Learning Keyword Phrases from Query Logs of USPTO Patent Examiners for Automatic Query Scope Limitation in Patent Searching

Wolfgang Tannebaum [a], Andreas Rauber [a *]

[a] *Vienna University of Technology, Institute of Software Technology and Interactive Systems, Favoritenstrasse 9-11/188, A-1040 Vienna, Austria*

Abstract

Professional search in patent repositories poses several unique challenges. One key requirement is to search the entire affected space of concepts, following well-defined procedures to ensure traceability of results obtained. Several techniques have been introduced to enhance query generation, preferably via automated query term expansion, to improve retrieval effectiveness. Currently, these approaches are mostly limited to computing additional query terms from patent documents based on statistical measures. For conceptual search to solve the limitation of traditional keyword search standard dictionaries, such as *WordNet*, or lexica, like *Wikipedia*, are used to provide synonyms and keyword phrases for query refinement. Studies show that these are insufficient in such highly specialized domains. In this paper, we present an approach to learn keyword phrases from query logs created during the validation procedure of the patent applications. This creates valuable domain-specific lexical databases for several specific patent classes that can be used to both expand as well as limit the scope of a patent search. This provides a more powerful means to guide a professional searcher through the search process. We evaluate the lexical databases based on real query sessions of patent examiners.

Keywords: Patent Searching, Query Term Expansion, Lexical Resources;

1.Introduction

Professional search in the patent domain poses several unique challenges. One key requirement is to search in the entire affected space of concepts. Virtually all patent search systems are based on Boolean retrieval and exact matching of the query terms. Several techniques have been introduced to assist patent searchers in expansion of the query terms, which relate to parts of the original text with (1) synonyms and equivalents to expand the query scope and (2) co-occurring terms or (3) keyword phrases to narrow the search. Currently, these approaches are mostly limited to computing additional query terms from patent documents based on statistical measures. For conceptual search to solve the limitation of traditional keyword search standard dictionaries are used to provide synonyms and keyword phrases for query refinement. For the patent domain specific lexical resources are not available to provide assistance in identifying these additional query terms to refine the search [1,2]. However, actual queries that have been posed by patent experts promise to be a valuable resource to learn such domain-specific and highly optimized lexical resources. This, in turn, can provide valuable assistance for patent searchers easing the process of query expansion.

In previous work, we analyzed query logs of the patent examiners of the United Patent and Trademark Office (USPTO), in particular the basic characteristics of the patent examiners' query logs, such as query and query log length, or the number of search sessions available for each patent application. Furthermore, we could show, that lexical knowledge can be extracted directly from the query logs and used for automated query expansion in patent searching. In particular, synonyms are learned based on the Boolean operator "OR" which is used in the queries [3,4]. Experiments have shown that the extracted specific lexical databases drastically outperform general-purpose sources, such as *WordNet* [5]. The learned lexical databases provided up to 8 out of 10 expansion terms used by the patent examiners, whereas *WordNet,* on average, suggested only 2 out of 10 expansion terms used by the examiners.

In this paper, we go beyond learning only synonyms from the query logs. We present an approach to learn keyword phrases, in particular search terms consisting of two words, from the query logs, which patent examiners created during the validation procedure of the patent applications. This creates valuable domain-specific lexical databases for several specific patent classes that can be used to both expand as well as limit the scope of a patent search. This provides a more powerful means to guide a professional searcher through the search process.

We collected a corpus of patent query logs (103,896 log files) making it the largest collection of query logs used for experiments in the patent domain.

The remainder of the paper is organized as follows. We first review related work on automatic query term expansion in patent search and based on query logs. In Section 3 we present our approach to extract keyword phrases from the query logs and the lexical databases learned for specific US patent classes. Experiments on

* Corresponding author. Tel: +43 1 58801 18826
Email addresses: tannebaum@ifs.tuwien.ac.at (Wolfgang Tannebaum), rauber@ifs.tuwien.ac.at (Andreas Rauber)

automatic query scope limitation are provided in Section 4, followed by conclusions and an outlook on future work in Section 5.

2. Related Work

2.1. Enhancing Query Generation in Patent Searching

Many techniques to enhance query generation have been introduced in the field of semantic search to improve retrieval effectiveness. Aiming at solving the limitation of traditional keyword search, which provides limited capabilities to capture the information need of the searchers, current works focus on conceptual search (searching by meanings rather than literal strings). Contextual semantic information, for example statistical properties of documents related to a given query, are computed for the initial query. Common techniques use ontologies to enable semantic search within digital libraries, or thesauri, which compute synonyms and keyword phrases for the initial query terms. These are used to refine keyword based queries to semantic queries.

Several techniques have been proposed in the patent domain to enhance query generation, preferably via automated query expansion. These techniques are mostly limited to computing co-occurring terms to the searchable features of the invention. Additional query terms are extracted automatically from the query 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), or from the translations of the claim sections [6,7,8,9]. Further, also whole documents or whole sections of the query documents, like the title, abstract, description or the claim section, at least clusters of keywords are used for query generation and query expansion [10,11]. In addition, recent research focuses on the usage of patent images, chemical information and cross-lingual information to enhance patent searchers in query generation [12,13,14].

