent offices and commercial operators. So we recommend using *PatNet* as a lexical resource for semi-automatic *QTE* in Boolean patent retrieval, where synonym expansion is particularly common to improve recall.

Kurzfassung

Patente sind Schutzrechte, welche seinem Inhaber das ausschließliche Recht geben, eine Erfindung für einen bestimmten Zeitraum und für ein begrenztes Gebiet zu benutzen, d.h. herzustellen, anzubieten oder zu vertreiben. Eine wichtige Aufgabe vor der Erstellung einer solchen Patentanmeldung, aber auch bei der Prüfung der Rechtsbeständigkeit eines erteilten Patentes, ist die Recherche nach relevanter Patentliteratur in den speziellen Datenbanken der Patentämter, welche die zu beanspruchende oder bereits beanspruchte Erfindung möglicherweise vorwegnimmt. Diese Aufgabe wird in der Regel durch Patentprüfer in den Patentämtern und Patentrechercheuren in privaten Unternehmen durchgeführt.

Nahezu alle Recherchetools, die hierzu den Experten zu Verfügung stehen, insbesondere jene von den Patentämtern und kommerzielle Systeme, verarbeiten hierbei Boolesche Suchabfragen. Diese ermöglichen dem Anwender eine Wiederholbarkeit der Abfrage, aber auch eine klare und einfache Nachverfolgbarkeit der Patentrecherche. Trotz der großen Bedeutung der Booleschen Suche in der Praxis, in der Suchbegriffe mittels Booleschen Operatoren miteinander kombiniert werden, wurden bis dato wenige wissenschaftliche Forschungsansätze publiziert, die die Patentexperten bei der Formulierung solcher Abfragen, speziell in der automatischen Erweiterung von Suchanfragen während der Gestaltung und Ausformulierung der Abfragen, unterstützen. Aktuelle Ansätze zur automatischen Erweiterung, in denen lexikalischer Begriffe, insbesondere Synonyme, automatisch vorzuschlagen werden, beschränken sich hierbei vor allem auf die Verwendung von Standardwörterbüchern, wie zum Beispiel *WordNet*, oder Lexika, wie zum Beispiel *Wikipedia*. Enthalten ist in diesen allgemeinen Quellen jedoch nicht das hochspezifische Fachvokabular, welches im Patentbereich zur Formulierung der Patentanmeldungen verwendet wird. Es ist nämlich gängige Praxis, dass Patentanmelder ihre eigenen Worte erfinden, um die Erfindung zu beschreiben, um so einen größtmöglichen Schutzumfang zu definieren.

In dieser Forschungsarbeit beschäftigen wir uns mit der automatischen Erweiterung von Suchanfragen in der Patentsuche mit dem Ziel relevante Begriffe, insbesondere Synonyme und Wortphrasen, für einen Suchbegriff halbautomatisch oder vollautomatisch vorzuschlagen. Zuerst wurden protokollierte Suchabfragen von Patentexperten analysiert, um einen Einblick in das Suchverhalten und den Aufbau derartiger Abfragen zu bekommen. Hierzu haben wir echte Abfragen von Patentprüfern des US Patent- und

Markenamtes (USPTO) verwendet. Die Analyse zeigt, dass die Formulierung dieser Suchabfragen in der Patentsuche sehr speziell, insbesondere domainspezifisch, ist und dass die protokollierten Abfragen der Patentprüfer des USPTO eine wertvolle Quelle zur Extraktion von lexikalischen Wissen, insbesondere für den speziellen Patentbereich, ist.

Nach der Analyse der Suchabfragen wurde sich mit der Extrahierung von lexikalischem Wissen aus den protokollierten Abfragen beschäftigt. Basierend auf den in den Suchabfragen verwendeten Booleschen Operatoren und Abstandsoperatoren, wurden Wortphrasen und Synonyme extrahiert und patentklassenspezifische, patentklassenbezogene sowie klassenunabhängige lexikalische Datenbanken erstellt. Experimente zeigen, dass die generierten lexikalischen Datenbanken Patentexperten in der Erstellung der Suchabfragen, insbesondere bei der Formulierung von Booleschen Abfragen, unterstützen kann.

In einem weiteren Schritt wurde versucht die Genauigkeit zu verbessern, in der die Synonyme für einen Suchbegriff in halbautomatischer oder vollautomatischer Weise vorgeschlagen werden. Hierzu berücksichtigen wir die US-Patentklassifikation, die Häufigkeiten der Synonyme und Wortphrasen in den Suchabfragen, sowie die Wortbedeutungen, um die lexikalischen Begriffe nach einer Rangliste zu ordnen und abhängig vom Rang vorzuschlagen. Die Evaluierung der vorgeschlagenen Ansätze erfolgte ebenfalls an Hand der protokollierten Suchabfragen der Patentprüfer vom USPTO, also an Hand echter Patentrecherchen. Die Ergebnisse zeigen, dass die domänenspezifischen lexikalen Datenbanken signifikant bessere Ergebnisse erzielen, als die bereits bekannten Ansätze, welche patentfremde Quellen nutzen.

Schließlich wurde die Auswirkung von der vollautomaischen Erweiterung von Suchabfragen mittels patentspezifischen Synonymen, insbesondere mittels Synonymen von USPTO Patentprüfern, in der Patentsuche untersucht. Experimente basierend auf das CLEF-IP 2010 Benchmark Datensatzes zeigen, dass die vollautomatische Erweiterung von Suchanfragen mittels patentspezifischen Synonymen dazu führt, dass über alle Abfragen hinweg die Effektivität der Suche sinkt oder nur leicht verbessert wird, ohne dass eine signifikante Verbesserung bemerkbar ist. Jedoch zeigt eine genauere Analyse der Retrievalergebnisse, dass das Erweitern der Suchabfragen mit Synonymen nicht generell einen negativen Effekt auf die Retrievaleffektivität hat. Recall wurde für Suchabfragen drastisch verbessert, bei annähernd gleichen Precision-Werten, bei denen die ursprünglichen Abfrage nur geringe Recall-Werte erreichte. Der vorgeschlagene Ansatz, nämlich die Verwendung von patentspezifischen Synonymen zur Erweiterung von Suchabfragen, ist somit insbesondere für die Patentsuche geeignet, in dem in der Praxis Systeme, die eine Wiederholbarkeit der Abfrage und eine klare Nachverfolgbarkeit der

Ergebnisse erlauben, verwendet werden. Wir empfehlen somit die lexikalische Datenbank *PatNet* als Quelle für die halbautomatische Erweiterung von Suchabfragen in der Booleschen Patentsuche zu verwenden, bei der die Erweiterung von Suchabfragen mit Synonymen gängige Praxis ist.

Contents

Acknowledgements .... **v**

Abstract .... **vi**

Kurzfassung .... **viii**

Contents .... **xi**

Tables .... **xiv**

Figures .... **xvii**

1 Introduction **1**

1.1 Patents and Patent search 1

1.2 Challenges in Patent search 3

1.3 Research Questions 5

1.4 Main Contributions 6

1.5 Thesis Overview 7

1.6 Publications 9

2 Related work **11**

2.1 Introduction 11

2.2 Information Retrieval 11

2.3 Traditional Query Generation in Patent Searching 13

2.4 Automatic Query Recommendation 14

2.5 Automatic Query Term Expansion in Patent Searching 15

2.5.1 Models computing co-occurring query terms 16

2.5.2 Methods to provide keyword phrases 19

2.5.3 Approaches to suggest synonyms and equivalents 20

2.6 Mining Query Logs for IR 22

2.6.1 Query Log Analysis 22

2.6.2 Acquiring lexical knowledge 24

2.7 Conclusions 26

3 Query Logs of the USPTO **28**

3.1 Introduction 28

3.2 Access to the Query Logs 28

3.3 Nature of the Logs Files 29

3.3.1 Format and Elements 30
3.3.2 Type of Queries 31
3.3.3 Search Operators 31
3.4 Setup 32
3.5 Query Log Analysis 35
3.5.1 Basic statistic 35
3.5.2 Search Operator Analysis 36
3.5.3 Vocabulary Analysis 37
3.6 Conclusions 40

4 Acquiring lexical knowledge 42
4.1 Introduction 42
4.2 Lexical Knowledge Extraction 42
4.2.1 Detecting Keyword Phrases 43
4.2.2 Detecting Synonyms and Equivalents 44
4.3 General workflow to acquire lexical knowledge 47
4.4 Lexical Databases 49
4.4.1 *PhraseNet* 49
4.4.2 *PatNet* 50
4.5 Conclusions 54

5 Automatic Query Term Expansion 55
5.1 Introduction 55
5.2 Overall scheme of our proposed *QTE* method 55
5.3 Experiment Setup 56
5.3.1 Evaluation Sets 58
5.3.2 Evaluation 60
5.4 Automatic Query Scope Expansion 61
5.4.1 Query Term Expansion based on query log collection size 61
5.4.2 Query Term Expansion compared to *WordNet* 65
5.5 Automatic Query Scope Limitation 68
5.5.1 Query Term Expansion based on US patent classification 69
5.5.2 Considering Query Log Length in Query Term Expansion 71
5.6 Conclusions 73

6 Query Term Expansion Strategies to improve precision 75
6.1 Introduction 75
6.2 Experiment Setup 75
6.3 Considering US patent classification 76
6.3.1 Query Term Expansion across different US patent classes 76

6.3.2 Successively suggesting the *ETs* based on class information.................... 79
6.3.3 Using *Relevance Feedback t*o suggest *ET*.. 80
6.4 Using frequency information.. 80
6.4.1 Suggesting the *ETs* based on frequency 81
6.4.2 Successively suggesting the *ETs* based on frequency 82
6.5 Considering word senses of the *ETs*.. 83
6.5.1 Suggesting *ETs* based on overlap of word senses 83
6.5.2 Using different context sizes for *WSD* 84
6.6 Conclusions 85

7 Effect of log based *QTE* on Retrieval Effectiveness 87
7.1 Introduction 87
7.2 Benchmark Evaluation in the Patent Domain 87
7.2.1 Benchmark Dataset... 88
7.2.2 Evaluation.. 89
7.2.3 Best Official Results of the CLEF-IP Challenges 91
7.3 Statistical Significance 91
7.4 Baseline runs 91
7.4.1 Query Generation 92
7.4.2 Retrieval Performance 92
7.5 Query Term Expansion.. 94
7.5.1 Expanded Query Sets 94
7.5.2 Effect on Retrieval Performance 95
7.5.3 Analysis of the retrieval results 98
7.6 Conclusions 100

8 Conclusions and Future Work.. 102
8.1 Summary and Contributions... 102
8.1.1 Analyzing Query Logs of the USPTO... 102
8.1.2 Acquiring lexical knowledge from the query logs 103
8.1.3 Automatic query term expansion based on query logs............................. 104
8.1.4 Suggesting *ETs* in a useful order 105
8.1.5 Evaluation in *IR*... 105
8.2 Future Directions.. 106

Appendix.. 109

Bibliography.. 111

Tables

Table 1. Experiment Setup 33

Table 2. Query Log Statistics 35

Table 3. Co-Occurring Terms based on Operator “OR” 38

Table 4. Co-Occurring Terms based on the Operator “AND” 38

Table 5. Co-Occurring Terms based on the Operator “ADJ(cent)” 39

Table 6. Query Log Term Analysis 40

Table 7. Semantic relations provided by the query logs 42

Table 8. Number of extracted keyword phrases based on confidence values CV_{1-5} 43

Table 9. Number of extracted *STR* based on confidence values CV_{1-5} 45

Table 10. Most frequently used nouns, adjectives and verbs 45

Table 11. Synonym Relations provided by the Search Operators “OR” and “ADJ” 46

Table 12. Detected Synonyms based on the Search Operators “OR” and “ADJ” 47

Table 13. Keyword Phrases provided by *PhraseNet* across all classes for the term “control” 50

Table 14. Synonym Relations provided by *PatNet* 51

Table 15. Example of an Invention Diagram 53

Table 16. Number of query documents and growth per year 58

Table 17. Evaluation sets 59

Table 18. Recall and precision provided by the class-specific and class-independent lexical databases 62

Table 19. Coverage provided by the class-specific and class-independent lexical databases 64

Table 20. Recall and Precision achieved by the class-independent lexical database, *WordNet* and the best performing class-specific lexical database 65

Table 21. Coverage provided by the class-independent and class-specific lexical databases and *WordNet* 66

Table 22. Recall achieved by the *ciDB, WordNet* and the best performing *csDBs* without excluding out-of-vocabulary words 67

Table 23. Query Scope Limitation based on US patent class 69

Table 24. Recall achieved by the *csDBs* and the *ciDB* without excluding out-of-vocabulary words 71

Table 25. Performance of *ciDB* when considering query log length 73

Table 26. Advantages and disadvantages of the extracted lexical databases 74

Table 27. Recall, Precision and Coverage achieved when using related lexical databases *crDB* for suggesting synonyms 77

Table 28. Recall, Precision and Coverage achieved when using related lexical databases *crDB* for suggesting keyword phrases 78

Table 29. Recall and Precision achieved when suggesting *ETs* based on class information 79

Table 30. Recall and Precision achieved when using *RF* to suggest *ETs* 80

Table 31. Recall, Precision and Coverage achieved when considering frequency of the *ETs* 81

Table 32. Recall and Precision achieved when successively suggesting the highest ranked *ETs* 82

Table 33. Recall and Precision achieved when suggesting the *ETs* based on overlap of sense definitions 84

Table 34. Recall and Precision achieved using different context sizes *n* for *WSD* 84

Table 35. Advantages and disadvantages of the expansion strategies 85

Table 36. Relevant documents for the query topic “PACt-1” 89

Table 37. The performance of the best official results on English test set 91

Table 38. Baseline query set 92

Table 39. Retrieval Results of the baseline queries compared to best official results of CLEF-IP 2010 93

Table 40. Expanded Query Sets for the CLEF-IP 2010 test set 94

Table 41. Retrieval Results when using *PatNet* for *QTE* and querying the baseline query sets and the expanded query sets simultaneously 95

Table 42. Retrieval Results when using *PatNet* for *QTE* and querying the baseline query sets and the expanded query sets separately 96

Table 43. Retrieval Results when considering patent classification of the query topics and the classes *PatNet* was extracted from 97

Table 44. Percentage of topics which show improved, unchanged, and degraded performance compared to QS-*BL* using QS-*BLE*. 98

Table 45. Recall achieved for improved, unchanged and degraded query topics 99

Table 46. Query topic, query and citation characteristics... 99

Table 47. Rank positions of the retrieved relevant documents provided by *QS-BL*..... 100

Table 48. Achieved recall measures of the *csDB* providing synonyms when used for the class they were based upon and for the other classes 109

Table 49. Achieved recall measures of the *csDB* providing keyword phrases when used for the class they were based upon and for the other classes............................ 109

Table 50. Number of synonyms provided by the lexical databases *csDB[1-5]* 110

Table 51. Number of keyword phrases provided by the lexical databases *csDB[1-5]*. 110

Figures

Figure 1. Invention diagram.......... 13

Figure 2. Query log for the USPTO patent application with number 10/519347.......... 29

Figure 3. Text query length analysis.......... 36

Figure 4. Search Operator Popularity 37

Figure 5: The general workflow of our proposed approach to acquire lexical knowledge from query logs of USPTO patent examiners.......... 48

Figure 6. Using *PhraseNet* for query scope limitation.......... 49

Figure 7. Provided single term to single term relations for term “*tube*” 51

Figure 8. Suggested single term to phrase relations for term “*airbag*” 52

Figure 9. Phrase to phrase relations for the phrase “*electromagnetic shield*” 52

Figure 10: Overall scheme of our proposed *QTE* method.......... 56

Figure 11: Vocabulary used for querying Keyword phrases.......... 57

Figure 12: Vocabulary used for querying Synonyms.......... 58

Figure 13. Test set length analysis........... 72

Figure 14: Example of a query topic selected from CLEF-IP 2010.......... 89

1 Introduction

1.1 Patents and Patent search

The importance of intellectual property rights, in particular of patent rights, is growing faster than ever in the world of e-economy. Private companies have recognized this and acquire, utilize, and manage these rights. They obtain and hold patents not only for image purposes. They use the intellectual property rights as a tool to gain an advantage over their competitors.

Generally, a patent document is a legal title granting its holder the exclusive right to make use of an invention for a limited area and time by stopping others from making, using or selling it without authorization in exchange for a detailed public disclosure of the claimed invention. Patent applications and granted patent documents have a well-defined structure. They include the following sections:

Front Page: The front page contains bibliographic data. This includes patent title, priority and filing date, grant date, the name of the inventor and applicant, the classes and subclasses assigned by the patent office to the document.

Abstract: The abstract provides a brief summary of the invention.

Description: The description is a lengthy written description of the underlying invention. It provides context for the invention and describes how persons of ordinary skilled in the art can make and use the invention.

Claims: The claims are most relevant. They define the scope of protection provided by the patent. The granted patent shows the claims allowed by the patent examiners.

Drawings: The drawings provide details of the claimed invention.

References: Both the applicant and the examiner may cite patent references as prior art. To find all references cited by the patent examiners the examiner’s office actions have to be red [37].

In preparing a patent application or judging the validity of an applied patent based on novelty and inventiveness, an essential task is searching patent databases for related patents that may invalidate the invention. National patent offices provide web interfaces, like “Depatisnet” of the German Patent and Trademark Office, “Espacenet” of the European Patent Office (EPO), or “PatFT” and “AppFT” of the US Patent and Trademark Office (USPTO) for searching their patent databases containing millions of patent documents (the original documents in PDF format or at least the corresponding metadata). Additionally, free providers, such as Google Patent, and commercial operators, usually add value to the information already provided by the non-commercial providers.

The data coverage of the patent databases goes back to the 19th century. For example, the European Patent Office archives patent information published by the German Patent and Trademark Office since 1877, or patents published by the USPTO since 1836. Until today, this historical patent information plays an important role in the decision whether to file a patent or to assess the patentability of an applied patent. For Example, at the end of the 20th century the automobile manufacturer “Mercedes” started selling a roaster having a retractable hardtop. Many people were amazed about this new type of convertible that forgoes a textile roof. But this idea was not particularly new. E.g. the patent document US 2,007,873 filed in December 1932, more than sixty years before, discloses a motor vehicle having a rigid moveable body top, which is made of one or more elements capable of being retractable.

The process of patent searching differs significantly, depending on the purpose of retrieval. Following we introduce the several types of searches used by patent searchers to help assess patentability, validity, infringement, clearance and state-of-the-art:

Patentability search is conducted prior to the filing of a patent application. The search helps patent attorneys to determine whether an invention can be patented.

Invalidity search is used to determine absolute novelty at the time of invention. The validity search is conducted after publication of the patent application. With this search the patent claims are validated against all prior art.

Infringement search is carried out to determine whether an enforceable patent claims the same matter as a concept or unpatented invention.

Clearance search is used to determine whether a party has “clearance” to make, use, and sell an inventive concept.

State-of-the-art search is a comprehensive search of all available patent and non-patent literature. The search does not focus a single invention, but assembles all references that relate to a defined technical field.

Patent landscape is a comprehensive state-of-the-art search. The patent landscape search is a deeper analysis of patent and non-patent references after completion of the state of the art [37].

In summary, patent searching is generally a patent-to-patent associative retrieval task and usually performed by examiners in a patent office and patent searchers in private companies. For interested readers, a more detailed introduction into patent searching is presented in [1] [57].

In the following section we explain our motivations for pursuing this line of research.

1.2 Challenges in Patent search

Information Retrieval in the patent domain poses several unique challenges. Patent retrieval specialists, in particular patent examiners from patent offices and patent searchers in private companies, have to search patent databases containing millions of patent documents to judge the validity of a patent applied for and patents that may invalidate a granted patent because of lack of novelty and inventiveness. Virtually all search systems of the patent offices and commercial operators process Boolean queries. This is not because this kind of retrieval is the most effective one. Rather, Boolean queries are easy for patent experts to manipulate and they provide a record of what documents were searched. Hence, the use of Boolean operators is one of the most important features to formulate effective queries from the patent experts point of view [46]. But despite the importance of Boolean retrieval in patent searching, there is not much work in current research assisting patent experts in formulating such Boolean queries, preferable via semi-automatic or fully automatic query term expansion (*QTE*). Hence, there is great demand in this area of research. An overview of further current challenges in patent searching, such as grasping the patent claim structure a presented in [74], is given in [58].

In this thesis we focus on *QTE*, which is a challenging task in the query generation process, in particular in formulating Boolean queries, and a critical step in patent searching.

The starting point of *QTE* in the query generation process of patent searching is a so-called *invention diagram* specifying the searchable features of an invention selected from the patent document (query document) or an invention report [37]. The challenge is now to brainstorm additional query terms, in particular synonyms and equivalents, to the searchable features of the invention to formulate a complete Boolean query set to search the patent databases. Equivalents are different from synonyms. Equivalents describe alternate parts or steps that will make the invention work the same way and serve the same purpose. They can be explained as alternate embodiments of the invention. For example, if the invention is an instrument for writing the substance for writing could be: “ink”, “graphite”, or “wax”. Glue may not be an appropriate equivalent for writing [37].

Following we list some of the reasons that make the creation of this invention diagram a challenging task:

- For the patent domain no lexical sources providing synonymous expansion terms, in particular patent domain specific lexica or thesauri, such as *WordNet* or *Wikipedia* for general information retrieval, are available to assist patent searchers in formulating Boolean queries. Hence, the probability to miss relevant expansion terms is high.
- To describe the inventions in the patent documents patent applicants use highly unusual vocabulary to avoid narrowing the scope of protection of their patent rights [37]. In addition, they are permitted to be their own lexicographers, i.e. they can define their own terminology. Brainstorming of synonyms and equivalents to expand the search is of great importance, but also a difficult task.
- Because of the highly unusual vocabulary used by the patent applicants and the missing lexical patent domain specific resources, the query term expansion process is time- and cost-intensive for patent offices and private companies having their own patent searchers.
- The goal of patent searching is to retrieve all relevant documents to query document, in particular to avoid infringement of patent rights. Hence, providing a complete set of expansion terms to the query terms is essential.

In addition, we mention some useful characteristics of the patent domain:

- Patent documents have a well-defined structure according to national patent laws. Most patents have a title page containing bibliographical data and the abstract, a description of the state of the art and the invention, and a claim section. Optionally, images or diagrams can be attached to the patents. The claim section is the focal point of a patent disclosures. The subject features in the claim text describe the invention and define the actual subject of legal protection.
- Patent documents are classified in technological domains. Hence, the patent classification system can be used to carry out class-specific *QTE*.

- Various techniques have been introduced in previous work for automatic *QTE* in the patent domain. These studies focus on suggesting expansion terms to the searchable features of the invention extracted from patent documents based on statistical measures. But there are further relevant documents available for the patent domain, which can be utilized as lexical resources for automatic query term expansion. For example, the national patent register, such as the one of the United States of America or the European Patent Register, publish the examination procedures, in particular the communications between the patent office and the patent applicants. Yet, these documents have not been considered in IR for the patent domain.

1.3 Research Questions

In this thesis we address the following research questions:

RQ1 How can we assist patent searchers in formulating Boolean queries? More detailed sub-questions are:

(1) What type of expansion terms and semantic relations are used by the patent searchers for query term expansion in real sessions?
(2) What are the most frequently used expansion terms and semantic relations?

RQ2 How can we leverage query logs of patent examiners for automatic query term expansion? This research question includes to the following detailed sub-questions:

(1) Can we use the query logs to extract lexical knowledge directly from the patent domain?
(2) How can we assist patent searchers in query term expansion to formulate Boolean queries based on this extract lexical knowledge?

RQ3 How does an automatic query expansion strategy based on query logs perform in query term expansion compared to the manual expansion performed by experts? This research question leads to the following sub-questions:

(1) Can we evaluate our query term expansion approach based on real query expansion scenarios?
(2) Does the query log based query expansion approach outperform standard dictionaries?

RQ4 How can we optimize the query log based query term expansion strategy to carry out effective query term expansion? We break this research question into following sub-questions:

(1) Are there weights available with the query logs to suggest expansion terms to a query term in a useful order?

(2) Does the involvement of information from past queries improve the query log based query term expansion model?

RQ5 Does a query log based expansion approach improve retrieval effectiveness in patent searching? More detailed sub-questions are:

(1) Does log based query expansion outperform standard approaches which are commonly based on terms selected from patent documents?

(2) Can a log based query expansion approach assist related expansion approaches to improve the retrieval performance?

1.4 Main Contributions

In this section we outline the main contributions of this thesis:

- We present an analysis of query sessions of patent experts to gain insights into the search behaviour and characteristic of patent searchers queries. We use query logs of patent examiners of the United Patent and Trademark Office (USPTO).
- We propose approaches to extract lexical knowledge from the query logs of the USPTO patent examiners. In particular, we extract from the query logs the query and expansion terms to the patent applications and detect keyword phrases and synonyms in the query logs.
- We present patent domain-specific, in particular US class-specific, class-related and class-independent lexical databases. The lexical databases provide patent domain specific vocabulary and semantic relations to assist patent searchers in formulating Boolean queries.
- We use the extracted lexical databases for automatic *QTE*. For the evaluation we automatically expand the query terms from real query sessions of patent examiners (gold standard). To the best of our knowledge no previous work has used query logs of patent experts to assist patent searchers in query term expansion.
- We propose query term expansion strategies for our expansion approach to improve precision in suggesting expansion terms in a semi-automatic or fully automatic manner.
- We present the effect of the query log based query term expansion approach, in particular when expanding queries with synonyms on retrieval effectiveness in patent searching.

1.5 Thesis Overview

This thesis consists of five main research chapters, in particular Chapters 4 to 8. Each chapter addresses the research questions as mentioned in Section 1.3. Chapters 1 to 3 serve to introduce the reader into the subject of patent searching and query log mining for information retrieval (IR). Furthermore, the reader will be familiarized with query logs of patent examiners of the United Patent and Trademark Office (USPTO). In Chapter 8 we present the final remarks. In particular:

Chapter 2. We provide an introduction to general IR and Query Log Mining. Further, we introduce patent searching, in particular the specific characteristics of *QTE* in a manual, semi-automatic and fully automatic manner and present the related work in this area of patent searching. Finally, we highlight the challenges that were not paid sufficient attention to in the related work.

Chapter 3. We present query logs of patent examiners of the USPTO, which serve as the basis of this thesis. We first explain where and how the log files have been made available. We describe the nature of the query logs, in particular the format and the elements of the log files. We define the search queries, in particular the kind of queries and available search operators. Specifically, we utilize the text queries to analyze the used vocabulary and search operators to find out what type of query and expansion terms and semantic relations are used by patent searchers for *QTE*. This chapter provides answers to the research question RQ1.

Chapter 4. We propose approaches to extract lexical knowledge from the query logs to support query term expansion in patent searching. We detect keyword phrases and synonyms, in particular several types of synonym relations, from the query logs based on the extensive usage of Boolean and proximity operators in the text queries. Our goal is to support patent searchers in the query generation process, in particular in generating the so-called invention diagram. To this end, we build two lexical databases, which we call *PhraseNet* and *PatNet* and which are based on the extracted lexical knowledge. We show that the lexical databases can support patent searchers in the query generation process, in particular in suggesting keyword phrases and synonyms or equivalents in a semi-automatic or fully automatic manner. This chapter provides answers to the research question RQ2.

Chapter 5. In this chapter our goal is to evaluate the performance of the proposed query term expansion approach based on real query expansion scenarios. To this end, we split the query log collection into a test set and a set for generating the lexical databases. The second set is further divided into subsets to extract multiple lexical databases for each class to evaluate size and class dependency characteristics in automatic query scope expansion and limitation. We show that the patent domain specific lexical databases drastically outperform general-purpose sources, such as *WordNet*. In addition, with a larger number of query logs for a specific patent US class available, the performance of the extracted lexical databases increases. Finally, we considered query log length in automatic *QTE*. We find out that the performance of the lexical databases is independent from the length of the query sessions. Research question RQ3 are addressed in this chapter.

Chapter 6. In this chapter our goal is to optimize the query log based query term expansion model to carry out effective query term expansion. In particular, we present approaches to improve precision in suggesting expansion terms in a semi-automatic or fully automatic manner by ranking the expansion terms. For that we consider (1) US patent classes and expand class-specific lexical databases with related classes, (2) frequencies of the semantic relations in the query logs, (2) successively suggest expansion terms according to their frequency in the query logs, and (3) use information from past queries to carry out word sense disambiguation of the expansion terms. Experiments show that the precision scores can be drastically improved and compared to the precision scores achieved by the standard query term suggestion approaches for patent searching and academic professional search the extracted lexical database achieves significantly better precision scores. This chapter provides answers to the research question RQ4.

Chapter 7. In this chapter we measure the effect of query term expansion using synonyms on retrieval effectiveness in patent searching. All experiments are performed on the CLEF-IP 2010 benchmark data set. The experiments show that the retrieval performance of the query generation and expansion models presented in this work is decreased or only marginally improved when using synonyms and equivalents for query term expansion. No significant improvement is recognized. But the analysis of the retrieval results shows that the query log-based query term expansion method does not have generally a negative effect on the retrieval effectiveness. Recall is drastically improved for query topics where the baseline queries achieve, on average, only low recall values. But we have not detected any commonality that allows us to characterize these queries. So we recommend to use synonyms and equivalents for semi-automatic *QTE* in Boolean patent retrieval, where synonym expansion is particularly common to improve recall (as shown in Chapter 3). This chapter provides answers to the research question RQ5.

Finally, the conclusions and a list of possible future directions are reported in the final Chapter 8.

1.6 Publications

This thesis is based on a number of published conference and journal papers. Chapter 3 in which an analysis of query logs of USPTO patent examiners is presented to gain insights into the search behaviour and characteristic of patent examiners queries is based on the following works:

- Tannebaum, W., Rauber, A. 2012. *Analyzing Query Logs of USPTO examiners to identify useful Query Terms in Patent Documents: A Preliminary Study*. In Proceedings of the Information Retrieval Facility Conference (IRFC 2012), Vienna, Austria.
- Tannebaum, W., Rauber, A. 2013. *Mining Query Logs of USPTO Patent Examiners*. In Proceedings of 4th International Conference of the CLEF Initiative (CLEF 2013), Valencia, Spain.

Chapter 4 which presents approaches to acquire lexical knowledge from the query logs and to create lexical databases from the query logs for automatic query term expansion in patent searching is based on the following studies:

- Tannebaum, W., Rauber, A. 2012. *Acquiring lexical knowledge from Query Logs for Query Expansion in Patent Searching*. In Proceedings of the 6th IEEE International Conference on Semantic Computing (IEEE ICSC 2012), Palermo, Italy.

Automatic query term expansion experiments, which are based on query expansion sessions done by patent examiners of the USPTO are presented in Chapter 5 and based on the following publications:

- Tannebaum, W., Rauber A. 2014. *Using Query Logs of USPTO Patent Examiners for automatic Query Expansion in Patent Searching.* In Information Retrieval, Volume 17, Issue 5-6, pp. 452-470.
- Tannebaum W., Rauber A. 2015 *Learning Keyword Phrases from Query Logs of USPTO Patent Examiners for Automatic Query Scope Limitation in Patent searching.* In World Patent Information, Volume 41.

Chapter 6 which presents approaches to improve the precision scores in suggesting expansion terms in a semi-automatic or fully automatic manner is described in:

- Tannebaum, W., Rauber. 2015. *PatNet: A lexical database for the patent domain.* In Proceedings of the 37th European Conference on Information Retrieval (ECIR 2015), Vienna, Austria.

Chapter 7 in which we measure the effect of log based query term expansion on retrieval effectiveness in patent searching using the CLEF-IP 2010 benchmark data set is based on the following publication:

- Tannebaum, W., Mahdabi, P. and Rauber, A. 2015. Effect of log-based Query Term Expansion on Retrieval Effectiveness in Patent Searching. In Proceedings of 6th International Conference of the CLEF Initiative (CLEF 2015), Toulouse, France.

This thesis also includes material that is not based on query logs of patent examiners [95]. The conference paper is based on patent documents published by the European patent office (*EPO)*:

- Tannebaum, W., Rauber, A. 2010. *Query Expansion for Patent Retrieval using Domain Specific Thesaurus*. In Proceedings of the 2010 Conference on the Interaction of Information Related Rights, Information Technology and Knowledge Management (KnowRight 2010), Vienna, Austria.

