Acquiring lexical knowledge from Query Logs for Query Expansion in Patent Searching

Wolfgang Tannebaum
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria
tannebaum@ifs.tuwien.ac.at

Andreas Rauber
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria
rauber@ifs.tuwien.ac.at

***Abstract*—Query expansion is a crucial step in recall-oriented domains such as Patent Searching. Currently, automatic query expansion in patent search is mostly based on statistical measures. Additional query terms are extracted from the query documents based on entropy measures. To automate query expansion in patent searching, we acquire lexical knowledge from Query Logs of USPTO Patent Examiners. Results show good performance in query expansion and patent searching using the lexical database. This will help improving (semi-) automated query expansion in patent searching.**

Lexical Knowledge, Patent Searching, Query Expansion, Query Logs;

I. INTRODUCTION

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. Just as general information retrieval, patent searching consists of three phases: Query Generation (1), Document Retrieval (2) and Document Reviewing (3). In the query generation process query terms are combined to form a comprehensive query set. In the document retrieval step the patent databases of the national patent offices or commercial operators are searched. Finally, in document reviewing documents are reviewed to select the relevant ones and reiterate the patent search with new queries [2, 3]. Because patent searching is keyword-based and patent applicants are permitted to be their own lexicographers, i.e. they can define their own terminology, the success of patent searching relies on the quality of the query terms used by the patent searchers. This process is very time-consuming and the probability of missing relevant search terms is high. Furthermore, in the patent domain no sources, such as patent domain specific lexica or thesauri, are available to assist the query expansion process. In this paper, we take a closer look at acquiring lexical knowledge from Query Logs of USPTO Patent Examiners for query expansion in patent searching.

II. RELATED WORK

A. Standard Query Expansion techniques

Currently, in automatic query expansion in patent searching additional query terms are extracted from the query documents based on statistical measures, such as term frequencies (tf) and a combination of term frequencies and inverted document frequencies (tfidf) [5, 10]. Further approaches use pseudo relevance feedback or citation analysis to expand the query terms from the query documents [7]. Missing terms are discovered from feedback documents or from the cited documents. Other approaches use existing domain specific ontologies, lexical databases, such as WordNet, translation dictionaries, machine translation systems, parallel corpora for query expansion in patent searching or acquire lexical knowledge directly from the patent domain using parallel translations, particularly of the title and claim sections [4, 7, 9]. All approaches use whole documents or whole sections, like the title, abstract, description or the claim section, of the documents for query generation, query expansion and dictionary learning [2, 4, 7]. Yet, little thought is given to the query expansion happening in real query sessions. Learning from actual queries submitted by experts could address this shortcoming of the automatic query generation approaches.

B. Learning Lexical Knowledge from query logs

In previous research for acquiring lexical knowledge using query logs, terms are extracted directly from the logs or from the retrieved documents. Relations between the query terms are learned by analyzing the clicked documents. If two queries are related with the same documents, these two queries are associated with each other and the terms in the queries and in the documents can be used for query expansion [1, 6]. Synonym relations are learned by using external sources, such as lexical databases like WordNet [8, 11]. All approaches depend on synonym relations provided by external sources such as lexica, glossaries, or databases. These approaches do not utilize relations between the query terms in the query logs. Yet, we need these for learning a lexical database, because in the patent domain no domain specific lexica or thesauri will be available for relation finding.

III. ACQUIRING LEXICAL KNOWLEDGE

A. Experiment Set up

For our experiments we use a collection of query logs of patent examiners freely available from the US Patent and Trademark Office Portal PAIR. The query logs called "Examiner`s search strategy and results" are published for most patent applications since 2003. We collected all patents that are listed under the International Patent Classification (IPC)

978-0-7695-4859-3/12 $26.00 © 2012 IEEE
DOI 10.1109/ICSC.2012.15

A61C1 (Dentistry Domain) for which the examination procedure is published. We downloaded 162 query documents and the corresponding query logs (346 Logs). We extracted from the logs 1780 unique text queries consisting of query terms and search operators (Boolean operators, Proximity operators and Truncation Limiters).

B. Lexical Knowledge Extraction

Our approach for acquiring lexical knowledge from the text queries of the logs based on the search operators to learn a domain specific lexical database works as follows: we first filter all 3-grams generated from the text queries in the form "X *b* Y", where *b* is an operator and X and Y are query terms and then generate the database using the query terms and the semantic relations provided by the operators. The semantic relations provided by the logs are presented in Tab. 1.