To provide synonyms for conceptual search, standard dictionaries, such as *WordNet*, or lexica, like *Wikipedia*, are used for query refinement [15]. To learn them directly from the patent domain, as described in [7,16], the claim sections of a European Patent Office (EPO) patent collection including the claims in English, German and French are aligned to extract translation relations for each language pair. Based on the language pairs having the same translation terms, synonyms are learned in English, French and German.

In the retrieval of keyword phrases for query refinement in patent searching, particularly to narrow a search, as well for automatic document categorization, keyword phrases are learned automatically from the query documents using natural language processing applications or statistical measures [17,18,19]. In the same way as for learning synonyms, standard dictionaries and lexica are used to learn the keyword phrases. To achieve translations of keyword phrases, particularly term to phrase translations, phrase to term translations and phrase to phrase translations, claim sections are aligned [16].

2.2. Enhancing Query Generation based on Query Logs

In several information retrieval applications, especially for web search, query logs are being intensively studied. Large-scale data sets of web queries, such as *AltaVista log* or *AOL log*, have been made publicly available [20]. 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 [21].

Most work is related to automatic query expansion based on the information learned from earlier query logs. The challenge is to extract semantic relations between the query terms to learn lexical knowledge. Techniques used to measure query similarity are based on, for example: (1) differences in the ordering of documents retrieved in the answers; (2) association rules (the query log is viewed as a set of transactions, in which a single user submits a sequence of related queries in a time interval); (3) click-through data information (i.e. using the content of clicked web pages, in particular to consider terms in the URLs clicked after a query), or (4) graph-based relations among queries [22,23]. A survey on the use of web logs to improve search systems is presented in [20].

A specific task in learning semantic relations from previous query logs is the extraction of keyword phrases to enhance searchers in narrowing their search. Standard approaches to the extraction of phrases are based on statistical measures. Query logs and their query terms are considered as free text. Every pair of non-function words are considered as a candidate phrase, particularly only those that occur with frequency above a given threshold in a relevant collection. Alternative methods involve the usage of co-occurrence frequency statistics [24]. Further approaches use tagging and linguistic information in order to identify phrases, or combine grammatical and statistical information to learn the keyword phrases [24,25]. External sources, such as lexica, glossaries or databases like *WordNet* are used for this [26,27]. For the patent domain dedicated external lexical resources, like patent domain specific lexica or thesauri, are not available.

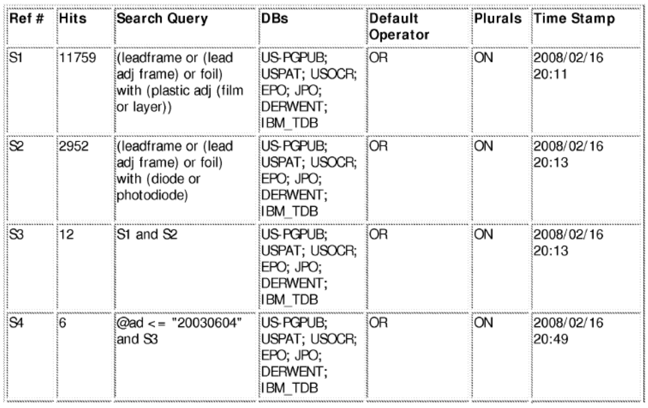
3. Extracting Lexical Databases

Finding query logs in the patent domain is a difficult task. Private companies and searchers are hesitant to make their query logs available as these would reveal their current R&D activities. The only source known to us which publishes the query logs of patent examiners is the USPTO.

A detailed analysis of the USPTO patent examiners query logs are presented in [28].

3.1. Experiment Setup

The query logs of USPTO patent examiners (called “Examiner`s search strategy and results”) are published for most patent applications since 2003 by the US Patent and Trademark Office Portal PAIR (Patent Application Information Retrieval)[1]. Since that time, the USPTO published about 2.7 million patent applications, which are classified into 473 classes each including several subclasses (about 6,000 patent applications per class). Each query log of the USPTO is a PDF file consisting of a series of queries. Figure 1 shows an example, particularly an extract of four text queries of such a query log.



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

Fig. 1 Example of a USPTO query log.

Each query has several elements: reference, hits, search query, database(s), default operator, plurals, and time stamp. Our focus is on the search query element including the text, non-text and reference queries formulated by the patent examiners. The text queries are used to query the documents or only sections of the patent documents, such as the title section (title search). As shown in query *S1* and *S2*, the text queries include the search operators (Boolean and Proximity operators) between the query terms. Furthermore, the text queries can include truncation limiters, such as “$”. Non-text queries are used to search patent document numbers, classifications, or application and publication dates, for example “@ad <= 20030604” for searching patent documents applied before 4th June of 2003, as shown in query *S4* of Figure 1. The reference query, which is a combination of earlier queries, for example “*S1* and *S2*”, i.e. re-using the terms of a previous query and expanding it with further elements, thus avoiding to have to re-type an earlier query. For our purposes we are specifically interested in the text queries including the Proximity operator “ADJ” to learn keyword phrases.