~ ~

2 Related work

2.1 Introduction

In this chapter we present the related work for this thesis. First, we provide an introduction to Information Retrieval (IR), in particular to the terminology of IR in Section 2.2. In Section 2.3 we describe the traditional query generation process in patent searching. Specifically, we take a closer look at the query expansion step. Then, we discuss previous approaches for *QTE* in a semi-automatic or fully-automatic manner for the patent domain in Section 2.4. In Section 2.5, we present existing work on mining query logs to enhance query generation for IR. Finally, in Section 2.6, we present our conclusions.

2.2 Information Retrieval

Information retrieval (IR) is part of computer science with the aim of obtaining information relevant to an information need from a collection of documents or other data. The information need of users to which the answer is a set of documents is expressed by queries. Search engines process the queries with the purpose to retrieve all the *relevant* documents in the document collection, in particular at the same time, and retrieving as few of the *non-relevant* as possible [101].

The success of the search engines depends primarily on the expression of the information need, in particular on formulating effective queries. But it is not always easy for users to formulate such effective queries representing their information need. Amongst others, the reasons are the following:

- Ambiguous queries retrieve documents which are not relevant for the current information need.
- Too precise queries will not cover the information need and relevant documents will not be retrieved by the search engine.
- Users are not always familiar with the terminology of the specific domain they are searching for.

Several approaches have been made available to help users to express their information need, in particular via automatic *QTE* to cover the information need, in query-based information retrieval. Additional query terms, in particular synonyms and co-occurring terms, will be added to the original query terms to increase the performance of the search engine. An overview of related techniques for automatic Query Expansion in IR is provided in [21]. These approaches can be roughly classified into the following five techniques:

Linguistic analysis: These techniques are based on dictionaries, thesauri, or other knowledge sources, such as *WordNet*. Expansion terms are generated independently from the query and the document collection. These techniques can be grouped into three main approaches: (1) using word stems, (2) ontology browsing (most of the related work has focused on the use of *WordNet*), and (3) syntactic analysis (relations between query terms are extracted, which then can be used to identify expansion terms) [18] [71] [103];

Corpus-specific (global) techniques: The techniques in this category analyze the contents of a database to identify correlations between pairs of terms by exploiting term co-occurrence in documents, paragraphs, or sentences. For example, thesauri are built using interlinked *Wikipedia* articles [81].

Query-specific (local) techniques: Query-specific techniques take advantage of the local context provided by the query. Query specific techniques typically make use of top-ranked documents. They preprocess the top retrieved documents for filtering out irrelevant features and for finding informative document representations.

Search log analysis: The idea is to mine query associations that have been implicitly suggested by users. There are two main techniques based on search logs. The first is to extract features from the user queries with or without making use of their associated retrieval results. The second technique consists of exploiting the relation of queries and retrieval results to provide additional or greater context in finding expansion terms [35].

Web data: A common web data source for automatic *QTE* is represented by anchor texts. Techniques use, for example, *Wikipedia* documents and hyperlinks, specific categories of *Wikipedia* articles, or other types of web data [21].

In the following sections, we will take a closer look at traditional query generation in patent searching and review previous approaches for automatic *QTE* in patent searching and based on query logs.

2.3 Traditional Query Generation in Patent Searching

Just as general information retrieval patent searching consists of three phases, in particular query generation, document retrieval, and document reviewing. At first, appropriate query terms have to be selected and combined to formulate a complete query set. Then, the patent databases of the national patent offices or commercial operators are searched. Finally, the retrieved documents are reviewed to select the relevant ones [37].

Specifically, to scope the search, patent searchers follow a strict scheme including the following three steps:

Identifying subject features: The first step is to compartmentalize the invention into searchable features. The searchable features selected from the source document, particularly from a patent document or an invention report, are used to create a so-called invention diagram.

Brainstorming additional query terms: The generated invention diagram serves as a template to brainstorm additional query terms. It is a way to capture all query terms that can be associated with the invention.

Preparing initial text queries: Finally, the query and expansion terms of the invention diagram and search operators are used to assemble initial search queries, which are modified throughout the search. Typically, search operators are Boolean operators, proximity operators and truncation limiters [37].

Figure 1 shows an example of an expanded invention diagram including the searchable features of an invention completed with expansion terms as they are used for query generation by the patent searchers.

Features	**Expansion Terms**
voice	audio, speech, mail, message, verbal, …
sensor	indicator, monitor, chemical, force, …
module	control, terminal, computer, station, …
transmitter	radio, infrared, ultrasonic, transmit, send, signal, …

Figure 1. Invention diagram

The first column includes the searchable features of the invention selected from the source document. The second column provides the corresponding expansion terms. The terms are (1) synonyms or equivalents, such as ‘‘speech’’ for ‘‘voice’’, (2) co-occur in the source document, for example ‘‘signal’’ with ‘‘transmitter’’, or limit a feature of the invention to a (3) keyword phrase, such as ‘‘force sensor’’ for ‘‘sensor’.

In particular, the expansion of the query terms which belong to parts of the original text with synonyms and equivalents to expand the query scope and keyword phrases to narrow the search is a crucial task. This process is very time-intensive and the probability to miss relevant expansion terms is high. Hence, it is essential to provide assistance in identifying these additional query terms to refine the search.

In the following section we present related work, specifically for the patent domain, which addresses the recommendation of queries, in particular the expansion of query terms with additional query terms in a semi-automatic or fully automatic manner.

2.4 Automatic Query Recommendation

After reviewing traditional query generation in patent searching, we now will give an overview of related approaches in the field of *Query Recommendation*. These approaches can be mainly grouped into *Query Expansion* and *Query Suggestion*:

Query Expansion consists of adding one or several terms to the original query, specifically to increase the precision of the search engine by narrowing the query scope.

Query Suggestion is mainly a way to provide a list of queries that have been proven to be effective for expert users. The search is expanded [86].

For example, in [29] is described one of the first approaches in the field of query expansion, which makes use of previous queries and is called past-query feedback. Expansion terms are selected from the resulting top scoring documents based on frequency information (tf/idf).

Pseudo-Relevance Feedback (PRF), as presented in [105], is a common technique used to expand queries. The top-ranked documents retrieved for a query are analyzed. Expansion terms are than selected based on co-occurrence with the query terms within the top-ranked documents.

Other approaches, such as presented in [25], consider the clicked documents and generate correlations among the query terms and the terms appearing the clicked documents. In addition, each link, which is generated between the document terms and the query terms is then weighted. In [9] a list of related queries is suggested based on a query clustering process, in which groups of semantically similar queries are identified. The clustering process uses the content of historical preferences, in particular user preferences in the form of clicks stored in the query logs of the search engine.

Most approaches to detect related queries for automatic query suggestion are based on measuring query similarity. For example, in [108] to find similar queries the approach consists of looking for those queries sharing query terms. Further approaches, such as described in [30], suggests those queries appearing frequently in the same query sessions to measure query similarity. Also the click-through data information is used to devise query similarity, as proposed for example in [109]. Whole query sessions have been considered for finding related queries for query suggestion. For example in [30], the basic idea is that if users issue a first query and a second query afterwards, the second query is suggested for the first query.

In [43], to measure query similarity for query suggestion, the interdependencies between query terms are measured. The process is based on a Log-Likelihood Ratio (LLR). Queries with high log likelihood ratio are related to each other.

2.5 Automatic Query Term Expansion in Patent Searching

In this subsection we discuss approaches for *QTE* in a semi-automatic or fully-automatic manner for the patent domain. A more general overview of the recent literature explaining document processing and retrieval methods for the patent domain is presented in [57].

Recent surveys of patent users show that *QTE* in the query generation of patent searching is seen as very important with respect to the information retrieval process [8] [42] [7]. Patent searchers spend hours and also days to find all possible relevant documents to a query topic. In particular, the use of Boolean operators is one of the most important features to formulate effective search queries [46]. Despite the importance of Boolean retrieval in patent searching, as also mentioned in Section 1.2., there is not much work in current research assisting patent experts in formulating such Boolean queries. Techniques proposed in the patent domain to enhance query generation, preferably via automated *QTE*, did not investigate Boolean query generation and expansion, while virtually all search systems of the patent offices and commercial operators process Boolean queries (but do not support *QTE*).

At first, we review *QTE* approaches, which are based on computing co-occurring query terms. Most of them are not able to expand Boolean queries. So these approaches are not easy to use for Boolean query term expansion. After that we discuss methods to provide keyword phrases followed by approaches to suggest synonyms and equivalents in a semi-automatic or fully-automatic manner.

2.5.1 Models computing co-occurring query terms

Several techniques have been proposed in the patent domain to enhance query generation, preferably via automated *QTE*. Currently, additional query terms are extracted automatically from the query documents, related documents, the feedback documents or from the cited documents based on statistical measures, such as term frequencies (tf) and a combination of term frequencies and inverted document frequencies (tfidf) [83]. Also, whole documents or whole sections of the query documents, like the title, abstract, description or the claim section are used for query generation and query expansion.

In particular, the whole patents were considered as the query [45] [76] [107]. The query documents had been split based on document length [76]. Due to the huge amount of text in a full patent, text from certain fields has been extracted to create the query from the patent topic [62]. Specifically, the short text fields of the patent applications, such as the title or the abstract, and the full patent description were used as the query. Also the various claim sections were used to formulate the queries, as the claim field is the legally important field [31] [36] [49]. Only the abstracts of the query documents were also used as the input query [17]. But in addition to the previous approaches (using the whole sections) the long input queries were broken into multiple shorter sub-queries comprising co-occurring query terms. Then the shorter queries were used with proximity operators in the retrieval process. The results of each sub-query were combined to a common final result list. In addition, queries are constructed by combining co-occurring terms from different sections of a query patent [22]. 20 or 30 query terms extracted from all fields of an original query patent based on values, in particular according their log(tf)idf values, are used to form a search query.

Experiments where user submits a whole patent as the query instead of selecting keywords, show that the summary of a patent is the most useful source of terms for generating a query, even though most previous work use the patent claims [36] [31] [49] [107]. Further experiments show that combining terms extracted from different fields of the query patent by giving higher importance to terms extracted from the abstract, claims, and description fields than to terms extracted from the title field is more effective than treating all extracted terms equally while forming the search query [22]. Best retrieval results were obtained by treating patents as a full document [62]. But 6 times the processing time is required. So the computational cost is much higher. More than 700% faster query response times can be achieved compared with the related methods for very long queries, when braking long input queries, in particular the abstract sections, into multiple shorter sub-queries [17]. Position information of the document terms can be computed during indexing, but distances between the query terms can only be computed at run time. So the response time of the system increases with the length of the queries. So it was proposed to use multiple shorter sub-queries and to eliminate computing of distances between the query terms [17].

The most frequent (top-10) terms in the query documents were used according to the (tfidf), as mentioned in [83], for query generation [14] [104] [106]. But compared to [104] in [73] [106] the sections of the patent documents have been considered: The extracted query terms have been ranked based on frequency scores in each section and ranked lists of query terms have been generated for each section of the query document. In particular, only the claim sections were used to extract co-occurring query terms based on statistical measures [90].

All the studies show that using frequency information to select query terms from the query documents, in particular to assign *Weight* to each query word, significantly improves retrieval performance. In particular, experiments in [106] show that for *Weight* the term frequency (*tf*) is the best weighting method. Highest recall measures were obtained in the experiments when using *tf* for weighting.

In addition, there are approaches where citation information was used to improve the text-based queries [31]. In particular, text and citation information were combined. First they perform the text-based retrieval and obtain top *N* documents. Then they compute a citation-based score for each of the *N* documents and combine the text-based and citation-based scores and re-sort the *N* documents. They assume that a cited patent is important when the patent is cited by a large number of other patents. As a citation-based method *PageRank*, which estimates the probability that a user surfing on the Web visits a document, and a topic-sensitive method was used. Here, only citations among the top *N* documents were used. In *PageRank*, the citation-based score is determined by the total votes. Experiments on a USPTO patent document collection show that the combination of the text-based and citation-based approaches improved the text-based method. The improvement was even greater when using the topic-sensitive citation-based method. This method provided, on average, best recall and precision measures. Experiments thus show that considering the citations of the top ranked documents help improving retrieval effectiveness.

As mentioned above also feedback documents were used as source to extract additional query terms. In this way, missing terms of query documents are retrieved from related documents. There is a mechanism specifically designed for patent search [47]. Experiments did not produce any significant improvement in retrieval results about the baseline queries [47] [61]. According to the authors the reason may be that all words from the feedback documents were used without any selection process. Only those terms which appear closely with terms of the initial queries in the same claim, paragraph, sentence or phrase were used to identify better patents for *PRF* as compared to using all terms of a query patent [16] [33].

But experiments show that an increase in the retrievability of individual patents can be obtained, when missing terms from query patents are discovered from feedback patents [13] [15]. Initial query terms were extracted from the claim sections of the query documents. All frequent terms (*minimum frequency* ≥ 3) were considered in the claim sections (two, three and four terms combinations were constructed for longer length queries). In addition, they constructed a related document set for each patent document in the collection using a k-nearest neighbor approach and generated an additional set of queries each including co-occurring query terms based on each of these sets of related documents.

To summarize, it can be noted that the related approaches to compute co-occurring query terms address the research question where to extract query and expansion terms, in particular from the query documents, the cited documents or from the feedback documents to improve retrieval effectiveness. They differ in terms of the document section, which should be used for query term extraction. What they have all in common is that they use patent documents, in particular patent applications, as lexical resources. They all propose to use frequency information, in particular term frequencies (tf) and a combination of term frequencies and inverted document frequencies (tfidf) to weight the query and expansion terms.

Because all these approaches are not able to generate Boolean queries, while we focus on supporting patent searchers in formulating Boolean queries, these approaches are not suitable for Boolean query term expansion. An approach which uses co-occurring terms to expand a query term, in particular to suggest Boolean queries is presented in [46]. Specifically, Boolean queries are generated based on unigrams or bigrams extracted from pseudo-relevant documents and based on the Boolean operators AND and NOT. Human judgments are used to evaluate the suggested expansion terms. Experiments showed that Boolean queries can be generated. In particle, about 200 queries are generated for each search topic, of which about only 10 queries are assessed as relevant.

Because the previous experiments show that it is difficult to formulate effective Boolean queries based on patent documents (only about 5% precision is achieved) and synonym expansion is scheduled for future work (no approach to detect synonyms in patent documents is provided), assisting patent experts in formulating Boolean queries requires further research. In particular, to improve precision in suggesting relevant Boolean query and expansion terms (as mentioned in RQ4 and addressed in Chapter 6) and to provide synonyms is required.

2.5.2 Methods to provide keyword phrases

Aiming at solving the limitation of traditional keyword search, which provides limited capabilities to capture the information need of the searchers, also in the patent domain current works focuses on semantic search (searching by meanings rather than literal strings) to improve retrieval effectiveness.

In the retrieval of keyword phrases for *QTE* in patent searching, particularly to narrow a search, as well as for automatic document categorization, keyword phrases are generally extracted automatically from the patent domain, specifically from the patent documents, using natural language processing applications and statistical measures. Additional phrases have been extracted from external lexical resources.

In particular, all words and noun phrases of the patent sections (title, abstract, description, and the claim section) have been extracted and ranked based on frequency scores in each section [106]. Experiments based on a USPTO patent collection demonstrated that the single best search feature is the combination of words and noun-phrases from the summary field. In further experiments all claim sentences of the patent document collection were Part-of-Speech tagged and noun phrases were extracted based on patterns including noun phrases with preposition „of" and participle used as adjectives [5]. In addition, the term frequency was used as weight technique. Further all extracted words were lemmatized via *WordNet* and patterns used in [5] have been reused with additional patterns including noun phrases with prepositions and participles used as adjectives in [6]. 2,288 multi-word phrases have been extracted.

Furthermore, patent corpus statistics and linguistic heuristics have been used for finding meaningful noun phrases [4] [64]. Candidate noun phrases with a length of at most 5 terms have been extracted from the query patent, with the help of the Stanford part of speech tagger. A Part-of-Speech Tagger was used to extract keyword phrases from patent documents and the detected keyword phrases, such as "speaker identification", were expanded with phrases, for example "speaker verification", using *Wikipedia* and *WordNet* [2]. Queries have been enriched utilizing the semantic annotations in *Wikipedia* pages [12]. Further, keyword phrases have been detected in Chinese patent documents based on statistical information and semantic knowledge is used from *HowNet*, a Chinese lexical database containing Chinese terms and their English equivalents, to detect the phrases in the patent documents [39]. The statistical approach is adopted to calculate the chosen value of the phrase in the patent document.

Experiments show that by using noun phrase queries an increase in performance in terms of MAP can be achieved [64]. Also the retrieval performance is improved in terms of MAP compared to state-of-the-art query expansion method, when using phrases for query term expansion, in particular when enriching queries with phrases disambiguating the original query words [2].

Other approaches use parallel corpora for *QTE* in patent searching, in particular to achieve translations of keyword phrases. Specifically, claim sections of granted patent documents from the European Patent Office (EPO) including the claims in English, German and French were aligned to achieve translations of keyword phrases, particularly term to phrase translations, phrase to term translations and phrase to phrase translations [40] [41].

Experiments showed that phrase translation seems to be more beneficial for French than for German, because German often uses single-term compounds instead of phrases, thus limiting the potential benefit of phrase to term and phrase to phrase translations. As noticed in [64] and [2], experiments show that MAP substantially improves over the baseline, but Recall decreases.

Finally, the studies show that either the patent documents or external resources, such as lexica or dictionaries, for example *WordNet* or *Wikipedia*, are used as lexical resources for query term expansion. They all show that the retrieval performance, in particular MAP is improved, when expanding query terms to keyword phrase. But they did all not consider Boolean queries, in particular suggest approaches for Boolean query term expansion.

We learn that expanding query terms with keyword phrases has to be carefully considered, as patent search is recall-oriented rather than precision-oriented, i.e. preferring a higher number of potentially irrelevant documents in a result set over a more limited result set missing relevant documents. Through the expansion of query terms with keyword phrases recall usually decreases in the related studies. So our primary focus in this work is to provide synonyms for Boolean query term expansion to improve recall.

2.5.3 Approaches to suggest synonyms and equivalents

Related experiments in *IR* commonly rely on the usage of standard dictionaries and lexica, such as *WordNet*, for query term expansion, in particular to provide synonyms and equivalents for semantic search [69]. Further, existing domain ontologies were used to expand the query scope in [89]. In particular, an ontology for the biomedical domain, was queried to expand the users query so that relevant and related documents that may not include the exact query terms can be retrieved.

For query expansion in the patent domain *WordNet* was used [61]. Experiments showed that the retrieval performance decreases over the baseline queries, when using *WordNet* for query term expansion. In particular, it was found that expanding the query terms with *WordNet* leads to a slight improvement in MAP, but significant degradation in PRES. This means that relevant documents are being moved higher in the ranked list, but a greater number of relevant documents are lost from the ranked list. For patent searching, this outcome is considered as a negative result.

Claim sections of granted patent documents from the European Patent Office (*EPO*) including the claims in English, German and French are aligned to extract translation relations for each language pair [61]. Based on the language pairs having the same translation terms, synonyms were learned in English, French and German. The lexical database extracted from the patent collection, called *SynSet*, was used to improve the retrieval effectiveness. Experiments showed that the retrieval performance, in particular recall and precision, decreases over the baseline queries, when expanding the baseline queries with synonyms provided by *SynSet*. As presented in [61], related studies for *QTE* in patent searching can also not achieve significant improvement of retrieval effectiveness when using co-occurring terms and keyword phrases from the patent documents, such as the query documents, cited documents or from the feedback documents. The reason for this is that the query documents frequently share few terms with the relevant documents [62].

Two approaches to detect synonyms in patent documents are presented in [68] and [23]. In particular, the semantic similarity between two terms is calculated based on the hypothesis that terms used in the same context are usually related to each other [68]. Each noun is expressed as a vector of verbs that modify the noun and the nouns whose verb vectors are similar to each other are extracted as related terms. For example, the noun "school" is expressed with a verb vector which includes "go", "enter" and “graduate". A term whose verb vector is similar to that of “school", such as “university", is extracted as a related term of “school". Parentheses in the patent documents have been considered to detect related terms. In particular, the authors assumed that terms just before a parenthesis and the term in the parentheses have a narrower relationship to each other. Preliminary experiments showed that synonym expansion is an effective method for improving recall and precision. But in their present experiments the authors noticed that synonym expansion contributed little to the retrieval results.

There is a small thesaurus constructed manually for automatic query expansion in patent searching [54]. This small thesaurus is comprised of 311 synonyms with 1,694 query and expansion terms. In order to evaluate the performance of the thesaurus, the authors expanded ten queries and found that the precision of search results was improved.

We show in this section that patent documents and external resources, such as *WordNet*, are used for detecting synonyms. All approaches show that the retrieval performance is improved (with one exception [68]), when expanding query terms with synonyms. Unexpectedly, all approaches notice that retrieval precision increases through the expansion of the query terms with synonyms. Recall degrades. This outcome is considered as a negative result for the recall-oriented patent search task. Further, they all did not consider Boolean query term expansion.

So we show in this section that little thought is given to *QTE* in patent searching happening in real query sessions, as mentioned in [44]. Learning from actual queries submitted by experts can address this shortcoming of the proposed query expansion approaches.

2.6 Mining Query Logs for IR

In information retrieval applications, especially for web searches, query logs are being intensively studied [110]. Large-scale data sets of web queries, such as *AltaVista log* or *AOL log*, have been made publicly available [100]. The purpose of most studies is to enhance either effectiveness or efficiency of searching based on knowledge discovered from the query logs, which contain information on past queries [3] [70]. For domain-specific search environments such query logs, in particular whole query sessions, are mostly not available. Thus most studies are based on web searches. A survey on the use of query logs to improve search systems based on query expansion and query suggestion is presented in [86].

Although the search behavior of patent professionals and the search systems of the patent offices and commercial operators are rather different from that of web searchers and web search systems (professionals rather prefer to find more relevant documents than retrieving a small number of relevant documents at the top ranks), we analyze in this section the related work for mining query logs in information retrieval, in particular the available studies referring to web searches.

We first review previous work related to the analysis of query logs. Then we take a closer look at related approaches to acquire lexical knowledge from the query log files. Finally, we discuss related approaches for *QTE* based on query logs in a semi-automatic or fully-automatic manner.

2.6.1 Query Log Analysis

The analysis of query logs is predominantly based on basic statistics that can be computed over query logs, such as: (1) query session length, (2) query and query term popularity, (3) co-occurring analysis, (4) querying activity and (5) categorizing queries into topics.

The first time a large set of web queries was analyzed in view of query session length (*number of queries per session*) [85]. The analysis of the web query sessions, in particular the number of queries submitted by the users in a query session, showed that about 77% of the query sessions end after the first query. In addition, the query length (*number of query terms per query)* of the web queries was analyzed [53]. The results showed that, on average, web search queries are quite short. An average web query contains 2,35 terms. Less than 4% of the queries have more than 6 query terms.

Also query and query term popularity (*submission frequency of the queries and the query terms*) were used to analyze the web search logs [67]. They explore the problem of caching of query results in order to reduce the computing requirements needed to support the functionality the web search engine. Results show that a significant percentage of the queries have been submitted more than once by the same or a different user and the average query is submitted 1.33 times. But their search engine would reach a hit rate of only 25%, when caching query results based on the average query popularity. But the study refers to the *AltaVista* trace, which mentions that the average query was submitted 3.97 times. Thus caching search engine results based on this average query popularity may reach a hit rate of up to 75%. Finally, the experiments show, that the query and query term popularity value is used to reduce computing requirements needed for the search system.

The classification of the queries into topics is a further task in mining query logs. The distribution of large-scale data sets across topics enables to retrieve domain specific characteristics, such as number of queries per query session or number of query terms per query. The characteristics can be used to support users with searching individually depending on the searched domain. But categorizing queries into topics is not a simple task. There are a number of papers showing techniques for the classification of the queries. An overview to this task is given in [86]. Due to the possibility of the usage of patent classification schemes for categorizing queries in the patent domain, we will not provide here an analysis of the query classification literature.

A further statistic to draw from query logs is how query terms co-occur in the query logs [88]. The fifty most frequently co-occurring query terms of over one million web queries submitted by users of the *Excite* search engine were analyzed. The analysis shows that most highly correlated terms are keyword phrases to narrow the search. Results show that user’s interactions with web search engines are short and limited. In particular, most people use few query terms (60% of all queries have only one or two query terms), visit few web pages (29% of the users examine only one page from the result list) and rarely use advanced search features (less than 5% of all queries used any Boolean operators).

A specific analysis of the query logs carried out in previous work is the querying activity (*submission frequency over time*). The distribution of the queries over time, variations of topics over time or distance between repetitions of queries over time, have been analyzed. Again, the analysis of web queries has shown that the frequency of querying varies considerably during the daytime. The querying activity is higher during the first hours of the day than the afternoon [75]. Some topics are more popular in an hour than in another. Because queries are formulated by patent examiners and private searchers during their daily working time, we will not provide here a detailed analysis of the relevant literature discussing the distribution of the queries over day time.

Further, the authors of the query logs are considered in mining query logs [99]. The aim is to identify differences between queries and sessions of different authors, for example between adults and children. Query length and domain rank data of the clicked domains are analyzed. Results show that queries that were used to retrieve information for children were significantly longer than the average of the queries in the whole query log. A greater use of questions in the queries was recognized.

In summary, the related work referring to the analysis of query logs shows that all studies analyze web queries. The purpose of the studies is to enhance either effectiveness or efficiency of web search engines. In particular, the studies show that web queries are short and limited. Most correlated terms are keyword phrases. Queries with multiple terms are generally used to narrow a search. Because professionals rather prefer to find more relevant documents, they usually add further query terms to expand the query scope. So the opposite happens in professional searches. Several studies show that advanced search features are not relevant in web searches. Less than 5% of the analyzed queries use any Boolean operators.

Despite the fact that the setup used for query log analysis in the related work seems to be completely different to the collection of query logs, which we use for our experiments (a highly professional search setting of patent examiners of the United Patent and Trademark Office), we will consider the popular statistics for our query log analysis study. In particular, we use query length analysis, co-occurring analysis, query and query term popularity to find out what type of expansion terms and semantic relations are used by the patent searchers for query term expansion in real sessions. Further, we want to detect the most frequently used expansion terms and semantic relations.

2.6.2 Acquiring lexical knowledge

The challenge of acquiring lexical knowledge from query logs is to detect semantic relations between the queries and specifically between the query terms in the queries for automatic *QTE*.

Techniques used to detect the relations between the queries are commonly based on analyzing the retrieved documents, particularly the clicked web pages of web searchers [35]. Queries were considered as a whole and it was studied how queries and clicks can be combined in determining relations between the queries [10]. To catch the relations URLs that have been clicked by the user were analyzed. URLs retrieved and clicked by two users indicates that these queries are semantically equivalent, such as “how to learn guitar” or “online guitar lessons”. Correlations among queries were extracted by analyzing the common documents the users selected for them [52]. It was shown that the approach to detect semantic relations based on the retrieved documents can also be used to detect synonym relations between query terms, in particular for web queries, which are often one-step single query term events [10].

A further approach to detect synonym relations is generally based on the usage of *WordNet* [51]. Terms are extracted directly from the query log collection. Word senses are extracted from *WordNet*.

An approach to acquire lexical knowledge is based on the analysis of the co-occurring terms in the query logs [84]. The basic idea was that query terms which appear in the same context might be good expansion terms. In particular, semantic categories (including about 200 categories and about 250,000 entries) and query logs are used for finding additional named entities. The list of entries is matched against the query logs and frequencies are counted in order to identify additional entities. The basic idea is that terms which appear with the context but which are not already category members might be good additions. The evaluation shows that query logs are great resources for finding additional entities for a category. The method achieves high accuracy in categorizing newly acquired terms. For the evaluation the new words were checked if they exist in *Wikipedia,* in the top 10 *Google* results or in *amazon.com*.

A further task is the extraction of keyword phrases from the query logs to support searchers in narrowing their search. Standard approaches to extract phrases are based on statistical measures or external sources, such as dictionaries [11] [26]. Free text input is used for the extraction of statistical phrases, which consider every pair of non-function word as a relevant phrase. Those that occur with a frequency above a given threshold in a collection are retained. An external source, in particular the Collins English-Spanish bilingual automatic machine readable dictionary, was used for the detection and translation of keyword phrases [11]. In [26], a training corpus of 16 million web queries and a probabilistic context-free grammar considering linguistic structure is used to extract phrasal query terms. The most probable parse for a user query, is parsed and used for phrase expansion. Experiments showed that the performance for short precision-biased queries, is improved.

An analysis of this related work concerning the extraction of lexical knowledge from query logs shows that all studies engage the extraction of knowledge from web queries. We learned that these queries are short and limited. Boolean queries were not considered. The purpose of all studies is to improve retrieval precision of web search engines. They differ in terms of the resource that is used to detect semantic relations. External sources, such as lexica like *WordNet*, statistical measures and linguistic approaches to extract lexical knowledge directly from the query logs, or the retrieved relevant web sites (combining queries and clicks to determine relations between queries) are used to detect lexical knowledge. Experiments show that through the expansion of the web queries with lexical knowledge retrieval precision can be improved, in particular when suggesting keyword phrases. But in professional search, in particular in the patent domain, recall is of interest. So the goal of the expansion approaches developed for web searchers is completely different to our task, in particular to improve recall based on Boolean query term expansion.

2.7 Conclusions

In this chapter we reviewed related work for automatic query recommendation, in particular *QTE* in patent searching and based on query logs.

As shown, there is not much work in current research assisting patent experts in formulating Boolean queries, despite the importance of Boolean retrieval in patent searching as mentioned in Section 1.2. related approaches to enhance *QTE* in patent searching focus on computing co-occurring terms and keyword phrases. The main literature referring to automatic query expansion in patent searching is presented in [61]. There are only few approaches to detect synonyms and equivalents. Hence, assisting patent experts in formulating Boolean queries requires further research. In particular, the challenge is to provide synonym and equivalents, in particular of query and expansion terms, and to extract these relations directly from the patent domain to assist patent searchers in formulating Boolean queries.

Further, the related work referring to the analysis of query logs shows that all studies analyze web queries. Advanced search features are not relevant in web searches. Although the setup seems to be completely different to our query log collection, we will consider the popular statistics for our query log analysis study. Based on the analysis of the query logs we develop approaches to extract lexical knowledge directly from the query logs including the query and expansion terms for the query patents.

We think that query logs of USPTO patent examiners are valuable resources to acquire lexical knowledge for semi-automatic or fully-automatic *QTE*, which still have not been considered in related work. We think that lexical knowledge, which is extracted directly from the patent domain, will outperform the standard dictionaries used in the related work for synonym expansion. We think that when using expansion terms from the query logs (and not from the patent documents as done in the related work) will improve retrieval effectiveness, as these terms are used by the examiners to retrieve the relevant documents, which share few terms with the query documents.

In the next chapter we discuss the query logs of USPTO patent examiners, which we use in the rest of this thesis for acquiring lexical knowledge from the patent domain and for semi-automatic and fully-automatic *QTE* in patent searching.

3 Query Logs of the USPTO

3.1 Introduction

In this chapter we analyze the search setting of patent examiners of the United Patent and Trademark Office (USPTO) as presented in [97] [98]. In particular, this chapter expands the results presented in section 2.3. We gain insight into the search behavior of patent examiners to explore ways for enhancing query generation, particularly query term expansion, in patent searching. At first, in Section 3.2, we present access possibilities to the query logs of the USPTO patent examiners. Then, in Section 3.3, we describe the nature of the query log files. In Section 3.4, we introduce the setup used for query log analysis. In Section 3.5 we present the results of our analysis of the actual queries being posed by the patent experts. Finally, in Section 3.6, we present our conclusions.

3.2 Access to the Query Logs

Finding query logs in the patent domain has been a difficult task due to the lack of publicly available logs. Private companies and searchers are not interested in making their logs available as these may include terms revealing their current R&D activities. The USPTO is the only source known to us which publishes the query logs of patent examiners. In [27] a detailed analysis of the USPTO patent examiners query logs is presented to reveal search strategies of patent examiners.