TABLE I. SEMANTIC RELATIONS PROVIDED BY THE QUERY LOGS

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

For acquiring lexical knowledge the operators shown in Table 1 can be assigned to specific semantic relations, namely synonym, co-occurrence and proximity relations. Characteristics of the resulting semantics are shown in Table 2.

TABLE II. LEARNED SEMANTICS

Type of Relations	Code	Semantic Relations	Terms
co-occurrence	CR	549	367
synonym	SR	500	380
proximity	PR	1365	1208
∑ unique relations and terms	-	**2414**	**975**

More than half of the relations (1365) are learned from the proximity operators (ADJ(acent), NEAR, SAME, WITH) to generate proximity relations between the terms. 500 synonym relations are generated from the search operator OR and 549 co-occurrence relations from the operator AND.

C. Lexical Database

We learned a lexical database, particularly a term network which resembles a thesaurus of English terms, for a specific patent domain. The English terms are grouped into synonym, proximity and co-occurring terms. More than half of the terms of the lexical database are nouns (64,93 %) followed by adjectives (16,75 %) and verbs (15,66 %). The other terms (2,66 %) are articles, adverbs or unclassified. The main relation for query expansion in patent searching among the terms is the synonym relation (20,71 %), as "drill" and "burr" or "tool" and "instrument". Terms that have the same meaning are grouped together. The second type of learned relation is the proximity relation (56,55 %). Terms that occur within a specific distance, particularly within a specific number of words or characters, are linked to each other, such as "foot" and "pedal" or "plastic" and "ring". This relation is used for phrase search. And the third type of learned relation is the co-occurrence relation (22,74 %). Terms that occur in the same document are grouped together.

IV. EVALUATION

A. Evaluation based on query logs

We split the query log collection (346 logs) which we used for learning the database in Section 3 in a training set (243 logs) for learning the lexical databases according to our approach and in a test set (100 logs) to apply the generated semantics. From the training set we generate five sub-sets for learning the lexical databases. Three lexical databases *S1* to *S3* are generated based on different sizes of the training sets (100, 175 and 243 logs), which we selected randomly from the query log collection. Further two databases are generated based on the publication dates of the query documents. We learned a first database *T1* from the query logs from query documents published from 2003 to 2006 and a second *T2* from query documents published from 2007 to 2010. From the test we generate a first test set *R1* (gold standard) including the 100 logs. Further we split the test set *R1* to generate the test sets *P1* (53 logs) and *P2* (47 logs) according to the periods of time we learned the lexical databases *T1* and *T2*. We evaluate the lexical databases based on Recall and Precision of the provided expansion terms (*ETs).* We query from each query log the document terms (*DTs)* to retrieve the *ETs* appearing in the logs. We learn that a part of the queried *DTs* of the test set *R1* are out of the vocabulary of the databases *S1* to *S3*. To calculate the recall scores we compare the suggested *ETs* from the lexical databases with the *ETs* to the provided *DTs* from the search logs. To compute precision we compare the *ETs* appearing in the query logs with all from the lexical databases suggested *ETs* to the *DTs*. We achieve the following evaluation results for the lexical databases *S1*, *S2* and *S3* using the test set R1 as shown in Table 3.

TABLE III. EVALATION RESULTS BASED ON TRAINING SET SIZE

Results	average Recall in %			average Precision in %		
Training Sets:	***S1***	***S2***	***S3***	***S1***	***S2***	***S3***
Synonym Semantics	12,68	30,48	**30,91**	21,43	40,58	**45,95**
Syntactic Semantics	19,59	25,35	**32,45**	7,02	7,44	**9,94**
Lexical Database	22,30	37,93	**60,93**	9,86	14,79	**16,35**

The results show that the recall increase with the size of the training sets. We assume that the recall score would further increase with the rise of the training set size. Best recall scores for the test set *R1* are provided by the learned lexical databases

learned from the training set *S3* including semantic relations and the biggest training set size. The lexical database *S3* provides on average 60,93 % of the relevant *ETs*. Further, the results show that the synonym semantics provide good precision scores. Best precision is provided by the synonym semantics learned from the training set *S3* with a value of 45,95 %. As expected the syntactic semantics learned from the training sets *S1* to *S3* achieves only low precision scores.