Google has begun crawling the USPTO's public PAIR sites and provides free download of all patent applications and their documents from the examination procedure published until now[2]. Google created single zip files for each patent application. The files can contain one or more query log files. Because patent searchers use the classification system to narrow the search, we selected fifteen classes for our experiments. In total, we downloaded 103,896 query logs available for the fifteen US classes. Table 1 shows the number of the selected classes, the titles of the classes, and the number of downloaded query logs for each class.

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

Table 1. Experiment Setup.

As shown in Table 1, we have sorted the classes according to their class size (number of documents) from 1,820 to 16,864 files. The ranking enables us to assess if the performance of the learned lexical databases increases with a larger number of query logs available for a specific US patent class. In addition, we have chosen US patent classes according to their topic. We selected classes which are topically related, for example classes 128, 433 and 623 from the *medical domain*, or classes 384 and 148 from the *mechanical domain*, to examine in how far lexical databases can be learned on a more generic level, and completely different classes, such as class 454 for *Ventilation* and class 398 for *Optical Communications*.

3.2. Detecting Keyword Phrases and Lexical Databases

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 learn semantic relations.

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

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

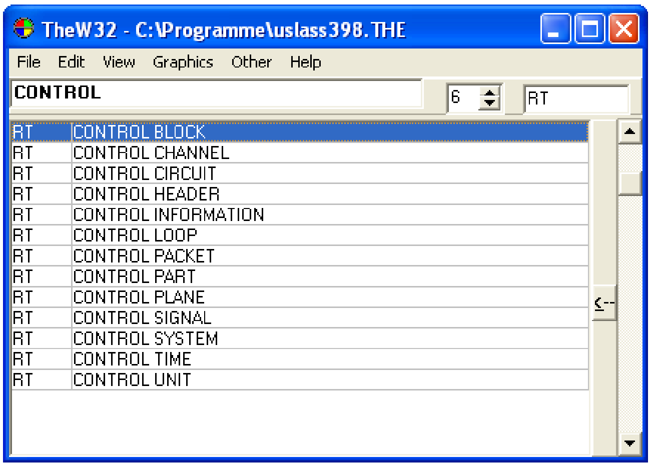
We extract keyword phrases from the query logs of the USPTO patent examiners based on the following process: Through OCR conversion and segmentation of the 103,896 PDF files, which were stored as images, we separate all text queries including the search operators between the query terms from the query log collection. We then filter all 3-grams generated from the text queries in the form "X *b* Y", where *b* is the proximity operator "ADJ" and X and Y are query terms. This results in 26,652 unique keyword phrases including 16,413 unique query terms. 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 support of *1*, greater than or equal to *2*, *3*, *4* and *5*. We notice for all classes only a minor decrease of the number of keyword phrases having a support of greater or equal to 2 (CV_2). The largest decrease in the number of keyword phrases is provided by the keyword phrases having a *CV* 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. Table 2 shows the number of keyword phrases extracted from the query logs and encountered at least two times, which we consider for learning the lexical databases, particularly the thesauri of English concepts.

Lexical Databases	**unique phrases**
Bearings	104
Ventilation	266
Metal treatment	309
Prosthesis	446
Stoves and furnaces	504
Dentistry	535
Disinfecting, Preserving, Sterilizing	704
Electric heating	768
Pumps	806
Motor vehicles	1,022
Land vehicles	1,023
Optical communications	1,373
Surgery	1,456
Telephonic communications	1,830
Electrical connectors	1,938
Σ	**11,145**

Table 2. Extracted Lexical Databases.

The lexical databases provide English keyword phrases for each specific patent class. Terms which in combination constitute a keyword phrase are linked to each other. For example, the query term "control" can be expanded using the domain specific lexical databases to limit the query scope as shown in Figures 3 and 4.

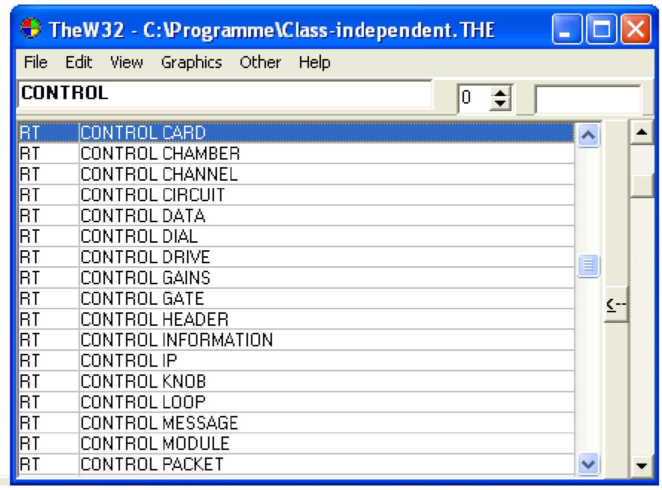
To query the lexical databases we use the open source thesaurus management software *TheW32* [29]. The resulting lexical resources can be used, particularly in each specific US patent class, for (semi-) automated query suggestion, particularly for query scope limitation. For example, the query term "control" can be expanded using the domain specific lexical databases to limit the query scope as shown in Figures 2 and 3.