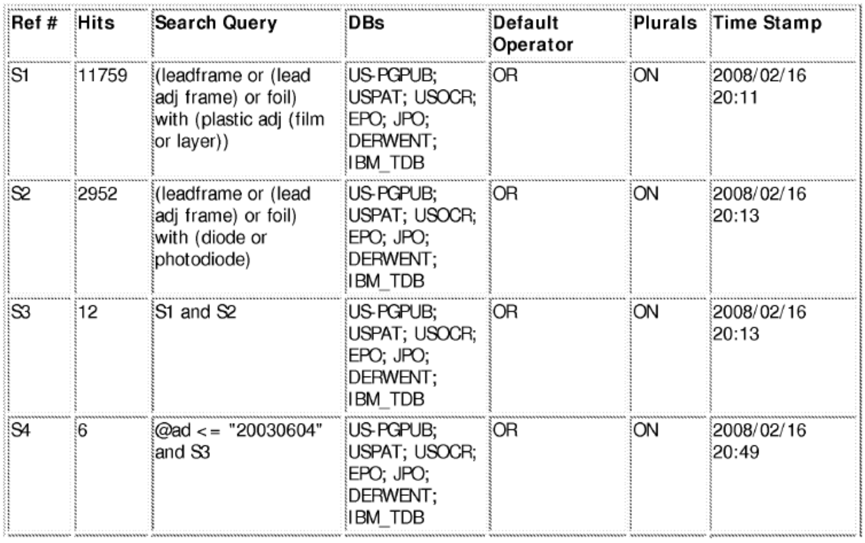
The query logs of USPTO patent examiners are called “Examiner`s search strategy and results” and are published for most patent applications since 2003 by the US Patent and Trademark Office Portal PAIR (Patent Application Information Retrieval). The log files are freely available from the US Patent and Trademark Office Portal.[1] The Portal PAIR download is limited by the USPTO. At first, for each patent application a verification code has to be entered. Then a document number, in particular an application, patent, publication, PCT, or control number, has to be selected. Then one of the subpages entitled “Application data”, “Transaction history”, “Image file wrapper”, “Foreign priority”, “Published documents”, “Address andAttorney/Agent”, “Display References” has to be selected to retrieve the respective information. The “Image File Wrapper” page is of concern to us here. This page can contain one or more query log files.

[1] http://www.uspto.gov/

Google has begun crawling the USPTO's public PAIR sites and provides free download of the documents concerning the examination procedures of patent applications.[2] Based on the URL “http://storage.googleapis.com/uspto-pair/applications/APP_NUM.zip”, where “APP_NUM” in the URL is an application number, a single zip file for each patent application is downloadable. The file contains multiple folders including information such as: Address and Attorney/Agent, Application Data, Continuity Data, Foreign Priority, Image File Wrapper, Patent Term Adjustments, Patent Term Extension History and Transaction History. The Image File Wrapper folder can contain one or multiple query log files. Each query log file of the USPTO is a PDF document consisting of a series of queries and having the ending “*SRNT.pdf*”. Also Reed Technology a contractor to the USPTO undertakes the task to provide free download of the documents concerning the examination procedures[3].

3.3 Nature of the Logs Files

Figure 2 shows an example, particularly an extract of four text queries, of a query log, in particular for the patent application with the number 10/519347, downloaded from the USPTO portal PAIR. The query log file consists of seven elements: Reference, Hits, Search Query, Database(s), Default Operator, Plurals and Time Stamp.



Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S1	11759	(leadframe or (lead adj frame) or foil) with (plastic adj (film or layer))	US-PGPUB; USPAT; USOCR; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2008/02/16 20:11
S2	2952	(leadframe or (lead adj frame) or foil) with (diode or photodiode)	US-PGPUB; USPAT; USOCR; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2008/02/16 20:13
S3	12	S1 and S2	US-PGPUB; USPAT; USOCR; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2008/02/16 20:13
S4	6	@ad <= "20030604" and S3	US-PGPUB; USPAT; USOCR; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2008/02/16 20:49

Figure 2. Query log for the USPTO patent application with number 10/519347

[2] http://www.google.com/googlebooks/uspto-patents.html

[3] http://patents.reedtech.com/Public-PAIR.php

3.3.1 Format and Elements

Each query log of the USPTO is a PDF file consisting of a series of queries. The original log files internally contains a picture of the document with no information about the text. But with Optical Character Recognition (OCR) it is possible with suitable software to extract the text of the query logs to make the log file searchable.

As mentioned before, each query log file is divided into the following columns:

Ref#: A reference number is assigned to each search query, such as *S1* or *S2* (*numbered consecutively*), which is shown in the reference element.

Hits: The hits element indicates the number of documents retrieved by the search query.

Search Query: The search query element shows the search query processed by the search system.

DBs: The selected and queried patent databases are shown in the database element.

Default Operator: The status of the default operator, which can be set to OR or AND, is presented in the default operator element.

Plurals: When a singular query term is searched, the plural forms of that term can also be searched. For this plurals must be activated. The Plurals element illustrates if plurals is set on ON or OFF.

Time Stamp: Each search query gets a time stamp, in particular a date and time, which is indicated by the time stamp element. Gaps between time stamps can help to determine which sets of documents retrieved have been reviewed by the patent examiners. The example in Figure 2 indicates that the examiner reviewed the documents retrieved by the query *S3*, which is a combination of query *S1* and *S2*.

In the following analysis of the query log files, we focus on the search query element showing the queries, in particular the text queries, formulated by the patent examiner of the USPTO.

3.3.2 Type of Queries

There are several kinds of queries. These can be grouped into text queries, non-text queries and reference queries:

Text queries: Text queries, such as queries *S1* and *S2* in Figure 2, are used for querying whole documents (fulltext search) using search operators and query terms. In addition, only sections of patent documents, such as the title section, can be searched when using section operators, such as “.ti.”, “.ab.”, “.clm.” for title search, abstract search or searching the claim sections.

Non-text queries: A non-text query is used for searching patent document numbers, classifications, or application and publications dates. For example, the non-text query “@ad <= 20030604” is used for searching patent documents applied before 4th June of 2003, as shown in query *S4* of Figure 2. Common operators used in these non-text queries include: classification operators for various classification schemes, in particular the current US Primary Classification with “.cor.” or “.ccor.”, the current US Cross Classification with “cxr.” or “.ccxr.” or the current US Classification with “.ccls.”, assignee operators, for example “.as.” or “.asn.”, publication number, such as “.pn.” or “.did.”, application date ”@ad<”, or publication date ”@pd>”.

Reference queries: The reference query is a combination of earlier queries, for example “*S1* and *S2*” as query *S3* shows in Figure 2, i.e. re-using the terms of a previous query and expanding it with further elements, thus avoiding to have to retype an earlier query.

Our focus is on the text queries in the query logs showing the queries, in particular the query terms and the search operators used by the patent examiner of the USPTO for querying the patent databases.

3.3.3 Search Operators

The types of search operators are Boolean, such as AND, OR and NOT, for finding intersections, unions, and subtractions from data sets, and proximity operators, such as ADJ, NEAR, WITH, and SAME for finding words within a defined perimeter of other words. In addition, truncation limiters are available, such as “$” for detecting varying derivatives of the same word ($ unlimited characters, $1 zero or 1 character) [37].

Below we present the most common search operators used by the USPTO patent examiners in the query logs, their purpose and an example:

AND: All terms in combination are in the document (Term_1 AND Term_2).

OR: One or the other or both terms are in the document (Term_1 OR Term_2).

ADJ (acent): Terms appear in the order specified next to one another or within a prescribed number of words of one another (Term_1 ADJ Term_2 or Term_1 ADJ3 Term_2).

NEAR: Terms appear in any order next to one another or within a prescribed number of words of one another (Term_1 NEAR Term_2 or Term_1 NEAR3 Term_2).

SAME: Terms appear in the same paragraph (Term_1 SAME Term_2).

WITH: Terms appear in the same sentence (Term_1 WITH Term_2).

In addition, patent examiners may also rely simply on a default operator, which can be set to OR or AND. As mentioned above, this is indicated by the default operator element in the query logs. In this case, no search operator is provided between the query terms in the query logs.

Using these Boolean, proximity and default operators query and expansion terms that were generated by brainstorming of the patent examiners are used to form initial search queries [37].

For our purposes, in particular to extract lexical knowledge from the query logs, we are specifically interested in the text queries including the Boolean operator OR which indicates that two query terms are synonyms, or can at least be considered as equivalents, and including the proximity operator ADJ, which indicates that two query terms can be considered as keyword phrases, in particular search terms consisting of two words.

3.4 Setup

The USPTO published about 2.7 million patent applications since 2003. The applications are classified into 473 classes each including hundreds of subclasses. Hence, on average, about 6,000 application documents are available for each class.

Because patent searchers use the classification system to narrow the search, we selected fifteen classes for our experiments. We selected classes that are topically related (e.g. classes 384 and 148; or classes 128, 433 and 623 from the medical domain) as well as completely disjunct classes. Furthermore, we selected classes having different numbers of query log files.

For our experiments, in particular to gain insights into the search behavior and characteristic of USPTO patent examiners queries, we downloaded from Google and preprocessed 103,896 query logs available for fifteen selected US classes, making it the largest collection of query logs used for experiments in the patent IR domain. In particular, through *OCR* conversion and segmentation of the 103,896 PDF files, which were stored as images, we separate all text queries including the search operators between the query terms from the query log collection.

Table 1. Experiment Setup

Nr.	US Class	Title	#Query Logs
1	454	Ventilation	1,820
2	384	Bearings	1,901
3	126	Stoves and furnaces	2,720
4	148	Metal treatment	2,877
5	219	Electric heating	3,926
6	433	Dentistry	4,025
8	180	Motor vehicles	5,205
7	417	Pumps	5,423
9	398	Optical communications	6,028
10	280	Land vehicles	7,905
11	128	Surgery	8,757
12	379	Telephonic communications	9,897
13	422	Chemical apparatus and process disinfecting, preserving, or sterilizing	11,842
14	439	Electrical connectors	14,706
15	623	Prosthesis (i.e., artificial body members)	16,864
Σ	-	-	**103,896**

Table 1 shows the number and title of the selected classes and the number of downloaded query logs for each class. As shown the number of query logs for the classes differs between 1,820 and 16,864 files.

Because OCR technologies do not have 100% correctness, we validate the correctness of the OCR outcome. We manually evaluate fifteen OCR-ed documents (randomly selected from the query log collection) character by character. This is done by proofreading the OCR-ed documents. We focus on the text query element containing the text queries formulated by the patent examiners and used for acquiring lexical knowledge.

The analysis of the OCR correctness shows that, on average, 98% accuracy rate in OCR can be achieved without any post-editing. 98% accuracy means that, on average, 2 out of 100 characters are recognized wrongly. Typical OCR accuracy rates exceed ninety-nine percent on high quality documents. Because the print quality of the scanned images (query log files) is not of high quality, we notice a slightly higher error rate during translating the scanned images into machine processable text.

In particular, the OCR error analysis shows that common and reiterating OCR errors are: (1) The character “*l*” is translated with the symbol ”|“, in particular for classification search, when translating the search operator “*ccls*”; (2) On the other side the symbol ”|“, which is used for querying multiple application or publication numbers simultaneously, is translated with the character “*I*”; (3) The character “*S*” used in the reference numbers is translated with “*$*” or “*5*”; As all these mentioned OCR errors do not occur in the text queries, which are of concern to us, these errors are acceptable for us. Considering only the text queries, on average, 99% accuracy rate in OCR can be achieved.

Because the scanned images are translated line by line, we further notice (4) that the multiple columns, in particular the elements (Reference, Hits, Search Query, and so on) are summarized. Text queries, which cover several lines, are sometimes separated, in particular when additional terms occur in the adjacent elements. Because we focus in our experiments on the search operators and the co-occurring query and expansion terms, this translation error is also tolerable.

Finally, (5) we notice that space characters between query terms and search operators are missing in the text queries. For example, “tube or conduit” is translated with “tube orconduit” in the analyzed query log set. Such errors being rather infrequent, these will mostly be eliminated by the frequency threshold settings applied the word pairings to determine their acceptance into the expansion dictionaries. Alternatives to detect errors, in particular the misspellings, and to correct them, such as domain-specific dictionaries, are not available. Further, when implementing the approach in a search system the queries could be exported directly from the system. The time-consuming translation process is not further needed. So our focus in the experiments using real query logs of patent examiners is not on the translation process. Rather, we concentrate in the analysis of the query logs, approaches to detect semantic relations, methods to expand query terms and test the approach in patent retrieval.

3.5 Query Log Analysis

In this section we present the results of our analysis of the actual queries posed by USPTO patent examiners. For the analysis we use the 103,896 text files available for the fifteen US patent classes. In particular, for vocabulary analysis and search operator analysis we use the text queries, which we extracted from the query log files.

3.5.1 Basic statistic

From the analysis of the layout of the query logs we sampled information related to the search query element, in particular to the search queries, for analysis of further general statistics of the query logs. In Table 2 we summarized some statistical properties.

Table 2. Query Log Statistics

Query Log Statistics	Median	Max.	Min.
Query Logs/ Query Document	2	32	1
Queries/ Query Log File	11	1,090	1
Text queries/ Query Log File	9	537	0
Unique *QLTs*/ Query Log File	17	304	0

As shown, on average, two query log files are available for each query document, i.e. the patent examiners searches on average two times for prior art in the examination procedure of a patent application. Each query log contains, on average, 11 queries. The maximum number of queries in a query log file is 1,090 queries (US 20050276411 A1 - Interaction between echo canceller and packet voice processing). The minimum number of queries in a query log file is one, while log files exist without any query in the standard format. In particular, one of the log files for the US20050051153 A1 - Wood burning stove having pivoting baffle and method – provides only the search results of the search system Patent Linguistics Utility System (PLUS) and the query "10761914", which is the application number of the query document. PLUS is a USPTO automated search system for U.S. Patents, in particular a query-by-example search system, which produces a list of patents that are most closely related linguistically to the application.

From the 11 queries, on average, 9 queries are text queries for full-text search or patent section search, such as searching the title, claim or description sections.

As shown in Table 2, on average, 17 unique query log terms (*QLTs)* are used by the examiners to express their information need. The maximum number of queries and text queries formulated for a query document are 1,090 queries and 537 text queries.

Figure 3 shows the distribution of query term lengths across queries. 96% of the text queries extracted from the query log collection have a length between 2 and 5 query terms. Most of the text queries (54%) formulated by the patent examiners have a length of two query terms followed by queries with three query terms (27%). 15% of the text queries have a length of four or five query terms. As mentioned before, only 4% of the text queries have only 1 query term or are longer than 5 query terms. Less than 1% of the text queries have a length of 1 query term. Only about 3% of the text queries are longer than 5 query terms.

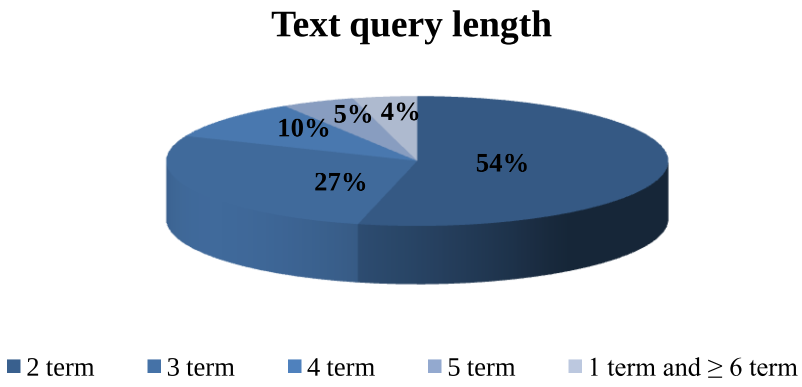


Figure 3. Text query length analysis

The length analysis shows that patent examiners formulate Boolean text queries with an average length of only three query terms. But we learned that reference queries, which are combinations of earlier queries are used to expand previous queries with further elements, thus avoiding to have to retype an earlier query. The analysis of the text queries show that about 27% of the queries include a reference to a previous query. So complex Boolean queries exist in the query log collection, but are formulated rarely by a single query in the query logs of USPTO patent examiners.

3.5.2 Search Operator Analysis

In this section we present some basic statistical properties on the used search operators. In particular, we analyze operator popularity.

Figure 4 shows the relative spread of the used operators for formulating Boolean and proximity queries. Nearly one half of the relations between two query terms are built using the Boolean or default operator "OR", nearly one third of the queries are generated using the "AND" operator. The remaining queries are built by the proximity operators "ADJ", "NEAR", "WITH", and "SAME" and by the Boolean operator "NOT".

Comparisons of the queries, particularly Boolean and proximity queries, show that two query terms can occur multiple times, but be connected by different operators. This would hint at conflicting usages, as two terms would be considered as synonyms and as phrases for more specific queries.

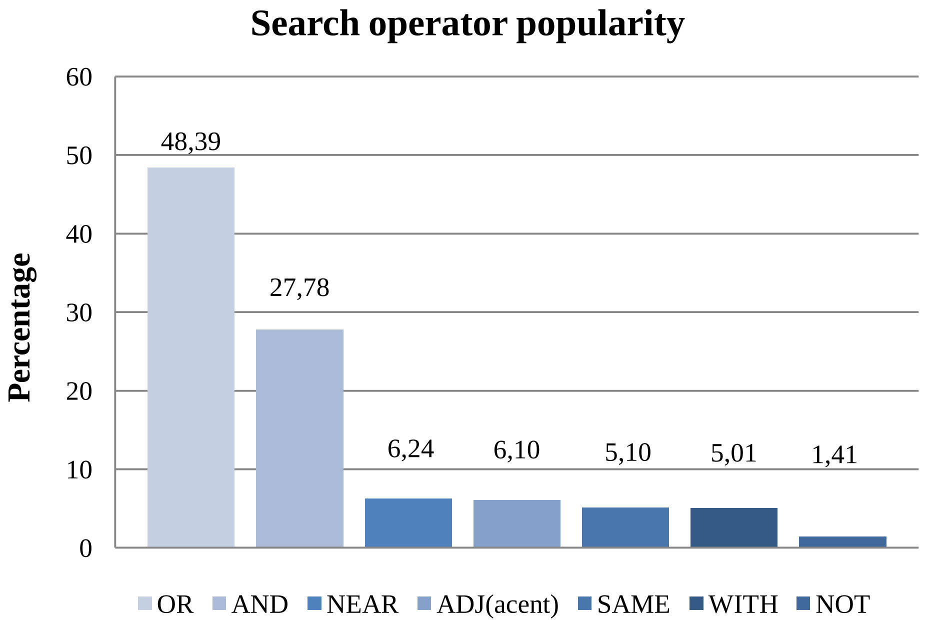


Figure 4. Search Operator Popularity

The query terms "drill" and "bit" for example, appearing in the US class 433, are used in a Boolean and a proximity query. The proximity query serves to search the keyword phrase "drill bit". On the other side, the Boolean query is used to search for the synonyms or equivalents "drill" or "bit".

3.5.3 Vocabulary Analysis

For vocabulary analysis we selected three US patent classes from the downloaded query log collection, in particular the classes called "Dentistry" (class 433), "Surgery" (class 128) and "Stoves and Furnaces" (class 126). We analyze the vocabulary of 15,502 query log files.

At first we learn from the USPTO query logs how terms co-occur in the query logs based on the Boolean and proximity operators. We preprocessed the query logs as follows: We extract all text queries including the search operators between the query terms from the query log collection. We then filter all 3-grams generated from the text queries in the form "X *b* Y", where *b* is the Boolean operator "OR", "AND" or the proximity operator "ADJ" and X and Y are query terms. We consider the correctly set parentheses, in particular we exclude 3-grams in the form "X b (Y" or "X) b Y". Further, we select all query logs containing the default operator element. We extract all text queries and considered those in which the default operator is set to “OR”. We then filter all bi-grams in the form “X Y”, where X and Y are query terms.

The analysis of the queries, in which the default operator is set to “OR” shows that even when the default operator is set to “OR” the patent examiners explicitly use the “OR” operator in the text queries. Hence, the majority of the “OR” relations are linked to each other by the Boolean operator “OR”.

In Table 3 we present the five most frequently co-occurring terms for the three US classes based on the search operator “OR”.

Table 3. Co-Occurring Terms based on Operator “OR”

Stoves and Furnaces	Dentistry	Surgery
tube pipe	tooth teeth	plurality plural
firewood fire	endodontic root	detection determination
hole opening	location position	motion movement
container pot	dental dentistry	stimulating stimulate
screen mesh	tube hose	hole opening

The majority out of the top-200 co-occurring terms are synonyms or equivalents at least for each specific US patent class. This shows that patent examiners use the Boolean operator “OR” to generate synonyms or equivalents to expand the query scope.

Table 4 shows the top-five co-occurring terms based on the Boolean operator “AND”.

Table 4. Co-Occurring Terms based on the Operator “AND”

Stoves and Furnaces	Dentistry	Surgery
radiant brooder	upper lower	first second
condensation glass	systems methods	scientific technical
glass door	first second	identify blood
mirror receiver	circuit speaker	controller electrical
fan stoker	blue dental	electrode anode

As shown in Table 4, the majority of the co-occurring terms have no semantic relation. So the patent examiners use the Boolean operator “AND” to narrow a search based on query terms, which occur in the same document, for example “fan” with “stoker”.

In addition, we show in Table 5 the top-five co-occurring terms based on the proximity operator “ADJ(cent)”.

Table 5. Co-Occurring Terms based on the Operator “ADJ(cent)”

Stoves and Furnaces	Dentistry	Surgery
heat exchanger	teeth caries	blood vessel
liquid propane	dental implant	respiratory device
solar collector	dental bracket	intra vascular
fuel type	tooth brush	mouth piece
temperature sensor	wireless lan	tissue image

In all classes studied the majority of term pairs are keyword phrases, in particular query terms consisting of two words. Hence, to narrow a search, particularly to limit the query scope, for example of the general query term “mouth”, a keyword phrase is generated by the patent examiners, such as “mouth piece”.

Further, we analyze the query terms of each US patent class w.r.t. the part of speech using the *CLAWS* part of speech tagger [34]. We identified 37,097 unique query terms for class 126, 76,868 terms for class 433 and 80,208 terms for class 128. We find out that in all classes about 70% of the classified terms are nouns followed by verbs (about 13%) and adjectives (about 10%). This shows that patent examiners use predominantly nouns to describe their information need, particularly to compartmentalize the invention into searchable features. This information, in particular that the examiners predominantly use nouns as query terms, has no direct impact on our work, as our approach is based on this lexical knowledge. But the information could be used, for example, for generating queries from query documents (extracting only nouns), using only nouns from external lexical sources (such as WordNet), or for an expansion strategy (suggesting nouns first followed by further terms).

We determine if the query terms used by the USPTO patent examiners are domain specific (the terms appear only in one specific US class). The class 128 for “Surgery” and class 433 for “Dentistry” have the most terms in common (3,673 terms, 4%) followed by the class 126 “Stoves” and US Class 433 “Dentistry” (having 3,483 common terms, 3%). Fewest common terms (1,751 terms, 2%) are shared between classes 126 and 128. Obvious, similar domains (classes 433 for “Dentistry” and 128 for “Surgery”) include more identical query terms than different classes. But we learn that patent searching is highly patent class specific. Less than 5% of the query terms of the specific US patent classes appear in the other classes, even across similar domains.

Finally, we analyze the source documents for which the queries are generated. Table 6 shows the analysis of the query log terms *QLTs* used by the USPTO patent examiners in view of the query documents.

Table 6. Query Log Term Analysis

Query Log Term Characteristics	avg. terms	%
QLTs per Query Log File	17	100,00
QLTs not in the Query Document	12	30.84
QLTs present in the Query Document	5	69,16

As shown, on average, the USPTO patent examiners selected 31% of the query terms from the query document. Hence, the majority of the *QLTs* (12 of 17) are expansion terms *ETs,* which do not appear in the query document. The other 31% of the *QLTs* come from the patent application. This means that the examiners expand on average the five *QLTs* from the query document with further 12 *ETs* by brainstorming.

We analyze the *ETs* which do not appear in the query documents. Therefore we queried the query log collection using the *ETs*. We find that 82% of the used vocabulary for query expansion appears in the query log collection, in particular in the specific US patent class. So the query log files appearing in the same US patent class are valuable resources to provide lexical knowledge for the patent domain.

3.6 Conclusions

In this chapter we introduced and analyzed query logs of USPTO patent examiners. The analysis shows patent examiners searches, on average, two times for prior art in the examination procedure of a patent application. Patent examiners formulate Boolean queries with an average length of only three query terms. But reference queries are used to expand previous queries with further elements.

The analysis of the search operators and vocabulary used by the USPTO patent examiners shows that means to enhance query generation in patent search, in particular to support patent experts in formulating Boolean queries, are to suggest (1) synonyms and equivalents indicated by the “OR” operator in the query logs, (2) co-occurring terms indicated by “AND” and (3) keyword phrases indicated by the “ADJ(cent)” operator. In particular suggesting synonyms is of particular importance. Nearly 50% of the query terms are expanded with synonyms or equivalents.

Further, the analysis shows that the majority of the *QLTs* are *ETs* which do not appear in the query document. So query terms selected from the query document are frequently expanded with *ETs* by brainstorming.

The analysis of the *ETs* shows that the majority of the used vocabulary for query expansion appear in the specific US patent class. Hence, the query log files appearing in the same US patent class and being posed by the patent examiners are valuable resources to provide lexical knowledge for the patent domain.

~ ~

4 Acquiring lexical knowledge

4.1 Introduction

In this section we present approaches to extract lexical knowledge from the query logs of USPTO patent examiners to assist patent searchers in formulating Boolean queries as presented in [96]. At first, in Section 4.2, we present our approaches to extract (1) keyword phrases consisting of two words and (2) synonym relations, in particular single term to single term, single term to phrase and phrase to phrase relations. In Section 4.3 we summarize the general workflow to acquire lexical knowledge from the query logs of USPTO patent examiners. In Section 4.4, we introduce the lexical databases *PhraseNet* and *PatNet*, which we extracted from the query expansion sessions done by patent examiners of the USPTO. The lexical databases can be used to both expand as well as limit the scope of a patent search and to guide a professional searcher through the query generation process. Finally, in Section 4.5 we present our conclusions.

4.2 Lexical Knowledge Extraction

As the analysis of the query logs of USPTO patent examiners has shown in Section 3, query generation in patent searching is highly domain specific. Patent examiners follow a strict scheme for generating text queries. They use the Boolean and default operator "OR" to expand the queries and the operator "AND" for querying co-occurring features of the invention. The proximity operators are used to narrow the search, particularly to limit a general query term to a keyword phrase using the proximity operator "ADJ(acent)". Table 7 shows the semantic relations provided by the query logs.

Table 7. Semantic relations provided by the query logs

Semantic Relations	Definition	Example
co-occurrence relation	X and Y	(scan) and (tooth)
synonym relation	X or Y	(drill) or (burr)
proximity relation	X near Y	(tool) near (gear)
proximity relation	X same Y	(plastic) same (ring)
proximity relation	X with Y	(drive) with (pin)
keyword phrase relation	X adj Y	(foot) adj (pedal)

As shown in Table 7, for acquiring lexical knowledge the operators "OR" and "ADJ" can be assigned to specific semantic relations. In the following subsections we use that to detect lexical knowledge in the query logs. We are aware that also the relations including the proximity operators "near", "same" and "with" are valuable resources for *QTE*. But in this work considering real query sessions we initially focus on the semantic relations, as these relations are commonly used by the examiners in the query sessions (85%). Further, the related work shows that approaches, in particular for synonym expansion, are needed.

4.2.1 Detecting Keyword Phrases

In patent search the proximity operator "ADJ(acent)" is used to narrow a search, particularly to limit the scope of a general query term, for example "mouth", to a keyword phrase, such as "mouth piece" in the medical domain concerning dentistry equipment. The Boolean operator "OR", on the other hand, is used to expand the scope of a search, specifying synonyms. We use the information provided by the proximity operator "ADJ", which indicates that two query terms can be considered as a keyword phrase, to detect semantic relations.

Table 8. Number of extracted keyword phrases based on confidence values CV_{1-5}

US Class	Title	CV_1	CV_2	CV_3	CV_4	CV_5
454	Ventilation	1,161	309	127	71	46
384	Bearings	818	196	97	44	21
126	Stoves and furnaces	2,162	749	315	192	112
148	Metal treatment	2,327	814	428	283	192
219	Electric heating	3,802	1,171	547	345	239
433	Dentistry	2,890	844	433	265	183
180	Motor vehicles	4,819	1,357	622	395	249
417	Pumps	5,643	1,506	719	437	285
398	Optical communications	10,454	3,125	1,530	974	675
280	Land vehicles	5,479	1,450	653	402	282
128	Surgery	7,957	2,615	1,342	876	626
379	Telephonic communications	12,733	4,254	2,238	1,454	1,009
422	Chemical apparatus	13,492	4,169	2,161	1,499	1,114
439	Electrical connectors	9,132	2,573	1,010	593	358
623	Prosthesis	10,523	2,811	1,364	895	619
Σ	-	72,482	**20,872**	9,812	6,280	4,227

Based on the approach to detect keyword phrases, as presented in Section 3.5, we extract keyword phrases from the query logs of the USPTO patent examiners.

For each class the number of unique keyword phrases and query terms learned from the query logs increases with the size of the query log collection. Because the USPTO publishes new query logs regularly for each class, the size of the collection keeps growing.

To exclude mismatches and misspellings, we utilize a confidence value *CV*. We measure the frequency of each keyword phrase in the specific class, i.e. that have a frequency of *1*, *2*, *3*, *4* and greater than or equal to *5*. We notice for all classes that the largest decrease in the number of keyword phrases is provided when moving to a required frequency of 3. Because patent searching is a recall oriented task, we consider those keyword phrases that were encountered at least two times as keyword phrases (CV_2) in the specific class to learn the lexical databases. This reduces spurious mismatches, but provides as many keyword phrases for query refinement as possible. So we retrieved 20,872 unique keyword phrases including 14,751 unique query terms.

Table 8 shows the number of keyword phrases extracted from the query logs and encountered at least two times, which we consider for the lexical databases, particularly the thesauri of English concepts. The total number of extracted keyword phrases based on CV_2 (20,872) is set in bold.

4.2.2 Detecting Synonyms and Equivalents

In patent searching the Boolean Operator “OR” is used to expand a query term with an expansion term, which has the same meaning, such as “drill” for “burr” or “tool” for “instrument” in the medical domain concerning dentistry equipment. We use that for automatically detecting synonyms (we distinguish three types of synonym relations) in the query logs based on the Boolean operator “OR”, which indicates that two query terms are synonyms, or can at least be considered as equivalents.

To detect the single term relations we use the process as described in Section 3.5. We extract 3-grams generated from the text queries in the form “X *b* Y”, where *b* is the Boolean or default operator “OR” and X and Y are query terms. Again, to exclude mismatches and misspellings and for ranking of the extracted synonyms according to their frequency in the specific classes, in particular for suggesting initially the synonyms having the highest frequency, we utilize the confidence value *CV*. Because the analysis of the query log collection shows that synonym expansion is used much more often by the patent examiners to expand a query term, we consider now those relations that were encountered at least three times as synonyms (CV_3). We retrieved 29,477 unique synonym relations including 18,804 unique query terms.

Table 9 shows for each class the number of unique synonyms extracted from the query logs, which increases with the size of the query log collection. The highest number of synonym relations can be extracted from the *large* class 422 for “Chemical apparatus and process disinfecting, preserving, or sterilizing”. The total number of extracted synonyms based on CV_3 (29,477) is set in bold.

Table 9. Number of extracted *STR* based on confidence values $CV_{1\text{-}5}$

US Class	Title	CV_1	CV_2	CV_3	CV_4	CV_5
454	Ventilation	2,595	826	383	215	135
384	Bearings	1,506	525	297	190	136
126	Stoves and furnaces	3,516	1,358	675	418	276
148	Metal treatment	6,427	2,903	1,813	1,312	980
219	Electric heating	6,523	2,729	1,582	1,103	812
433	Dentistry	7,543	3,280	1,713	1,184	851
180	Motor vehicles	7,819	2,998	1,675	1,067	711
417	Pumps	6,460	2,345	1,251	774	541
398	Optical communications	9,883	4,209	2,501	1,762	1,280
280	Land vehicles	6,534	2,377	1,279	776	524
128	Surgery	19,483	7,765	4,168	2,733	1,826
379	Telephonic communications	26,338	11,021	6,388	4,369	3,065
422	Chemical apparatus	35,221	14,055	7,776	5,288	3,816
439	Electrical connectors	10,190	3,967	2,157	1,452	989
623	Prosthesis	29,885	11,534	6,385	4,203	2,771
Σ	**-**	161,566	64,750	**29,477**	24,487	17,111

Furthermore, using the *CLAWS* part of speech tagger for English terms [34] we identified the synonyms w.r.t. part of speech and find out that more than half of the terms are nouns (69.61%) followed by adjectives (15.53%) and verbs (14.87%).