In further experiments we analyze if the vocabulary used for query expansion chances with time. Therefore we evaluate the two learned lexical databases *T1* and *T2* each generated for a separate period of time for the same and for the other time period using the test sets *P1* and *P2*. Again we learn that a part of the queried *DTs* of the test sets *P1* and *P2* are out of the vocabulary of the databases *T1* and *T2*. Also for these experiments we evaluate the performance of the databases in view of query expansion of the *DTs* which appear in the lexical databases to retrieve the *ETs* appearing in the logs. Table 4 shows the achieved results.

TABLE IV. RESULTS BASED ON TRAINING AND TEST SET TIME

Results:	**average Recall in %**				**average Precision in %**			
Training Sets:	***T1/ P1***	***T2/ P1***	***T1/ P2***	***T2/ P2***	***T1/ P1***	***T2/ P1***	***T1/ P2***	***T2/ P2***
Synonym Semantics	8,52	**12,43**	8,60	**12,09**	**36,67**	27,50	40,00	**40,24**
Syntactic Semantics	**11,97**	11,88	14,97	**22,34**	10,68	**11,46**	13,29	**23,11**
Lexical Database	**20,49**	15,47	23,57	**30,04**	**16,54**	14,72	17,55	**27,17**

The results show that for the period *P1* the lexical database *T1* achieves better recall and precision measures than the database *T2*. For the other time period *P2* the lexical database *T2* obtains better recall and precision scores than the lexical database *T1*.

B. Evaluation based on patent searches

To evaluate the lexical database on patent searches, we evaluate the lexical database *S3* in terms of the success of query expansion, particularly how well the provided *ETs* work in retrieving the relevant documents cited by the examiners (gold standard) based on the citations to the query documents of the test set *R1* and the text queries from the query logs. We use only those text queries including the *DTs* from the query documents and the *ETs* provided by the lexical database *S3*. To calculate the average recall score for each query document of the test set we queried the cited documents using the text queries. The results show that using the *DTs* from the query documents and the *ETs* provided by lexical database, we retrieve on average 76,45 % of the cited documents to the query documents of the test set *R1*. Hence, also in recall orientated patent searching, particularly through analysis how well the lexical database helps to retrieve the cited documents (gold standard) the database shows good performance.

V. CONCLUSIONS

In this paper we presented a new approach to automate query expansion in patent searching. The experiments show that in view of query expansion done by the patent examiners (gold standard) our approach achieves good recall and precision scores. The database *S3* suggests on average nearly two of three relevant *ETs* which are used by the patent examiners for query expansion. Further, the results show that the synonym semantics provide good precision scores. On average, nearly one of two suggested *ETs* to a query term is used by the patent examiners for query expansion. Further, we learn that patent examiners create permanently new terms for query expansion and patent searching. The lexical databases achieves for the same periods of time from which they are generated better evaluation results than for the other time periods. Also in recall orientated patent searching the database shows good performance. We retrieved on average 76,45 % of the citations using the *DTs* and the *ETs* provided by the lexical database. Future work will focus on enriching the lexical database with further semantics. Query logs from patent examiners are only available from the USPTO, and even these only since 2003. Therefore we want to use granted European patents including the translations of the claim sections for learning further translation and synonym semantics.

REFERENCES

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

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

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

[4] C. Jochim, C. Lioma, H. Schütze, S. Koch and T. Ertl. „Preliminary study into query translation for patent retrieval", In Proceedings of PaIR 2011, Toronto, Canada, pp. 57–66., 2011.

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

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

[7] W. Magdy and G. Jones. "A Study of Query Expansion Methods for Patent Retrieval", In Proceedings of PaIR 2011, Glasgow, Scotland, pp. 19–24, 2011.

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

[9] S. Taduri, G. T. Lau, K. H. Law and J. P. Kesan. " Retrieval of Patent Documents from Heterogeneous Sources Using Ontologies and Similarity Analysis", In Proceedings of the 5th International Conference on Semantic Computing (ICSC 2011). IEEE Computer Society, Washington, DC, USA, pp. 538-545, 2011.

[10] X. Xue W. Croft. "Automatic query generation for patent search", In Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009). Hong Kong, pp. 2037-2040, 2009.

[11] J. Zhang, M. Xiong and Y. Yu. "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, 2006.