CONTROL
6
RT

RT CONTROL BLOCK
RT CONTROL CHANNEL
RT CONTROL CIRCUIT
RT CONTROL HEADER
RT CONTROL INFORMATION
RT CONTROL LOOP
RT CONTROL PACKET
RT CONTROL PART
RT CONTROL PLANE
RT CONTROL SIGNAL
RT CONTROL SYSTEM
RT CONTROL TIME
RT CONTROL UNIT

Figure 2. Using the lexical database *Optical communications* for query scope limitation.



CONTROL
0

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 3. Using the class-independent lexical database for query scope limitation.

Figures 2 and 3 show the suggested expansion terms to limit the query scope for the query term "*control*". Twelve of the fifteen classes for which we learned the lexical databases provide expansion terms for this specific term. The class-specific lexical database *Optical Communications* provides most expansion terms. As shown in Figure 3, the lexical database suggests thirteen expansion terms, which refine the general query term to a keyword phrase. In total, the class-independent lexical database provides 41 unique expansion terms. Figure 4 shows only an extract from the class-independent lexical database.

On average, the lexical databases suggest 12 expansion terms to limit a query term to a keyword phrase. The maximum number of query suggestions for a query term, in particular for the term "*power*", are 96 expansion terms. The general query term "*power*" is limited to the following keyword phrases, for example: "*power source*", "*power signal*", "*power tool*", "*power supply*", "*power distribution*", "*power transfer*" and so on.

4. Semi-automatic Query Term Expansion

In this section we learn several class-specific lexical databases and one class-independent lexical database from the query log collection and evaluate them based on real query sessions of patent examiners (gold standard). In particular, 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 [30], we evaluate the lexical databases based on real query expansions carried out by the patent examiners in the search sessions. For this we split for each class the query log collection in a training set for learning the lexical databases and in a test set to evaluate the learned lexical databases on real query sessions. Specifically, we have ordered the query logs by time of application of the invention. In each class we use the most recent 500 query logs for testing, whereas the oldest query logs are used for training purposes. Table 3 shows the generated test and training sets for the several classes.

US Class	Test Sets	Training Sets
454	500	1,000
384	500	1,000
126	500	2,000
148	500	2,000
219	500	2,500
433	500	3,500
180	500	4,500
417	500	4,500
398	500	5,000
280	500	5,000
128	500	8,000
379	500	9,000
422	500	10,000
439	500	10,000
623	500	10,000

Table 3. Test and Training Sets.

In each class the test sets are built based on 500 query log files. The size of the training sets depends on the class size. The training sets, which we use for learning the lexical databases, comprise between 1,000 and 10,000 log files. This particular way of splitting training and test sets aims at creating a realistic evaluation setting where first lexical databases used in operational settings can only be trained on earlier query sessions. Secondly, these databases will usually not be updated with every single patent application that has been processed. Rather, we assume a re-creation of a database with every 500 new applications per class.

To measure the performance of the lexical databases, we calculate recall, precision, and coverage scores. We compare the suggested terms from the lexical databases with the terms used for generating keyword phrases by the examiners as available in the query logs. In particular, to calculate the recall scores, we compare the suggested terms from the lexical databases with the keyword phrases from the test sets, which were used by the patent examiners for searching. To compute precision we compare the keyword phrases used by the examiners in the test sets with all expansion terms suggested by the lexical databases. Furthermore, for calculating the recall and precision measures we excluded keyword phrases in the test sets that are out of the vocabulary of the lexical databases, i.e. terms that did not appear in any earlier query log. For this reason we additionally measure coverage of the lexical databases by determining the number of out-of-vocabulary words, i.e. expansion terms that were used later in time that the lexical database could not learn. This indicates how the vocabulary drifts over time.

Our focus is on the recall scores, as users will be able to choose from a variety of possible expansion terms and can easily reject ones that are useless for their current search. However, in approaches where expansion terms 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. Current work for automatic query term suggestion obtain precision scores of about 17% in academic professional search and about 5% in patent searching as shown in [31] [32]. Both evaluate their query term suggestion approaches based on user simulations and real examples. In particular, human judgments are used to evaluate the suggested expansion terms [31]. On average, up to 2 out of 10 suggested terms are judged as relevant. Specifically, for patent searching and for the medical domain a Boolean query term suggestion approach is provided, which is based on pseudo-relevance feedback. About 200 queries are generated for each search topic, of which about 10 queries are assessed as relevant. To improve the precision scores they suggest to place these queries at top ranks. They assume that the user always selects the highest-ranked expansion terms [32].