Table 10. Most frequently used nouns, adjectives and verbs

Stoves and Furnaces			**Dentistry**			**Surgery**		
adjective	*verb*	*nouns*	*adjectives*	*verbs*	*nouns*	*adjectives*	*verbs*	*nouns*
solar	rotate	burner	dental	rotate	workspace	medical	detect	preparation
automated	exchange	system	orthodontic	detect	tooth	biological	bond	device
prepared	adjust	oven	virtual	control	treatment	synthetic	form	image
thermal	mount	temperature	digital	guide	method	therapeutic	shape	pressure
open	duct	heater	technical	shape	preparation	magnetic	transmit	method

In Table 10 we present the five most frequently used nouns, adjectives and verbs used in each US patent class. Expanding the approach used for single synonym term detection, we further rely on the extensive usage of Boolean and proximity operators in the query logs. We use the proximity operator “ADJ” to detect keyword phrases and the Boolean operator “OR” to extract synonyms thereto.

Table 11 shows the synonym relations provided by the search operators “OR” and “ADJ” and for each type of relation an example. The search operators are used to formulate single term synonyms to retrieve documents containing any of the words, such as “drill” or “burr”. Single term to phrase relations are used to retrieve documents containing either the single term or the phrase, such as “blackberry” or “digital assistant”. Finally, phrase to phrase relations are used to retrieve documents containing any of the phrases, such as “force sensor” or “force detector”. As shown, there are multiple ways to formulate such Boolean queries, in particular to formulate the single term to phrase and phrase to phrase relations.

Table 11. Synonym Relations provided by the Search Operators “OR” and “ADJ”

Type	Definition	Example
single term	term OR term	drill OR burr
single term to phrase	(term ADJ term) OR term	(digital ADJ assistant) OR blackberry
	term OR (term ADJ term)	transponder OR (data ADJ carrier)
phrase to phrase	term ADJ (term OR term)	force ADJ (sensor OR detector)
	(term OR term) ADJ term	(control OR instrument) ADJ panel
	(term ADJ term) OR (term ADJ term)	(duty ADJ cycle) OR (band ADJ width)

The process to detect single term to phrase and phrase to phrase relations, we filter all 5-grams generated from the text queries in the form “X *b* Y *p* Z” and “ X *p* Y *b* Z”, and all 7-grams in the form “X *p* Y *b* Z *p* W“, where X, Y, Z and W are query terms, *p* the proximity operator “ADJ” and *b* the Boolean or Default operator “OR”. We consider the correctly set parentheses, in particular we exclude n-grams in the form “X) *b* Y *p* Z” or “ X *p* (Y *b* Z”.

Table 12. Detected Synonyms based on the Search Operators “OR” and “ADJ”

Type of Relation	Code	#Relations	#Terms
single term to phrase	STPR	920	1,523
phrase to phrase	PPR	530	984
Σ	-	**1,450**	**2,507**

As shown in Table 12, the query logs of USPTO patent examiners are a rich source to detect synonym relations from and for the patent domain. In addition to the single term relations we extracted 1,450 single term to phrase and phrase to phrase relations. As expected, the majority of the detected synonym relations are single term relations.

4.3 General workflow to acquire lexical knowledge

Figure **5** shows the general workflow to acquire lexical knowledge from the query logs of USPTO patent examiners. In advance, we collected all application numbers of the published patent applications for the fifteen classes and generate a list of download links for each class based on the download URL “http://storage.googleapis.com/uspto pair/applications/APP_NUM.zip”, where we replace “APP_NUM” in the URL with the application numbers. Google created a single zip file for each patent application containing all documents as the USPTO makes them publicly available for the patent application.

At first, we harvest the zip files via *Wget*[4] a free software package for retrieving files from web servers. Next, we unzip and filter the files using the file name ending “*SRNT.pdf*” to retrieve the query logs called “Examiner`s search strategy and results”. Then we carry out *OCR* conversion using *ABCocr*[5] a product to extract text from images on a Windows 7 platform and converted the *PDF* files to *TXT* files. Subsequently, all terms were fed into the extraction process. Following, in the extraction process we generate n-grams, in particular 3-grams, 5-grams, 7-grams, and 9-grams, from the extracted text using *AntConc*[6] a free n-gram extraction tool and filter the n-grams according to our approach to extract semantic relations, in particular the term expansions, as presented in subsection 4.2. In particular, we filter the 3-grams in the form “X *b* Y” and “X *p* Y”, the 5-grams in the form “X *b* Y *p* Z” and “ X *p* Y *b* Z”, and all 7-grams in the form “X *p* Y *b* Z *p* W“, where X and Y are the query terms and *p* is the proximity operator “ADJ”

[4] http://www.gnu.org/software/wget/
[5] http://www.websupergoo.com/abcocr-1.html
[6] http://www.laurenceanthony.net/antconc_index.html

and *b* the Boolean or Default operator “OR”. We exclude all other n-grams providing no semantic relations between the terms. In addition, we measure the frequency of the extracted relations to rank them according to their frequency in the specific classes. To exclude mismatches and misspellings, we consider those relations that were encountered at least two times. To query the ranked expansion lists, we load them into the open source thesaurus management software *TheW32* [28].

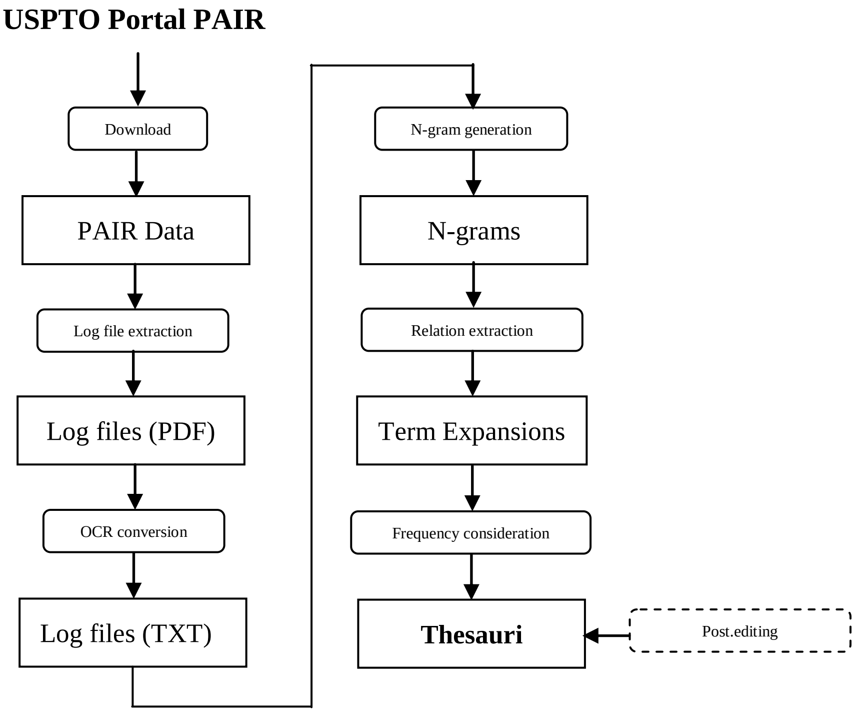


Figure 5: The general workflow of our proposed approach to acquire lexical knowledge from query logs of USPTO patent examiners

Finally, for the application of the lexical databases in real query sessions, we propose to implement a human judgement step to evaluate, in particular to post-edit, the suggested expansion terms. As mentioned in Chapter 3.4 in detail, the OCR conversion step, which is needed because no access to the search system of the USPTO is available to export the queries directly, leads to mistakes in the term expansions. When implementing the approach in a search system, the post-editing step is unnecessary. In the following experiments we first consider all expansions, as our focus is on recall then we carry out experiments to improve precision, in particular to exclude OCR errors and to improve efficiency of the proposed approach. We are aware that the precision measures can be further improved when post-editing the extracted relations or exporting the relations directly from a search system.

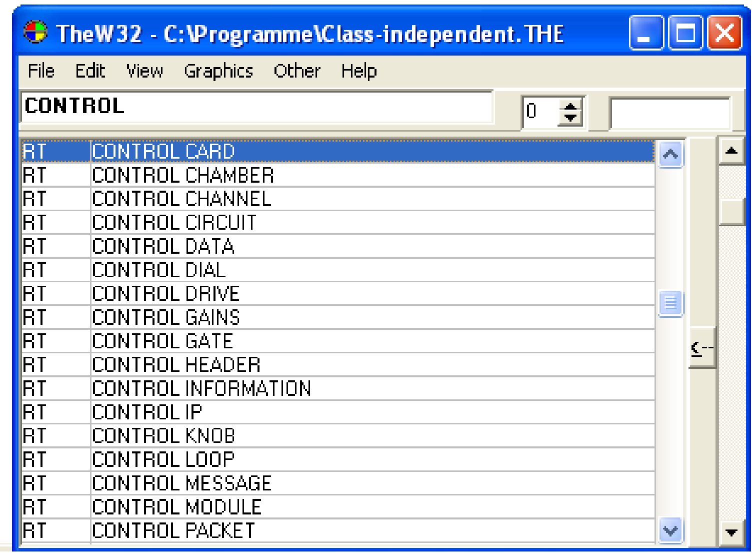
4.4 Lexical Databases

In this section we build two patent domain specific lexical databases, which we call *PhraseNet* and *PatNet*. We use the detected keyword phrases and synonym relations as presented in and marked in bold in Table 8 and Table 9. The lexical databases *PhraseNet* and *PatNet* resemble thesauri of English concepts that can be used for semi-automatic *QTE*, in particular for expanding as well for limiting the query scope.

4.4.1 *PhraseNet*

The lexical database *PhraseNet* provides English keyword phrases for the patent domain, in particular across all classes selected for the experiments. Terms which constitute a keyword phrase are linked to each other. In total, the lexical database *PhraseNet* provides 20,872 unique keyword phrases including 14,751 unique query terms. For example, the query term “*control*” can be expanded using the domain specific lexical database *PhraseNet* to limit the query scope as shown in Figure 6.

PhraseNet suggests, for example 41 expansion terms for the term “control”, which refine the general query term to a keyword phrase, such as “*control card*”, “*control chamber*”, “*control channel*”, “*control circuit*” and so on. Figure 6 shows only an extract from this expansion by the lexical database, in particular expansion terms judged by human experts.



CONTROL

RT	CONTROL CARD
RT	CONTROL CHAMBER
RT	CONTROL CHANNEL
RT	CONTROL CIRCUIT
RT	CONTROL DATA
RT	CONTROL DIAL
RT	CONTROL DRIVE
RT	CONTROL GAINS
RT	CONTROL GATE
RT	CONTROL HEADER
RT	CONTROL INFORMATION
RT	CONTROL IP
RT	CONTROL KNOB
RT	CONTROL LOOP
RT	CONTROL MESSAGE
RT	CONTROL MODULE
RT	CONTROL PACKET

Figure 6. Using *PhraseNet* for query scope limitation

The extracted keyword phrases can be also used only for specific US patent classes for (semi-) automated query suggestion, particularly for class-specific query scope limitation. Table 13 shows the keyword phrases available in each US patent class and the title of the class. Twelve of the fifteen US patent classes provide expansion terms for the specific term "*control*".

Table 13. Keyword Phrases provided by *PhraseNet* across all classes for the term "control"

class ID	Expansion term
433	box, channel, pad, section, valve
454	panel, tower, unit, valve
126	knob, loop, panel, valve
219	beam, circuit, panel, unit
180	Arm, circuit, data, drive, gains, module, panel, quadrant, unit, valve
417	board, button, card, chamber, circuit, dial, module, panel, rod, unit, valve
398	block, channel, circuit, information, header, loop, packet, part, plane, signal, system, time, unit
280	arm, assembly, box, rod, unit
128	agent, circuit, gate, panal, point, signal, tower, unit
379	data, ip, message, point, signal, station, unit
422	unit
439	apparatus, dial, module, unit

The class-specific lexical database *Optical Communications* (*US class 398*) provides most expansion terms. The lexical database suggests thirteen expansion terms, which refine the general query term to a keyword phrase.

On average, the lexical database *PhraseNet* suggests 12 expansion terms to limit a query term to a keyword phrase. The maximum number of query suggestions for the query term "power" is 96 expansion terms: "*power source*", "*power signal*", "*power tool*", "*power supply*", "*power distribution*", "*power transfer*" and so on.

4.4.2 *PatNet*

The lexical database *PatNet* provides English synonyms for the patent domain. Terms that have the same meaning are linked to each other. The lexical database provides

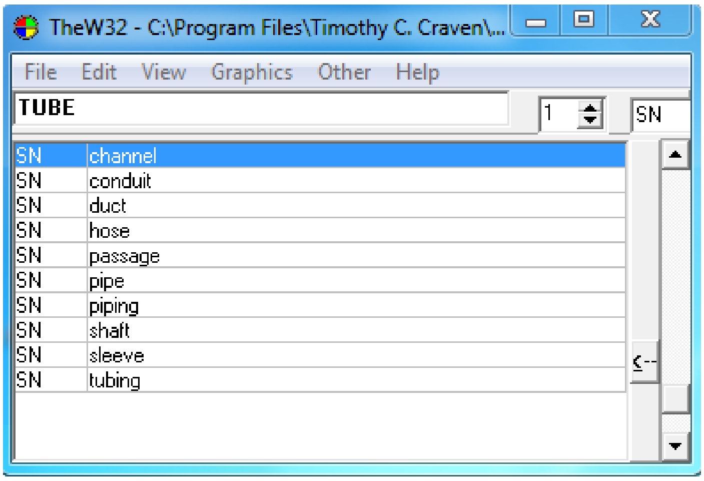
30,927 unique synonym relations and 19,040 unique query terms in total. *PatNet* suggests to a single query term: (1) single synonym terms, (2) synonym phrases, and (3) single terms, which in combination with the query term constitute a keyword phrase and finally suggests a synonym phrase thereto. Table 14 shows the synonym relations and unique query terms provided by *PatNet*.

Table 14. Synonym Relations provided by *PatNet*

Type of Relation	Code	#Relations	#Terms
single term	STR	29,477	18,804
single term to phrase	STPR	920	1,523
phrase to phrase	PPR	530	984
Σ	-	**30,927**	**19,040**

Figure 7 shows how *PatNet* can be used for semi-automatic *QTE*. For example the single query term "*tube*" can be expanded using the domain specific lexical database to expand the query scope.

PatNet provides ten synonymous expansion terms judged by human experts, in particular single terms, for the query term "*tube*", in particular *"channel", "conduit", "duct", "hose", "passage", "pipe", "piping", "shaft", "sleeve", and "tubing*" to expand the query scope.



TUBE

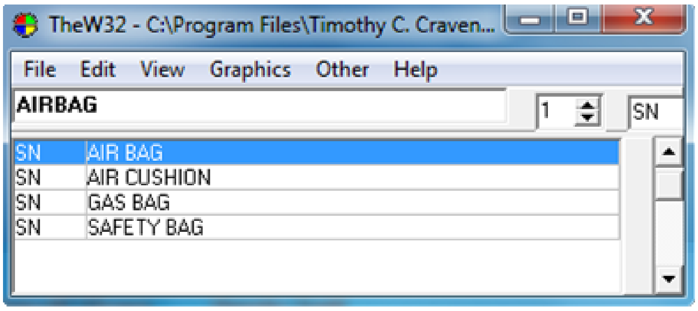
1

SN

SN	channel
SN	conduit
SN	duct
SN	hose
SN	passage
SN	pipe
SN	piping
SN	shaft
SN	sleeve
SN	tubing

Figure 7. Provided single term to single term relations for term "*tube*"

Figure 8 shows that the single term "*airbag*" can be expanded with synonymous keyword phrases to expand the query scope. *PatNet* suggests "*air bag*", "*gas bag*", "*safety bag*" and "*air cushion*".

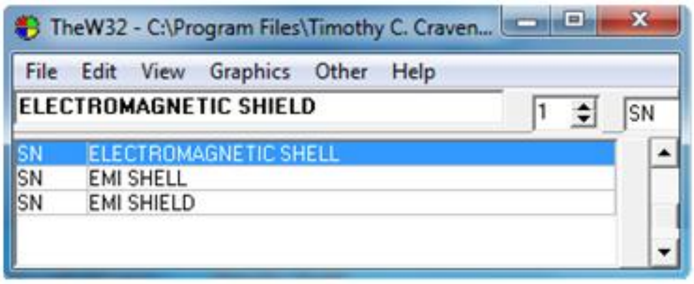


AIRBAG 1 SN

SN	AIR BAG
SN	AIR CUSHION
SN	GAS BAG
SN	SAFETY BAG

Figure 8. Suggested single term to phrase relations for term “*airbag*”

Finally, as shown in Figure 9, to expand the query scope of the keyword phrase “*electromagnetic shield*”, the lexical database suggests “*EMI shield*”, “*EMI shell*” and “*electromagnetic shell*”.



ELECTROMAGNETIC SHIELD 1 SN

SN	ELECTROMAGNETIC SHELL
SN	EMI SHELL
SN	EMI SHIELD

Figure 9. Phrase to phrase relations for the phrase “*electromagnetic shield*”

Again the extracted relation can be used only for specific US patent classes for (semi-) automated query suggestion, particularly for class-specific query scope expansion. Table 9 shows the number of synonyms available in each US patent class via the class-specific lexical databases.

Finally, in Table 15 we show a continuous example, in particular an example of an expanded invention diagram, which can be used by the patent searchers for the query terms “voice” and sensor” for generating Boolean queries. The invention diagram includes in a first column the searchable features of the invention, for example selected from a source document, particularly from a patent application or an invention report, and in a second column the corresponding *ETs* suggested by *PatNet.*

PatNet suggests for the query terms “*voice*” and “*sensor*” single terms (*STR*), keyword phrases (*STPR*), and single terms, which in combination with the query term constitute a keyword phrase and finally suggests synonym phrases (*PPR*). In particular, *PatNet* suggests for the single query term “*voice*”: (1) single synonym terms, such as “*acoustic*”, “*audio*”, or “*speech*”, (2) synonym phrases, such as “*voice mail*” or “*voice message*”, and (3) single terms, which in combination with the query term constitute a keyword phrase, such as “*voice print*” or “*voice sample*” and a synonym phrase thereto, for example “*speech sample*” for the phrase “*voice sample*”.

Table 15. Example of an Invention Diagram

Term	**Type of Relation**			
	STR	***STPR***	***PPR***	
voice	acoustic	voice exchange	voice mail	machine **mail**
	audio	voice **mail**	voice print	speech recognition
	sound	voice **message**	voice sample	speech sample
	speak	voice print	-	-
	speech	voice response	-	-
	telephony	voice sample	-	-
	verbal	-	-	-
sensor	airsensor	chemical sensor	force sensor	force detector
	indicator	weather sensor	weather sensor	**rain** sensor
	IRsensor	force sensor	-	
	monitor	-	-	-
	photodetector	-	-	-
	photosensor	-	-	-
	pyrometer	-	-	-
	detector			
	transducer			
	measur			
	biosensor			
	…			

In the second example, *PatNet* suggests synonym phrases, such as “*force sensor*” or “*weather sensor*. The examples show that *PatNet* can support patent searchers in the query generation process, in particular in generating the invention diagram in a semi-automatic manner.

Further, we demonstrate the performance of *PaNet* based on two real examples, in particular based on the query logs for the patent applications with the number 14/640554 and 14/640554, which do not appear in the test set. We marked in bold the terms used by the examiners as synonyms for the term “voice” in the query log 14/640554 and the term “sensor” in the query log 14/640554. Further, we indicated the terms, in particular “*mail*”, “*message*” and “*rain*”, used in combination with these query terms and suggested by *PatNet.*

PatNet provides all expansion terms used by the examiners in the query logs as synonyms for the term “v*oice*” and “*sensor*”, which corresponds to 100% *Recall*. In view of *Precision* the comparison shows that *PatNet* suggest in addition to the used synonyms additional expansion terms, which are not used by the examiner in these query logs, despite they are all relevant expansion terms. In particular, *PatNet* suggest for the common term “*sensor*” 92 expansion terms.

So in the next section we have to evaluate how well this approach to extract lexical databases from query logs of patent examiners works over a larger test set. Further, if necessary, depending on the results, we have to carry out experiments to improve precision, in particular to avoid time-consuming term selection, as best shown on the second example.

4.5 Conclusions

In this section we presented a new approach to detect keyword phrases and several types of synonym relations in query logs, which patent examiners of the USPTO created during the validation procedure of the patent applications. We built two lexical databases, in particular *PhraseNet* and *PatNet,* to support patent experts in formulating Boolean queries, preferable via semi-automatic *QTE*. The lexical databases suggest keyword phrases to narrow a search, particularly to limit the scope of a general query term, and provide synonym relations, in particular (1) single terms to single term, (2) single term to phrase and (3) phrase to phrase relations, to expand the query scope.

In addition, we have shown that the lexical databases can support patent searchers in the query generation process, in particularly in generating the invention diagram, which is used by the searchers for generating Boolean queries.

In the next section we evaluate how well this approach to extract lexical databases from query logs of patent examiners works, i.e. how complete and correct a set of suggestions is.

~ ~

5 Automatic Query Term Expansion

5.1 Introduction

In this section we use the query logs of the USPTO patent examiners for automatic query term expansion as presented in [91] [94]. For the evaluation we automatically expand the query terms of the queries from query sessions of patent examiners (gold standard) with single synonym terms and keyword phrases. We evaluate the lexical databases based on the collection size for each class, across different US patent classes and compared to *WordNet*. We considered characteristics of the query logs, in particular the query log length.

This chapter is organized as follows: At first, we present the overall scheme of our proposed *QTE* method in Section 5.2.. Then we introduce the Experiment Setup used for the experiments in Section 5.3. Following, in Sections 5.4 and 5.5 we present the experiments and the results of our automatic query scope expansion and limitation approaches. Finally, in Section 5.6, we present our conclusions.

5.2 Overall scheme of our proposed *QTE* method

Figure 10 illustrates the overall scheme of our proposed *QTE* method using class-specific and class-independent lexical databases, which we created from query logs of USPTO patent examiners, for automatic *QTE*.

In the first step the system receives a query term, which shall be expanded to extend or limit the query scope. In the example we use the query term "*drill*". In step 2, a specific US patent class is selected, for example the class 433 called "Dentistry", based on the class from which the query term is selected. The kind of *QTE*, in particular computing synonyms or keyword phrases, is selected in steps 3 and 4. The example shows that the systems computes synonymous single term relations *STR*.

Following in steps 5 and 6 the query term is expanded with the terms extracted first from the class-specific, then from the class-independent lexical databases. The lexical databases suggests class-specific expansion terms, such as "*burr*", "*reamer*" or "*powerdrill*". More generic and class-independent expansion terms, in particular "*tool*", "*instrument*", "*device*" or "*cutter*", are provided by the class-independent lexical databases.

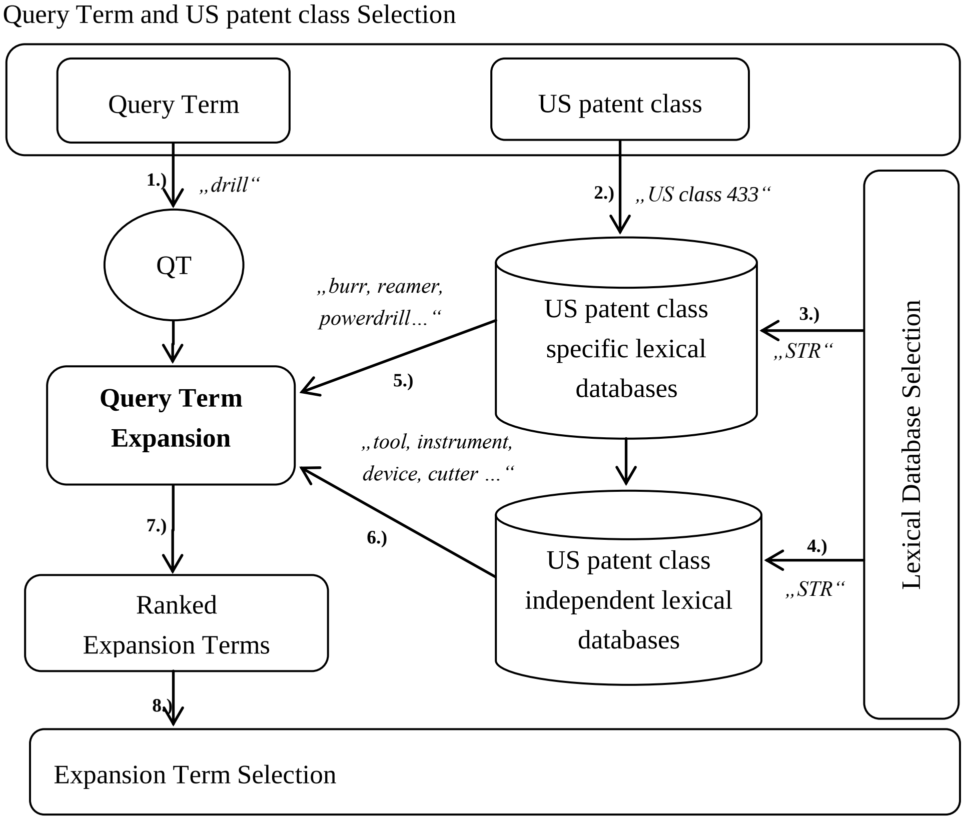

Query Term and US patent class Selection
Query Term
US patent class
1.)
„drill“
2.)
„US class 433“
QT
„burr, reamer, powerdrill…“
US patent class specific lexical databases
3.)
„STR“
Lexical Database Selection
5.)
Query Term Expansion
„tool, instrument, device, cutter …“
7.)
6.)
US patent class independent lexical databases
4.)
„STR“
Ranked Expansion Terms
8.)
Expansion Term Selection

Figure 10: Overall scheme of our proposed *QTE* method

In step 7, a ranked list of expansion terms is generated. Finally, expansion terms can be selected from the ranked list manually by the user or automatically by a search system in step 8.

5.3 Experiment Setup

For each US patent class we split the query log collection, which we used to acquire lexical knowledge in Chapter 4.2, in a test set for evaluation and a set for generating the lexical databases. Specifically, having the query logs ordered by time of application of the patent, we use the set of earlier query logs of each class for generating the lexical databases.

The test set being created from the chronologically last set of query logs in each class. This particular way of splitting the query log collection aims at creating a realistic evaluation setting, where lexical databases used in operational settings can only be trained on earlier query sessions.

We conceptually grouped the classes according to their size: (*small*) having less than 4,000 query logs, (*medium*) having up to 8,000 files and (*large*) having more than 8,000 logs. The grouping allows us to assess, in how far the performance of class specific lexical databases depends on the class size (number of query logs) and whether a minimum number of query logs is needed to achieve a certain level of performance in automatic *QTE*.

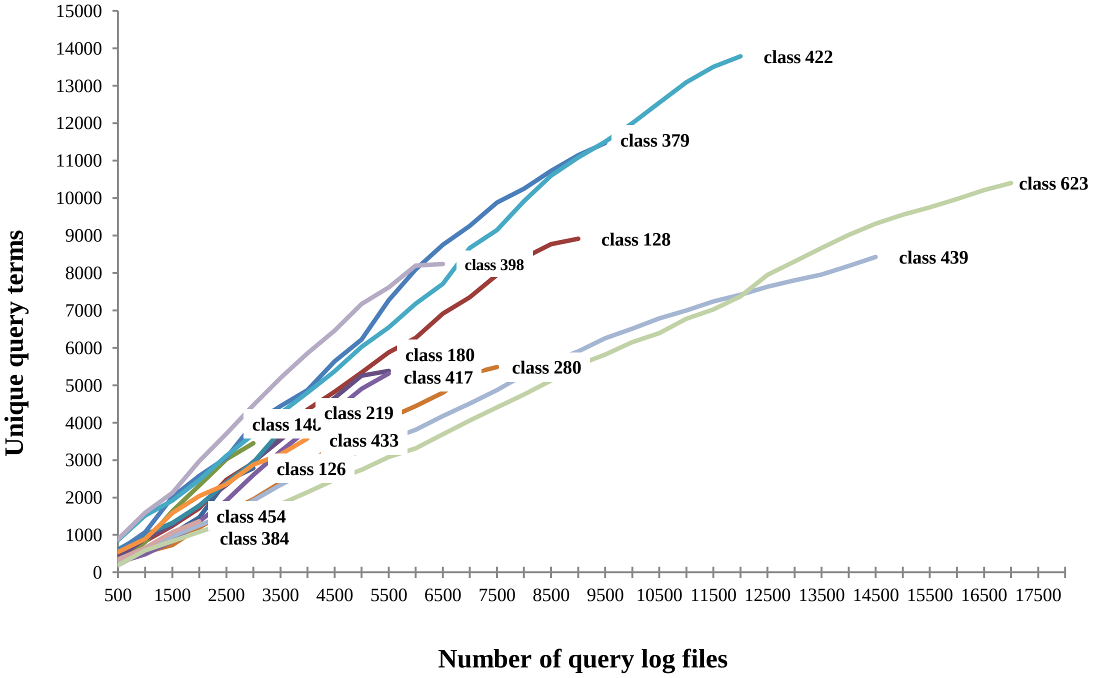


Figure 11: Vocabulary used for querying Keyword phrases

Figure 11 and Figure 12 show the increase of unique query terms used by the patent examiners, in particular to formulate keyword phrases and synonym relations, based on the number of available query logs for each specific US patent class.

In particular, we notice that the number of unique query terms continuously increases with the rise of the collection size. There is no decrease in the unique number of unique query terms with the increase of the query log files, in particular for classes where more than 8,000 query log files are available. So we assume that with rise of the query log collection the number of unique query terms will further increase with even larger query log collections.

We are aware that the number of unique query terms will not grow infinitely. In particular, Heaps' Law states that when the size of corpus grows new words occur, but the number of new words decreases while the size of the corpus increases. Further, when the corpus is small the number of new words will increase very rapidly, but continue to increase at a slower rate for larger corpus.

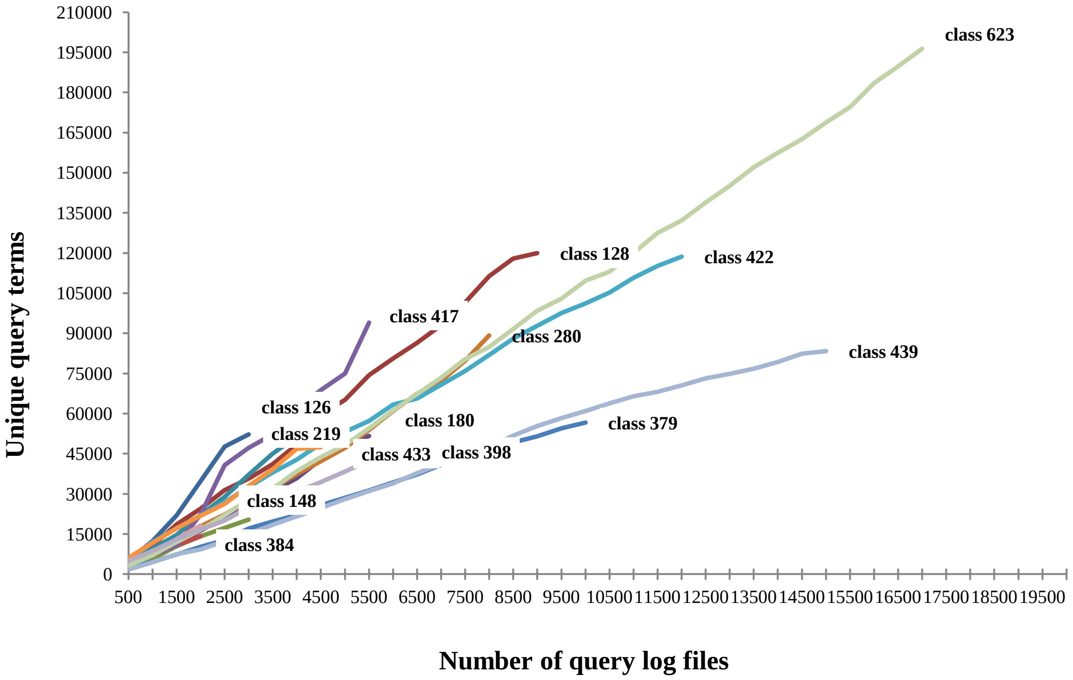


Figure 12: Vocabulary used for querying Synonyms

Table 16 shows the average growth per year of the patent applications for all 473 classes since 2003 and for the US class 623, for example, which we used to build the query log collection. As shown, the collection size is continuously rising (+30%).

Table 16. Number of query documents and growth per year

US class	query documents	avg. growth per year
623	15,535	+30%
all	2.7 Mio.	+31%

Further, the collection can be drastically expanded when considering further US classes. We use in these experiments only fifteen of 473 US classes. So for further experiments and the application of the proposed approach in real query sessions, the query log collection can be considerably increased, whereby the number of unique query terms will further increase.

5.3.1 Evaluation Sets

In each US patent class we use the most recent 500 query logs for testing, whereas the older query logs are used for generating the lexical databases. In particular, in each specific class the query logs collections are further divided into sub-sets to generate multiple lexical databases for each US patent class to evaluate size and time dependency characteristics. For each class we generate up to five sets (*TS1* to *TS5*) starting with the oldest logs. The size of these sub-sets depends on the class size. In particular, for the five classes having less than 4,000 query logs (grouped as *small*) we generate sets having between 500 and 2,500 query logs in increments of 500. For the *medium* grouped classes we create 20 sets having between 3,000 and 5,000 query logs. Finally, for the *large* classes we generate 21 sets having between 6,000 to 10,000 log files. In total, for all classes we generate 59 sets based on specific class and size.

Table 17 shows the generated sets used for building US patent class specific lexical databases. Based on our approach to acquire lexical knowledge from the query logs as mentioned in Chapter 4, we generate up to five class-specific lexical databases (*csDB[1-5])* for each specific US patent class. These lexical databases are based on 500 up to 10,000 query log files. Furthermore, we generate a class-independent lexical databases (*ciDB).* For this we use the largest sets of each specific class. Table 17 shows the selected sets in bold. The sets comprise between 1,000 and 10,000 log files. The class-independent lexical databases is still domain-specific in the sense that it is based on patent query logs, yet it stretches across class boundaries and is thus less specific.

Table 17. Evaluation sets

US Class	TS1	TS2	TS3	TS4	TS5	unused	Test Sets
454	500	**1,000**	-	-	-	320	500
384	500	**1,000**	-	-	-	401	500
126	500	1,000	1,500	**2,000**	-	220	500
148	500	1,000	1,500	**2,000**	-	377	500
219	500	1,000	1,500	2,000	**2,500**	926	500
433	3,000	**3,500**	-	-	-	25	500
180	3,000	3,500	4,000	**4,500**	-	205	500
417	3,000	3,500	4,000	**4,500**	-	423	500
398	3,000	3,500	4,000	4,500	**5,000**	528	500
280	3,000	3,500	4,000	4,500	**5,000**	2,405	500
128	6,000	7,000	**8,000**	-	-	257	500
379	6,000	7,000	8,000	**9,000**	-	397	500
422	6,000	7,000	8,000	9,000	**10,000**	1,342	500
439	6,000	7,000	8,000	9,000	**10,000**	4,206	500
623	6,000	7,000	8,000	9,000	**10,000**	6,364	500
ciDB					**78,000**	18,396	7,500

5.3.2 Evaluation

Because the success of keyword-based search depends on contextual factors, such as the individual search behavior (individual query formulation or reviewing of the retrieved documents) and influence of the search system (search interface, search engine, or ranking methods), as shown in the thesaurus evaluation literature [48], we evaluate the lexical databases based on query expansions carried out by the patent examiners in the search sessions.

To measure the performance of the lexical databases we compare the suggested terms from the lexical databases with the terms used for query expansion by the examiners as available in the query logs. In particular, we define for the expansions in the form ”X *b* Y”, where b is the operator “OR” or “ADJ”, that the first term in the relation X is the query term and Y the expansion term. This reflects real query expansion scenarios, where initially the query term is typed into the search system followed by expansion terms suggested by an expansion tool and is necessarily for suggesting keyword phrases.

We use the standard measures in IR, in particular Recall, Precision and Coverage, to measure the performance of the lexical databases. The definition of Recall and Precision is:

Recall. The fraction of relevant documents in response to a query that are retrieved [24].

$$Recall = \frac{\{\text{relevenat documents}\} \cap \{\text{retrieved documents}\}}{\{\text{relevant documents}\}} \tag{5.1}$$

Precision. The fraction of retrieved documents in response to a query that are relevant [24].

$$Precision = \frac{\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}}{\{\text{retrieved documents}\}} \tag{5.2}$$

For calculating Recall, Precision and Coverage in query term expansion, we apply the standard measures, as follows:

Recall. We measure how many of the *ETs* used by patent searchers were suggested by the lexical databases:

Precision: We measure how many of the suggested *ETs* were used by the patent searchers.

Coverage: In addition, we determine the number of out-of-vocabulary words, because we excluded *ETs* in the test set, which are out of the vocabulary of the lexical databases (the *ET*s do not appear in the lexical databases) for calculating recall and precision.

Our focus is on the recall scores, as users will be able to choose from a variety of possible *ETs* and can easily reject ones that are useless for their current search. However, in approaches where *ETs* are added to a query in a full-automatic manner, without prior approval by the users, precision is more important, because non-relevant expansion terms in the queries can degrade the performance. We compare the performance of the lexical databases. To check for statistical significant difference between the lexical databases, we use the t-test [87].

For calculating the overall average scores, we use the macro-average method. We use the recall, precision and coverage measures of the lexical databases on the different test sets. We selected macro-averaging, because this method gives equal weight to every class.

5.4 Automatic Query Scope Expansion

In this section we evaluate the performance of our proposed query term expansion approach based on size and class dependency characteristics. As a reference baseline we use *WordNet*, which constitutes a defacto-standard for evaluating the performance of a lexical database.

5.4.1 Query Term Expansion based on query log collection size

First we use the sets *TS1* to *TS5* of each US patent class to extract class-specific lexical databases *(csDB[1-5])* and the class-independent lexical database *ciDB* to evaluate the performance of the lexical databases based on the size of the query log collection and for each class. Number of semantic relations and unique query terms provided by the class-specific lexical databases are given in the Appendix in Table 50. The number of unique query terms and relations increases with the rise of the collection size. Most query terms and relations are provided by the *large* classes. In particular, the class-specific lexical database extracted from the set *TS5* for class 433 provides 10,411 synonym relations and 3,794 keyword phrases.

Table 18 shows the achieved recall and precision measures. For almost all classes the recall measures increase with the increase in set size. Specifically, we can assume that the recall scores will further increase with even larger sets. We marked in bold the highest performing class-specific lexical databases per row and the class where the class-independent lexical database performed best.

A strange behavior that can be noted is the decrease in recall for some classes with increasing query log collection size. Because we excluded synonyms that are out of the vocabulary, particularly for US patent classes 180 and 417, with larger sets the recall scores go down. The reason for that is, with the larger sets more synonyms and equivalent terms appear in the lexical databases (the *ETs* from the test set are not out of vocabulary any more) but not necessarily as synonyms, i.e. they were not co-occurring in queries at least twice.

Table 18. Recall and precision provided by the class-specific and class-independent lexical databases

	Recall						**Precision**					
US Class	*TS1*	*TS2*	*TS3*	*TS4*	*TS5*	*ciDB*	*TS1*	*TS2*	*TS3*	*TS4*	*TS5*	*ciDB*
454	45.83	**48.39**	-	-	-	54.10	20.00	**21.13**	-	-	-	1.88
384	18.18	**26.09**	-	-	-	58.21	25.00	**39.58**	-	-	-	5.68
126	29.41	29.63	30.30	**32.56**	-	51.43	**31.50**	17.68	14.76	10.74	-	3.32
148	35.16	39.22	39.62	**46.28**	-	61.39	**22.01**	17.23	13.09	18.78	-	5.82
219	40.74	41.46	36.21	35.71	**40.54**	66.67	**40.13**	18.95	21.88	19.70	21.84	2.74
433	43.75	**45.31**	-	-	-	59.60	10.15	**10.51**	-	-	-	2.33
180	**68.00**	67.31	63.16	62.07	-	47.42	**13.34**	13.07	12.60	12.24	-	5.35
417	54.17	60.00	**64.00**	62.96	-	**78.00**	**12.40**	10.57	10.09	9.83	-	5.43
398	52.63	50.85	50.85	51.67	**56.67**	59.21	**12.94**	12.37	11.58	10.90	10.70	5.63
280	50.00	44.00	46.15	48.28	**51.72**	56.52	12.74	11.38	14.33	**16.66**	12.24	4.48
128	30.06	29.35	**32.09**	-	-	45.21	21.64	20.30	**44.17**	-	-	**7.14**
379	60.76	64.89	66.91	**67.91**	-	72.26	**7.98**	6.64	5.62	5.66	-	4.15
422	56.88	58.56	56.52	58.26	**59.83**	70.00	8.25	**8.27**	7.79	7.06	7.22	6.29
439	60.00	57.14	59.10	59.10	**59.10**	68.97	**7.31**	6.46	6.73	6.13	6.02	2.66
623	64.29	64.29	64.29	66.67	**73.33**	61.90	25.52	25.25	**25.53**	19.87	22.05	3.79
avg.	47.32	48.43	50.77	53.77	56.87	**60.73**	**18.06**	15.96	15.68	12.51	13.35	4.45

Best recall measures are provided, on average, by the databases generated from the sets of the large US patent classes with a size larger than 6,000 query logs (with one exception, class 128). In particular, best recall is provided by the lexical database based on the set *TS5* for the class 623. The lexical database provides a recall of 73%.

The precision values show that with increasing set sizes, in particular for classes 128, 379 and 439, that the achieved precision scores decrease as the number of suggested synonyms increases. The lexical database detected from the set *TS3* of class 128 provides with a score of 44% best precision, while having a recall of only 32%.

Considering the lexical database providing the best recall performance (73% based on set *TS5* for class 623) on average only 2 out of 10 terms suggested are used by the patent examiners for query expansion. Note that the lower precision may not be a serious impediment for deployment of the *QTE*: as patent search is recall-oriented rather than precision-oriented, i.e. preferring a higher number of potentially irrelevant documents in a result set over a more limited result set missing relevant documents, especially when suggested by a system for manual deployment rather than performed in a fully autonomous manner, may be assistive rather than harmful. Furthermore, precision is likely to be under-estimated: Certain suggested expansion terms may still be correct and useful (as they are used in the query log collection by patent examiners), but simply have not been used by the patent examiners in the test set. To verify whether there are also incorrect relations (based on OCR errors for example, as mentioned in 3.4), we propose to use frequency information instead of a prohibitively time-intensive human evaluation. We will address this in the following chapter. Once again, please note when implementing the approach in a search system the queries could be exported directly from the system and OCR errors will not appear.

For the class-independent lexical database *ciDB* in almost all classes the recall measures of the class-specific lexical databases are improved. In particular, best recall is provided for class 417 with a recall of 78%. As explained above, in two cases the recall decreases because of synonyms that appear as terms in *ciDB*, but not as synonyms.

Mirroring the trend observed with the class-specific databases, the *ciDB* achieves on average precision measures about 4% across all classes, peaking at 7% for class 128.

In addition, we measure coverage of the respective lexical database by determining the number of out-of-vocabulary words, i.e. expansion terms that were used later in time that the databases can not include. This provides an indication on the comprehensiveness of the respective *DBs* suggested.

Table 19 shows the coverage of the class-specific and class-independent lexical databases. Again we marked in bold the highest performing class-specific lexical databases per row and the class where the class-independent lexical database performed best.

With the increasing class size the coverage of the lexical databases obviously increases. Best coverage scores are provided by the databases generated from the *large* classes. In particular, the lexical database generated for class 379 provides 81% of the query terms from the test set. For all classes, on average, *ciDB* provides a coverage of 88%.

The experiments show that for almost all classes the recall and coverage measures of the class-specific databases rise with the class size and can be further improved using the class-independent lexical database. On the other hand, the class-specific lexical databases achieve much better precision scores than *ciDB*. In this case, query terms are expanded in a certain context.

Table 19. Coverage provided by the class-specific and class-independent lexical databases

	Coverage					
US Class	*TS1*	*TS2*	*TS3*	*TS4*	*TS5*	*ciDB*
454	42.45	**52.83**	-	-	-	90.57
384	23.76	**44.55**	-	-	-	**96.04**
126	39.82	48.67	55.75	**64.60**	-	95.58
148	34.62	45.45	49.65	**56.64**	-	80.42
219	27.08	43.75	55.21	61.46	**62.50**	92.08
433	65.88	**66.47**	-	-	-	90.59
180	51.47	53.43	56.37	**56.86**	-	82.35
417	62.89	65.98	67.01	**68.04**	-	91.75
398	65.49	60.90	67.61	69.01	**69.72**	85.21
280	51.04	54.17	56.26	61.46	**61.46**	88.54
128	62.86	65.15	**66.60**	-	-	82.57
379	78.31	78.84	80.42	**81.48**	-	84.13
422	75.41	77.46	78.69	78.69	**79.51**	85.25
439	65.57	68.85	72.13	72.13	**72.13**	85.25
623	73.95	73.95	74.79	75.63	**75.63**	88.24
avg.	54.71	60.03	65.04	67.82	70.16	**87.90**

To provide lexical databases for automatic *QTE* achieving high recall/ coverage and precision scores, either (1) the recall measures of the class-specific databases or (2) the precision scores of the class-independent lexical database have to be improved. We address this issue in Chapter 6.

5.4.2 Query Term Expansion compared to *WordNet*

As mentioned in Chapter 2, most approaches use standard dictionaries for automatic query expansion, particularly for finding synonyms. In this section we evaluate the performance of our approach compared to the dictionary *WordNet* [69]. In particular, we test the best performing class-specific lexical databases, *ciDB* and the dictionary *WordNet* based on the test sets generated for each specific class.

For the expansion of the query terms we use all lexical relations included in the patent domain specific databases and in *WordNet*. We will not consider the meaning of the query terms, as our main focus is on the recall score in automatic *QTE*. Thus, *WordNet* should benefit from higher recall due to the large number of synonyms added without a potentially harmfuly limitation to specific word senses. We are aware that the precision measures can be improved when considering the word senses.

In spite of this rather defensive assumption, *ciDB* achieves better recall measures than the standard dictionary *WordNet* across all classes, as shown in Table 20. The highest performing lexical database per class is marked in bold.

Table 20. Recall and Precision achieved by the class-independent lexical database, *WordNet* and the best performing class-specific lexical database

	Recall			**Precision**		
US Class	*csDBs*	*ciDB*	*WordNet*	*csDBs*	*ciDB*	*WordNet*
454	48.39 *(TS2)*	**54.10**	**43.06**	**21.13**	1.88	1.60
384	26.09 *(TS2)*	**58.21**	22.54	**39.58**	5.68	4.49
126	32.56 *(TS4)*	**51.43**	27.27	**10.74**	3.32	1.96
148	46.28 *(TS4)*	**61.39**	27.52	**18.78**	5.82	1.96
219	40.54 *(TS5)*	**66.67**	32.37	**21.84**	2.74	1.91
433	45.31 *(TS2)*	**59.60**	19.13	**10.51**	2.33	1.12
180	**68.00** *(TS1)*	47.42	19.69	**13.34**	5.35	1.00
417	64.00 *(TS3)*	**78.00**	16.67	**10.09**	5.43	2.89
398	56.67 *(TS5)*	**59.21**	16.16	**10.70**	5.63	5.49
280	51.72 *(TS5)*	**56.52**	33.33	**12.24**	4.48	1.18
128	32.09 *(TS3)*	**45.21**	26.30	**44.17**	7.14	1.80
379	67.91 *(TS4)*	**72.26**	15.98	**5.66**	4.15	1.71
422	59.83 *(TS5)*	**70.00**	15.85	**7.22**	6.29	2.50
439	59.10 *(TS5)*	**68.97**	30.77	**6.02**	2.66	2.67
623	**73.33** *(TS5)*	61.90	17.57	**22.05**	3.79	1.75
avg.	51.45	**60.73**	22.06	**16.93**	4.45	2.15

We highlighted the best recall measure of *WordNet*. A comparable performance is only achieved for class 454. Over all classes, *WordNet* provides, on average, only a recall of 22%. Comparing the precisions measures, *WordNet* achieves as expected, like *ciDB*, only weak precision across all classes, peaking at 5% for class 398.

We also measure coverage of the respective lexical databases by determining the number of out-of-vocabulary words. Table 21 shows the vocabulary covered by the lexical databases. We highlighted the best coverage measures of the lexical databases.

WordNet being the most comprehensive thesaurus provides best coverage followed by *ciDB*. Best coverage is provided by *WordNet* for the class 128 at 99%. The class-independent lexical database has the highest coverage for class 384 at 96%.

Table 21. Coverage provided by the class-independent and class-specific lexical databases and *WordNet*

US Class	***csDBs***	***ciDB***	***WordNet***
454	52.83 *(TS2)*	90.57	93.40
384	44.55 *(TS2)*	96.04	93.07
126	65.00 *(TS4)*	95.58	93.81
148	56.64 *(TS4)*	80.42	97.55
219	62.50 *(TS5)*	92.08	96.35
433	66.47 *(TS2)*	90.59	95.88
180	51.47 *(TS1)*	82.35	96.57
417	67.01 *(TS3)*	91.75	92.78
398	69.72 *(TS5)*	85.21	95.07
280	61.46 *(TS5)*	88.54	92.71
128	66.60 *(TS3)*	82.57	98.55
379	81.48 (*TS4)*	84.13	96.21
422	79.51 *(TS5)*	85.25	97.13
439	72.13 *(TS5)*	85.25	88.52
623	75.63 *(TS5)*	88.24	94.12
avg.	64.87	87.90	94.78

In addition, we measure the performance of our approach compared to the dictionary *WordNet* without excluding out-of-vocabulary words. This affects the achieved recall values. In the experiments before, we have not considered out-of-vocabulary, because expansion terms that were used later in time cannot be extracted from the query log collection.

Table 22 shows the achieved recall measures achieved by the *ciDB*s, *WordNet* and the best performing *csDB* without excluding out-of-vocabulary words.

Table 22. Recall achieved by the *ciDB, WordNet* and the best performing *csDBs* without excluding out-of-vocabulary words

	csDBs		***ciDB***		***WordNet***	
US Class	*Recall*	*Change*	*Recall*	*Change*	*Recall*	*Change*
454	19.44	-60%	45.83	-15%	39.24	-9%
384	8.45	-68%	54.93	-6%	19.51	-13%
126	18.18	-44%	46.75	-9%	23.08	-15%
148	25.69	-44%	44.50	-28%	24.10	-12%
219	36.25	-11%	58.99	-12%	30.20	-7%
433	25.22	-44%	51.30	-14%	16.18	-15%
180	22.05	-68%	36.22	-24%	15.82	-20%
417	26.67	-58%	65.00	-17%	13.51	-19%
398	34.34	-39%	45.45	-23%	14.29	-12%
280	26.32	-10%	45.61	-10%	27.94	-16%
128	17.34	-46%	39.60	-12%	24.07	-8%
379	53.85	-21%	58.58	-19%	14.14	-12%
422	41.68	-30%	55.49	-21%	13.68	-14%
439	33.33	-44%	51.28	-26%	25.53	-17%
623	33.78	-54%	52.70	-15%	14.77	-16%
avg.	28.17	-45%	50.15	-18%	19.21	-14%

Over all classes, the best performing class-specific lexical databases *csDBs* provide now a recall of 28%, instead of 51% when excluding out-of-vocabulary words.

As shown in Table 22, a lot of vocabulary is not covered by the class-specific lexical databases. However, the class-independent lexical database *ciDB* still achieves over all classes a recall of 50%, compared to a value of 61% when excluding out-of-vocabulary words. The class-independent lexical database *ciDB* covers most query and expansion terms of the test sets.

WordNet now provides, on average, a recall of 19%, instead of 22% when excluding out-of-vocabulary words. We notice similar recall values, because *WordNet* provides best coverage.

So the analysis of the recall performance with and without excluding out-of-vocabulary words shows that with a larger number of query logs available the recall performance of the extracted lexical databases increases.

Through the analysis of the failed synonym relations provided by *WordNet* we learn that (1) patent examiners expand class-specific query terms using general terms. For example, in the class 379 called "*Telephonic communications*" they expand the specific query term "*cellphone*" using the general expansion term "*device*". A further example is the expansion of the class-specific term "*camper*" with the general term "*vehicle*". Further, (2) the examiners expand query terms w.r.t. part of speech, such as "*burn*" for "*burning*" or "*coat*" for "*coating*" or (3) they relate terms, which have the same meaning in specific classes, such as "*portable*" for *handheld*", in particularly for the class 379 called "*Telephonic communications*". The analysis of the relations provided by *WordNet* and failed in the lexical databases shows, that 82% of the vocabulary used in these relations appears in the lexical databases, but not yet in a synonym relation. Because the previous experiments show that the recall and coverage measures of the lexical databases increase with the class size, we assume that also these relations will be provided by the lexical databases with the rise of the collection size.

Additionally, through analysis of the vocabulary, which is not covered by *WordNet*, we find out that patent examiners (4) use popular trademarks, such as "*iphone*", "*ipad*" or "*blackberry*" for *QTE*. Finally, (5) the patent applicants are allowed to create their own terms, such as "*pocketpc*" for "*notebook*", "*watergas*" for "*steam*" or "*passcode*" for "*password*". Because of these highly specific expansions of query terms in the patent domain, standard dictionaries, such as *WordNet*, achieve only low performance. In these standard dictionaries, such patent domain specific relations are not included. These kinds of synonyms, equivalents and relations between the vocabulary are needed for automatic *QTE* in the patent domain. Using our approach to extract lexical databases from the patent domain and directly from the query logs of patent examiners fulfills the requirements of this highly domain specific query expansion.

5.5 Automatic Query Scope Limitation

In this section we automatically expand the query terms to limit the scope of the queries by keyword phrases. We use our proposed query term expansion approach, in particular the keyword phrases detected in the query logs based on the proximity operator "ADJ", and measure the performance of our new approach in automatic query scope limitation.

For the experiments we again consider the US patent classification to detect class characteristics. Furthermore, we consider characteristics of the query logs, in particular the query log length (*number of strings*).

5.5.1 Query Term Expansion based on US patent classification

We now use the relations of the lexical databases to evaluate class and size dependency characteristics of the class-specific and class-independent lexical databases. Table 23 shows the achieved Recall, Precision, and Coverage scores. We marked in bold the highest recall, precision and coverage measure per column.

Best recall measures of the class-specific lexical databases are provided, on average, by the lexical databases generated from the US patent classes with a size larger than 6,000 query logs (with one exception, class 623). In particular, best recall is provided by the lexical database generated for the class 422. The lexical database suggests, on average, 7 of 10 keyword phrases, which are used by the patent examiners for query expansion.

Table 23. Query Scope Limitation based on US patent class

	Recall		**Precision**		**Coverage**	
US Class	*csDB*	*ciDB*	*csDB*	*ciDB*	*csDB*	*ciDB*
454	35.00	66.67	15.91	4.99	27.03	89.19
384	60.00	71.93	**50.00**	5.68	25.86	**98.28**
126	35.14	56.77	22.03	5.69	45.12	94.51
148	41.11	**77.94**	40.00	7.59	52.78	94.44
219	33.65	67.08	22.29	5.15	60.47	93.60
433	55.56	50,00	**41.67**	8.79	5.94	94.12
180	54.87	65.69	12.53	5.66	80.71	97.86
417	44.14	66.91	19.44	4.07	79.86	97.84
398	56.20	59.75	5.97	3.83	80.59	93.53
280	47.69	68.42	13.84	5.85	66.33	96.94
128	50.00	61.29	22.88	4.39	78.26	89.86
379	60.17	61.54	7.91	5.30	**86.76**	95.59
422	**70.00**	74.00	13.21	5.92	76.92	96.15
439	55.00	55.88	8.89	4.80	82.19	93.15
623	27.27	50.00	27.27	**9.46**	52.38	66.67
avg.	48.39	**63.59**	**21.59**	5.81	60.08	**92.78**

The precision values of the class-specific lexical databases show that with increasing set sizes the achieved precision scores decrease, as the number of suggested terms increases. Considering the lexical database providing the best recall performance (class 422), on average, only 1 out of 10 terms suggested by the lexical database is used by the patent examiners for query expansion with an “ADJ” operator.

Best coverage scores of the class-specific lexical databases are also provided by the lexical databases generated from the classes having more than 6,000 query log files. In particular, for class 379, the class-specific lexical database provides 87% of the phrases from the test set.

Furthermore, the experiments show that for almost all classes the recall measures of the class-specific lexical databases can be further improved using the class-independent lexical database. Only in class 433 the recall decreases, because we excluded query terms which are out of the vocabulary to calculate the recall measures. While more query terms appear in the *ciDB*, they do not necessarily form keyword phrases. The query terms are not out of vocabulary any more. Best recall is provided for class 148. The class-independent lexical database suggests, on average, 8 out of 10 keyword phrases used by the patent examiners.

Compared to the class-specific lexical databases, the precision scores achieved by *ciDB* are obviously lower. For example, in class 422, on average, only 6 out of 100 terms suggested by *ciDB* are used by the patent examiners for query expansion.

The coverage scores of the *ciDB* decrease with larger class size. In particular, for class 384, *ciDB* provides 98% of the query terms from the test set. Considering all classes the class-independent lexical database provides, on average, a coverage of 94%. But compared to *ciDB*, the class-specific databases again achieve much better precision scores. Considering all classes the class-specific lexical databases provide, on average, a precision of 18%.

Again, we measure the performance of our approach without excluding out-of-vocabulary words. Table 24 shows the achieved recall and precision measures achieved by the *csDBs* and the *ciDB* without excluding out-of-vocabulary words.

Considering all classes, the *csDBs* provide now a recall of 32%, instead of 48% when excluding out-of-vocabulary words (-33%). Otherwise the class-independent lexical database *ciDB* still achieves over all classes a recall of 59%, compared to a value of 64% when excluding out-of-vocabulary words (-7%). Again we notice that the class-independent lexical database *ciDB* covers most query and expansion terms.

Table 24. Recall achieved by the *csDBs* and the *ciDB* without excluding out-of-vocabulary words

	csDB		*ciDB*	
US Class	*Recall*	*Change*	*Recall*	*change*
454	9.46	-73%	59.46	-11%
384	15.52	-74%	70.69	-2%
126	15.85	-55%	53,.66	-5%
148	22.22	-46%	73.61	-6%
219	20.35	-40%	62.79	-6%
433	29.41	-47%	47.06	-6%
180	44.29	-19%	64.29	-2%
417	35.25	-20%	65.47	-2%
398	45.29	-19%	55.88	-6%
280	31.63	-34%	66.33	-3%
128	39.13	-22%	55,.07	-10%
379	52.12	-13%	58.82	-4%
422	53.85	-23%	71.15	-4%
439	45.21	-18%	52.05	-7%
623	14.29	-48%	33.33	-33%
avg.	31,.59	-33%	59.31	-7%

5.5.2 Considering Query Log Length in Query Term Expansion

In the previous experiments we measured the performance of the class-specific and class-independent lexical databases based on query expansions of patent examiners in real query sessions. Yet, we have not considered characteristics of the query logs used for evaluation.

In this section we evaluate whether the performance of the lexical databases depends on the length of the query logs/ of the search sessions. With the increase of the query log length more detailed queries and query terms are included in the query logs, which might be harder to expand. This experiment allows us to validate, whether the proposed expansion strategy should be applied only in the earlier stages of a search process when the “easier” i.e. more common expansion terms should be suggested, or whether it will also work in the later stages when more specific expansion terms are to be considered.

For the experiments we use *ciDB* providing best recall scores for the recall oriented patent search task. Further, we use the test sets of all US patent classes i.e. 7,500 query log files. We divide the test set in multiple subsets to create multiple evaluation sets. Specifically, we order the query logs by length (in particular by the number of strings) and group them into 10 bins. The resulting ten evaluation sets called *length*$_1$ up to *length*$_{10}$ each comprise 750 log files with a length of 3 up to 35,144 strings.

Figure 13 shows the average number of strings per query log file for each subset. As shown, the average numbers of strings per query log file steadily increases from the subset *length*$_1$ with an average number of 42 strings up to *length*$_9$ with an average number of 1010 strings. For the subset *length*$_{10}$ we notice an above average increase of text characters per query log file with an average number of 2795 strings.

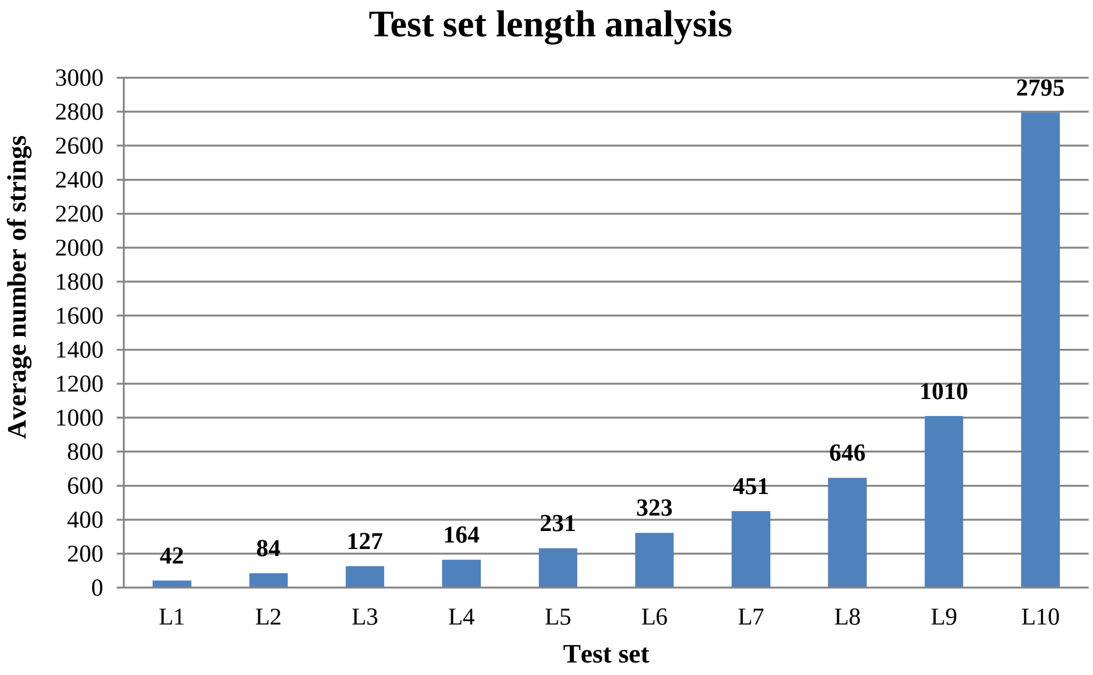


Figure 13. Test set length analysis.

Table 25 shows the subsets and the recall, precision and coverage measures achieved by the lexical database *ciDB*. We marked in bold the highest and lowest recall, precision and coverage value.

The class-independent lexical database *ciDB* achieves for all subsets *length*$_1$ to *length*$_{10}$ equivalent recall and coverage scores. In particular, *ciDB* suggests for all subsets, between 7 up to 8 out of 10 keyword phrases, which are used by the patent examiners for query expansion. Furthermore, *ciDB* provides between 89% and 95% of the phrases used in the subsets.

Only weak precision measures can be achieved across all subsets, as only 4 to 6 out of 100 terms suggested by the *ciDB* are used by the patent examiners for query expansion.

Table 25. Performance of *ciDB* when considering query log length

Subsets	Length	Recall	Precision	Coverage
length$_1$	3 - 51	68.92	5.78	94.87
length$_2$	59 - 91	69.09	5.88	94.83
length$_3$	103 - 123	75.76	**6.46**	**89.19**
length$_4$	146 - 179	**67.86**	4.02	91.80
length$_5$	209 - 225	71.28	5.40	**94.95**
length$_6$	237 - 349	74.60	4.28	91.30
length$_7$	372 - 406	**82.09**	5.56	94.37
length$_8$	446 - 545	74.85	**3.63**	90.56
length$_9$	662 - 768	72.57	5.18	91.51
length$_{10}$	1,181 – 35,144	72.17	5.41	93.35

Finally, the evaluation shows that the performance of *ciDB* is independent from the query log length. The lexical database thus helps in automatic query scope limitation during the whole search process independent from the number of previously submitted queries.

5.6 Conclusions

In this chapter we used lexical databases extracted from query expansion sessions done by patent professionals for automatic query expansion. In particular, US patent class-specific and class-independent lexical databases were used to suggest synonym expansion terms and keyword phrases.

We evaluated these lexical databases based on query expansion done by patent professionals in real sessions (gold standard). The experiments have shown that our approach to generate lexical databases from the patent domain, specifically directly from the query logs, helps in automatic *QTE*.