4.1. Considering patent classification in automatic Query Scope Limitation

At first we use the patent classification system to learn lexical databases for each specific patent class. This allows us to evaluate (1) class and (2) size dependency characteristics. In total, we learn fifteen class-specific lexical databases *csDB* from the training sets according to our approach presented in section 3.2. In addition, we learn a class-independent lexical database, which we call *PhraseNet*. For this we use the training sets of all classes. This lexical database is still domain-specific because it is based on patent query logs. But it is less specific, because it is learned from multiple patent classes. *PhraseNet* provides 22,492 unique keyword phrases. To measure the performance of the learned class-specific and class-independent lexical databases we use the test sets of each specific class. Table 4 shows the achieved Recall, Precision, and Coverage scores.

	Recall		**Precision**		**Coverage**	
US Class	*csDB*	*PhraseNet*	*csDB*	*PhraseNet*	*csDB*	*PhraseNet*
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

Table 4. Query Scope Limitation based on patent class.

Best recall measures of the class-specific lexical databases are provided, on average, by the lexical databases learned from the US 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 learned for the class 422. The lexical database suggests with a recall of 70%, on average, 7 of 10 keyword phrases, which are used by the patent examiners for query expansion. The precision values of the class-specific lexical databases show, that with increasing training set sizes the achieved precision scores decrease as the number of suggested outterms increases. Considering the lexical database providing the best recall performance (class 422), on average, 1 out of 10 terms suggested by the lexical database is used by the patent examiners for query expansion. Best coverage scores of the class-specific lexical databases are also provided by the lexical databases learned 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 query terms 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 *PhraseNet*. Only in class 433 the recall decreases, because we excluded query terms which are out of the vocabulary to calculate the recall measures. The reason for that is, more query terms appear in the lexical database *PhraseNet*, but not necessarily as 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, which are used by the patent examiners for query expansion. Compared to the class-specific lexical databases, the precision scores achieved by *PhraseNet* further decreases. For example, in class 422, on average, only 6 out of 100 terms suggested by *PhraseNet* are used by the patent examiners for query expansion. The coverage scores of *PhraseNet* obviously increases with the rise of the class size. In particular for class 384, *PhraseNet* provides 98% of the query terms from the test set. Considering all classes the class-independent lexical database *PhraseNet* provides, on average, a coverage of 94%.

Finally, the experiments indicate that for almost all classes the recall and coverage measures of the class-specific lexical databases can be further improved using the class-independent lexical database *PhraseNet* providing, on average, 8 out of 10 keyword phrases that are used by the patent examiners for query expansion. But compared to *PhraseNet*, the class-specific databases achieve much better precision scores. Considering all classes the class-specific lexical databases provide, on average a precision of 18%. On average, only 2 out of 10 terms suggested by the class-specific lexical databases are used by the patent examiners for query expansion. Compared to the precision scores achieved by the related approaches (17% precision in academic search and about 5% in patent searching) the lexical databases achieves accurate precision scores.

To provide valuable lexical databases for semi-automatic query expansion achieving high recall/ coverage and precision scores, either (1) the recall measures of the class-specific databases or (2) the precision scores of *PhraseNet* have to be improved. To this end, we carry out the following experiments.

4.2.Considering Confidence Values in automatic Query Scope Limitation

In this section, we introduce a confidence value *CV* for learning the keyword phrases and to optimize the achieved precision scores of the class-independent lexical database *PhraseNet* providing the lowest precision scores, as shown in section 4.1.

We consider only those expansion terms that were encountered up to ten times (CV_1 to CV_{10}) as keyword phrases in the training sets of *PhraseNet*, i.e. that have a support >10 in the training set. Based on the CV_1 to CV_{10} we learn nine further class-independent lexical databases $PhraseNet_2$ to $PhraseNet_{10}$. The number of unique keyword phrases in the learned lexical databases is strongly decreasing with the rise of the confidence value. While *PhraseNet* provides 22,492 unique keyword phrases, only 9,717 phrases appear in $PhraseNet_2$. The number of unique phrases is further decreasing. The lexical databases $PhraseNet_6$ to $PhraseNet_{10}$ provide less than 2,000 unique keyword phrases.

We evaluate the lexical databases using a test set of 7,500 query logs, which we built from the class-specific test sets from Table 4. The resulting scores of the lexical databases are provided in Table 5.

Lexical database	***CV***	**Recall**	**Precision**	**Coverage**
PhraseNet	1	**61.83**	**5.29**	**94.36**
$\mathit{PhraseNet}_{2}$	2	55.46	8.61	89.18
$\mathit{PhraseNet}_{3}$	3	51.05	11.72	84.62
$\mathit{PhraseNet}_{4}$	4	47.54	14.20	80.14
$\mathit{PhraseNet}_{5}$	5	44.41	17.53	76.74
$\mathit{PhraseNet}_{6}$	6	42.03	19.87	73.72
$\mathit{PhraseNet}_{7}$	7	40.69	22.11	69.71
$\mathit{PhraseNet}_{8}$	8	39.72	24.71	66.54
$\mathit{PhraseNet}_{9}$	9	38.90	25.78	64.37
$\mathit{PhraseNet}_{10}$	10	**38.04**	**27.83**	**62.36**

Table 5. Considering Confidence Values *CV* for learning *PhraseNet*.