In particular, the experiments show for the class-specific databases that recall and coverage measures increase with the availability of a larger set of query logs and can be further improved when using the class-independent databases. The class-independent lexical databases suggest, on average, up to 8 out of 10 *ETs*, which are used by the examiners for query expansion.

Expectedly, the class-specific databases achieves better precision scores than the class-independent databases. Query terms are expanded in a certain context (US patent class). Table 26 shows the advantages and disadvantages of the different lexical databases.

To strike reasonable balance between increasingly higher recall/ coverage and lower precision, one approach could be to suggest initially all the *ETs* from the class-specific databases providing high precision scores, followed by more generic terms from the class-independent databases achieving higher recall measures later-on.

Furthermore, we considered characteristics of the query logs which we used for evaluation, in particular the length of the query logs/ search sessions. We find that the *ciDB* achieves equivalent recall, precision, and coverage scores for all subsets having different query log lengths. Hence, the performance of the *ciDB* is independent from the length of the query sessions showing that it can be used both through initial as well as later stages of the search process.

Table 26. Advantages and disadvantages of the extracted lexical databases

Lexical Databases	+	-
class-specific (*csDB*)	Query terms are expanded within a certain context (highest precision).	Relevant expansion terms from related domains are missing (lowest recall).
class-independent *(ciDB)*	All possible expansion terms are suggested at once (best recall).	Too many non-relevant expansion terms are suggested (lowest precision).

Finally, the experiments show that specific lexical databases drastically outperform general-purpose sources such as *WordNet.* The standard dictionary *WordNet* achieves for all US patent classes only low performance in recall. This may be attributed to the fact that patent searchers, (1) expand class-specific query terms using general terms, (2) expand query terms w.r.t. part of speech, (3) relate terms, which have the same meaning in a specific class, (4) use popular trademarks and (5) patent applicants are allowed to create their own terms for query expansion.

~ ~

6 Query Term Expansion Strategies to improve precision

6.1 Introduction

In this chapter we present *QTE* strategies as described in [92]. Generally, the goal is to suggest a complete list of *ETs* for creating the invention diagram (assisting brainstorming of possible *ETs*). Searchers will be able to choose from the list of suggested *ETs* and can reject ones that are useless for their current search. But we learned that compared to the literature (starting with an invention diagram) searches often begin with a very narrow set of *QTs* and *ETs* and the searches are incrementally expanded based on what is found. This leads to the question whether we can devise means to suggest *ETs* in a useful order to avoid time-consuming term selection. In particular, for the generic synonymous single term relations the lexical database *PatNet* provides, on average, 11 *ETs*. But the maximum number rise up to 92 terms, i.e. for common terms, such as “sensor”.

The chapter is organized as follows: First, we present the experiment setup in Section 6.2. Then we present and discuss the results of the experiments when using the US patent classification to improve recall and coverage of the class-specific lexical databases in Section 6.3., followed by the results when using frequency information in Section 6.4. and when considering word senses in Section 6.5 for automatic *QTE*. Finally, in Section 6.6, we present our conclusions.

6.2 Experiment Setup

For the experiments we use the whole setup (103,896 files) from section 3.4. Specifically, for each US patent class we kept the most recent 500 query log files as a hold-out set for testing resulting in 7,500 log files and the oldest query logs (96,396 files) are used for generating the lexical databases.

We refrain from testing generic lexica, *such as WordNet* – the only ones available for this kind of term expansion, as experiments in Section 5.4.2 have shown that these lexica achieve only low performance (about 22% recall) in this specific domain.

Again, we use the standard measures in *IR* to evaluate the efficiency of the query term expansion strategies: recall, precision and coverage.

6.3 Considering US patent classification

For semi-automatic *QTE* precision scores between 5% (in patent searching) and 17% (in professional academic search) are considered as acceptable, because users will be able to reject ones which are not relevant for their current search [46]. We initially use the US patent classification to improve recall and coverage of the class-specific lexical databases providing higher precision scores.

6.3.1 Query Term Expansion across different US patent classes

We evaluate the performance of the class-specific lexical databases when used for patents from other US patent classes. We assume that this will help to detect classes where cross-domain applications might be useful, in particular to improve recall and coverage of the class-specific lexical databases providing higher precision scores.

At first, we carry out experiments to improve the recall and coverage measures of the class-specific lexical databases. For that, we measure the performance of the best performing *csDB* when used for patents from other classes. In particular, to detect related classes, we calculate the recall measures. Again, we exclude query terms in the test sets that are out of the vocabulary of the lexical databases.

Table 48 and Table 49 in the appendix show the achieved recall measures of the *csDBs* providing synonyms and suggesting keyword phrases when used for the class they were based upon and the recall measures when used for other classes. We marked in bold the detected class-specific lexical databases that achieved best recall measures in other classes.

As shown, the class-specific lexical databases achieve respectable recall measures in other classes. For example, the lexical database providing keyword phrases for the class 280 called "*Land vehicles*" achieves a recall of almost 36% for class 180 called "*Motor vehicles*". The lexical database generated for class 623 called "*Prosthesis*" provides a recall measure of 37% for class 384 called "*Bearings*". (Note: The movement of two components against each other is common in both classes, for prosthesis as well as for bearings.) But such cross-class improvement is not necessarily reciprocal. We notice, for example, the lexical database of class 422 called "*Dentistry*" achieves at 43% a better recall measure for class 128 called "*Surgery*" than the corresponding lexical database extracted from class 128 when applied to class 422 with a recall score of only 24%.

We thus use the detected class-specific lexical databases that achieved best recall measures in other classes (marked in bold) to expand, in particular combine, the class-specific lexical databases. This leads to fifteen related class-specific lexical databases (*crDB*) providing synonyms and fifteen related class-specific lexical databases suggesting keyword phrases.

To measure the performance of the related class-specific lexical databases (*crDB*), we again use the test sets of each specific class and calculate recall, precision and coverage, as shown in Table 27 and Table 28.

Table 27. Recall, Precision and Coverage achieved when using related lexical databases *crDB* for suggesting synonyms

	Recall			**Precision**			**Coverage**		
US Class	*csDBs*	*crDB*	*ciDB*	*csDBs*	*ciDB*	*ciDB*	*csDB*	*crDB*	*ciDB*
454	48.39	45.76	54.10	21.13	13.63	1.88	52.83	64.78	90.57
384	26.09	23.40	58.21	39.58	29.83	5.68	44.55	67.14	96.04
126	32.56	29.69	51.43	10.74	5.28	3.32	64.60	84.21	95.58
148	46.28	47.33	61.39	18.78	12.15	5.82	56.64	69.27	80.42
219	40.54	47.32	66.67	21.84	8.26	2.74	62.50	69.40	92.08
433	45.31	34.56	59.60	10.51	9.45	2.33	66.47	66.66	90.59
180	68.00	58.16	47.42	13.34	12.21	5.35	51.47	59.84	82.35
417	64.00	65.65	78.00	10.09	8.95	5.43	67.01	78.33	91.75
398	56.67	57.12	59.21	10.70	9.04	5.63	69.72	69.93	85.21
280	51.72	53.56	56.52	12.24	6.99	4.48	61.46	65.45	88.54
128	32.09	37.98	45.21	44.17	10.64	7.14	66.60	74.86	82.57
379	67.91	61.11	72.26	5.66	5.45	4.15	81.48	82.70	84.13
422	59.83	63.08	70.00	7.22	6.53	6.29	79.51	79.75	85.25
439	59.10	66.15	68.97	6.02	4.30	2.66	72.13	69.23	85.25
623	73.33	65.76	61.90	22.05	7.87	3.79	75.63	80.82	88.24
Overall	51.45	49.78	60.73	16.94	10.04	4.45	64.84	72.16	87.90

Because we excluded synonyms that are out of the vocabulary with larger lexical databases, in particular for the class-related lexical databases, the recall scores go down in some of the classes. More synonyms and equivalent terms appear in the lexical databases but not as synonyms. But with a further increase of the lexical databases the recall scores increase, in particular for the class-independent lexical database. Now the synonyms and equivalent terms appear in the lexical database as synonyms.

Best improvement in recall is achieved for class 439. Recall increases from 59% achieved by the *csDB* to 66% provided by the *crDB*. Overall, the precision measures of the class-specific lexical databases degrade from 17% up to 10% when using *crDB* and further trop to 5% when using the *ciDB*.

Table 28. Recall, Precision and Coverage achieved when using related lexical databases *crDB* for suggesting keyword phrases

	Recall			**Precision**			**Coverage**		
US Class	*csDB*	*crDB*	*ciDB*	*csDB*	*crDB*	*ciDB*	*csDB*	*crDB*	*ciDB*
454	35.00	51.28	66.67	15.91	7.97	4.99	52.70	52.70	89.19
384	60.00	56.41	71.93	50.00	30.14	5.68	25.86	67.24	98.28
126	35.14	41.88	56.77	22.03	11.89	5.69	45.12	71.34	94.51
148	42.11	58.00	77.94	40.00	20.71	7.59	52.78	69.44	94.44
219	33.65	66.98	67.08	22.29	12.31	5.15	60.47	61.63	93.60
433	55.56	50.00	50.00	41.67	17.65	8.79	52.94	70.59	94.12
180	54.87	58.87	65.69	12.53	9.14	5.66	80.71	88.57	97.86
417	44.14	49.18	66.91	19.44	9.71	4.07	79.86	87.77	97.84
398	56.20	58.70	59.75	5.97	5.19	3.83	80.59	81.18	93.53
280	47.69	57.69	68.42	13.84	9.59	5.85	66.33	79.59	96.94
128	50.00	55.17	61.29	22.88	9.33	4.39	78.26	84.06	89.86
379	60.17	60.16	61.54	7.91	6.95	5.30	86.76	90.44	95.59
422	70.00	70.00	74.00	13.21	11.24	5.92	76.92	76.92	96.15
439	55.00	55.56	55.88	8.89	7.19	4.80	82.19	86.30	93.15
623	27.27	33.33	50.00	27.27	33.33	9.46	52.38	57.14	66.67
Overall	48.45	54.88	63.59	21.59	13.49	5.81	64.92	74.99	92.78

While recall of the lexical databases providing keyword phrases extracted from related classes obviously is still lower than from the class-independent lexical database, the recall values improve strongly over the class-specific databases.

Best improvement in recall is achieved for class 219, where the related lexical database *crDB* provides a recall of 67% compared to 34% for the class-specific *csDB*. This is virtually identical to the recall offered by *ciDB*, yet at a much higher precision (12% as opposed to 5% for *ciDB*). Overall, the recall measures of the class-specific lexical databases can be improved from 48% up to 55%, while precision drops from 22% to 13%.

Obviously, *ciDB* achieves best coverage for all classes. But the coverage of the class-specific lexical databases *csDB* can be significantly improved using the related lexical databases *crDB*, specifically for the classes where few query logs are available. In particular, for class 384 the coverage can be improved by about 41%.

To sum up, the experiments show for the lexical databases providing synonyms and suggesting keyword phrases that through the expansion of the class-specific lexical databases with related classes, recall and coverage increase considerably, while offering only a moderate drop in precision. This provides valuable expansion opportunities, starting first from class-specific lexical databases, followed by expansions using related classes, up to the most generic lexical database extracted from the entire corpus, i.e. *ciDB*. This is specifically valuable for smaller classes, i.e. for classes where few query logs are available. In the following section we will measure the performance of this expansion strategy.

6.3.2 Successively suggesting the *ETs* based on class information

From the previous section we notice that the lexical databases for suggesting synonyms and the databases for providing keyword phrases achieve similar results. Because we learned that synonym expansion is used much more often by the patent examiners to expand a query term (about 50% of the *ETs* are synonyms), we focus in the following experiments on synonymous query term expansion.

We carry out three expansion steps (*Step$_1$ to Step$_3$*). Initially, we use the US patent classes of the *QTs* and expand the terms with class-specific *ETs* using *csDBs* (*Step$_1$*). Following, we expand the *QTs* with further *ETs* appearing in related classes using *crDBs* (*Step$_2$*). Finally, we expand the *QTs* with additional *ETs* from all other classes using the *ciDB* (*Step$_3$*). Table 29 shows the achieved recall and precision scores. The highest and lowest recall and precision measures are marked in bold.

Table 29. Recall and Precision achieved when suggesting *ETs* based on class information

Expansion Step	*Expansion Terms*	Recall	Precision	avg. #terms
Step$_1$	*class-specific*	**49.38**	**18.50**	8
Step$_2$	*class-related*	56.33	9.90	25
Step$_3$	*class-independent*	**60.73**	**4.45**	79

As shown in *Step$_1$*, about half of the used *ETs* are provided by the class-specific *ETs* with the best precision score (19%). When providing further *ETs* from related US patent classes in *Step$_2$*, the recall measure can be further improved (up to 56%), while precision decreases (9%). In *Step$_3$*, when suggesting *ETs* from all other US patent classes, precision decreases to 12% and recall rises to 61%.

To further improve the precision scores, in particular of the class-related and class-independent lexical databases, we suggest to use the idea behind Relevance Feedback *RF* to take the *ETs* that are initially suggested (in *Step$_1$*) for a *QT* and to use information about whether or not those are relevant, in particular used by the examiners for *QTE* in the test set, to perform a new expansion step (*Step$_2$* and *Step$_3$*) in the following experiments.

6.3.3 Using *Relevance Feedback* to suggest *ET*

As mentioned before, we now use the US patent classification, in particular the class-specific, class-related and class-independent lexical databases, and Relevance Feedback *RF* to suggest possible *ETs* to the *QTs* from the test set.

Again, we carry out three expansion steps (*Step$_1$ to Step$_3$*). At first, we consider the US patent classes and expand the query terms with class-specific *ETs* (*Step$_1$*). Then, we expand the *ETs* used by the examiners in the test set from *Step$_1$* with further *ETs* appearing in related classes (*Step$_2$*). Finally, we expand the relevant *ETs* from *Step$_2$* with additional *ETs* from all other classes (*Step$_3$*). Table 30 shows the achieved recall and precision scores. The highest and lowest recall and precision measures are marked in bold.

Table 30. Recall and Precision achieved when using *RF* to suggest *ETs*

Expansion Step	***Expansion Terms***	**Recall**	**Precision**	**avg. #terms**
Step$_1$	*class-specific*	**49.38**	**18.50**	8
Step$_2$	*class-related*	50.86	17.37	10
Step$_3$	*class-independent*	**54.99**	**12.21**	21

Table 30 shows that, as before, after *Step$_1$* about half of the used *ETs* are provided by the class-specific *ETs* with the best precision score (19%) the recall measure can be further improved in *Step$_2$*, when suggesting additional *ETs* from related classes, while we notice only a minor decrease in precision (17%). Finally, in *Step$_3$* precision fall to 12% (still exceeding precision as mentioned in [46]) and recall rises to 55%.

Compared to suggesting *ETs* based on the patent classification without *RF*, as shown in the section before, precision can be further improved, but recall further decreases.

6.4 Using frequency information

In the previous section, we use the patent classification information to improve the precision scores. We now suggest using frequency information of the synonym relations in the query log collection to suggest possible *ETs* in the order of their frequency in the query log collection.

6.4.1 Suggesting the *ETs* based on frequency

For the experiments we utilize the confidence value *CV*. We measure the frequency of each synonym relation, i.e. that have a frequency of at least 1, greater than or equal to 2, 3, 4 and so on. As shown in Table 9, we observe that the number of the extracted unique synonym relations is strongly decreasing with the rise of the frequency. While 64,750 unique single term relations are provided with a frequency of at least 2, only 29,477 synonyms have a frequency of at least 3. The number of unique synonym relations is further decreasing. Less than 7,533 unique synonym relations exist with a frequency greater than 7.

The resulting scores based on the test set, when suggesting the *ETs* based on frequency are provided in Table 31. We marked in bold the highest and lowest recall and precision measures.

Table 31. Recall, Precision and Coverage achieved when considering frequency of the *ETs*

CV	Recall	Precision	Coverage
1	**69.54**	**1.21**	**86.49**
2	55.96	1.54	84.50
3	50.96	2.48	80.62
4	47.52	3.08	78.29
5	45.26	4.17	73.64
6	44.94	6.63	68.99
7	42.68	6.90	63.57
8	41.56	8.26	59.69
9	41.10	10.26	56.59
10	**39.44**	**13.41**	**55.04**

As shown, when considering a *CV* for the synonymous *ETs* a significant increase in precision can be observed. Best precision with a value of 13% is achieved when suggesting *ETs* having a *CV* of at least 10. But the recall score decreases considerably from 70% to 40%.

The experiments show that we can use the frequency information to iteratively suggest an increasing number of *ETs* as the search evolves. This allows the system to strike a reasonable balance between increasingly higher recall/ coverage by suggesting additional *ETs* that have a lower support in the set at the cost of lower precision after having initially suggested the most likely, highest-precision *ETs*.

To further improve the precision scores, we suggest in the following experiments to rank the *ETs* according to their frequency in the query log collection and successively suggest five expansion terms to each query term. We assume that in real query expansion scenarios patent searchers are willing to select *ETs* for a query term from a list of maximum five terms.

6.4.2 Successively suggesting the *ETs* based on frequency

First we rank the extracted synonym relations according to their frequency in the evaluation set. Then we carry out five expansion steps (*Step$_1$ to Step$_5$*) that appears to be an entirely realistic value in real query expansion sessions. We start with the top-5 *ETs* (having the highest ranking *r$_1$*) in *Step$_1$* followed by additional *ETs* based on the rankings *r$_2$* to *r$_5$* in *Step$_2$ to Step$_5$*. In particular, in *Step$_1$* we expand each *QT* from the test set with the *ETs* having the rank *r$_{1,}$* in *Step$_2$* we expand the *QTs* with *ETs* having the ranking *r$_2$* and so on.

For each expansion step we calculate recall and precision. For recall we consider the obtained recall scores of the previous expansion steps. Table 32 shows the achieved recall and precision scores. The highest and lowest recall and precision measures are marked in bold. Further, we indicate the performed expansions in each expansion step, in particular the percentage of queries for which 5, 10, 15, 20, and 25 expansion terms are suggested.

Table 32. Recall and Precision achieved when successively suggesting the highest ranked *ETs*

Expansion Step	***Ranking***	***Positions***	**Recall**	**Precision**	**Expansions**
Step$_1$	*r$_1$*	*1 – 5*	**38.46**	23.10	55%
Step$_2$	*r$_2$*	*6 – 10*	48.72	**24.81**	27%
Step$_3$	*r$_3$*	*11 – 15*	55.38	22.31	13%
Step$_4$	*r$_4$*	*16 – 20*	58.38	20.45	3%
Step$_5$	*r$_5$*	*21 - 25*	**62.54**	**20.00**	2%

As shown in Table 32, in each of the five expansion steps (*Step$_1$ to Step$_5$*), on average, 1 out of 5 terms that are suggested as synonyms were used by the examiners for query expansion (on average 22% precision). Further, already after *Step$_2$* about half of the used *ETs* (49% recall) are provided. In view of recall and precision achieved, when suggesting all possible *ETs* in one single step (on average 70% recall and 5% precision), there is a drastic increase in precision (up to 25%) and only a minor decrease in recall (63%). Coverage (86.49%) will not change as only the maximum number of *ETs* to a query term is limited. Hence, splitting the query term expansion process into multiple expansion steps helps to overcome the limitation in precision as noted in Chapter 5.

6.5 Considering word senses of the *ETs*

In the sections before, we used the US patent classification and frequency information to suggest the *ETs* in a useful order. We achieved precision scores, on average, up to 22%. Now we perform word sense disambiguation (*WSD*) to suggest the most suitable *ETs* and remove spurious expansions.

To determine the sense of an *ET* we consider the surrounding words (defining a window size of content words around each term) and measure the number of common words in the content words (overlap), as indicated in [72]. Window sizes range from n-grams, specifically unigrams, bigrams, and trigrams, to a full sentence or paragraph containing the target word. In addition, several positions of the surrounding words can be considered (to the left or right of the target word). We consider the *QTs*, which appear before the single term relations in the evaluation sets (reflecting query expansion scenarios, where information from immediately preceding queries can be used).

Because the number of surrounding words depends on the position of the single term relation in the query log collection (*ETs* used at the beginning in the query logs have no words, which appear before the *ETs* in the evaluation sets), we consider for the experiments only those relations which have at least a context size of n = 10 words (95% of all relations). For the experiments we use various context sizes of up to 20, 30, 40, and 50 words. We will not consider lager context sizes, because less than 6% of the relations have a size of n = 50 words.

6.5.1 Suggesting *ETs* based on overlap of word senses

At first, we rank the *ETs* according the number of common words (highest overlap) and initially suggest the highest ranked *ETs* (starting with at least 5 common terms) in *Step*$_1$ followed by additional ones (having 4, 3, 2 and 1 common terms) in *Steps*$_2$ to *Step5*. For the experiments we use a context size of n = 20 words. Again, we calculate recall and precision as shown in Table 33. We marked in bold the highest and lowest recall, precision and coverage measures.

Compared to the expansion strategies applied in the sections before, there is an increase in precision (up to 44% in *Step*$_1$). On average, almost half of the terms suggested in *Step*$_1$ are used by the patent examiners in the test set. In the following expansion steps (*Step*$_1$ *to Step*$_5$) the precision scores fall to 11% in *Step*$_5$*,* but still exceed the precision as obtained in Chapter 5.

However, with the usage of *WSD* also a decrease in recall and coverage has to be noticed. In particular, recall measures already decrease from 70% to 30%, when considering only one common term in the context words having a context size of n = 20 words.

Table 33. Recall and Precision achieved when suggesting the *ETs* based on overlap of sense definitions

Expansion Step	*Ranking*	*Overlap*	**Recall**	**Precision**	**Coverage**
$Step_1$	*r_1*	*≥ 5*	**6.06**	**44.44**	**4.58**
$Step_2$	*r_2*	*4*	9.09	37.50	11.01
$Step_3$	*r_3*	*3*	12.12	36.36	20.18
$Step_4$	*r_4*	*2*	18.18	19.35	30.28
$Step_5$	*r_5*	*1*	**30.30**	**11.24**	**47.71**

To overcome the limitation in recall as noted in the current experiments, we consider in the following experiments different context sizes (up to 50 terms) and measure the performance in automatic query term suggestion.

6.5.2 Using different context sizes for *WSD*

Again, we rank the *ETs* according the number of common words (highest overlap). Now we only suggest the highest ranked *ETs* (having at least 5 common terms) providing best precision scores (44.44%) as shown in the experiments before. We start with a context size of n = 20 words in *$Step_1$* and rise the context size (having up to 20, 30, 40, and 50 words) in *$Steps_2$* to *$Step_5$*. Thus, additional information from previous queries in the log files is successively used.

Table 34 shows achieved recall and precision when using different context sizes for *WSD*. The highest and lowest recall and precision measures are marked in bold.

Table 34. Recall and Precision achieved using different context sizes *n* for *WSD*

Expansion Step	***n***	**Recall**	**Precision**
$Step_1$	*20*	**6.06**	**44.44**
$Step_2$	*30*	26.61	33.52
$Step_3$	*40*	42.20	28.85
$Step_4$	*50*	**55.96**	**19.93**

The experiments show that the recall measures increase when expanding the context size. Best recall (56%) is achieved based on the largest context size. But with the expansion of the context size, we also notice a decrease in precision. While a precision score of 44% is achieved in $Step_1$, only 20% precision is obtained in $Step_4$. But the experiments also show that still a considerable decrease in recall has to be noticed (from 70% to 56%) compared to the recall measures achieved when suggesting all *ETs* without *WSD*.

The expansion approach as presented in Section 6.4.2, in particular to suggest *ETs* based on their frequency and successively suggest the *ETs* to each *QT*, provides still the best recall and precision performance (22% precision and 63% recall).

6.6 Conclusions

In this section we applied *QTE* strategies to improve the precision measures of the extracted lexical databases. We (1) used the US patent class-specific and class-related *ETs*, (2) suggested *ETs* based on their frequency in the set, and (3) suggested *ETs* based on overlap of sense definitions.

The experiments showed that the achieved precision scores (up to 25%) significantly exceed the scores achieved in related work for patent searching (about 5%) and are comparable to numbers reported for professional academic search (about 17%) [46] [102].

Table 35. Advantages and disadvantages of the expansion strategies

Expansion strategy	Section	+	-
class information (related classes) -	6.3.1	Precision increases from 6% up to 19%.	Recall decreases from 70% to 49%.
class information (successively)	6.3.2		
relevance feedback (*RF*)	6.3.3		
frequency information	6.4.1		
frequency information (successively)	6.4.2	Precision increases up to 25%.	**Minor decrease in recall (63%).**
word senses	6.5.1	Precision increases up to 44%.	Recall decreases to 56%.
word senses (context sizes)	6.5.2		

In particular, we notice only a minor decrease in recall (from 70 to 63%), when considering frequency of the extracted relations and successively suggesting the highest ranked *ETs* (while precision can be improved up to 22%), as presented in Section 6.4.2.

This expansion strategy fits very well with the recall-oriented patent search task and with query term expansion scenarios (as they occur in patent searching), where search sessions extend over many queries that are gradually refined. The recall measures of *PatNet* (about 70%) rise automatically, because the USPTO publish new query logs regularly. Table 35 shows the advantages and disadvantages of the different expansion strategies.

As regards the question whether we can devise means to suggest *ETs* in a useful order to avoid time-consuming term selection from a complete list of *ETs* or invention diagram, we recommend to guide users through the query expansion process through successively suggesting the highest ranked *ETs*, instead of limiting the number of suggested *ETs*. The latter had the effect that relevant *ETs* (available in *PatNet*) are not suggested. So users can decide by themselves, which *ETs* proposed for a query term are relevant for their current search - in any case they are all appropriate *ETs*, as patent examiners used them at least two times for expanding the query term. Furthermore, we do not recommend anyone of the other *QTE* strategies, because these strategies carry the risk that relevant *ETs* are not suggested.

∼ ∼

7 Effect of log based *QTE* on Retrieval Effectiveness

7.1 Introduction

In this chapter we study the impact of *QTE* using synonyms on patent document retrieval as described in [93]. We learned in Chapter 3 that synonym expansion is the most popular *QTE* method (about 50% of the expansions of the patent examiners are synonym expansions) in the patent domain. Synonyms are used to expand the query scope, in particular to improve recall. Limiting the query scope based on keyword phrases is rarely used. Only 6% of the expansions are used to generate keyword phrases. Further, up to now there has been little research on the effect of synonym expansion on retrieval effectiveness in the patent domain. Otherwise related experiments using keyword phrases for *QTE* are well studied and show that the retrieval performance, in particular precision, can be drastically improved, but recall decreases [2] [6]. Because patent search is a recall-oriented search task, we focus in the experiment on synonym expansion. We use the class-independent lexical database *PatNet* extracted from the query logs of USPTO patent examiners, which provide synonyms for a query term.

We conduct two experiments. First, we measure the performance of the lexical databases in automatic *QTE,* in particular if the retrieval performance can be improved compared to the baseline runs. Then we measure the retrieval performance of the lexical databases when used with related *QTE* approaches. This will show, if *PatNet* can assist other *QTE* approaches to improve the retrieval performance. All experiments will be performed on the CLEF-IP 2010 benchmark data set.

This chapter is organized as follows: At first, we explain the benchmark data set and the evaluation metrics used for evaluating the performance of *PatNet* in Section 7.2.. Then we present the used test to check statistical significance in Section 7.3. Following, we present the results achieved by the baseline runs in Section 7.4. In Section 7.5. we present the retrieval performance, when using *PatNet* for automatic *QTE*. Finally, we present in Section 7.6 our conclusions.

7.2 Benchmark Evaluation in the Patent Domain

To evaluate the retrieval effectiveness of an *IR* approach, benchmark evaluation is particularly common. Benchmark data sets contain at least one document collection, a set of query topics, and a set of query relevance judgments. For the patent domain the following initiatives provide such benchmark evaluation datasets, on which patent retrieval tasks have been carried out:

The main academic research in patent searching started after the third and fourth NTCIR workshop in 2003 and 2004, where the first test collections (containing full text Japanese patents published between 1998-1999 and Japanese and English exactly translated abstracts) have been made available [32] [38].

In 2009, a further initiative concerning patent *IR*, namely the CLEF Intellectual Property (CLEF-IP)[7] initiative, has been started. There have been various tasks in the workshops from 2009 to 2014, such as: *Prior Art Candidate Search Task (*find patent documents that are likely to constitute prior art to a given patent application), *Classification Task* (classify a given patent document according to the IPC), *Image-based Patent Retrieval* (find patent documents relevant to a given patent document), *Image-based Classification* (categorize given patent images into pre-defined categories of images) [77] [78] [79] [80]; Further task have been: *Flowchart Recognition, Chemical Structure Recognition, Passage retrieval starting from claims* (topics in this task are based on the claims in patent application documents)*, and Structure Recognition* [82];

A further initiative, in particular for evaluating chemical *IR* tools, has started in 2009. The TREC Chemical IR track focuses on evaluation of search technologies for retrieval and knowledge discovery from chemical patents and academic journal articles on chemistry [56].

7.2.1 Benchmark Dataset

For our experiments, we use in this thesis the data set of the CLEF-IP initiative, namely the CLEF-IP 2010 datasets. The data set consists of a document collection (2.6 Million documents) and a test set. All documents from the data set were obtained from the European Patent Office and are presented in XML format with annotations about different textual fields, such as title, abstract, description and claims, and metadata, such as inventors, assignees, and priority dates. The document collection consists of patents in three different languages, namely English, French and German.

The test set of the CLEF-IP data includes a subset of 1,348 English patent topics also referred to as query patents. Each topic is a full patent application including a title, an abstract, a description and a claims section. Figure 14 shows an example of such a query topic.

[7] http://ifs.tuwien.ac.at/ clef-ip/

```
<topic>
<num>PACt-1</num>
<narr>Find all patents in the collection that potentially invali-
date
patent application EP-1752549-A1.</narr>
<file>PACt-1_EP-1752549-A1.xml</file>
</topic>
```

Figure 14: Example of a query topic selected from CLEF-IP 2010

As ground truth data the patent citations of the query patents have been used. The citations were extracted by the organizers of the CLEF-IP from different sources: (1) the patent search reports, (2) the opposition procedures, and (3) from the patent documents themselves; In addition, query relevance judgments (qrels) have been built for the citations of the documents in the test set of the CLEF-IP data set. Two different relevance scales have been selected indicating the source of the cited document: Scale 1 indicates that the cited documents is disclosed by the applicant or cited by the examiner in the patent search report, Scale 2 indicates that the citation is mentioned in an opposition procedure. Table 36 shows the list of qrels for a topic selected from the training set of CLEF-IP 2010.

Table 36. Relevant documents for the query topic “PACt-1”

Query Topic	Relevant documents	Relevance Scale
PACt-1	EP-1473371-B1	1
PACt-1	EP-1473371-A3	1
PACt-1	EP-1473371-A2	2
PACt-1	EP-1356126-B1	2
PACt-1	EP-1356126-A2	1
PACt-1	EP-0484904-B1	1
PACt-1	EP-0484904-A3	1

7.2.2 Evaluation

To evaluate the performance of *IR* approaches the metrics *Precision, Recall, Average Precision* (*AP*) and *Mean Average Precision* (MAP) are particular common. We use these metrics to measure the retrieval performance of our query log based *QTE* approach. In Section 5.3.2 we already used these metrics to measure the performance of the lexical database in query term suggestion. We compared the suggested terms with the terms used by the examiners. Now to evaluate the retrieval performance, we compare the retrieved documents with the relevant documents, in particular with the documents cited by the patent examiners. We use the standard measures in IR, in particular

Recall and Precision. Further, the averaging over all the queries in the test set is performed to allow the reporting of the performance of a retrieval system over the full test set:

Average Precision (AP). Precision score is calculated at each position in the ranked list where a relevant document is retrieved, and then these precision scores are averaged [20].

$$AP_i = \frac{1}{|R_i|}\sum_{r\in R_i} P@rank(q_i, r) \tag{7.1}$$

where AP_i denotes the average precision for the *i*th query, *R* denotes a ranked list, *r* denotes a relevant document, and *P* denotes the precision.