Table 5 shows that an increase in precision compared to the values provided in Table 4 can be observed. The limited lexical database $\mathit{PhraseNet}_{10}$ provides with a value of 28% best precision. On average, 3 out of 10 terms that are suggested by the lexical database as keyword phrases were actually used by the examiners for query expansion in the test set. But the recall and coverage scores achieved with the more limited lexical databases $\mathit{PhraseNet}_{2}$ *to* $\mathit{PhraseNet}_{10}$ decreases considerably. Best recall is provided by *PhraseNet*. The lexical database provides with a recall of 62%, on average, 6 of 10 keyword phrases, which are used by the patent examiners for query expansion in the test set. Best coverage of the learned lexical databases is also provided by the lexical database *PhraseNet.* In particular, the lexical database provides, on average, 94% of the query terms from the test set.

Finally, the confidence value is a valuable resource to (1) strike reasonable balance between increasingly higher recall/ coverage and lower precision, and to (2) iteratively suggest an increasing number of terms to limit the query scope as needed for searching. After having initially suggested the most likely and highest-precision terms (using $\mathit{PhraseNet}_{10}$), additional terms that have a lower support and lower precision (using *PhraseNet*) can be suggested by an automatic query expansion system.

4.3. Cross-domain Applications in automatic Query Scope Limitation

In the previous section, we addressed the low precision performance of the class-independent lexical database *PhraseNet* achieving highest recall and coverage scores. In this section we carry out experiments to improve the recall and coverage measures of the class-specific lexical databases, which achieve the highest precision scores. For that, we initially measure the performance of the class-specific lexical databases *csDB* when used for patents from other classes. This will help to detect classes where cross-domain applications might be useful. Specifically, as patent classes are organized in a hierarchical structure, we assume that intermediate dictionaries spanning several classes of a more generic category may be used to learn category-specific dictionaries. We apply the learned class-specific lexical databases across all test sets from the other classes. Again, for calculating the recall measures we excluded keyword phrases in the test sets that are out of the vocabulary of the lexical databases. Table 6 shows the achieved recall measures of the class-specific *csDB* lexical databases when used for the class they were based upon and the best recall measures when used for the other classes.

Recall			
US Class	*csDB*	*US Class*	*csDB*
454	**35.00**	180	27.27
384	**60.00**	126	26.32
126	35.14	454	**39.29**
148	**42.11**	422	31.58
219	33.02	148	**41.46**
433	**55.56**	180	28.00
180	54.87	280	**55.00**
417	44.14	384	**42.86**
398	**55.80**	623	50.00
280	**47.69**	180	35.71
128	50.00	623	**60.00**
379	**60.17**	398	27.27
422	**70.00**	219	43.90
439	**55.00**	454	37.78
623	27.27	384	**37.04**

Table 6. Achieved recall measures of the class-specific lexical databases when used for the class they were based upon and the best recall measures when used for the other classes.

As shown, the learned class-specific lexical databases achieve respectable recall measures in other classes. For example, the lexical database learned for the class 280 called “Land vehicles” achieves a recall of almost 36% for class 180 called “Motor vehicles”. The lexical database learned for class 623 called “Prosthesis” provides a recall measure of 37% for class 384 called “Bearings”. 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 learned from class 128 when applied to class 422 with a recall score of only 24%.

Finally, we use the class-specific lexical databases that achieved best recall measures in the other classes to expand the class-specific lexical databases of these classes. We learn further fifteen class-related lexical databases *crDB* .

To measure the performance of the class-related lexical databases we use again the test sets of each specific class and calculate recall, precision and coverage, as shown in Table 7.

Compared to the class-specific lexical databases *csDB* learned from the training sets of each class and to the lexical databases *crDB* learned from the related training sets, *PhraseNet* achieves still the best recall measures for almost all classes, but also the lowest precision scores. But through the expansion of the class-specific lexical databases with training sets of related classes, in particular for classes where few query logs are available (e.g. classes 454 and 219), the recall measures can be significantly improved. Best improvement in recall is achieved for class 219. The related lexical database *crDB* for class 219 provides a recall of 67% - considerably better than the recall achieved by the class-specific lexical database (34%) almost at the level of *PhraseNet* (67%).

	Recall			**Precision**		
US Class	*csDB*	*crDB*	*PhraseNet*	*csDB*	*crDB*	*PhraseNet*
454	35.00	**51.28**	66.67	15.91	**7.97**	4.99
384	**60.00**	56.41	71.93	50.00	**30.14**	5.68
126	35.14	**41.88**	56.77	22.03	**11.89**	5.69
148	42.11	**58.00**	77.94	40.00	**20.71**	7.59
219	33.65	**66.98**	67.08	22.29	**12.31**	5.15
433	**55.56**	50.00	50.00	41.67	**17.65**	8.79
180	54.87	**58.87**	65.69	12.53	**9.14**	5.66
417	44.14	**49.18**	66.91	19.44	**9.71**	4.07
398	56.20	**58.70**	59.75	5.97	**5.19**	3.83
280	47.69	**57.69**	68.42	13.84	**9.59**	5.85
128	50.00	**55.17**	61.29	22.88	**9.33**	4.39
379	**60.17**	60.16	61.54	7.91	**6.95**	5.30
422	70.00	70.00	74.00	13.21	**11.24**	5.92
439	55.00	**55.56**	55.88	8.89	**7.19**	4.80
623	27.27	**33.33**	50.00	27.27	**33.33**	9.46

Table 7. Recall and Precision achieved when using related lexical databases *TS(r)*.

Considering the precision scores of *PhraseNet*, significantly better precision can be achieved with the related lexical databases *crDB* in all classes. For example for class 280, while the recall can be improved from 48% to 58%, the precision only decrease from 14% to 10%. For class 280 *PhraseNet* achieves a precision of only 6%.

Again, we measure coverage of the respective lexical databases. Table 8 shows the achieved coverage measures.

US Class	**TS(l)**	**TS(r)**	***PhraseNet***
454	52.70	52.70	89.19
384	25.86	**67.24**	98.28
126	45.12	**71.34**	94.51
148	52.78	**69.44**	94.44
219	60.47	**61.63**	93.60
433	52.94	**70.59**	94.12
180	80.71	**88.57**	97.86
417	79.86	**87.77**	97.84
398	80.59	**81.18**	93.53
280	66.33	**79.59**	96.94
128	78.26	**84.06**	89.86
379	86.76	**90.44**	95.59
422	76.92	76.92	96.15
439	82.19	**86.30**	93.15
623	52.38	**57.14**	66.67

Table 8. Coverage achieved when using related lexical databases.

As Table 8 shows, *PhraseNet* achieves best coverage for all classes compared to the lexical databases learned from the training sets of each class and to the related training sets. 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 about 41%.

Hence, the experiments show that through the expansion of the class-specific lexical databases with related classes recall and coverage obviously increases considerably. And compared to the precision values provided by *PhraseNet*, significantly better precision can be achieved with the class-related lexical databases. This provides valuable expansion opportunities specifically for smaller classes, i.e. for classes where few query logs are available.

4.4. Considering Query Log Length in automatic Query Scope Limitation

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 learned lexical databases depends on the length of the query logs/ of the search sessions. In particular, if with the increase of the query log length more detailed queries and query terms are included in the query logs, which are harder to expand. For the experiments we use the learned lexical database *PhraseNet* providing best recall scores for the recall oriented patent search task. Further, we use the test set from section 4.2 including 7,500 query log files. We divide the test set in several subsets to create multiple evaluation sets. Specifically, we order the query logs by length (in particular by the number of character strings) and divide the collection by the factor of ten. So we generate ten evaluation sets called *length*$_1$ up to *length*$_{10}$ (each comprises 750 log files) including query logs with a length of 3 up to 35,144 character strings. Table 9 shows the several subsets and the recall, precision and coverage measures achieved by the lexical database *PhraseNet*.

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

Table 9. Performance of *PhraseNet* when considering query log length.

PhraseNet achieves for all subsets *length*$_1$ to *length*$_{10}$ accurate and equivalent recall and coverage scores. In particular, *PhraseNet* 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, *PhraseNet* provides between 89% and 95% of the query terms used in the subsets.

Again, the precision values of *PhraseNet* show that only weak precision measures can be achieved across all subsets. However, on average, only 4 up to 6 out of 100 terms suggested by *PhraseNet* are used by the patent examiners for query expansion.

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

5. Conclusions and Future Work

In this paper we presented an approach to learn keyword phrases and lexical databases from query logs, which patent examiners created during the validation procedure of the patent applications, to support patent searchers in query generation, particularly in limiting the query scope.

We extracted keyword phrases from a collection of 103,896 query logs based on the proximity operator "ADJ(acent)", which is used to narrow a search, particularly to limit the scope of a general query term. To learn the lexical databases, we measured the frequency of each keyword phrase in the specific class and considered those keyword phrases that were encountered at least two times to exclude mismatches and misspellings. Finally, we learned fifteen class-specific lexical databases, particularly thesauri, providing English keyword phrases. To query the lexical databases we used an open source thesaurus management software.

Furthermore, we evaluated our approach based on real query sessions of patent examiners (gold standard). We learned multiple class-specific, class-independent, and class-overlapping lexical databases from the query log collection. The experiments have shown that our approach to learn lexical databases from the patent domain, specifically directly from the query logs, helps in semi-automatic query term expansion. The class-independent lexical database *PhraseNet* provide up to 8 out of 10 keyword phrases for the recall oriented patent search task. On average, 2 out of 10 suggest expansion terms provided by the class-specific lexical databases are used by the patent examiners for query expansion. Compared to the precision scores achieved by the related approaches the class-specific lexical databases achieve accurate precision scores. Furthermore, the precision scores can be further improved up to almost 28%, when considering a confidence value.

In addition, the expansion of the class-specific lexical databases with related classes leads to an increase in recall and coverage, while also a drastic increase in precision compared to the values provided by *PhraseNet* can be observed. This provides valuable expansion opportunities specifically for smaller classes.