Mean Average Precision (MAP). MAP measure is the mean of AP_i, where *Q* is the number of queries.

$$MAP = \frac{\sum_{q=1}^{Q} \; AP(\mathrm{q})}{Q} \tag{7.2}$$

In addition, we use the *Patent Retrieval Evaluation Score* (*PRES*), which is especially designed for recall-oriented applications [60]. The metric combines recall and the user's search effort in one single score. Following equation shows how *PRES* is calculated:

$$PRES = 1 - \frac{\frac{\sum r_{\mathrm{i}}}{n} - \frac{n+1}{2}}{N_{\mathrm{max}}} \tag{7.3}$$

where *Nmax* is the number of documents to be checked by the user (cut-off value), *n* is the number of relevant documents, and Σr_i is the summation of ranks of relevant documents, which is shown in the following equation:

$$\sum r_i = \sum_{i=1}^{nR} r_i + nR(N_{\mathrm{max}} + n) - \frac{nR(nr-1)}{2} \tag{7.4}$$

where *R:* Recall (number of relevant retrieved docs. in the first N_{max} documents).

Equation 7.4 shows the direct calculation of the summation of ranks of relevant documents in the general case, when some relevant documents are missing in the top N_{max} documents [60].

7.2.3 Best Official Results of the CLEF-IP Challenges

Table 37 shows the results, in particular MAP, Recall and PRES, of the best official results of CLEF-IP 2010 on the English subset of the test set [79].

Table 37. The performance of the best official results on English test set

Method	Rank	MAP	Recall	PRES
Humb [55]	1	**0.2264**	**0.6946**	**0.6149**
Dcu [59]	2	0.1807	0.6160	0.5167

In this chapter we will compare our *QTE* method with these approaches. We selected the two best official results on the English test set of CLEF-IP 2010 from the evaluation report.

7.3 Statistical Significance

In this chapter we compare the effect of *QTE* on the retrieval effectiveness based on the CLEF-IP benchmark data set. In order to check for statistical significant difference between the various *QTE* approaches, we run a t-test. The test allows us to conclude that there were a statistically significance ($p < 0.05$) [87]. The † symbol indicates that the improvement over the baseline is statistically significant.

7.4 Baseline runs

To test the effect of the lexical databases in patent searching, initial query terms that can be expanded have to be extracted from the query patents of the CLEF-IP test set. Several approaches have been presented to reduce the query patents of the test set to effective queries, as shown in Chapter 2. The goal of these approaches is to determine useful query terms for each patent document.

In the following subsections, we present the baseline queries and the expanded queries, which we used for *QTE*. Further, we present the retrieval performance of these queries based on the CLEF-IP 2010 benchmark document collection. Both the baseline queries and the retrieval results of the baseline queries have been made available by the authors of [66].

7.4.1 Query Generation

For our experiments we use the queries (1,348 queries) generated from the query topics based on the query model (*QS-BL)* as presented in [66]. *QS-BL* estimates the importance of each term according to a weighted log-likelihood based approach comparing the foreground (query patent) and background (collection) language models. Terms with high similarity to the foreground language model and low similarity to the background language model are used as query terms representing the specific terminology of the query patent. Top *k* terms (96 terms) with higher weights are selected as query terms from this query model. All fields of the query patents are considered in the query estimation process and *k* is experimentally set to 100.

To find out if our log-based query term expansion approach can assist other query expansion approaches to improve the retrieval performance, we use in addition to the baseline queries expanded, in particular reformulated query sets. The initial query set *QS-BL* is expanded using the information available in the citations of the query patent. Two different weighting algorithms are used for calculating query weights while taking into account the citation information. The first approach (*QS-PR*) uses PageRank scores to identify influential documents in the citation graph of a query patent and then uses those documents for drawing expansion terms. The second approach (*QS-TPR*) uses a time-aware decay function to give importance to newer documents in the citation graph and penalize older documents. Again, the top *k* terms (96 #terms) with higher weights are selected as query terms. Further explanations on the used queries can be found in [63]

Table 38 shows the number of unique query terms i.e. that have a frequency of at least 1 available in the query sets.

Table 38. Baseline query set

Query Set Name	*QS-BL*	*QS-PR*	*QS-TPR*
avg. #query terms/ topic	16,848	30,234	14,418

In total, we use for our experiments three query sets for the CLEF-IP 2010 test set. In the next subsection, we present the retrieval performance of the baseline queries over the CLEF-IP 2010 benchmark data set.

7.4.2 Retrieval Performance

We now present in this subsection the retrieval performance of the baseline query set and the expanded query sets with the best official results of the CLEF-IP 2010 benchmark data set. For the retrieval process the *Terrier* toolkit was used.[8] *Terrier* is an information retrieval system, which implements state of-the-art retrieval functionalities.

Table 39 shows the evaluation results using the CLEF-IP 2010 corpora in terms of MAP, Recall, and PRES at cut-off value of 1000. The effectiveness of the query sets is evaluated according to the performance of the final ranked list. We marked in bold the highest and lowest MAP, Recall and PRES value per column.

Table 39. Retrieval Results of the baseline queries compared to best official results of CLEF-IP 2010

Method	Rank	MAP		Recall		PRES	
		value	*change*	*value*	*change*	*value*	*change*
Humb [55]	1	**0.2264**	NA	**0.6946**	NA	**0.6149**	NA
Dcu [59]	2	0.1807	NA	**0.6160**	NA	0.5167	NA
QS-TPR	3	**0.1391**	+1.7%	0.6305 †	+1.4%	0.5123 †	+1.1%
QS-PR	4	0.1392	+1.7%	0.6302 †	+1.4%	0.5121 †	+2.1%
QS-BL	5	0.1368	NA	0.6215	NA	**0.5067**	NA

The results show that both query sets *QS-PR* and *QS-TPR* reformulated based on citation information obtained better performance compared to the baseline query set in view of recall and PRES. Further, the citation query model, which is based on citation information together with the publication dates, and the citation query model using Page Rank scores achieve similar performance in view of recall and PRES.

Further, Table 39 shows the position of the baseline and expanded baseline queries with respect to the best official results on the English subset of the test set of CLEF-IP 2010 participants.[9]

We can see that the approaches to build the baseline queries from the query topics and the approaches to expand these queries achieve similar results, in particular in view of Recall and PRES, as the approaches presented in [55] and in [59]. Further, it can be seen that the Humb method proposed by [55] and the Dcu method proposed by [59] still performed best in view of PRES and MAP. With reference to recall, the query models to build the baseline queries, which we use for *QTE*, can be considered as the second best methods on CLEF-IP 2010. In view of MAP the approaches to build the baseline

[8] available at http://ir.dcs.gla.ac.uk/terrier/

[9] http://www.ifs.tuwien.ac.at/clef-ip/pubs/CLEF-IP-2010-IRF-TR-2010-00003.pdf

queries are the third best methods on CLEF-IP 2010. The overall spread in MAP of CLEF-IP 2010 results is 0,0048 to 0,2264.

7.5 Query Term Expansion

In this section we use the lexical database *PatNet* to expand the query terms of the baseline query sets selected from the query documents and from the citation information with synonymous *ETs*. We use the most likely *ETs*, which are commonly used by patent examiners of the USPTO for *QTE*. For each expansion we use the highest ranked, in particular the most frequent, *ET* provided by the lexical database for a query term. Specifically, we replace the query terms in the baseline query sets with synonymous expansion terms for which *PatNet s*uggest *ETs*. We maintain the weights assigned to the query terms in the baseline query sets. We generate five additional query sets for the query topics.

In the following subsections, we present the expanded query sets used for measuring the effect of *PatNet* on retrieval effectiveness. Then we present the retrieval performance of these queries based on the CLEF-IP 2010 benchmark document collection.

7.5.1 Expanded Query Sets

Table 40 shows the number of queries, the number of query terms for each query and the average number of expansion terms for each query topic in the three expanded additional query sets.

Table 40. Expanded Query Sets for the CLEF-IP 2010 test set

Query Set Name	*QS-BLE*	*QS-PRE*	*QS-TPRE*
#queries in the query set	1,348	1,348	1,348
query terms/ query	96	96	96
avg. expansion terms/ query	55	35	43

For each of the baseline query sets one additional query set was built based on the expansion terms provided by *PatNet*. Most expansion terms (57%) are provided by *PatNet* for the baseline query set *QS-BL,* where the query terms were selected only from the English query documents. For the expanded baseline query sets *QS-PR* and *QS-TPR,* 36% and 45% of the query terms can be expanded with *PatNet*. The reason for the lower number of expansion is that the additional query terms from the citation information used in these query sets are no longer in English anymore. These query sets include also query terms in other languages as patent searching is multilingual, such as German, for example "*Kupplungsteil*" or "*Speichermedium*". However, *PatNet* is an English lexical database.

In the next subsection, we present the retrieval performance of the expanded query sets over the CLEF-IP 2010 benchmark data set.

7.5.2 Effect on Retrieval Performance

We now compare in this subsection the retrieval performance of the expanded query sets *QS-BLE, QS-PRE* and *QS-TPRE* with the baseline query sets and the expanded baseline queries using citation information.

Table 41 shows the evaluation results using the CLEF-IP 2010 corpora in terms of MAP, Recall, and PRES at cut-off value of 1000 for the baseline query set and for the expanded query set, when using *PatNet* for *QTE*. The results for the expanded query set *QS-BLE* are obtained by querying the baseline queries in combination with the additional expanded queries. We highlighted the best MAP, Recall and PRES values.

The results show that, when querying *QS-BL* in combination with *QS-BLE*, the retrieval performance drastically decreases. In particular, recall goes down (-19%) from 62% to 50%. Further, PRES decreases by 24% from 51% to 38%. This result means that a great number of relevant documents are moved lower and are lost from the ranked list (cut-off value of 1000), when using *PatNet* for *QTE* (as only the top ranked documents were considered). Also MAP decreases from 14% to 8%. Through the expansion many additional non-relevant documents were retrieved and appear in the ranked list. So through the usage of *PatNet* for automatic *QTE,* in particular to improve the retrieval performance (specifically the recall measure through expanding the query scope based on synonyms and equivalents), the opposite of what it was intended has been achieved.

Table 41. Retrieval Results when using *PatNet* for *QTE* and querying the baseline query sets and the expanded query sets simultaneously

Query Model	MAP		Recall		PRES	
	value	*change*	*value*	*change*	*value*	*change*
QS-BLE	0.0848 †	-38%	0.4983 †	-19%	0.3835 †	-24%
QS-PRE	**0.1390**	-0.1%	**0.6307**	+0.1%	**0.5123**	+0.1%
QS-TPRE	0.0066 †	-95%	0.1871 †	-70%	0.1238 †	-76%

Further Table 41, shows the retrieval results of the expanded query sets (*QS-TPRE and QS-PRE*) when using citation information and *PatNet* for *QTE*. Although the difference is not statistically significant, Recall and PRES can be slightly improved, while precision decreases slightly (-0.1%), when using *PatNet* with *QS-PR*. In combination with the second expansion approach (*QS-TPR*) the retrieval performance drastically decreases.

Generally, the experiments show that *PatNet* has the same effect on retrieval effectives, in particular on recall and precision, as the standard dictionary *WordNet* and the lexical database *SynSet* extracted from patent documents, as shown in [61]. Both lexical databases have also been tested on the CLEF-IP 2010 benchmark data set. MAP and PRES were lower than the baseline runs.

The previous experiments showed the retrieval performance, when querying the baseline query set and the expanded query sets simultaneously. Even though additional relevant documents can be retrieved the retrieval performance decreases compared to the baseline query set. We now test an expansion procedure, which reflects real query expansion scenarios, where initially the documents retrieved by the document terms are reviewed followed by further documents retrieved by an expanded query sets to address the problem as mentioned before: We build a combined ranked list by taking the first 500 retrieved results from *QS-BL* and then appending the first 500 documents from the list provided by *QS-BLE* (the queries which were expanded using *PatNet)*. Documents retrieved in the first ranked list have been ignored in the second ranked listed.

Table 42 shows the evaluation results in terms of MAP, Recall, and PRES at a cut-off value of 1000 and the difference to the baseline runs *QS-BL*, *QS-PR*, *QS-TPR*. Again, we highlighted the best MAP, Recall and PRES values.

Table 42. Retrieval Results when using *PatNet* for *QTE* and querying the baseline query sets and the expanded query sets separately

Query Model	MAP		Recall		PRES	
	value	*change*	*value*	*change*	*value*	*change*
QS-BLE	0.1360	-0.6%	0.5809 †	-6.5%	**0.4955** †	-2.2%
QS-PRE	**0.1382**	-0.7%	**0.5817** †	-7.7%	0.4985 †	-2.7%
QS-TPRE	**0.1382**	-0.6%	0.5769 †	-15%	0.4962 †	-9.5%

Again the retrieval performance of the combined list is inferior to the performance of the original query model *QS-BL*. In particular, recall decreases (-6,5%) from 62% to 58 and PRES decreases by 2,2% from 51% to 49%. MAP is stable by 14%. Additional documents can be retrieved, but initially retrieved documents (ranked in the second 500 retrieved documents of the initial queries) are not retrieved anymore.

Also the analysis of the rank positions of the additional retrieved documents (provided by *PatNet*) shows that only half of the additional provided documents have been considered in this expansion scenario (ranked in the top 500 retrieved documents). Hence, this method is not suitable to combine both retrieval results.

It has been shown in [8] that the average number of documents to be checked by a patent examiners is 100. So we calculate recall of the query models at cut-off value 100 in addition to the cut-off value 1000, which was specified by the track organizers of the CLEF-IP 2010 challenge. The results show at a cut-off value 100 that, when querying *QS-BL* in combination with *QS-BLE*, the retrieval performance also drastically decreases. In particular, recall goes down from 34% to 23%. Also recall of the expanded query sets (*QS-PRE* and *QS-TPRE*) goes down from 34% to 28% and from 28% to 5%, when querying the baseline query sets and the expanded query sets simultaneously and considering a cut-off value 100. We further analyze the recall measures between the cut-off value 100 and 1000 in intervals of 100. In each case recall decreases, when expanding the baseline query sets.

Because *PatNet* was extracted from query logs of specific patent US classes, we now consider only those query topics in the test set (659 query topics), which were classified in the same classes as *PatNet* was extracted from. Again, we compare the retrieval performance of the expanded query sets with the baseline query sets.

Table 43 shows the evaluation results in terms of MAP, Recall, and PRES at a cut-off value of 1000.

Table 43. Retrieval Results when considering patent classification of the query topics and the classes *PatNet* was extracted from

Method	MAP		Recall		PRES	
	value	*change*	*value*	*change*	*value*	*change*
QS-BL	0,1613	NA	0,5232	NA	0,5338	NA
QS-PR	0,1605	+0.5%	0,6475 †	+23%	0,5411	+1.6%
QS-TPR	0,1633	+1.2%	0,6492 †	+24%	0,5429	+1,7%
QS-BLE	0,0964 †	-40%	0,5233 †	-18%	0,4109 †	-23%
QS-PRE	0,1631	+0,8%	0,6503	+0.18%	0,5433	+0.1%
QS-TPRE	0,0066 †	-95%	0,1787 †	-72%	0,1190 †	-78%

The results show for the baseline query model *QS-BL* that also for the query topics, which were classified in the same classes as *PatNet* was extracted from, the retrieval performance drastically decreases. Recall decreases (-18%) from 64% to 52%. PRES goes down by 23% from 53% to 41%. Also MAP decreases from 16% to 10%. Also for these topics many additional non-relevant documents were retrieved through the expansion and appear in the ranked list.

Further the retrieval results of the expanded query set *QS-PRE* show that recall, PRES and precision can be slightly improved, even though the difference is not statistically significant. Again, in combination with the second expansion approach (*QS-TPR*) the retrieval performance drastically decreases.

As the experiments carried out on the CLEF-IP 2010 benchmark data set show that there was no improvement in the retrieval effectiveness using *PatNet* for fully-automatic *QTE*, we now analyze in the next subsection the results per topic (1348 topics) to validate whether there are certain characteristics that indicate when the approach comes in useful.

7.5.3 Analysis of the retrieval results

Table 44 shows the percentage of topics for which the retrieval performance is improved, remains unchanged, or is degraded, when expanding the baseline query set *QS-BL* with synonymous *ETs.*

Table 44. Percentage of topics which show improved, unchanged, and degraded performance compared to QS-*BL* using QS-*BLE*.

QS-BLE	**Recall**	**MAP**	**PRES**
improved	13.96%	23.46%	24.20%
unchanged	36.90%	2.52%	2.52%
degraded	49.15%	74.02%	73.27%

Expanding the baseline queries with synonyms improves the recall of 14% of query topics. That shows that the log-based *QTE* method can be useful for some of the topics. But for about 49% of the topics recall decreases. The proposed method has strong influence on the retrieval performance (only about 37% of the topics are unchanged). As expected through the expansion of the query scope, precision decreases for a large number of topics (74%). Through the expansion a lot of additional non-relevant documents are retrieved. Precision rises only for 23% of the topics.

Table 45 shows the recall measures achieved for improved, unchanged and degraded query topics, when expanding the baseline query set *QS-BL* with synonymous *ETs.*

Recall can be significantly improved (+34%) for query topics, which achieve, on average, only low recall (44%). Otherwise, recall drastically degrades (from 64% to 35%) when queries, which already achieve good recall measures, are expanded with *PatNet* (initially retrieved relevant documents are lost from the ranked list). Further, *PatNet* significantly outperform the related expansion approaches *QS-PR* and *QS-TPR,* which achieve only moderate recall for these query topics. The retrieval performance of these query topics are apparently difficult to improve with all types of expansion approaches tested.

Table 45. Recall achieved for improved, unchanged and degraded query topics

Recall	*QS-BL*	*QS-PR*	*QS-TPR*	*QS-BLE*
Avg.	0.6215	0.6302	0.6305	0.4983
improved	0.4407	0.5263	0.5298	0.5911
unchanged	0.6635	0.6726	0.6724	0.6635
degraded	0.6411	0.6289	0.6285	0.3478

To characterize for which queries the expansion performs better, we now try to detect commonalities. First, we consider the patent classifications of the query topics and the cited documents, and the classes *PatNet* was extracted from. We measure the overlap of the classes based on the queries for which the retrieval performance is improved, remains unchanged, or is degraded. The analysis shows for the query topics as well as for the citations that in each case (for improved, unchanged or degraded topics) about half of the query patents and citations are classified in the same classes as *PatNet* was extract from. So the patent classification is no criteria to detect queries for which the expansion performs better. Next, we evaluate whether the performance of the lexical database depends on the number of provided *ETs (n)*, the query topic length *(l)* or on the number of retrieved relevant documents *(c)*.

Table 46 shows the number of provided *ETs*, the query topic length, and the number of retrieved relevant documents for the improved, unchanged and degraded query topics.

Table 46. Query topic, query and citation characteristics

QS-BL		Avg.	Max.	Min.
improved	*n*	51	74	28
	l	14,959	133,762	1,280
	c	13	76	0
unchanged	*n*	56	75	26
	l	11,174	110,506	1,513
	c	11	57	0
degraded	*n*	56	77	19
	l	12,370	102,371	1,509
	c	16	85	1

The performance of *PatNet* is independent from the number of provided *ETs* and from the query topic length. We consider the number of character strings of each query patent. For improved, unchanged or degraded query topics virtually the same number of *ETs* are used for query expansion. Further, query topics have, on average, virtually equivalent topic lengths showing that *PatNet* can be used both for shorter topics and for longer query topics. Also the number of retrieved relevant documents is no criteria to detect when *PatNet* comes in useful.

Table 47 shows the rank positions of the relevant documents provided by the baseline query set to detect whether it is an issue of being to generic or not found via the given query terms. The latter would argue for extending the query scope using synonyms.

Table 47. Rank positions of the retrieved relevant documents provided by *QS-BL*

QS-BL	**1 - 250**	**251 - 500**	**501 - 750**	**751 - 1000**
improved	65%	16%	11%	8%
unchanged	79%	12%	6%	3%
degraded	63%	18%	11%	8%

As shown more than two-thirds of the retrieved relevant documents appear in the ranked lists under the top 250 documents. Less than 8% appear in the last 250 documents. These distributions of the documents speaks for extending the query scope using synonyms. But the experiments indicate just the opposite.

Finally, we consider the patent conventions and countries the relevant documents have been filed to detect, whether it is an issue that *PatNet* was extracted only from US patents. In each case about half of the relevant documents are *EP* or *WO* patents and about one third of the topics are US patents. There is no increase of US patents for improved query topics.

7.6 Conclusions

The experiments show that the retrieval performance of the automatic query generation and expansion models decreases or can only be slightly improved, when using *PatNet* for fully-automatic *QTE*. So synonym expansion has generally no positive effect on the retrieval performance.

In particular, through the expansion of the initial query terms with synonyms and equivalents the query scope of the query topics is radically expanded. Many additional relevant and non-relevant documents will be retrieved. While for some of the topics the retrieval performance can be improved, a great number of relevant documents are moved lower or are lost from the ranked list (cut-off value of 1000), when using *PatNet* for *QTE*. As only the top ranked documents were considered, the retrieval performance decreases (despite additional relevant documents being retrieved). Hence, *PatNet*, in particular expanding query terms with synonyms and equivalents, is not a suitable method for expanding queries for patent searching in a fully-automatic manner.

But the analysis of the retrieval results, in particular Table 44, shows that the query log-based *QTE* method does not have generally a negative effect on the retrieval effectiveness. Recall is drastically improved (+34%) for query topics, where the baseline queries achieve only low recall values (44%), as shown in Table 45. But we have not detected any commonality that allows us to characterize these queries. So we recommend using *PatNet* as a lexical resource for semi-automatic *QTE* in Boolean patent retrieval, where synonym expansion is particularly common to improve recall and tracking of the results is possible to expand baseline queries achieving only low recall values.

8 Conclusions and Future Work

In this thesis we investigated the problem of *QTE* in the query generation step of patent searching with the goal of suggesting relevant expansion terms, in particular synonyms and equivalents, to a query term in a semi-automatic or fully automatic manner for Boolean retrieval.

We analyzed query sessions of patent examiners of the United Patent and Trademark Office and proposed approaches to extract lexical knowledge from the query log files. We presented patent domain-specific, in particular US class-specific, class-related and class-independent lexical databases, which provide patent domain specific vocabulary and semantic relations to assist patent searchers in formulating Boolean queries. Further, we applied the detected lexical databases to query sessions including real query expansions of patent examiners to evaluate our query term expansion approaches. Finally, we evaluated our query log-based query term expansion approach in patent searching using benchmark data sets for the patent domain.

In this chapter, we initially conclude our thesis in Section 8.1 and then point out possible directions for future work in Section 8.2.

8.1 Summary and Contributions

Now we summarize the individual chapters of this thesis in view of the research questions RQ1 to RQ5 as presented in Chapter 1:

8.1.1 Analyzing Query Logs of the USPTO

In Chapter 3, we analyzed query logs of patent examiners to assist patent searchers in formulating Boolean queries, in particular to expand the searchable features of an invention diagram, as shown in Figure 1, with additional query terms in a semi-automatic or fully-automatic manner (RQ1). In particular, we provide answers to the research questions RQ1(1) *What type of expansion terms and semantic relations are used by the patent searchers for query term expansion in real sessions?* and RQ1(2) *What are the most frequently used expansion terms and semantic relations?*

First, we introduced and analyzed query logs of USPTO patent examiners. We showed that query generation in patent searching is highly domain specific. In particular, the results of the analysis of the query logs show that the majority of the query terms are expansion terms, which do not appear in the query document. So query terms selected from the query document are frequently expanded with *ETs* by brainstorming. But the majority of the used vocabulary for *QTE* appears in the specific US patent class. So the query log files of the same US patent class are valuable resources to provide lexical knowledge for the patent domain.

Further, we tried to find out what type of expansion terms and semantic relations are used by the patent searchers for *QTE* in real sessions. The results show that means to enhance query generation in patent search, in particular to support patent experts in formulating Boolean queries, are to suggest (1) synonyms and equivalents, (2) co-occurring terms and (3) keyword phrases. In particular, suggesting synonyms is of particular importance, as almost 50% of the used *ETs* are synonyms or equivalents.

8.1.2 Acquiring lexical knowledge from the query logs

In Chapter 4, we studied how we can leverage the query logs of patent examiners of the United States Patent and Trademark Office for automatic query term expansion (RQ2). In particular, we provide answers to the research questions RQ2(1) *Can we use the query logs to extract lexical knowledge directly from the patent domain?* and RQ2(2) *How can we assist patent searchers in query term expansion to formulate Boolean queries based on this extract lexical knowledge?*

We extracted lexical knowledge from the query logs of USPTO patent examiners for automatic query term expansion. We detected keyword phrases and synonym relations in query logs, which patent examiners of the USPTO created during the validation procedure of the patent applications.

Further, we generated two lexical databases, in particular *PhraseNet* and *PatNet,* to support patent experts in formulating Boolean queries. The lexical databases suggest keyword phrases to narrow down a search, particularly to limit the scope of a general query term, and provides several types of synonym relations, in particular (1) single terms to single term, (2) single term to phrase and (3) phrase to phrase relations, to expand the query scope.

Based on multiple examples, we have shown that the lexical databases can help patent searchers in the query generation process, in particularly in generating the invention diagram in a semi-automatic manner, which is used by the searchers for generating Boolean queries.

8.1.3 Automatic query term expansion based on query logs

In Chapter 5, we explored how an automatic query expansion strategy based on query logs perform in query term expansion compared to the manual expansion performed by experts (RQ3). In particular, we provide answers to the research questions RQ3(1) *Can we evaluate our query term expansion approach based on real query expansion scenarios?* and RQ3(2) *Does the query log based query expansion approach outperform standard dictionaries?*

First, we tried to find out if the performance of our query term expansion approach depends on the query log collection and class size, and if there any advantages in using the classification system to build class-specific lexical databases. Further, we calculated whether the query log based query term expansion approach outperform standard dictionaries.

For the experiments we used the query logs to extract patent domain-specific, in particular US class-specific and class-independent lexical databases. Further, we present a set of real query sessions including real query expansions of patent examiners to evaluate query term expansion approaches.

The evaluation of the extracted lexical databases has shown that recall and coverage measures increase with the availability of a larger set of query logs. On average, up to 8 out of 10 *ETs*, which are used by the examiners for query term expansion, are suggested by the class-independent lexical databases. Expectedly, the class-specific and class-related databases achieve better precision scores than the class-independent databases. On average, 1 out of 20 suggested *ETs,* which were suggested by the class-independent databases, were used by the examiners for *QTE*. This is similar to numbers achieved in related work for patent searching (about 5%) [46].

We also considered characteristics of the query logs that we used for evaluation, in particular the length of the query logs/ search sessions. We found that equivalent recall, precision, and coverage scores can be achieved for all subsets having different query log lengths. Hence, the performance of our log based expansion approach is independent from the length of the query sessions.

Finally, the results of the evaluation show that the specific lexical databases drastically outperform the general-purpose source *WordNet.* The standard dictionary *WordNet* achieves for all US patent classes only low performance in recall. The patent domain specific relations are not included in the dictionary. Examiners expand query terms with: general terms, w.r.t. part of speech, popular trademarks, terms which have the same meaning in specific classes, and terms created by themselves.

8.1.4 Suggesting *ETs* in a useful order

We studied in Chapter 6, how the query log-based query term expansion model can be optimized to carry out effective *QTE* (RQ4). In particular, we provide answers to the research questions RQ4(1) *Are there weights available with the query logs to suggest expansion terms to a query term in a useful order?* and RQ4(2) Does the involvement of information from past queries improve the query log based query term expansion model?

We proposed query term expansion strategies for our modeled expansion approach to avoid time-consuming expansion term selection. In particular, we (1) used patent US class-specific and class-related *ETs,* (2) successively suggested *ETs* based on their frequency in the evaluation set, and (3) suggested *ETs* based on overlap of sense definitions.

The results of the experiments showed that the achieved precision scores (about 20%) significantly exceed the scores achieved in related work for patent searching (about 5%) and are comparable to numbers reported for professional academic search (about 17%) [102].

In particular, only a minor decrease in recall (from 70 to 63%) has been noticed, when considering frequency of the extracted relations and successively suggesting the highest ranked *ETs* (while precision can be improved up to 22%). This expansion strategy fits very well with the recall-oriented patent search task and with query term expansion scenarios (as they occur in patent searching), where search sessions extend over many queries that are gradually refined (RQ4).

To avoid time-consuming term selection from a complete list of *ETs* or invention diagram, we recommend to guide users through the query expansion process, instead of limiting the number of suggested *ETs*. The latter had the effect that relevant *ETs* (available in *PatNet*) are not suggested.

8.1.5 Evaluation in *IR*

In Chapter 7, we studied whether a query log based expansion approach improve retrieval effectiveness in patent searching (RQ5). More detailed sub-questions were: RQ5(1) *Does log based query expansion outperform standard approaches, which are commonly based on terms selected from patent documents?*and RQ5(2) *Can a log based query expansion approach assist related expansion approaches to improve the retrieval performance?*

In particular, we evaluated our log based query term expansion approach based on a benchmark data for patent search. We used the CLEF-IP 2010 benchmark data set and measured the effect of synonymous query term expansion on retrieval effectiveness in patent searching.

The experiments show that the retrieval performance decreases or can only be slightly improved, when using *PatNet* for fully-automatic *QTE*. No significant improvement can be recognized. Through the expansion of the initial query terms with synonyms and equivalents the query scope of the query topics is radically expanded. Many additional relevant and non-relevant documents will be retrieved. So synonym expansion in fully automatic manner has generally no positive effect on the retrieval performance.

But recall is drastically improved for query topics, where the baseline queries achieve, on average, only low recall values. So the query log-based *QTE* method does not have generally a negative effect on the retrieval effectiveness. On the contrary, the approach works fine for queries, where the expansion of the query scope is needed. Unfortunately, we could not detected any commonality that allows us to characterize these queries to use the approach in a fully automatic manner. So we recommend using *PatNet* as a lexical resource for semi-automatic *QTE* in Boolean patent retrieval, in particular in the search systems used by the patent offices and commercial operators, where synonym expansion is particularly common to improve recall and tracking of the results is possible to selectively expand the baseline queries with synonyms.

8.2 Future Directions

Following we propose some possible future work for this line of research:

Multilingual Thesauri. Patent retrieval is an essentially multilingual search task. Virtually all search systems of the patent offices and commercial operators enable for searching patents in different languages. Hence, providing a multilingual lexical resource for automatic *QTE* to search patents in different languages is essential.

Yet, our query log-based query term expansion approach is limited to suggesting possible expansion terms, in particular synonyms and equivalents, in English. As the USPTO is the only source known to us which publishes the query logs for the patent domain, future work could try to compute translations to the extracted lexical knowledge.

One interesting work would be to find the translations directly in the patent domain, as our experiments have shown that the highly specific vocabulary used in the settings of patent applications is not included in standard dictionaries. For example, patent family members or translations of parts of the patent documents, as shown in [40] [41], could be valuable resources to compute translations to the query and expansion terms extracted from the query log files. In particular, the claim sections of granted European patents include the claims in English, German and French.

Boolean Query Formulation. As mentioned above, in the patent domain Boolean retrieval is particularly common. Virtually all search systems of the patent offices and commercial operators process Boolean queries. But despite the importance of Boolean retrieval, there is not much work in current research assisting patent experts in formulating such queries. In addition, the amount of work concerning the design of an interface for the patent domain, in particular for issuing Boolean queries, is particularly limited.

Hence, an interesting work would be to find an appropriate interface for automatic Boolean query term expansion. For example, similar to an invention diagram a user could be allowed to enter the query terms and the system automatically expands and assembles search queries.

Because the majority of systems used by the patent searchers are commercial ones and contain their own search interfaces varying in the searchable fields of the patent application considerable, a further interesting work would be to find an automatic Boolean query formulation model, which assembles effective search queries dependent on the searchable fields of the selected search system. For example, the search system of the German patent office allows searching the whole patent applications. However, the system of the European patent office enables only title and abstract search. To formulate effective queries, the queries have to be adapted to the search interfaces involving time-consuming query formulation, in particular query reformulation.

Academic professional search. Professional academic search is a form of search that is carried out by scientists. Like patent search it takes place in a specific domain and it tends to be recall-oriented. Further, multiple queries are needed to satisfy one information need. Hence, providing additional query and expansion terms is also in academic search essential [102].

Also in previous work for academic search, search engine query logs are used as sources for possible expansion terms. Especially, query terms of other users and user's own previous queries are used as a source for query term suggestion.

One interesting work could be to find out, if the lexical knowledge extracted from query logs (created by patent experts) could assist scientist in professional academic search, in particular for automatic *QTE*, as the goal of professional academic search is similar to the keyword-based patent search task. In favor, for example, the extracted lexical databases *PatNet* and *PhraseNet* could be used in real query sessions, as we used them for our query term expansion experiments in Section 5.