Finally, 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 *PhraseNet* achieves equivalent recall, precision, and coverage scores for all subsets having different query log lengths. In particular, *PhraseNet* suggests for all subsets, between 7 up to 8 out of 10 keyword phrases, which are used by the patent examiners for query expansion. Hence, the performance of *PhraseNet* is independent from the length of the query sessions.

In future work we want to use the query log collections to learn further semantic relations that are needed for automatic query expansion in patent searching. In particular, we aim to learn synonyms of keyword phrases and synonym terms to keyword phrases from the text queries. To this end, we will rely on the extensive usage of the Boolean and proximity operators in query formulation of patent examiners. Expanding the approach currently used for individual terms, we will use the proximity operator "ADJ" to detect keyword phrases and the Boolean operator "OR" to learn the synonyms.

References

[1] Azzopardi, L., Vanderbauwhede, W. and Joho, H. 2010. Search system requirements of patent analysts. In Proceeding of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010). Geneva, Switzerland, pp. 775-776.

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

[3] Tannebaum, W., Rauber, A. 2012. Acquiring Lexical Knowledge from Query Logs for Query Expansion in Patent Searching. In the IEEE Sixth International Conference on Semantic Computing (IEEE ICSC 2012), Italy, Palermo, pp. 336-338.

[4] Tannebaum, W., Rauber, A. 2014. Using Query Logs of USPTO Patent Examiners for automatic Query Expansion in Patent Searching. In Information Retrieval. Published Online First: February 2014.

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

[6] Cetintas, S. and Luo Si, L. 2012. Effective query generation and postprocessing strategies for prior art patent search. J. Am. Soc. Inf. Sci. Technol. Volume 63, Issue 3, pp.512-527.

[7] Jochim, C., Lioma, C., Schütze, H. 2011. Expanding queries with term and phrase translations in patent retrieval. In Proceedings of the Second International Conference on Multidisciplinary Information Retrieval Facility (IRFC 2011), Vienna, Austria, pp. 16-29.

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

[9] Xue, X., Croft, W. 2009. Automatic query generation for patent search. In Proceedings of CIKM 2009, Hong Kong, China, pp. 2037-2040.

[10] Zhongquan Xiea, Z, Miyazaki, K. 2013. Evaluating the effectiveness of keyword search strategy for patent identification. In World Patent Information, Volume 35, Issue 1, pp. 20–30.

[11] Xue, X., Croft, W. 2009. Transforming patents into prior-art queries. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, USA, pp. 808-80.

[12] Vrochidis, S., Moumtzidou, A., Kompatsiaris, I. 2012. Concept-based patent image retrieval. In World Patent Information, Volume 34, Issue 4, pp. 292–303.

[13] Lupua, M, Huangb, J., Zhuc, 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.

[14] Farag Saada, F., Nürnberger, A. 2012. Overview of prior-art cross-lingual information retrieval approaches. In World Patent Information, Volume 34, Issue 4, pp. 304–314.

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

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

[17] Andersson, L., Mahdabi, P., Hanbury, A., Rauber, A. 2013. Exploring patent passage retrieval using nouns phrases. In Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR'13), Moscow, Russia, pp. 676-679.

[18] Jin, B., Teng, H., Shi, Y., Qu, F. 2007. Chinese Patent Mining Based on Sememe Statistics and Key-Phrase Extraction. In Proceedings of the 3rd international conference on Advanced Data Mining and Applications (ADMA '07), Harbin, China, pp. 516-523.

[19] 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.

[20] 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.

[21] Clough, P., Berendt, B. 2009. Report on the Treble CLEF query log analysis workshop 2009. In SIGIR Forum 43, pp. 71-77.

[22] Baeza-Yates, R., Tiberi, A. 2007. Extracting semantic relations from query logs. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07), San Jose, California, USA, pp. 76-85.

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

[24] De Lima, E., Pedersen, J. 1999. Phrase recognition and expansion for short, precision-biased queries based on a query log. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1999), Berkeley, Canada, pp. 145-152.

[25] Feuer, A., Savev, S., Aslam, J. 2009. Implementing and evaluating phrasal query suggestions for proximity search. In Information Systems, Volume 34, Issue 8, pp.711-723.

[26] Sekine, S., Suzuki, H. 2007. Acquiring Ontological Knowledge from Query Logs. In Proceedings of the 16th International Conference on World Wide Web (WWW 2007), Banff, Canada, pp. 1223-1224.

[27] Zhang, J., Xiong, M., Yu, Y. 2006. Mining Query Log to Assist Ontology Learning from Relational Database. In Proceedings of the 8th Asia Pacific Web Conference (APWeb 2006), Harbin, China, pp. 437–448.

[28] De Marco, D. 2011. Plumbing the Depths of Examiner Search (il)-Logic: A Patent Searching Perspective. Presentation given at PIUG 2011 Northeast Conference, New Brunswick (New Jersey), USA,http://www.demarcoip.com/wordpress1/wpcontent/uploads/2013/07/ExaminerSearch.pdf.

[29] 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.

[30] 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.

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

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