~ ~

Appendix

Table 48. Achieved recall measures of the *csDB* providing synonyms when used for the class they were based upon and for the other classes

US class based on used for	454	384	126	148	219	433	180	417	398	280	128	379	422	439	623
454	**48,4**	4,3	5,7	11,9	11,5	12,0	10,3	15,0	6,3	16,7	6,9	17,4	16,7	**41,7**	0,0
384	12,5	**26,1**	12,5	18,8	11,1	29,4	6,3	26,3	0,0	0,0	10,4	0,0	22,6	22,2	22,2
126	22,9	0,0	**32,6**	8,1	7,1	21,4	12,5	12,5	6,3	0,0	0,3	0,0	8,5	20,0	12,5
148	10,0	**21,6**	0,0	**46,3**	18,6	**54,5**	4,1	7,1	**25,0**	0,0	10,5	0,0	22,2	28,6	9,1
219	**41,2**	2,9	5,7	10,0	**41,5**	14,3	10,3	15,8	6,3	11,1	3,5	**30,8**	9,8	35,7	0,0
433	22,2	8,7	8,3	15,2	14,4	**45,3**	5,1	28,6	6,7	13,0	17,8	13,1	20,0	17,6	21,3
180	28,3	5,9	19,0	7,7	14,3	28,9	**68,0**	20,7	13,5	**25,0**	7,7	6,3	18,5	27,3	13,0
417	26,2	7,9	10,0	10,0	12,9	15,8	5,3	**64,0**	14,8	22,7	4,7	10,3	**32,3**	15,0	15,4
398	17,4	0,0	9,1	5,7	13,6	28,6	7,4	27,3	**56,7**	7,1	7,1	17,6	10,0	5,0	7,1
280	27,3	7,0	16,1	7,9	23,4	23,5	**30,4**	15,0	4,2	**51,7**	7,9	1,5	12,5	15,0	14,3
128	19,6	7,1	14,0	4,7	21,9	24,7	7,4	27,9	**25,0**	18,2	**32,1**	8,0	12,4	32,0	**38,9**
379	21,4	0,0	6,3	8,5	16,9	27,5	21,1	35,0	24,1	18,8	19,6	**67,9**	6,8	0,0	5,9
422	29,1	17,3	**25,4**	20,5	**30,8**	29,1	13,9	**40,0**	19,3	19,4	**32,0**	18,6	**59,8**	19,2	26,0
439	30,3	19,1	12,9	8,2	13,7	13,7	4,0	30,0	7,9	12,5	1,8	12,5	9,8	**60,0**	3,8
623	18,4	5,1	14,3	**28,1**	12,6	30,2	12,1	20,0	16,3	21,4	19,5	10,3	15,4	28,6	**73,3**

Table 49. Achieved recall measures of the *csDB* providing keyword phrases when used for the class they were based upon and for the other classes

US class based on used for	454	384	126	148	219	433	180	417	398	280	128	379	422	439	623
454	**35.0**	0.0	**39.3**	9.5	30.3	15.2	22.2	28.1	25.9	16.7	27.5	22.6	42.9	**37.8**	10.7
384	8.3	**60.0**	23.5	30.8	21.7	26.1	32.3	**42.9**	6.7	38.1	32.0	16.7	34.5	34.3	**37.0**
126	20.5	**26.3**	**35.1**	13.6	20.8	11.9	23.7	22.4	15.9	24.1	18.0	24.1	32.0	22.9	19.7
148	15.4	8.3	22.2	**42.1**	**41.5**	24.2	16.0	34.5	20.0	17.4	22.5	26.7	41.4	25.6	17.9
219	18.0	6.3	30.0	13.8	**33.0**	21.6	21.2	30.7	27.1	19.2	39.4	23.6	**43.9**	31.5	21.9
433	0.0	25.0	20.0	0.0	25.0	**55.6**	42.9	33.3	0.0	28.6	14.3	33.3	14.3	0.0	20.0
180	27.3	15.4	12.0	10.0	18.2	**28.0**	**54.9**	30.1	18.6	**35.7**	16.1	16.7	17.7	24.7	15.4
417	13.7	3.3	19.6	10.3	21.4	23.0	34.4	**44.1**	18.0	19.3	20.3	13.3	34.0	22.5	15.2
398	**34.5**	0.0	7.5	5.6	12.9	8.3	9.3	10.5	**55.8**	5.8	14.0	**27.2**	16.7	16.5	4.4
280	26.3	5.9	4.4	12.0	20.0	11.1	**55.0**	25.0	13.6	**47.7**	21.7	29.4	25.6	26.7	19.5
128	17.4	0.0	12.5	5.6	31.3	20.0	22.9	29.7	26.9	20.0	**50.0**	25.0	43.2	20.5	15.2
379	7.4	8.3	0.0	9.1	14.3	15.0	5.6	7.7	22.9	5.4	14.1	**60.2**	17.0	16.4	3.1
422	7.7	12.5	30.8	**31.6**	27.8	21.7	9.5	22.7	5.6	0.0	23.5	10.0	**70.0**	10.7	8.3
439	10.0	0.0	30.8	27.3	34.6	27.6	29.0	24.4	36.4	17.5	17.1	24.1	28.2	**55.0**	11.5
623	0.0	0.0	0.0	0.0	0.0	11.1	0.0	0.0	**50.0**	0.0	**60.0**	0.0	40.0	33.3	**27.3**

Table 50. Number of synonyms provided by the lexical databases *csDB[1-5]*

US Class	$csDB_1$		$csDB_2$		$csDB_3$		$csDB_4$		$csDB_5$	
	#rel	#terms	#rel	#terms	#rel	#terms	#rel	#terms	#rel	#terms
454	331	419	683	784	-	-	-	-	-	-
384	197	279	424	516	-	-	-	-	-	-
126	305	344	635	655	968	946	1,421	1,327	-	-
148	434	459	989	952	1,376	1,268	1,938	1,673	-	-
219	206	285	490	620	805	930	1,134	1,247	1,158	1,267
433	2,702	2,657	3,062	2,938	-	-	-	-	-	-
180	717	1,882	2,369	2,180	2,717	2,430	2,925	2,573	-	-
417	1,519	1,476	1,816	1,732	2,172	1,983	2,424	2,193	-	-
398	1,715	1,577	1,962	1,751	2,183	1,920	2,424	2,092	2,653	2,245
280	944	951	1,209	1,171	1,437	1,353	1,689	1,531	1,876	1,676
128	5,550	4,866	6,257	5,387	7,180	5,922	-	-	-	-
379	5,642	3,726	6,731	4,285	7,670	4,746	8,492	5,111	-	-
422	6,317	5,076	7,451	5,770	8,458	6,398	9,670	7,123	10,411	7,532
439	1,830	1,556	2,258	1,864	2,814	2,251	3,110	2,457	3,334	2,616
623	4,819	3,780	5,297	4,090	5,727	4,326	6,575	4,815	7,240	5,193

Table 51. Number of keyword phrases provided by the lexical databases *csDB[1-5]*

US Class	$csDB_1$		$csDB_2$		$csDB_3$		$csDB_4$		$csDB_5$	
	#rel	#terms	#rel	#terms	#rel	#terms	#rel	#terms	#rel	#terms
454	205	191	275	262	-	-	-	-	-	-
384	98	97	207	202	-	-	-	-	-	-
126	72	66	183	178	334	313	493	452	-	-
148	148	145	235	174	376	362	531	494	-	-
219	308	284	559	504	703	638	851	769	1,045	941
433	1,170	1,006	1,245	1,107	-	-	-	-	-	-
180	1,289	1,062	1,503	1,144	1,694	1,332	1,832	1,508	-	-
417	950	821	1,115	952	1,309	1,101	1,455	1,217	-	-
398	1,556	1,248	1,786	1,381	2,035	1,547	2,272	796	2,531	1,785
280	973	853	1,085	931	1,214	1,036	1,362	1,150	1,460	1,217
128	2,158	1,897	2,505	2,173	2,848	2,439	-	-	-	-
379	2,501	1,811	2,944	2,094	3,328	2,312	3,560	2,457	-	-
422	2,365	1,912	2,881	2,295	3,158	2,471	3,539	2,723	3,794	2,891
439	2,455	1,811	2,814	2,031	3,124	2,233	3,353	2,270	3,594	2,416
623	1,328	1,144	1,491	1,279	1,736	1,480	1,990	1,677	2,099	1,771

Bibliography

[1] Alberts, D., Yang, C., Fobare-DePonio, D., Koubek, K., Robins, S., Rodgers, M., Simmons, E., De Marco, D. 2011. Introduction to Patent Searching. In Current Challenges in Patent Information Retrieval. The Information Retrieval Series. Volume 29, pp. 3-43.

[2] Al-Shboul, B., Sung-Hyon Myaeng, S. 2014. Wikipedia-based query phrase expansion in patent class search. In Information Retrieval, Volume 17, Issue 5-6, pp. 430-451.

[3] Amitay, E., Broder, A. 2008. Introduction to special issue on query log analysis: Technology and Ethics. In ACM Transactions on the Web, Volume 2, Issue 4, Article 18.

[4] Andersson, L., Lupu, M., Palotti, J., Hanbury, A., Rauber, A. 2016. When is the Time Ripe for Natural Language Processing for Patent Passage Retrieval? In Proc. of the 25th Int. Conf. on Inf. and Knowledge Man. (CIKM2016), Indianapolis,USA, pp. 1453-1462.

[5] Andersson, L., Mahdabi, P., Hanbury, A., Rauber, A. 2012. Report on the CLEF-IP 2012 Experiments: Exploring Passage Retrieval with the PIPExtractor. In Proc. of CLEF (Notebook Papers/LABs/Workshops).

[6] Andersson, L., Mahdabi, P., Hanbury, A., Rauber, A. 2013. Exploring patent passage retrieval using nouns phrases. In Proc. of the 35th European Conf. on Advances in Information Retrieval (ECIR 2013), Moscow, Russia, pp. 676-679.

[7] Atkinson, K. 2008. Toward a more rational patent search paradigm. In Proc. of the 1st ACM workshop on Patent Information Retrieval (PAIR 2008), pp. 37–40.

[8] Azzopardi, L., Vanderbauwhede, W. and Joho, H. 2010. Search system requirements of patent analysts. In Proc. of the 33rd Int. Conf. on Research and Development in Information Retrieval (SIGIR2010). Geneva, Switzerland, pp. 775-776.

[9] Baeza-Yates, R., Hurtado, C. , Mendoza, M. 2004. Query Recommendation Using Query Logs in Search Engines. In Proc. of the 2004 Int. Conf. on Current Trends in Database Technology (EDBT 2004), Crete, Greece, pp. 588-596.

[10] Baeza-Yates, R., Tiberi, A. 2007. Extracting semantic relations from query logs. In Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2007), San Jose, California, USA, pp. 76-85.

[11] Ballesteros, L., Croft, W. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In Proc. of the 20th Int. Conf. on Research and Development in Information Retrieval (SIGIR 1997), Philadelphia,USA, pp. 84-91.

[12] Bashar, A., Myaeng, S. 2011. Query phrase expansion using wikipedia in patent class search. In Proc. of the 7th Asia Conf. on Information Retrieval Technology (AIRS'11), Dubai, United Arab Emirates, pp. 115-126.

[13] Bashir, S., Rauber, A. 2009. Identification of Low/High Retrievable Patents using Content-Based Features. In Proc. of 2nd Int. Workshop on Patent Information Retrieval (PaIR 2009), Hong Kong, pp. 9-16.

[14] Bashir, S., Rauber, A. 2009. Improving Retrievability and Recall by Automatic Corpus Partitioning. In Transactions on Large-Scale Data- and Knowledge Centered Systems, pp. 122-140.

[15] Bashir, S., Rauber, A. 2009. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In Proc. of the 18th Int. Conf. on Information and Knowledge Management (CIKM 2009), Hong Kong, China, pp.1863-1866.

[16] Bashir, S., Rauber, A. 2010. Improving Retrievability of Patents in Prior-Art Search. In Proc. of European Conf. on Information Retrieval (ECIR2010), Milton Keynes, United Kingdom, pp. 457–470.

[17] Bhatia, S., Qi He, B.H., Spangler, S. 2012. A scalable approach for performing proximal search for verbose patent search queries. In Proc. of the 21st Int. Conf. on Information and Knowledge Management (CIKM 2012). Maui, Hawai, pp. 2603-2606.

[18] Bhogal, J., Macfarlane, A., Smith, P. 2007. A review of ontology based query expansion. In Information Processing Management, Volume 43, Issue 4, pp. 866-886.

[19] Billerbeck, B., Scholer, F. , Williams, H. E., Zobel, J. 2003. Query expansion using associated queries. In Proc. of the 12th Int. Conf. on Information and Knowledge Management (CIKM 2003), New Orleans, USA, pp. 2-9.

[20] Buckley, C., Voorhees, E. 2000. Evaluating evaluation measure stability. In Proc. of the 23rd Int. Conf. on Research and Development in Information Retrieval (SIGIR 2000), Athens, Greece, pp. 33–40.

[21] Carpineto, C., Giovanni Romano, G. 2012. A Survey of Automatic Query Expansion in Information Retrieval. In ACM Computing Surveys, Volume 44, Issue 1, Article 1.

[22] Cetintas, S. and Luo Si, L. 2012. Effective query generation and postprocessing strategies for prior art patent search. In Journal of the American Society for Information Sience and Technology, Volume 63, Issue 3, pp. 512-527.

[23] Chen, X., Peng, Z., Zeng, C. 2012. A co-training based method for chinese patent semantic annotation. In Proc. of the 21st Int. Conf. on Information and Knowledge Management (CIKM 2012). Maui, Hawai, pp. 2379-2382.

[24] Croft, W., Metzler, D., Strohman, T. 2010. Search Engines - Information Retrieval in Practice. Addison Wesley.

[25] Cui, H., Wen, J.-R., Nie, J.-Y. , Ma, W.-Y. 2002. Probabilistic query expansion using query logs. In Proc. of the 11th Int. Conf. on World Wide Web (WWW 2002), Honolulu, Hawaii, pp. 325-332.

[26] De Lima, E., Pedersen, J. 1999. Phrase recognition and expansion for short, precision-biased queries based on a query log. In Proc. of the 22nd Int. Conf. on Research and Development in Information retrieval (SIGIR 1999), Berkeley, Canada, pp. 145-152.

[27] De Marco, D. 2011. Plumbing the Depths of Examiner Search (il)-Logic: A Patent Searching Perspective. Presentation given at Patent Information Users Group (PIUG 2011) Northeast Conf., New Brunswick, USA.

[28] De Vorsey, K., Elson, C., Gregorev, N., Hansen, J. 2006. The Development of a local thesaurus to improve access to the anthropological collections of the American Museum of Natural History. In D-Lib Magazine, Volume 12, Issue 4.

[29] Fitzpatrick, L., Dent, M. 1997. Automatic feedback using past queries: social searching? In Proc. of the 20th Int. Conf. on Research and Development in Information Retrieval (SIGIR 1997), Philadelphia, USA, pp. 306-313.

[30] Fonseca, B. M., Golgher, P. B., de Moura, E. S., Ziviani, N. 2003. Using association rules to discover search engines related queries. In Proc. of the First Conf. on Latin American Web Congress (LA-WEB '03), Santiago, Chile, pp. 66.

[31] Fujii, A. 2007. Enhancing patent retrieval by citation analysis. In Proc. of the 30th Int. Conf. on Research and Development in Information Retrieval (SIGIR2007), Netherlands, Amsterdam, pp. 793-794.

[32] Fujii, A., Iwayama, M., Kando, N. 2004. Overview of patent retrieval task at NTCIR-4. In Proc. of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, Tokyo, Japan, pp. 6-9.

[33] Fujita, S. 2007. Technology survey and invalidity search: An comparative study of different tasks for Japanese patent document retrieval. In Information Processing and Management, An Int. Journal, Volume 42, Issue 5, pp. 1154-1172.

[34] Garside, R., Smith, N. 1997: A hybrid grammatical tagger: CLAWS4. In Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121.

[35] Hang, C., Ji-Rong, W., Jian-Yun, N., Wei-Ying, M. 2002. Probabilistic query expansion using query logs. In Proc. of the 11th Int. Conf. on World Wide Web (WWW 2002), Hawaii, USA, pp. 325-332.

[36] Herbert, B., Szarvas, G., Gurevych, I. 2009. Prior art search using international patent classification codes and all-claims-queries. In Proc. of the 10th Cross-lang. Eval. Conf. on Multiling. Inf. Access Eval. (CLEF2009), Corfu, Greece, pp. 452-459.

[37] Hunt, D., Nyugen, L., Rodgers, M. 2007. Patent Searching: Tools & Techniques. John Wiley & Sons, Inc.

[38] Iwayama, M., Fujii, A., Kando, N., Takano, A. 2003. Overview of the third NTCIR workshop. In Proc. of the ACL-2003 Workshop on Patent corpus processing, Tokyo, Japan, pp. 24–32.

[39] Jin, B., Teng, H., Shi, Y., Qu, F. 2007. Chinese Patent Mining Based on Sememe Statistics and Key-Phrase Extraction. In Proc. of the 3rd Int. Conf. on Advanced Data Mining and Applications (ADMA 2007), Harbin, China, pp. 516-523.

[40] Jochim, C., Lioma, C., Schütze, H. 2011. Expanding queries with term and phrase translations in patent retrieval. In Proc. of the Second Int. Conf. on Multidisciplinary Information Retrieval Facility (IRFC 2011), Vienna, Austria, pp. 16-19.

[41] Jochim, C., Lioma, C., Schütze, H., Koch, S., Ertl, T. 2010. Preliminary study into query translation for patent retrieval. In Proc. of the Patent Information Retrieval Workshop (PaIR 2011), Toronto, Canada, pp. 57–66.

[42] Joho, H., Azzopardi, L., Vanderbauwhede, W. 2010. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In Proc. of the 3rd Sym. on Inf. interaction in context (IIiX 2010), New Brunswick, USA, pp. 13-14.

[43] Jones, R., Rey, B., Madani, O., Greiner, W. 2006. Generating query substitutions. In Proc. of the 15th Int. Conf. on World Wide Web (WWW 2006), Edinburgh, Scotland, pp. 387-396.

[44] Jürgens, J., Hansen P., Womser-Hacker, C. 2012. Going beyond CLEF-IP: The ’Reality’ for Patent Searchers? In the Proc. of the third Int. Conf. of the CLEF Initiative (CLEF2012), Rome, Italy, pp. 30-35.

[45] Kim, W., Jang, H., Kim, H., Kim, D. 2016. A document query search using an extended centrality with the word2vec. In Proc. of the 18th Int. Conf. on Electronic Commerce: e-Commerce in Smart connected World (ICEC2016), Suwon, Korea, Article 14.

[46] Kim, Y., Seo, J., Croft, W.B. 2011. Automatic Boolean query suggestion for professional search. In Proc. of the 34th Int. Conf. on Research and development in Information Retrieval (SIGIR2011), Beijing, China, pp. 825-834.

[47] Kishida, K. 2003. Experiments on Psuedo Relevance Feedback Method Using Taylor Formula at NT4CIR-3 Patent Retrieval Task. In Proc. of NTCIR-3 Workshop Meeting, Tokyo, Japan.

[48] Kless, D., Milton, S. 2010. Towards Quality Measures for Evaluating Thesauri. In Metadata and Semantic Research. Communications in Computer and Information Science, Volume 108, pp. 312-319.

[49] Konishi, K. 2005. Query terms extraction form Patent Documents for invalidity search. In Proc. of NTCIR-5 Workshop Meeting, Tokyo, Japan.

[50] Koster, C., Beney, J., Verberne, S., Vogel, M. 2011. Phrase-Based Document Categorization. In Current Challenges in Patent Information Retrieval, The Information Retrieval Series, Volume 29, pp. 263-286.

[51] Kotis, K., Papasalouros, P., Maragoudakis, M. 2011. Mining query-logs towards learning useful kick-off ontologies: an incentive to semantic web content creation.

In Int. Journal of Knowledge Engineering and Data Mining. Volume 1, Issue 4, pp. 303-330.

[52] Kunpeng, Z., Xiaolong, W., Yuanchao, L. 2009. A new query expansion method based on query logs mining. In Int. Journal on Asian Language Processing, Volume 19, pp. 1-12.

[53] Lempel, R., and S. Moran, S. 2003. Predictive caching and prefetching of query results in search engines. In Proc. of the 12th Int. Conf. on World Wide Web (WWW'03), Budapest, Hungary, pp. 19-28.

[54] Lim, S., Jung, S., Kwon, H. 2004. Improving patent retrieval system using ontology. In Proc. of the 30th Annual Conf. of IEEE, Volume 3, pp. 2646-2649.

[55] Lopez, P., Romary, L. 2010. Experiments with citation mining and key-term extraction for prior art search. In Proc. of CLEF (Notebook Papers/LABs/Workshops).

[56] Lupu, M, Huang, J., Zhu, J., Tait, J. 2011. TREC chemical information retrieval – An initial evaluation effort for chemical IR systems. In World Patent Information, Volume 33, Issue 3, pp. 248–256.

[57] Lupu, M., Hanbury A. 2013. Patent retrieval. In Foundations and Trends in Information Retrieval, Volume 7, Issue 1, pp.1-97.

[58] Lupu, M., Mayer, K., Tait, J., Trippe, A. 2011. Current Challenges in Patent Information Retrieval. Springer.

[59] Magdy, W. Jones, G. J. F. 2010. Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. In Proc. of CLEF (Notebook Papers/LABs/Workshops).

[60] Magdy, W., Jones, G. J. F. 2010. PRES: A score metric for evaluating recall oriented information retrieval applications. In Proc. of the 33rd Int. Conf. on Research and Developement in Information Retrieval (SIGIR2010), Geneva, Switzerland pp. 611-618.

[61] Magdy, W., Jones, G.J.F. 2011. A Study of Query Expansion Methods for Patent Retrieval. In Proc. of PaIR 2011, Glasgow, Scotland, pp. 19-24.

[62] Magdy, W., Leveling, J., Jones, G.J.F.: Exploring structured documents and query formulation techniques for patent retrieval. In Proc. of the 10th Cross-language Eval. Forum Conf. on Multilingual Inf. Access Eval. (CLEF 2009), Corfu,

Greece, pp. 410-417.

[63] Mahdabi P., Crestani, F. 2014. Patent Query Formulation by Synthesizing Multiple Sources of Relevance Evidence. In Trans. on Inf. Systems, Volume 32, Issue 4, Article No. 4.

[64] Mahdabi, P., Andersson, L., Keikha, M., Crestani, F. 2012. Automatic Refinement of Patent Queries using Concept Importance Predictors. In Proc. of the 35th Int. Conf. on Research and Development in IR (SIGIR2012), Portland, USA, pp. 505-514.

[65] Mahdabi, P., Crestani F. 2014. The effect of citation analysis on query expansion for patent retrieval. In Information Retrieval, Volume 17, Issue 5-6, pp. 412-419.

[66] Mahdabi, P., Keikha, M., Gerani, S., Landoni, M., Crestani, F. 2011. Building queries for prior-art search. In Proc. of the Second Int. Conf. on Multidisciplinary Information Retrieval Facility (IRFC 2011), Vienna, Austria, pp. 3-15.

[67] Markatos, E.P. 2000. On caching search engine query results. In Computer Communications, Volume 24, Issue 1, pp. 137-143.

[68] Mase, H., Matsubayashi, T., Ogawa, Y., Iwayama, M., Oshio, T. 2005. Proposal of two-stage patent retrieval method considering the claim structure. In ACM Transactions on Asian Language Information Processing (TALIP), Volume 4, Issue 2, pp. 190–206.

[69] Miller, G. 1995. WordNet: A Lexical Database for English. In Communications of the ACM, Volume 38, No. 11, pp. 39-41.

[70] Murray, G. C., Teevan, J. 2007. Query log analysis: Social and technological challenges. In SIGIR Forum 41, pp. 112-120.

[71] Nanba, H. 2007. Query expansion using an automatically constructed thesaurus. In Proceedings of NTCIR-6 Workshop Meeting, Tokyo, Japan, pp. 414-419.

[72] Navigli, R. 2009. Word sense disambiguation: A survey. In ACM Computing Surveys, Volume 41, Issue 2, Article 10.

[73] Oh, S., Lei, Z., Lee, W., Mitra, P., Yen, J. 2013. CV-PCR: a context-guided value-driven framework for patent citation recommendation. In Proc. of the 22nd Int. Conf. on Information & Knowledge Management (CIKM 2013). Burlingame, USA, pp. 2291-2296.

[74] Okamoto, M., Shan, Z., Orihara, R. 2017. Applying Information Extraction for

Patent Structure Analysis. In Proc. of the 40th Int. Conf.on Research and Development in Information Retrieval (SIGIR2017), Tokyo, Japan, pp. 989-992.

[75] Ozmutlu, S., Spink, A. Ozmutlu H. C. 2004. A day in the life of web searching: an exploratory study. In Information Processing and Management,Volume 40, Issue 2, pp. 319-345.

[76] Piroi, F., Lupu, M., Hanbury, A. 2012. Effects of Language and Topic Size in Patent IR: An Empirical Study. In the Proc. of the third Int. Conf. of the CLEF Initiative (CLEF 2012), Rome, Italy, pp. 54-66.

[77] Piroi, F., Lupu, M., Hanbury, A. 2013. Overview of CLEF-IP 2013 Lab. In Information Access Evaluation. Multilinguality, Multimodality, and Visualization Lecture Notes in Computer Science, Volume 8138, pp. 232-249.

[78] Piroi, F., Lupu, M., Hanbury, A., Zenz, V. 2011. CLEF-IP 2011: Retrieval in the intellectual property domain. In Proc. of CLEF (Notebook Papers/Labs/Workshop).

[79] Piroi, F., Tait, J. 2010. CLEF-IP 2010: Retrieval experiments in the intellectual property domain. In CLEF-2010 (Notebook Papers/LABs/Workshops).

[80] Roda, J., Tait, J., Piroi, F., Zenz, V. 2009. CLEF-IP 2009: Retrieval experiments in the intellectual property domain. In Proc. of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009), Corfu, Greece, pp. 385–409.

[81] Sahlgren, M., Hansen, P., Karlgren, J. 2003. English-Japanese cross-lingual Query Expansion using Indexing of aligned bilingual text data. In Proc. of the Third NTCIR Workshop, Kista, Schweden.

[82] Salampasis, M., Paltoglou, G., Giahanou, A. 2012. Report on the CLEF-IP 2012 experiments: Search of topically organized patents. In Proc. of CLEF (Notebook Papers/LABs/Workshops).

[83] Salton, G., Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. In Information Processing and Management, Volume 24, pp. 513–523.

[84] Sekine, S., Suzuki, H. 2007. Acquiring Ontological Knowledge from Query Logs. In Proc. of the 16th Int.l Conf. on World Wide Web (WWW 2007), Banff, Canada, pp. 1223-1224.

[85] Silverstein, C., Marais, H. Henzinger, M., Moricz, M. 1999. Analysis of a very large web search engine query log. In SIGIR Forum 33, pp. 6-12.

[86] Silvestri, F. 2010. Mining Query Logs: Turning Search Usage Data into Knowledge. In Foundations and Trends in Information Retrieval, Volume 4, Issue 1-2, pp. 1-174.

[87] Smucker, M., Allan, J., Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. In Proc. of the 16th Int. Conf. on Information and Knowledge Management (CIKM 2007), Lisbon, Portugal, pp. 623-632.

[88] Spink, A., Wolfram, D., Jansen, M. B. J., Saracevic, T. 2001. Searching the web: the public and their queries. In Journal of the Association for Information Science and Technology, Volume. 52, pp. 226-234.

[89] Taduri, S., Lau, G. T., Law, K. H., Kesan, J. P. 2011. Retrieval of Patent Documents from Heterogeneous Sources Using Ontologies and Similarity Analysis. In Proc. of the 5th Int. Conf. on Semantic Computing (ICSC 2011), pp. 538-545.

[90] Tanioka, H., Yamamoto, K. 2007. A passage retrieval system using Query Expansion and Emphasis. In Proc. of NTCIR-6 Workshop Meeting, Tokyo Japan, pp. 428-432.

[91] Tannebaum W., Rauber A. 2015. Learning Keyword Phrases from Query Logs of USPTO Patent Examiners for Automatic Query Scope Limitation in Patent searching. In World Patent Information, Volume 41, pp. 15-22.

[92] Tannebaum W., Rauber, A. 2015. PatNet: A lexical database for the patent domain. In Proc. of the 37th European Conf. on Information Retrieval (ECIR 2015), Vienna, Austria, pp. 550-555.

[93] Tannebaum, W., Mahdabi, P. and Rauber, A. 2015. Effect of log-based Query Term Expansion on Retrieval Effectiveness in Patent Searching. In Proc. of 6th Int. Conf. of the CLEF Initiative (CLEF2015), Toulouse, France, pp. 300-305.

[94] Tannebaum, W., Rauber A. 2014. Using Query Logs of USPTO Patent Examiners for automatic Query Expansion in Patent Searching. In Information Retrieval, Volume 17, Issue 5-6, pp. 452-470.

[95] Tannebaum, W., Rauber, A. 2010. Query Expansion for Patent Retrieval using Domain Specific Thesaurus. In Proc. of the 2010 Conf. on the Interaction of Inf. Related Rights, Information Technology and Knowledge Management (KnowRight 2010),Vienna, Austria.

[96] Tannebaum, W., Rauber, A. 2012. Acquiring lexical knowledge from Query Logs for Query Expansion in Patent Searching. In Proc. of the 6th IEEE Int.Conf. on Semantic Computing (IEEE ICSC 2012), Palermo, Italy, pp. 336-338.

[97] Tannebaum, W., Rauber, A. 2012. Analyzing Query Logs of USPTO examiners to identify useful Query Terms in Patent Documents: A Preliminary Study. In Proc. of the Information Retrieval Facility Conf. (IRFC 2012), Vienna, Austria, pp. 127-136.

[98] Tannebaum, W., Rauber, A. 2013. Mining Query Logs of USPTO Patent Examiners. In Proc. of 4th Int.l Conf. of the CLEF Initiative (CLEF 2013), Valencia, Spain, pp. 136-142.

[99] Torres, S., Hiemstra, D., Serdyukov, P. 2010. Query log analysis in the context of information retrieval for children. In Proc. of the 33th Int. Conf. on Research and Development in Information Retrieval (SIGIR 2010), Geneva, Switzerland, pp. 847-848.

[100] Tyler, S., Teevan, J. 2010. Large scale query log analysis of refinding. In Proc. of the third ACM Int. Conf. on Web search and data mining (WSDM 2010), New York, USA, pp. 191-200.

[101] Van Rijsbergen, C. J. 1979. Information Retrieval. Butterworths.

[102] Verberne, S., Sappelli, M., Kraaij, W. 2014. Query Term Suggestion in Academic Search. In Proc. of the 36th European Conf. on IR Research (ECIR 2014), Amsterdam, The Netherlands, pp. 560-566.

[103] Voorhees, E. 1994. Query expansion using lexical-semantic relations. In Proc. of the 17th Int. Conf. on Research and Development in Information Retrieval (SIGIR 1994), Dublin, Ireland, pp. 61–69.

[104] Wanagiri, M.Z. Adriani, M. 2010. Prior art retrieval using various patent document fields contents. In Cross-Language Evaluation Forum (Notebook Papers/LABs/Workshops).

[105] Xu J., Croft, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. In ACM Trans. Inf. Syst., Volume 18, Issue 1, pp. 79-112.

[106] Xue, X., Croft, W. 2009. Automatic query generation for patent search. In Proc. of the 17th Int. Conf. on Information and Knowledge Management (CIKM 2009), Hong Kong, China, pp. 2037-2040.

[107] Xue, X., Croft, W. 2009. Transforming patents into prior-art queries. In Proc. of the 32th Int. Conf. on Research and Development in Information Retrieval (SIGIR 2009), Boston, USA, pp. 808-80.

[108] Zaiane, O. R., Strilets, A. 2002. Finding similar queries to satisfy searches based on query traces. In Proc. of the Workshops on Advances in Object-Oriented Information Systems (OOIS 2002), Montpellier, France, pp. 207-216.

[109] Zhang, Z. Nasraoui, O. 2006. Mining search engine query logs for query recommendation. In Proc. of the 15th Int. Conf. on World Wide Web (WWW 2006), Edinburgh, Scotland, pp. 1039-1040.

[110] Zhang, Z., Yang, M., Li, S., Qi, H., Song, C.: Sogou Query Log Analysis: A Case Study for Collaborative Recommendation or Personalized IR. In Proc. of the 2009 Int. Conf. on Asian Language Processing (IALP 2009), Singapore, pp. 304-307.